E-ISSN: 2602-3563

ACTA INFOLOGICA (ACIN)

JUNE 2025
Volume 9 • Issue 1

ISTANBUL UNIVERSITY, INFORMATICS DEPARTMENT

acin.istanbul.edu.tr
dergipark.org.tr/tr/pub/acin

İSTANBUL UNIVERSITY PRESS

E-ISSN: 2602-3563

Acta Infologica

Volume 9 • Issue 1 • June 2025

İSTANBUL UNIVERSITY PRESS

Indexing & Abstracting

Web of Science Core Collection – Emerging Sources Citation Index (ESCI)
TÜBİTAK ULAKBİM TR Index
DOAJ (Directory of Open Access Journals)
ERIH PLUS
EBSCO Applied Sciences Source Ultimate
Cabells Journalytic
Gale Cengage
Bielefeld Academic Search Engine (BASE)
OpenAIRE
ResearchBib
ASOS Index
Ulrichsweb

Owner
Prof. Dr. Sevinç GÜLSEÇEN
İstanbul University, Faculty of Science, Department of Computer Science, İstanbul, Turkiye

Responsible Manager
Prof. Dr. Sevinç GÜLSEÇEN
İstanbul University, Faculty of Science, Department of Computer Science, İstanbul, Turkiye

Correspondence
İstanbul University, Department of Informatics
Kalenderhane Mahallesi, 16 Mart Şehitleri Caddesi, No: 8, Vezneciler, Fatih, İstanbul, Türkiye
+90 212 440 00 00 / 10037
acin@istanbul.edu.tr
https://acin.istanbul.edu.tr/
https://dergipark.org.tr/tr/pub/acin

Publisher
İstanbul University Press
İstanbul University Press, İstanbul University Central Campus, 34452 Beyazit, Fatih, İstanbul, Türkiye
+90 (212) 440 00 00
iupress@istanbul.edu.tr
https://iupress.istanbul.edu.tr/

Publication Type
Periodical

Authors bear responsibility for the content of their published articles.
The publication language of the journal is English.
This is a scholarly, international, peer-reviewed and open-access journal published biannually in June and December.

Editorial Management

Editor in Chief

Prof. Dr. Sevinç GÜLSEÇEN
İstanbul University, Faculty of Science, Department of Computer Science, İstanbul, Turkiye
gulsecen@istanbul.edu.tr

Co-Editor in Chief

Prof. Dr. Çiğdem EROL
İstanbul University, Faculty of Science, Department of Computer Science, İstanbul, Turkiye
cigdems@istanbul.edu.tr

Section Editors

Res. Assist. Serra ÇELİK
İstanbul University, Faculty of Economics, Department of Management Information Systems, İstanbul, Turkiye
serra.celik@istanbul.edu.tr

Assoc. Prof. Emre AKADAL
İstanbul University, Faculty of Economics, İstanbul, Turkiye
emre.akadal@istanbul.edu.tr

Assoc. Prof. Fatma Önay KOÇOĞLU
İstanbul University, Faculty of Computer and Information Technologies
fonayk@istanbul.edu.tr

Ethics Editor

Prof. Dr. Mehpare TİMOR
İstanbul University, School of Business
timorm@istanbul.edu.tr

Editorial Board

Prof. Dr. Malgorzata PANKOWSKA
University of Economics in Katowice, Department of Informatics, Poland
malgorzata.pankowska@ue.katowice.pl

Prof. Dr. Mehpare TİMOR
İstanbul University,School of Business,Department of Business Administration, İstanbul, Turkiye
timorm@istanbul.edu.tr

Prof. Dr. Meltem ÖZTURAN
Boğaziçi University, School of Applied Disciplines, Department of Management Informations Systems, İstanbul, Turkiye
meltem.ozturan@boun.edu.tr

Prof. Dr. Orhan TORKUL
Sakarya University, Faculty of Engineering, Department of Industrial Engineering, Sakarya, Turkiye
torkul@sakarya.edu.tr

Prof. Dr. Selim YAZICI
İstanbul University, Faculty of Political Sciences, Department of Business Administration, İstanbul, Turkiye
selim@istanbul.edu.tr

Prof. Dr. Sushil K. SHARMA
Ball State University, Miller College of Business, USA
ssharma@bsu.edu

Prof. Dr. Türksel KAYA BENSGHİR
Hacı Bayram Veli University, Faculty of Economics and Administrative Sciences, Department of Business Administration, Ankara, Turkiye
t.bensghir@hbv.edu.tr

Prof. Dr. Üstün ÖZEN
Atatürk University, Faculty of Economics and Administrative Sciences, Department of Management Informations Systems, Erzurum, Turkiye
uozen@atauni.edu.tr

Prof. Dr. Vesselina NEDEVA
Trakia University, Faculty of Engineering and Technologies, Department of Electrical Engineering, Electronics and Automation, Bulgaria
veselina.nedeva@trakia-uni.bg

Prof. Dr. Yacine LAFİFİ
University 8 May 1945 Guelma, Faculty of Mathematics, Computer Science and Materials Science, Department of Computer Science, Algeria
lafifi.yacine@univ-guelma.dz

Prof. Dr. Elzbieta MAGDALENA WASIK
Mickiewicz University, Department of Old Germanic Languages, Poznan, Poland
elawasik@amu.edu.pl

Prof. Dr. Cem SÜTÇÜ
Marmara University, Faculty of Communication, Department of Journalism, İstanbul, Turkiye
csutcu@marmara.edu.tr

Assoc. Prof. Dr. Tuncay ÖZCAN
İstanbul Technical University, Faculty of Management, Department of Management Engineering, İstanbul, Turkiye
tozcan@itu.edu.tr

Assoc. Prof. Dr. Jan GUNCAGA
Comenius University in Bratislava,Faculty of Education, Department of Didactics of Natural Sciences in Primary Level, Slovakia
guncaga@fedu.uniba.sk

Assist. Prof. Dr. Zerrin AYVAZ REİS
İstanbul University-Cerrahpaşa, Hasan Ali Yücel Faculty of Education, Department of Computer Education and Instructional Technology, İstanbul, Turkiye
ayvazzer@iuc.edu.tr

Dr. Luis Miguel CARDOSO
University of Lisbon, Polytechnic Institute of Portalegre, Centre for Comparative Studies, Portugal
lmcardoso@ipportalegre.pt

Assoc. Prof. Dr. Tetiana BONDARENKO
Ukrainian Engineering Pedagogics Academy, Department of Information Computer Technologies and Mathematics, Kharkov, Ukrain
bondarenko_tc@uipa.edu.ua

Prof. Dr. Jagdish KHUBCHANDANI
New Mexico State University,Department of Public Health, USA
jagdish@nmsu.edu

Assoc. Prof. Dr. Natalija LEPKOVA
Vilnius Gediminas Technical University, Faculty of Civil Engineering, Department of Construction Management and Real Estate, Lithuania
natalija.lepkova@vilniustech.lt

Table of Contents

Research Article
Multi-Category E-Commerce Insights via Social Media Analysis using Machine Learning and BERT **01**
Meryem Gürbüz, Muhammed Kotan

Research Article
DFCEA: Deep Forest Classifier-Based Employee Attrition Prediction **19**
Mahmut Tokmak

Research Article
Classification of Ventricular Septal Defect Disease Using Deep Learning **34**
Kadir Barut, İhsan Pençe, Özlem Çetinkaya Bozkurt, Melike Şişeci Çeşmeli

Research Article
A Novel and Robust LSTM Model for Customer Churn Analysis Using Deep, Machine Learning, and Ensemble Learning: A Telecommunications Case **55**
Muhammet Sinan Başarslan, Aslıhan Ünal, Fatih Kayaalp

Research Article
Optimizing the Sentiment Recognition in Spotify Playlists Through Ensemble-Based Approaches **74**
M. Erdem İsenkul

Research Article
Evaluating the Accessibility, Usability, and Security of Corporate Investor Relations Web Pages: A Case Study of the Turkish Stock Market **90**
Merve Garip, Serhat Peker

Research Article
YOLOv8 for Road DamageRecognition: Deep Learning-Based Segmentation and Detection System **112**
Mehmet Tevfik Ağdaş, Kaan Arık

Research Article
Artificial Intelligence and Cost Reduction Strategies for Healthcare Management: Opportunities and Limitations **133**
Canan Bulut

Research Article
Generative AI in Higher Education: Students' Perspectives on Adoption, Ethical Concerns, and Academic Impact **147**
Gökçe Karahan Adalı, Asiye Bilgili

İSTANBUL UNIVERSITY PRESS

Acta Infologica
Volume 9 • Issue 1 • June 2025

Research Article
An Investigation into the Impact of Influencers Producing Technological Content on Instagram on Customers' Purchasing Decisions **167**
Kamil Çelik, Ahmet Taş

Research Article
Designing a Large Language Model-Based AI System for Dynamic Difficulty Adjustment in Digital Games **183**
Onur Aşkın

Research Article
Investigation of User Comments on Videos Generated by Deepfake Technology **208**
Sertaç Kaya

Research Article
Voices from the Lungs: An Innovative Approach to Asthma Diagnosis using Machine Learning **223**
Murat Gezer, Mehmet Atilla Uysal, Neval Alagöz, Can Ortak, Elif Yelda Niksarlıoğlu

Research Article
The Impact of Social Networks on the Stock Market Using Sentiment Analysis and Machine Learning: Application to the Turkish Stock Market **253**
Mustafa Kemal Mayuk, Farid Huseynov

Research Article
Enhancing Software Requirement Classification via Dataset Fusion and Machine Learning **275**
Muhammad Owais Raza, Vajeeha Mir, Jawad Rasheed, Mirsat Yeşiltepe, Shtwai Alsubai

Research Article
Predicting Industry Maturity Index Using Machine Learning Methods **293**
Ayhan Doğan, Cihan Ünal

Research Article
Factors Influencing Consumer Adoption of Smart Home Systems: A Socio-Demographic Perspective **314**
Onur Muço, Üstün Özen

Research Article
Data Governance for Businesses: Challenges, Recommendations, and Critical Success Factors **339**
Vildan Ateş, Atacan Garip

Review Article
Engineering Education Meets the Metaverse **359**
Veselina Nedeva, Snejana Dineva, Svetoslav Atanasov

Preface to June 2025 Issue

Dear readers of Acta INFOLOGICA,

Today, the applicability of data-driven approaches in a wide range from institutional structures to social networks, from educational technologies to health systems; further increases the importance of the symbiotic relationship between artificial intelligence, informatics and computer science. In this context, theoretical and applied studies addressed with interdisciplinary perspectives bring not only technical developments but also ethical, social and philosophical issues to the agenda.

The current issue of our journal aims to open space to current research, experiences and discussions conducted at the intersection of artificial intelligence-informatics-computer science. We would like to thank our valuable researchers, editorial team and everyone who contributed; we hope that this issue will open new horizons for the literature and relevant stakeholders.

I wish to thank all authors, readers and referees for their invaluable efforts.

Sevinç GÜLSEÇEN

Chief-Editor

İSTANBUL UNIVERSITY PRESS

Acta Infologica,
ACIN 2025, 9 (1): 1–18

https://doi.org/10.26650/acin.1483488
Submitted: 14.05.2024
Revision Requested: 07.12.2024
Last Revision Received: 24.12.2024
Accepted: 31.12.2024
Published Online 03.02.2025

Acta Infologica

Research Article

Open Access

Multi-Category E-Commerce Insights via Social Media Analysis using Machine Learning and BERT

Meryem Gürbüz [1] & Muhammed Kotan [2]

[1] Sakarya University, Institution of Natural Sciences, Sakarya, Türkiye
[2] Sakarya University, Department of Information Systems Engineering, Sakarya, Türkiye

Abstract The burgeoning prevalence of Internet and social media usage has empowered consumers to effortlessly share their opinions about products and services on social media platforms and websites. Consequently, recent research has focused on using machine learning, text mining, and sentiment analysis techniques to extract valuable insights. These insights can then be employed to support businesses in enhancing customer satisfaction and making informed operational and strategic decisions. In this study, a dataset of 5806 Trendyol user reviews was collected from X using the X API within a specified time frame. The dataset was preprocessed and categorized into five predefined categories: product, support, logistics, advertising, and off-topic. Subsequently, the test set was classified using eight machine learning techniques and compared. Finally, sentiment analysis was performed using the pretrained BERTurk model to evaluate user satisfaction and dissatisfaction levels. By integrating machine learning and BERT, this study extracted a general assessment profile of social media users, particularly for e-commerce platforms, and examined social media perspectives on a multi-category basis.

Keywords Sentiment Analysis · Machine Learning · Transformers · User Reviews · E-commerce

Citation: Gürbüz, M. & Kotan, M. (2025). Multi-category e-commerce insights via social media analysis using machine learning and BERT. *Acta Infologica*, *9*(1), 1-18. https://doi.org/10.26650/acin.1483488

This work is licensed under Creative Commons Attribution-NonCommercial 4.0 International License.
2025. Gürbüz, M. & Kotan, M.

Corresponding author: Muhammed Kotan mkotan@sakarya.edu.tr

Introduction

With the increasing use of the internet and social media, users have started to prefer communication methods established on these platforms rather than traditional ones. Nowadays, people can constantly share their opinions and thoughts on various topics through social media platforms in text, image, or audio format. In this context, X (formerly known as Twitter) allows users to share long textual tweets for longer posts with a character limit of 25,000 characters (X Help Center, 2024). Thanks to this feature, X generates massive datasets, and it is subject to extensive data analysis studies. These vast data volumes have become a significant resource opportunity for various sectors, such as health, education, shopping, production, services, tourism, food, media, finance, and textiles. With the increase in online shopping, people's sharing of opinions, comments, support, or complaints about the products or services they have purchased through social platforms has also increased.

Consumers interact for various purposes, such as sharing their opinions and emotions, providing feedback or complaints about products and services, seeking information, requesting support for after-sales issues, and resolving problems related to their purchases. These interactions produce valuable data that businesses analyze to assess customer satisfaction and to inform strategic and operational decision-making.

Classification of text and reviews in the context of e-commerce is of significant importance for businesses seeking to gain a comprehensive understanding of customer satisfaction and enhance the quality of their offerings. The collection of customer reviews and ratings is a valuable means of obtaining invaluable insights. To effectively process data, various machine learning techniques are frequently used for this purpose in the literature. This study uses eight machine learning techniques to classify social media users' comments about Trendyol into five predefined categories, and the results are compared.

In addition, sentiment analysis plays a significant role in analyzing customers' attitudes and requirements from their reviews. It allows businesses to extract insights from user-generated content like product reviews. The main objective of this endeavor is to obtain and integrate personal viewpoints, emotions and subjective elements from written content, thereby enabling the categorization of content into positive, negative or neutral classifications (Grandi et al., 2015). Businesses prioritize researching and interpreting consumer behaviors to sustain their existence. Therefore, they value customer feedback, opinions, and suggestions on online platforms. Sentiment analysis is an effective method and opportunity for businesses to understand and enhance customer satisfaction, leading to increased competitive advantage and profit margins. This analysis can be executed at different levels, including the document, sentence, and aspect levels, to provide a more detailed understanding of the sentiment (Kotan, 2022). In this study, the Turkish comments of X users were evaluated. The categorical-based sentiment analysis of the comments categorized using machine learning techniques was performed using the Turkish BERT (BERTurk) (Schweter, 2020).

One of the prominent examples of e-commerce platforms in Turkey is Trendyol. Therefore, customers continuously share their opinions, comments, support inquiries, and shipment experiences about Trendyol on various social platforms related to its sales and post-sales processes. These shared posts serve for individuals to communicate directly with each other or with the company. This study analyzes and categorizes tweets to determine the extent to which customers view the offered service policies positively or negatively. The outcomes of this analysis provide strategic insights for decision-making processes.

In contrast to typical sentiment analysis research that applies polarity classification across entire datasets, the proposed method employs a category-based analytical framework. First, we classify user-generated content into a set of predefined categories that reflect key aspects of e-commerce. Then, we perform sentiment analysis at the category level rather than treating all posts as a monolithic dataset. This two-stage approach addresses a common challenge in e-commerce-focused social media analysis, where off-topic or tangential discussions often dilute or misrepresent the overall sentiment. By isolating relevant domains, the proposed method produces precise and sensitive sentiment evaluations. Consequently, stakeholders gain better actionable insights for strategic decision-making because sentiments are linked to meaningful e-commerce categories rather than an aggregated sentiment score spanning disparate or irrelevant comments.

Furthermore, e-commerce platforms can use these results and their analysis to adopt new strategic policies and activities in areas that require improvement, ultimately increasing customer satisfaction and profit margins. Users can freely express their opinions and share their experiences by revealing their identities or remaining anonymous, allowing for a genuine understanding of what users think about the given service, product, or business. Given these reasons, this study holds significance in understanding consumer behaviors and communication preferences for the retail sector.

The remainder of this study will first discuss similar works in the literature. The third section introduces the methodology used in the study. Section 4 discusses the findings. Finally, conclusions and future work are presented in the last section.

Literature Review

Text classification and sentiment analysis are critical tasks in e-commerce, allowing businesses to interpret customer feedback and enhance user satisfaction. In e-commerce, user comments and feedback are analyzed through machine learning algorithms to extract actionable insights that can improve customer experience, marketing strategies, and product development (Qin et al., 2024; George & Baskar, 2024). Traditional machine-learning models are recognized for their interpretability, scalability, and effectiveness in text classification tasks (Shanto et al., 2023).

Traditional machine learning algorithms are particularly well-suited to classifying e-commerce reviews. These models simplify the extraction of structured insights and highlight trends in customer satisfaction and product feedback, assisting companies in identifying key product features that resonate with customers and optimizing product offerings and targeted marketing strategies (Saleem et al., 2019; He et al., 2022). Algorithms such as Random Forest (RF), Naïve Bayes (NB), and Support Vector Machines (SVM) can efficiently process and classify text data when paired with natural language processing techniques like tokenization, stemming, and vectorization. These methods excel with labeled data, making them ideal for binary or multi-class analysis and often require less data to train effectively.

Dharrao et al. (2023) conducted a detailed study on e-commerce product review classification using Amazon reviews, categorizing ratings from one to five. NB, SVM, and Decision Trees (DT) were employed, with DT achieving the highest accuracy and precision. Deniz et al. (2022) created a dataset of over 50,000 Turkish e-commerce reviews, each labeled across multiple categories. The proposed method achieved promising results in terms of identifying nuanced customer opinions by applying multi-label machine learning techniques. They used embedding methods such as Term Frequency–Inverse Document Frequency (TF-IDF), Word2Vec, Global Vectors for Word Representation (GloVe), and Bidirectional Encoder Representations from Transformers (BERT). The machine learning algorithms include RF, SVM, NB, Multi-label k-nearest Neighbors

(Ml-kNN), and various One-vs-Rest (OvsR) classifiers. Polat and Ağca (2022) compared machine learning methods to classify emotional tendencies in TripAdvisor's Turkish and English hotel reviews. Using algorithms such as DT and RF, the analysis revealed that machine-learning approaches outperformed dictionary-based methods in terms of classification accuracy. Yıldız (2020) applied machine learning algorithms, such as RF and Gradient Boosting (GB), to analyze sales effectiveness across product groups in Turkish e-commerce. By examining stock differences based on e-commerce data, this study revealed that the predictive success of algorithms varies depending on the product category, which suggests that factors influencing sales performance may shift across different product groups. Xu et al. (2020) proposed a continuous NB framework for classification that adapts to new domains by using past knowledge and demonstrated robust performance on large-scale multi-domain e-commerce reviews from Amazon products and movies. Kubrusly et al. (2022) applied tree-based methods, such as Classification Tree (CT), RF, GB, and XGBoost (XGB), to a women's clothing e-commerce review dataset. The proposed methods achieved consensus regarding the key classification terms. Chaubey et al. (2023) compared various machine learning techniques for predicting customer purchasing behavior, including supervised classifiers like Logistic Regression (LR), DT, kNN, NB, SVM, RF, Stochastic Gradient Descent (SGD), Artificial Neural Network (ANN), AdaBoost (AB), and XGB, as well as hybrid models like SvmAda, RfAda, and KnnSgd. The hybrid KnnSgd classifier achieved the highest accuracy by combining the kNN and SGD techniques in an ensemble stacking approach. Tang et al. (2015) presented a neural network (NN) approach for predicting review ratings by integrating user-specific information and demonstrated the potential of NNs.

Sentiment analysis can be considered a text classification task involving binary or multi-class categorization and is an important area in e-commerce studies. This enables businesses to interpret customer emotions and opinions expressed in text data, thereby facilitating a better understanding of consumer behavior and enhancing user satisfaction. Huang et al. (2023) conducted a comprehensive study of current sentiment analysis techniques on e-commerce platforms and explored future directions. By analyzing 54 experimental papers selected from 271 papers identified through specific keywords, the research provides valuable insights for researchers. Marong et al. (2020) provided an overview of sentiment analysis and related methodologies for the e-commerce sector. Beyhaqy et al. (2018) analyzed social media users' sentiment toward e-commerce platforms using tweets from Tokopedia and Bukalapak on X. Among the DT, kNN, and NB classifiers, the NB classifier yielded the best results with an accuracy of 77%. Akter et al. (2021) analyzed the sentiment in Bangla text reviews collected from the Bangladeshi e-commerce platform Daraz. The study evaluated several machine learning models, including the RF, LR, SVM, kNN, and XGB models, with the kNN model achieving the highest performance. Demircan et al. (2021) applied SVM, RF, DT, LR, and kNN to classify the sentiments in Turkish product reviews and found that SVM and RF effectively classified the sentiments into three categories. Taşçı et al. (2024) conducted a sentiment analysis of Amazon product reviews using Kaggle data. Applying TF-IDF vectorization and evaluating algorithms like Passive Aggressive (PA), SVM, RF, AB, kNN, and XGB, RF achieved the highest accuracy of 96.13%. Kumar et al. (2016) focused on mining product reviews from websites like Amazon, using algorithms such as NB, LR, and SentiWordNet to classify reviews as positive or negative, with NB proving the most efficient. Shehu and Tokat (2020) examined Turkish tweets by collecting a dataset of 13,000 tweets via the X API. After preprocessing for noise reduction, 10,500 tweets were retained. The hybrid approach combining SVM and RF classifiers achieved the highest accuracy. Rumelli et al. (2019) addressed data imbalance by undersampling and selecting equal numbers of positive and negative samples while discarding neutral reviews. Sentiments were classified using NB, RF, SVM, and kNN. The NB and kNN methods were the fastest, while the SVM and RF methods took longer. Despite the preprocessing

challenges and misleading user ratings, NB achieved the highest accuracy. Haque et al. (2018) used Amazon product reviews focusing on electronics, cell phones, accessories, and musical instruments, with a dataset of approximately 48,500 reviews. They applied classifiers, including SVM, NB, SGD, RF, LR, and DT, and they found that linear SVM achieved the best classification performance. Zada and Albayrak (2023) analyzed user comments from a long-standing Turkish online shopping platform. Results were compared with user ratings using RF and AB for classification. Stephenie et al. (2020) used sentiment analysis on Tokopedia product reviews to gauge customer responses to product quality and seller interactions. The optimized RF model achieved a high accuracy of 97.38%. Using a dataset of product reviews from the AliExpress platform, Arobi et al. (2022) employed eight machine learning algorithms to classify the sentiments, and they found that LR was the most effective approach for this dataset. Jabbar et al. (2019) highlighted the use of SVM for constructing a sentiment analysis framework executed on an e-commerce application. This study utilizes online product reviews from Amazon and conducts sentiment analysis at both the review and sentence levels. Şimşek et al. (2023) used Turkish X data to examine users' sentiments on the Hepsiburada e-commerce site. After text mining and evaluation, the tweets were categorized into positive and negative sentiments. Savci and Das (2023) created new Turkish, English, and Arabic datasets, performed comparative sentiment analysis on texts in all three languages, and tested the performance of pretrained models.

The rise of transformer-based language models has dominated NLP studies, becoming a new paradigm in the field (Yildirim & Asgari-Chenaghlu, 2021). Guven (2021a) compared the performance of BERT models against traditional machine learning methods, like RF, NB, and LR, in classifying sentiment in Turkish tweets. BERT achieved a notable accuracy of 98.75%, with LR as the top-performing traditional method at 98.4%, highlighting the efficacy of BERT models for the Turkish language. Koksal et al. (2022) developed a category classification model that uses user comments and product descriptions from an online sales platform. The BERT model trained for the Turkish language provided the best results, forming the basis for a search engine optimization and intent analysis engine specific to the e-commerce sector. Guven (2021b) performed sentiment analysis on product reviews from Hepsiburada to assist consumers in making informed decisions. Initially, RF, NB, and LR were evaluated, with NB achieving the highest accuracy at 89.95%. The study then compared these results with the language models BERT, ELECTRA, and ALBERT and found that ELECTRA outperformed all, with a 92.54% accuracy. Köksal and Özgür (2021) proposed manually annotated Turkish sentiment analysis datasets from X, the BounTi dataset, containing Turkish tweets about specific universities in Turkey. They also evaluated the performance of transformer models, such as MBERT, XLM-Roberta, and BERTurk.

The literature confirms the significance of machine learning techniques in the e-commerce industry, highlighting their practicality and effectiveness in deriving insights that lead to strategic business advantages (Chen, 2022). Traditional machine learning algorithms are widely used due to their high efficiency and interpretability. The development of transformer-based language models represents a paradigm shift in natural language processing. Advanced models, such as BERT and its variants, have demonstrated superior performance in classification tasks across various languages. The proposed method bridges the gap between generalized sentiment analysis and the need for category-specific insights. Many previous studies have focused on sentiment analysis across entire datasets or single-stage classification approaches; however, this study introduces a two-stage methodology to address the limitations of such approaches. In the first stage, user-generated content is classified into predefined categories tailored to e-commerce. In the second stage, sentiment analysis is performed for each category to provide more precise and context-aware evaluations. This dual-stage process tackles the common issue of off-topic or irrelevant content prevalent in social media

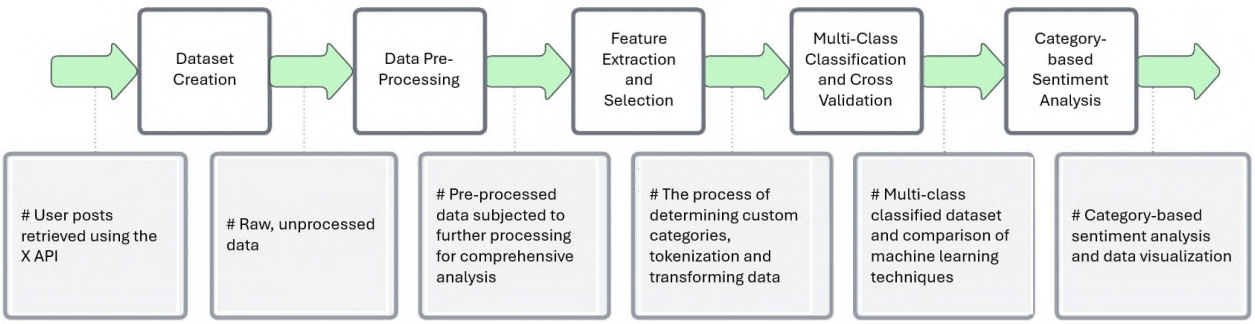
datasets, which often distorts the overall sentiment analysis. By isolating and analyzing relevant domains, the proposed approach ensures that sentiment evaluations are linked to meaningful e-commerce categories rather than aggregated scores spanning disparate or unrelated comments.

Methodology

In this study, the social media (X) posts of e-commerce customers (Trendyol) were manually labeled with five predefined categories, and the performance of eight machine learning methods for multi-class classification was examined. Subsequently, sentiment analysis was conducted on the categorized comments using the BERT method tailored to the Turkish language. Figure 1 illustrates the key steps of the application process, from data collection and preprocessing to category classification and sentiment analysis. In this study, we used Orange Data Mining software.

Figure 1

Steps of the application process.



Dataset Creation

The vast information sharing on the internet and through social media platforms has led to the accumulation of abundant data on the web. Data extraction and scraping involve retrieving data from websites or other platforms. The dataset used in this study consisted of comments about Trendyol collected through the X API. As a result of this process, 5,806 comments were obtained and stored in .csv format along with their respective dates. Word clouds visually represent text data, which allows researchers to quickly identify the most frequent and important words and phrases. This facilitates exploring datasets and uncovering hidden patterns. The word cloud is presented in Figure 2.

Figure 2
Dataset Wordcloud.

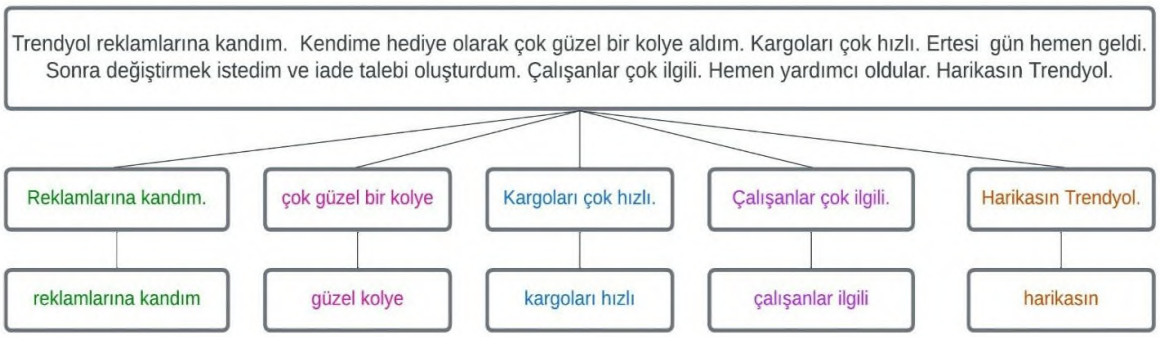

deneyim
özür
indirim
adına
olduğunuz
trendyol
merhaba
dileriz
kontrolünü
olumsuz
yaşamış
reklamı
ürün
artık
bi
kargo
ben
yok
var
ilgili
kadar
siparişinizin
sipariş
reklam
ekiplerimize
trendyol'dan
alışveriş
yardımcı
geldi
yeter
teslimat
sağlayıp
bana
yine
sonra
benim
reklamları
bilgilendirme
müşteri
iade
indirimi
yaşadığınız
ürünün
sürekli
görmekteyiz
dönüş
aldım
geri
mi
bugün
yılın
başka
kısa
ederim
ederiz
teslim
gelmedi
yapıldığını
bile
gün
oldu
önce
değil
kargoya
para
no
param
sürede
devam
ürünleri
takip
youtube
zaten
zaman
olsun
güzel
hafta
görmek
ürünler
yüzünden
trendyoldan
bekliyorum
hepsiburada
mesaj
arkadaşlar
gerçekten
kontrol
paramı
tane
haber
senin
geliyor
hediye
ettim
sizden
linki
ilk
başladı
yeni
varsa
size
lütfen
sana
iptal
bende
gündür
için
market
trendyol'da
olarak
aydır
uzak
teşekkür
rica
hala
ürünü
karşıma
beni
son
bitti
bunu
konu
tek
bak
asla
falan
tekrar
sadece
ptt
iade
mail
artık
yerde
hangi
aldığım
olup
sen
böyle
bize
bütün
ilgili
şimdi
aynı
allah
farklı
rağmen
indirimi
gelen
fırsatları
olan
kardeşim
siparişim
keşke
indirime
yerdim
bizi
lira
üzerinden
takibimizde
versin
bence
reklamlarından
istiyorum
istemiyorum
burada
almak
yerden
arada
kargom
teşekkürler
ediyorum
firma
içinde
biktım
siparişiniz
özel
kaç
ye
efsane
tl
sağlayacağız
aralık
hemen
büyük
be
ay
ürünlerin
olsa
batağına
sorun
cevap
çıkıyor
tam
düştüm
indirimleri
karşıma
ürünü
rica
teşekkür
aydır
uzak
olarak
bitti
trendyol'da
için
gündür
iptal
bende
sana
lütfen
başladı
ilk
linki
yeni
reklamları
ettim
hediye
sizden
geliyor
görmek
yüzünden
hafta
ürünler
güzel
olsun
zaten
zaman
youtube
takip
ürünleri
sürede
devam
müşteri
eksik
sağlayıp
yine
sonra
benim
sal
param
yanlış
siparişi
para
no
reklamı
olumsuz
iade
ptt
tekrar
sadece
değil
önce
yerde
mail
artık
kargoya
öyle
hangi
aldığım
oldu
gün
ürün
merhaba
dileriz
kontrolünü
yaşamış
bile
yok
artık
trendyol
bi
kargo
ben
son
yeter
teslimat
hala
ürünü
karşıma

Text Preprocessing

Text preprocessing is the process of cleaning raw data to prepare them for further analysis. These procedures include correction, filling in missing data, removing duplicate data, data transformation, cleaning, normalization, integration, and dimension reduction. Although these processes may vary depending on the specific study, data preprocessing is a crucial step that needs to be carefully executed for more efficient and accurate analysis results. The processes conducted during the data preprocessing stage are illustrated in Figure 3. The following steps were performed during data preprocessing:

- Conversion of uppercase letters to lowercase letters
- Cleaning the punctuation marks and numbers
- Removal of special characters and emojis
- Removal of stop words
- Deletion of empty lines

Figure 3
Sample data preprocessing outcomes.



Feature Extraction and Selection

In this step, TF-IDF was selected. TF-IDF is a widely used technique in text processing and information retrieval to evaluate the importance of a term within a document relative to a collection of documents. The TF-IDF value increases proportionally to the number of times a term appears in a document and is offset by the frequency of the term in the collection. Using the TF-IDF, frequent terms in a document but rare in the overall collection are given higher weights, indicating their importance in that specific document. In contrast, common terms across all documents were penalized in their TF-IDF scores. Figure 4 illustrates the conversion of clear text data to numeric scores using the TF-IDF algorithm.

Figure 4

Example of TF-IDF transformation.



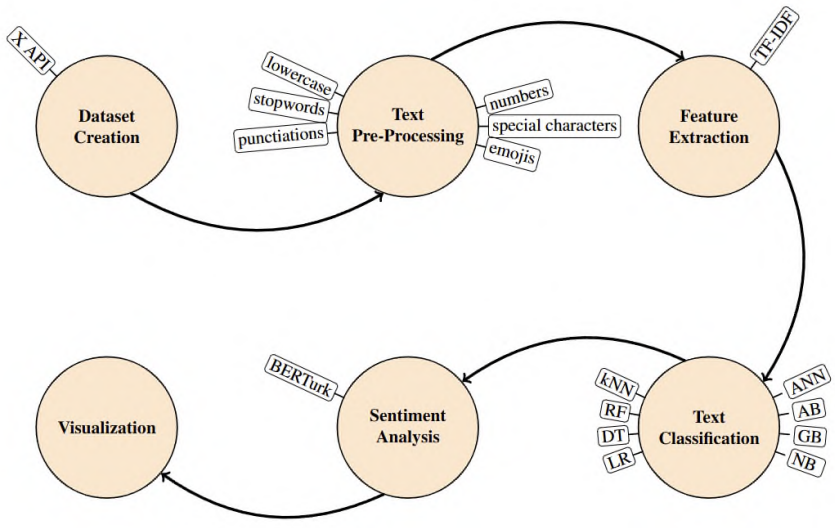
Term	Term ID	TF-IDF Score
kargolamaları	201123	0,56325
iyi	301212	0,89653
satış	125632	0,63254
sonrası	126547	0,31256
destekleri	201123	0,78541
iyi	301212	0,89632
tavsiye	254631	0,79856
ederim	145875	0,35623

Machine-Learning-based Multi-Class Classification

The text classification task assigns predefined categories to the text data. Machine learning algorithms are crucial for automating this process, as they enable computers to analyze large volumes of text and make accurate predictions. The algorithm is presented with a labeled dataset during training, where each text sample is associated with categories. The algorithms learn the patterns and relationships in the data and construct a model that can distinguish different categories. Once trained, the model can be used to predict the category of new, unseen text samples. In this study, AdaBoost (AB), Naïve Bayes (NB), Decision Tree (DT), Artificial Neural Network (ANN), k-Nearest Neighbor (kNN), Logistic Regression (LR), Random Forests (RF) and Gradient Boosting (GB) models were used. The general system structure is summarized in Figure 5.

Figure 5

Overview of the system architecture.



To compare the performance of the methods, a 70%-30% train and test dataset was created. The parametric information of the models and evaluation methods used in this study are presented in Table 2.

Sentiment Analysis using BERTurk

Sentiment analysis is a fundamental aspect of NLP, providing valuable insights for various applications, ranging from analyzing social media content to forecasting stock market trends. Transformer-based BERT (Devlin, 2018) models have demonstrated high accuracy and success in sentiment analysis tasks and have been used in sentiment analysis across various fields, including e-commerce (Sayeed, 2023).

Table 2

Overview of model and parameter settings

MODEL	
k Nearest Neighbor	k = 5; metric: Euclidean
AdaBoost	# of estimators: 50; learning rate: 1.0
Neural Network	# of neurons: 100; max iteration: 200
Decision Tree	Min number of instances in leaves: 2; maximal tree depth: 100
Random Forest	# of trees: 10
Gradient Boosting	# of trees: 100; learning rate: 0.1
Logistic Regression	
Naïve Bayes	
VALIDATION	
Train-Test	%70-%30 (4065 – 1741)

BERT employs different pre-training tasks, such as Masked Language Modeling (MLM) and Next Sentence Prediction (NSP), to increase the model's understanding of the relationship between words and context. This study used the BERTurk (Schweter, 2020) model and a fine-tuned version (Yildirim, 2024), which is available for Turkish texts. The current version of the model is trained on a filtered and sentence-segmented version of the Turkish OSCAR corpus, a recent Wikipedia dump, various OPUS corpora, and a special corpus (Poyraz, 2022). Yildirim's (2024) fine-tuned BERTurk model is trained with base settings for many downstream tasks and evaluated with a Turkish Benchmark dataset. The study outperformed some baseline approaches for Named-Entity Recognition, Sentiment Analysis, Question Answering, and Text Classification in Turkish Language (Yildirim,2024).

Experimental Results

We comprehensively analyzed e-commerce reviews for Trendyol on X. Machine learning techniques were used to classify user comments by considering predefined classes. Then, a category-based sentiment analysis was conducted. Initially, we employed supervised machine learning algorithms to construct predictive classification models to assess the content of reviews. The manual labeling of the tweets allowed us to train the algorithms effectively. Table 3 summarizes the performance metrics of each model on the multi-classification task.

Table 3

Multi-class classification results

Model	AUC	Accuracy	Precision	Recall	F1 Score
kNN	0.695	0.527	0.702	0.527	0.462
DT	0.835	0.681	0.675	0.681	0.670
RF	0.900	0.729	0.740	0.729	0.720
NN	0.874	0.707	0.715	0.707	0.701
NB	0.907	0.686	0.731	0.686	0.698
LR	0.893	0.731	0.731	0.731	0.726
GB	0.901	0.730	0.760	0.730	0.719
AB	0.897	0.726	0.735	0.726	0.721

The results show that the NB, GB, and RF models achieved the highest performance, with an AUC ≥0.90. The LR model demonstrated competitive performance, closely following with an accuracy of 0.73. The confusion matrices of all models are presented in Table 4. The category-based ROC curves of the LR model are also presented in Figure 6.

The confusion matrices in Table 4 provide a comprehensive view of the models' performances across the predefined categories: S, L, A, O, and P. Each matrix highlights the number of correctly classified instances (diagonal values) and misclassified instances (off-diagonal values), providing insights into the strengths and limitations of each model. The confusion matrices demonstrate that category O consistently achieved the highest correct classification counts across all models, which indicates that its features are relatively distinct. In contrast, category P consistently demonstrated the lowest correct classification counts and was frequently misclassified, which highlights the need for improved feature representation and dataset rebalancing. This analysis provides valuable insights into refining the classification pipeline and optimizing model performance.

Table 4

Confusion matrices of all models

LR

Actual \ Predicted	S	L	A	O	P	∑
S	278	7	1	51	10	347
L	31	129	3	31	9	203
A	7	1	279	81	5	373
O	23	17	55	519	24	638
P	15	10	14	73	68	180
∑	354	164	352	755	116	1741

GB

Actual \ Predicted	S	L	A	O	P	∑
S	267	4	1	68	7	347
L	15	135	0	45	8	203
A	2	0	236	128	7	373
O	20	4	16	589	9	638
P	14	3	5	114	44	180
∑	318	146	158	944	75	1741

RF

Actual \ Predicted	S	L	A	O	P	∑
S	263	8	3	65	8	347
L	25	131	2	33	12	203
A	3	0	259	102	9	373
O	20	3	26	569	20	638
P	18	6	9	99	48	180
∑	329	148	299	868	97	1741

AB

Actual \ Predicted	S	L	A	O	P	∑
S	271	8	5	47	16	347
L	32	125	3	31	12	203
A	6	0	248	104	15	373
O	28	6	30	550	24	638
P	21	5	9	75	70	180
∑	358	144	295	807	137	1741

NN		Predicted					
	Category	S	L	A	O	P	∑
Actual	S	262	8	2	63	12	347
	L	24	112	6	52	9	203
	A	6	0	262	99	6	373
	O	25	12	63	525	13	638
	P	12	8	15	75	70	180
	∑	329	140	348	814	110	1741

NB		Predicted					
	Category	S	L	A	O	P	∑
Actual	S	219	48	4	27	49	347
	L	12	153	3	13	22	203
	A	2	5	292	41	33	373
	O	9	28	77	425	99	638
	P	2	16	19	37	106	180
	∑	244	250	395	543	309	1741

DT		Predicted					
	Category	S	L	A	O	P	∑
Actual	S	272	11	4	52	8	347
	L	34	121	3	31	14	203
	A	13	2	246	104	8	373
	O	44	25	49	502	18	638
	P	18	24	15	79	44	180
	∑	381	183	317	768	92	1741

kNN		Predicted					
	Category	S	L	A	O	P	∑
Actual	S	188	1	1	157	0	347
	L	8	13	3	178	1	203
	A	1	0	79	290	3	373
	O	4	0	9	621	4	638
	P	2	0	3	159	16	180
	∑	203	14	95	1405	24	1741

We further conducted sentiment analysis on the classified test set reviews using BERTurk specifically pre-trained on Turkish text to understand the linguistic nuances of the language. The proposed model was employed to predict the sentiment polarity of each review in the test dataset. To evaluate the performance of the sentiment analysis, we compared the predicted sentiments against the ground truth labels in the test dataset. The results indicate that BERTurk achieved an overall accuracy of 77.1%, demonstrating its effectiveness in capturing the sentiment context within Turkish text. Accuracy reflects the model's ability to correctly predict sentiments in most cases. However, it also highlights some challenges, such as language ambiguities and the complexity of specific reviews. The positive-negative ratio of the test set was found to be 34%-66%. Figure 7 presents the sentiment analysis results of the LR-categorized reviews. Figure 7a shows the category-based positive review distributions, and Figure 7b shows the negative distributions. For comparison, the sentiment analysis distributions of the previously classified test dataset based on human-categorized reviews are presented in Figure 8. Figure 8a shows the category-based positive review distributions, and Figure 8b shows the negative distributions.

Figure 6

Category-based ROC curves of the LR model.

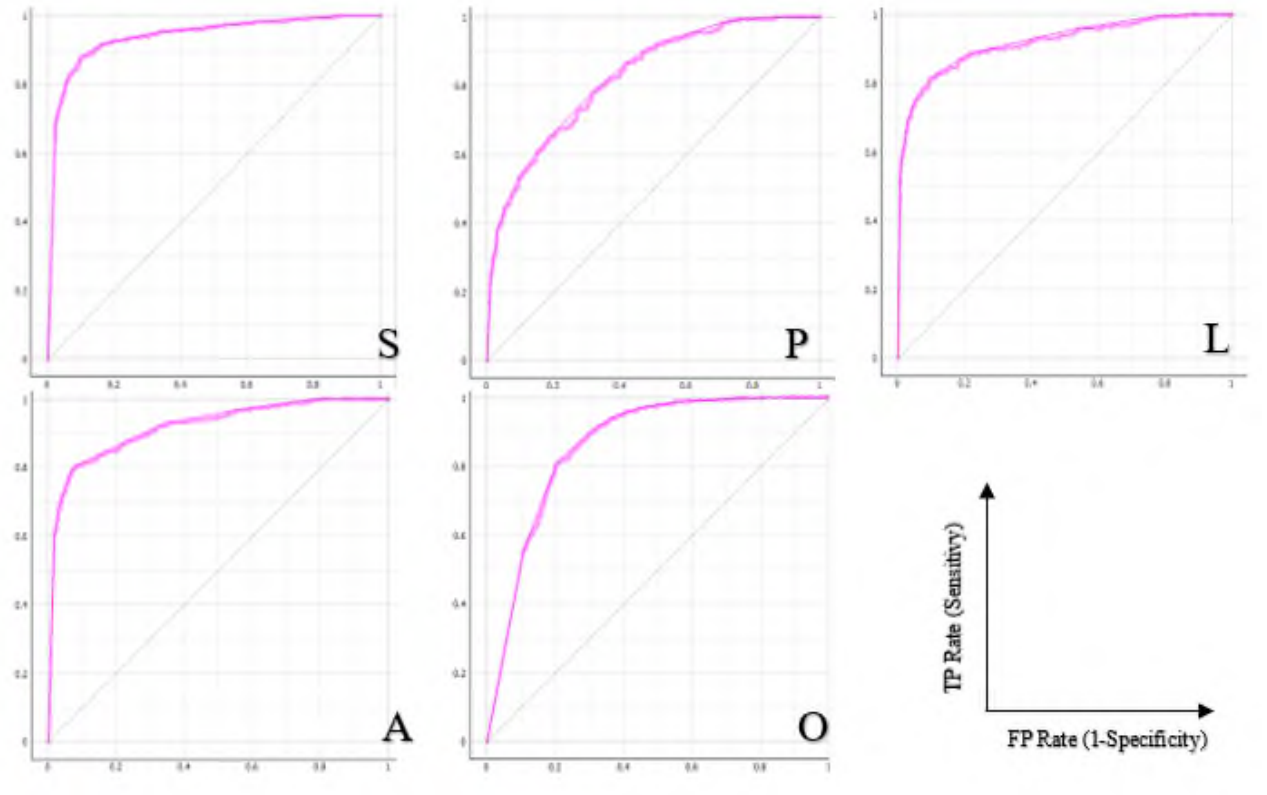


Figure 7

Category-based sentiment distributions of LR-based classification.

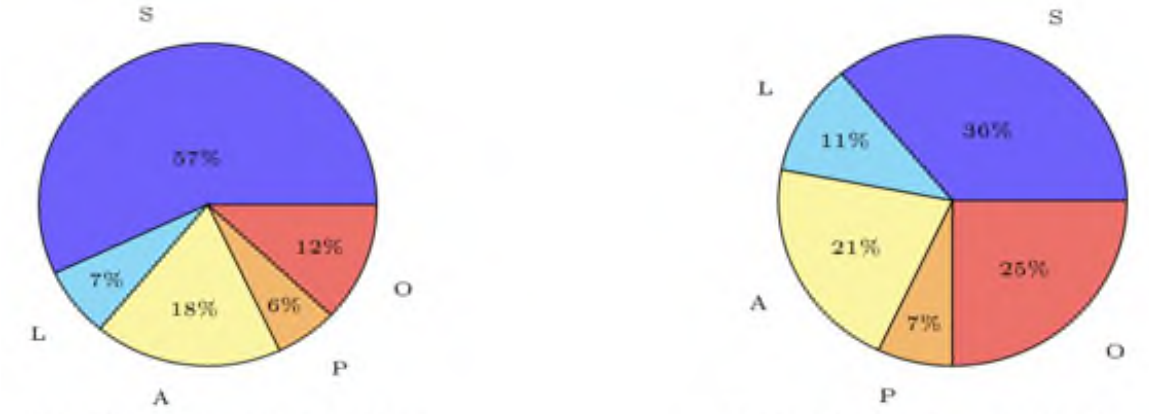
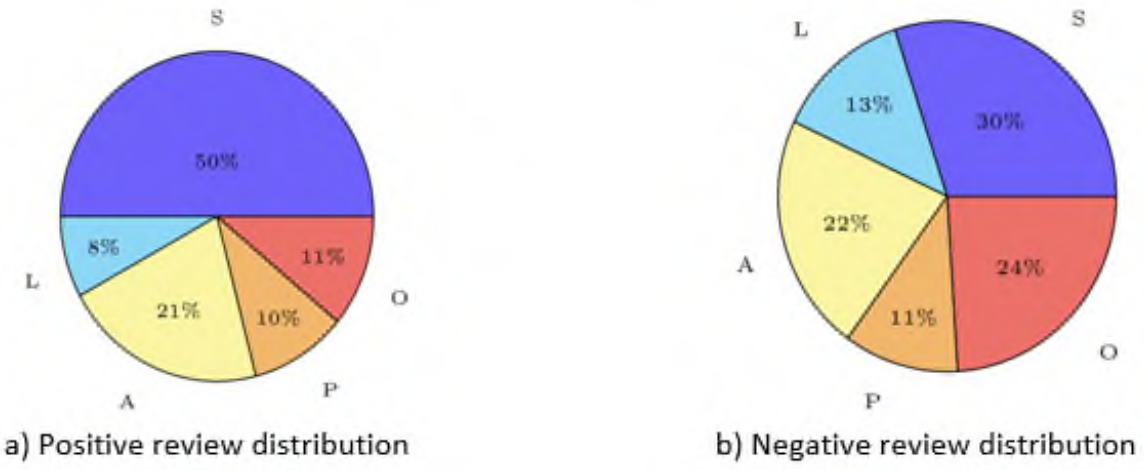


Figure 8

Category-based sentiment distributions of human-based classification.



According to the analysis results, comments shared on X about Trendyol have been examined separately, and customer satisfaction has been created for each category. This process attempts to determine the highest and lowest customer satisfaction scores and examine the data. In this context, comments in categories with low customer satisfaction can lead to customer loss. The business can prevent this ongoing loss by implementing improvement and transformation processes within relevant departments, focusing on areas highlighted by negative feedback to increase satisfaction scores.

The results offer several practical benefits for e-commerce platforms. By categorizing user feedback into predefined categories, businesses can better understand the specific aspects of their operations that require improvement. The integration of category-based sentiment analysis also enables businesses to tailor their marketing strategies. By isolating positive sentiment related to specific categories, such as advertising, businesses can identify and replicate successful campaign elements in future promotions. Conversely, identifying dissatisfaction in categories like product or support allows for targeted actions to address customer concerns and enhance brand reputation. Operationally, this approach supports more informed decision-making processes. These insights can also help refine post-sales service strategies, ensuring a seamless customer experience and increasing overall profitability. By offering actionable insights through category-based sentiment analysis, this study bridges the gap between raw data and strategic decision-making. It enables e-commerce platforms to adopt data-driven approaches to enhance their services, customer engagement, and market positioning.

The experimental results demonstrate the effectiveness of machine learning models in classifying e-commerce reviews. In addition, integrating BERTurk for sentiment analysis provided deeper insights into the sentiment distribution in the classified reviews. These findings underscore the potential of combining machine learning algorithms and advanced natural language processing techniques to analyze consumer sentiment on e-commerce platforms.

Discussion and Conclusion

In conclusion, this study investigated the potential of machine learning and sentiment analysis techniques to extract valuable insights from online customer reviews. As businesses increasingly recognize

the importance of understanding and responding to customer feedback, the ability to analyze such data effectively has become essential for maintaining competitiveness.

The study successfully employed eight machine learning algorithms for classification tasks using a dataset of 5,806 Trendyol user reviews categorized into five distinct classes. In addition, sentiment analysis using the pretrained BERTurk model provided valuable information about user satisfaction and dissatisfaction levels. This combined approach highlights the effectiveness of integrating machine learning and sentiment analysis to capture a comprehensive assessment profile of social media users about e-commerce platforms. By analyzing user reviews categorized by specific aspects like product, support, advertisement, logistics, and off-topic comments, this study offers a multi-faceted perspective on social media feedback. These insights can help businesses improve customer service, identify areas for improvement, and optimize marketing strategies.

Using traditional machine learning models for multi-category classification and integrating a pretrained BERTurk model for sentiment analysis makes a unique contribution to e-commerce text analysis. Although our dataset from Trendyol presents unique characteristics, such as five distinct thematic categories, the consistency in model performance observed in this study suggests that combining traditional models for category classification with transformer models for sentiment analysis can be a practical framework for similar e-commerce applications. This integrated approach aligns with current research trends and may provide a versatile model structure that can be adapted to other regional or platform-specific datasets, further validating the utility of traditional and transformer-based models in multidimensional text analysis.

The findings demonstrate several practical implications. Categorizing user feedback by specific operational areas allows businesses to pinpoint areas for improvement, such as addressing negative sentiment in logistics or identifying successful advertising strategies. This approach supports strategic decision-making, enabling businesses to optimize resource allocation, refine marketing strategies, and enhance customer satisfaction. By offering a granular perspective on sentiment, this study bridges the gap between raw social media data and actionable business insights, helping e-commerce platforms improve customer engagement and market positioning.

Our study introduces a unique approach by combining traditional machine learning models for category classification with a pretrained BERTurk model for sentiment analysis; however, some limitations must be acknowledged. A larger dataset would provide a broader spectrum of language patterns, thereby enhancing model robustness and generalizability. The traditional machine learning models used for classification were not extensively optimized, suggesting opportunities for further improvement through hyperparameter tuning or ensemble techniques. Our study also relies on different methodologies for category classification and sentiment analysis, with traditional machine learning models and a transformer model. Although this layered approach allows us to use the strengths of both techniques, it may increase the complexity of the classification process. Future research could explore using a single, unified model to handle category and sentiment classification, such as fine-tuning transformer-based models to perform both tasks simultaneously. This approach could streamline the process and potentially yield improved classification accuracy. In addition, although BERTurk performed well on e-commerce-specific data, fine-tuning the model could improve its sensitivity to domain-specific nuances. Addressing these limitations in future studies could enhance our approach's accuracy, adaptability, and scalability, further validating its effectiveness across larger and more diverse datasets.

Future research could focus on integrating classical and advanced machine learning techniques to address data preprocessing and class imbalance challenges and extend the methodology to multilingual datasets. Incorporating advanced models, such as RoBERTa or multilingual BERT variants, may further enhance classification accuracy and the depth of sentiment analysis. In addition, tracking the evolution of user sentiment over time and using fine-grained sentiment analysis techniques can provide richer insights into consumer behavior across diverse markets. Applying this framework to a broader range of e-commerce platforms and product categories could uncover more nuanced feedback patterns, ultimately extending the utility and applicability of this approach to understanding customer preferences and improving e-commerce strategies.

Peer Review Externally peer-reviewed.
Author Contributions Conception/Design of Study- M.G., M.K.; Data Acquisition- M.G.; Data Analysis/Interpretation- M.G., M.K.; Drafting Manuscript- M.G., M.K.; Critical Revision of Manuscript- M.K.; Final Approval and Accountability- M.G., M.K.; Technical or Material Support- M.G., M.K.; Supervision- M.K.
Conflict of Interest The authors have no conflict of interest to declare
Grant Support The authors declared that this study has received no financial support.

Author Details
Meryem Gürbüz
[1] Sakarya University, Institution of Natural Sciences, Sakarya, Türkiye
0000-0003-1354-9951

Muhammed Kotan
[2] Sakarya University, Department of Information Systems Engineering, Sakarya, Türkiye
0000-0002-5218-8848 mkotan@sakarya.edu.tr

References

Arobi, R., Rijon, R. H., Haque, N., Safkat, T. I. M., & Safoan, S. (2022). Sentiment analysis on E-commerce based product reviews using machine learning algorithms (Doctoral dissertation, Brac University).

Anjaria, M., & Guddeti, R. M. R. (2014). Influence factor based opinion mining of Twitter data using supervised learning. *In 2014 sixth international conference on communication systems and networks (COMSNETS)* (pp. 1-8). IEEE.

Bayhaqy, A., Sfenrianto, S., Nainggolan, K., & Kaburuan, E. R. (2018, October). Sentiment analysis about E-commerce from tweets using decision tree, K-nearest neighbor, and naïve bayes. *In 2018 international conference on orange technologies (ICOT)* (pp. 1-6). IEEE.

Chaubey, G., Gavhane, P. R., Bisen, D., & Arjaria, S. K. (2023). Customer purchasing behavior prediction using machine learning classification techniques. *Journal of Ambient Intelligence and Humanized Computing,* 14(12), 16133-16157.

Chen, N. (2022). Research on E-Commerce Database Marketing Based on Machine Learning Algorithm. *Computational Intelligence and Neuroscience,* 2022(1), 7973446.

Demircan, M., Seller, A., Abut, F., & Akay, M. F. (2021). Developing Turkish sentiment analysis models using machine learning and e-commerce data. *International Journal of Cognitive Computing in Engineering, 2,* 202-207.

Deniz, E., Erbay, H., & Coşar, M. (2022). Multi-label classification of e-commerce customer reviews via machine learning. *Axioms, 11*(9), 436.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint* arXiv:1810.04805.

Dharrao, D., Deokate, S., Bongale, A. M., & Urolagin, S. (2023, March). E-commerce product review classification based on supervised machine learning techniques. *In 2023 9th International Conference on Advanced Computing and Communication Systems (ICACCS)* (Vol. 1, pp. 1934-1939). IEEE.

George, A. S., & Baskar, T. (2024). Leveraging Big Data and Sentiment Analysis for Actionable Insights: A Review of Data Mining Approaches for Social Media. *Partners Universal International Innovation Journal,* 2(4), 39-59.

Grandi, U., Loreggia, A., Rossi, F., & Saraswat, V. (2015). A borda count for collective sentiment analysis. *Annals of Mathematics and Artificial Intelligence, 77*(3-4), 281-302. https://doi.org/10.1007/s10472-015-9488-0

Gound, R. S., Tikone, P. V., Suryawanshi, S. S., & Nagpal, D. (2018). Twitter Data Sentiment Analysis and Visualization. *International Journal of Computer Applications,* 975, 8887.

Guven, Z. A. (2021a). The effect of bert, electra and albert language models on sentiment analysis for turkish product reviews. *In 2021 6th international conference on computer science and engineering (UBMK)* (pp. 629-632). IEEE.

Guven, Z. A. (2021b). Comparison of BERT models and machine learning methods for sentiment analysis on Turkish tweets. *In 2021 6th International Conference on Computer Science and Engineering (UBMK)* (pp. 98-101). IEEE.

Haque, T. U., Saber, N. N., & Shah, F. M. (2018, May). Sentiment analysis on large scale Amazon product reviews. *In 2018 IEEE international conference on innovative research and development (ICIRD)* (pp. 1-6). IEEE.

He, H., Zhou, G., & Zhao, S. (2022). Exploring e-commerce product experience based on fusion sentiment analysis method. *IEEE Access*, 10, 110248-110260.

Huang, H., Asemi, A., & Mustafa, M. B. (2023). Sentiment analysis in e-commerce platforms: A review of current techniques and future directions. *IEEE Access*.

Jabbar, J., Urooj, I., JunSheng, W., & Azeem, N. (2019). Real-time sentiment analysis on E-commerce application. *In 2019 IEEE 16th international conference on networking, sensing and control (ICNSC)* (pp. 391-396). IEEE.

Koksal, D., Alacan, M. M., Olgun, E., & Sakar, C. O. (2022). Natural Language Processing-Based Product Category Classification Model for E-Commerce. *In 2022 30th Signal Processing and Communications Applications Conference (SIU)* (pp. 1-4). IEEE.

Köksal, A., & Özgür, A. (2021). Twitter dataset and evaluation of transformers for Turkish sentiment analysis. *In 2021 29th Signal Processing and Communications Applications Conference (SIU)* (pp. 1-4). IEEE.

Kotan, M. (2022). Duygu Analizi ve Dijital Dönüşüm Üzerine Etkileri. *In Book: Dijital Etkileşimler ve Sektörel Yansımaları II*. Gazi Kitapevi.

Kubrusly, J., Neves, A. L., & Marques, T. L. (2022). A statistical analysis of textual e-commerce reviews using tree-based methods. *Open Journal of Statistics,* 12(3), 357-372.

Kumar, K. S., Desai, J., & Majumdar, J. (2016, December). Opinion mining and sentiment analysis on online customer review. *In 2016 IEEE international conference on computational intelligence and computing research (ICCIC)* (pp. 1-4). IEEE.

Marong, M., Batcha, N. K., & Mafas, R. (2020). Sentiment Analysis in E-Commerce: A Review on The Techniques and Algorithms. *Journal of Applied Technology and Innovation,* 4(1), 6.

Polat, H., & Ağca, Y. (2022). Tripadvisor Kullanıcılarının Türkçe Ve İngilizce Yorumları Kapsamında Duygu Analizi Yöntemlerinin Karşılaştırmalı Analizi. *Abant sosyal bilimler dergisi*, 22(2), 901-916.

Poyraz, Y. Z., Tugcu, M., & Amasyali, M. F. (2022). Improving BERT Pre-training with Hard Negative Pairs. *In 2022 Innovations in Intelligent Systems and Applications Conference (ASYU)* (pp. 1-6). IEEE.

Rumelli, M., Akkuş, D., Kart, Ö., & Isik, Z. (2019, October). Sentiment analysis in Turkish text with machine learning algorithms. *In 2019 Innovations in intelligent systems and applications conference (ASYU)* (pp. 1-5). IEEE.

Qin, Z., Dong, K., & Xie, B. (2024). What affects customers' online shopping behavior, research that applied machine learning to Amazon product reviews. *Applied and Computational Engineering*, 76, 48-64.

Saleem, H., Muhammad, K. B., Nizamani, A. H., Saleem, S., & Aslam, A. M. (2019). Data science and machine learning approach to improve E-commerce sales performance on social web. *International Journal of Computer Science and Network Security (IJCSNS)*, 19.

Savci, P., & Das, B. (2023). Prediction of the customers' interests using sentiment analysis in e-commerce data for comparison of Arabic, English, and Turkish languages. *Journal of King Saud University-Computer and Information Sciences, 35*(3), 227-237.

Sayeed, M. S., Mohan, V., & Muthu, K. S. (2023). BERT: A Review of Applications in Sentiment Analysis. *HighTech and Innovation Journal,* 4(2), 453-462.

Schweter, S. (2020). Berturk—Bert models for turkish. *Zenodo.* https://doi.org/10.5281/zenodo.3770924

Shanto, S. S., Ahmed, Z., & Jony, A. I. (2023). Mining User Opinions: A Balanced Bangla Sentiment Analysis Dataset for E-Commerce. *Malaysian Journal of Science and Advanced Technology*, 272-279.

Shehu, H. A., & Tokat, S. (2020). A hybrid approach for the sentiment analysis of Turkish Twitter data. *In Artificial Intelligence and Applied Mathematics in Engineering Problems: Proceedings of the International Conference on Artificial Intelligence and Applied Mathematics in Engineering (ICAIAME 2019)* (pp. 182-190). Springer International Publishing.

Şimşek, İ., Kökçam, A. H., Demir, H. I., Erden, C. (2023). Sentiment Analysis of Twitter Data of Hepsiburada E-commerce Site Customers with Natural Language Processing. *In International Symposium on Intelligent Manufacturing and Service Systems* (pp. 567-578). Singapore: Springer Nature Singapore.

Tang, D., Qin, B., Liu, T., & Yang, Y. (2015, June). User modeling with neural network for review rating prediction. *In Twenty-fourth international joint conference on artificial intelligence*.

Taşcı, M. E., Rasheed, J., & Özkul, T. (2024, January). Sentiment Analysis on Reviews of Amazon Products Using Different Machine Learning Algorithms. *In International Conference on Forthcoming Networks and Sustainability in the AIoT Era* (pp. 318-327). Cham: Springer Nature Switzerland.

Stephenie, Warsito, B., & Prahutama, A. (2020). Sentiment analysis on tokopedia product online reviews using random forest method. *In E3S Web of Conferences* (Vol. 202, p. 16006). EDP Sciences.

X Help Center. (2024). About different types of posts. Retrieved from https://help.twitter.com/en/using-x/types-of-posts

Xu, F., Pan, Z., & Xia, R. (2020). E-commerce product review sentiment classification based on a naïve Bayes continuous learning framework. *Information Processing & Management*, 57(5), 102221.

Yang, L., Li, Y., Wang, J., & Sherratt, R. S. (2020). Sentiment analysis for E-commerce product reviews in Chinese based on sentiment lexicon and deep learning. *IEEE Access, 8*, 23522-23530.

Yildırım, S. (2024). Fine-tuning Transformer-based Encoder for Turkish Language Understanding Tasks. *arXiv preprint* arXiv:2401.17396.

Yıldırım, S., & Asgari-Chenaghlu, M. (2021). Mastering Transformers: Build state-of-the-art models from scratch with advanced natural language processing techniques. Packt Publishing Ltd.

Yıldız, H. (2020). The Role Of Machine Learning Algorithms In Determining Product Sales In E-Commerce: A Case Study For Turkey. *Current and Historical Debates in Social Sciences: Field Studies and Analysis*, 175.

Zada, A. J. J., & Albayrak, A. (2023). Duygu Analizi ve Topluluk Öğrenmesi Yaklaşımları ile Kullanıcı Yorumlarının Analizi. *Düzce Üniversitesi Bilim ve Teknoloji Dergisi*, 11(4), 1725-1732.

İSTANBUL UNIVERSITY PRESS

Acta Infologica,
ACIN 2025, 9 (1): 19–33

https://doi.org/10.26650/acin.1486319
Submitted: 18.05.2024
Revision Requested: 20.08.2024
Last Revision Received: 16.09.2024
Accepted: 31.12.2024
Published Online 03.02.2025

Acta Infologica

Research Article

Open Access

DFCEA: Deep Forest Classifier-Based Employee Attrition Prediction

Mahmut Tokmak [1]

[1] Burdur Mehmet Akif Ersoy University, Bucak Zeliha Tolunay School of Applied Technology and Management, Department of Management Information Systems, Burdur, Türkiye

Abstract For companies, employee attrition is an important issue because human resources are the most important resources of a company. In companies, employee attrition can have different causes. However, human resource managers must recognize employee attrition indicators in the early stages. Employee attrition can lead to organizational losses for various reasons, such as interruption of work, interruption of tasks that need to be performed, the cost of re-employment and retraining, and the risk of information leakage. Therefore, in this study, DFCEA: Deep Forest Classifier-Based Employee Attrition prediction model is proposed to predict employee attrition. Thus, this study aimed to help company managers take measures to prevent the loss of human resources. The IBM HR Analytics Employee Attrition & Performance dataset was used in this study. The dataset was subjected to data cleaning, data encoding, data normalization, and data balancing preprocessing. The model was then trained and tested using the Deep Forest algorithm. With the proposed method, 98.8% accuracy and 98.8% f1 score were obtained. The obtained performance metrics are compared with known machine learning methods and other studies, and the performance power of the proposed method is demonstrated. The results demonstrate that the proposed DFCEA framework is highly effective in predicting employee attrition. Therefore, the framework presented in this study can help researchers, organization leaders, and human resource professionals predict employee attrition and contribute to the development of new prediction models.

Keywords Artificial Intelligence · Machine Learning · Deep Forest · Employee Attrition

Citation: Tokmak, M. (2025). DFCEA: Deep forest classifier-based employee attrition prediction. *Acta Infologica*, *9*(1), 19-33. https://doi.org/10.26650/acin.1486319

This work is licensed under Creative Commons Attribution-NonCommercial 4.0 International License.

© 2025. Tokmak, M.

Corresponding author: Mahmut Tokmak mahmuttokmak@mehmetakif.edu.tr

Introduction

In the current, intensely competitive business landscape, enterprises encounter a perpetual hurdle in attracting and retaining skilled personnel. However, many organizations deal with the issue of employee attrition (Wardhani & Lhaksmana, 2022). Employee attrition refers to the phenomenon of employees voluntarily or involuntarily leaving an organization, decreasing the number of employees. It has become a significant concern for organizations as the loss of skilled and experienced employees can negatively impact productivity, increase costs, and diminish competitive advantage (Alharbi et al., 2023; Raza, Munir, Almutairi, Younas, & Fareed, 2022).

Employee attrition has various effects on companies. The most important of these effects is the departure of qualified and experienced employees. This can lead to a loss of organizational knowledge and experience, disruption of customer relationships, and jeopardize the organization's overall ability to compete. In addition, employee attrition can have a negative impact on morale, motivation, and team spirit among existing staff. In addition, employee attrition can have a significant financial impact on businesses due to the costs associated with recruitment, training, and new hires (Alshiddy & Aljaber, 2023; Frye, Boomhower, Smith, Vitovsky, & Fabricant, 2018; Jain, Jain, & Pamula, 2020).

Machine Learning (ML) technologies are becoming increasingly common in modern technologies. Due to recent advances in machine intelligence, it is now possible to gain insights to guide companies using employee attrition models. Using datasets, organizations can use ML algorithms to find predictive indicators and underlying trends related to employee attrition. By analyzing various data sources, such as survey responses, employee demographics, performance metrics, and sentiment analysis, these algorithms can identify hidden patterns and signs of employee attrition risk (Chung, Yun, Lee, & Jeon, 2023).

To predict the risk of attrition among current employees, ML models can be trained on historical employee data. Examples of these models include Logistic Regression (LR), Naïve Bayes (NB), Decision Trees (DT), Extra Trees (ET), Random Forests (RF), Support Vector Machines (SVM), Artificial Neural Networks (ANN), and deep neural networks (DNN). Using these models, organizations can proactively identify individuals who are at higher risk of leaving and implement different strategies for employee attrition (Al-Darraji et al., 2021; Chaurasia, Kadam, Bhagat, Gauda, & Shingane, 2023; Chung et al., 2023; Fukui et al., 2023).

In addition, ML algorithms can provide valuable insights into the factors that drive employee attrition. This information can guide the development of tailored interventions, such as improving compensation, enhancing career development programs, or implementing work-life balance initiatives, to address the identified concerns and increase employee attrition (Chaurasia et al., 2023; Meraliyev, Karabayeva, Altynbekova, & Nematov, 2023).

In addition to predicting and understanding attrition, ML by analyzing employee attrition data can identify the characteristics and traits of employees who have demonstrated long-term commitment to the organization. This knowledge can help organizations identify candidates who are more likely to stay and thrive within the company (Chung et al., 2023).

One machine learning method, the Deep Forest algorithm, is inspired by deep neural networks (Zhou & Feng, 2019). DNNs have been used to solve many problems and have also been used in employee attrition prediction (Al-Darraji et al., 2021; Metlek, 2021). However, using DNNs has some disadvantages. First, DNNs are often difficult to train because they have too many hyperparameters, and the hardware capacity required to solve these problems increases as the model becomes deeper. Second, a large amount of data is typically

required to train DNNs. Depending on the topic under investigation, it is sometimes difficult to obtain the required data. For these reasons, Zhou and Feng developed a deep forest algorithm that consists of a layered structure, similar to deep neural networks. The deep forest model requires fewer hyperparameters than DNNs and does not incur high computational costs. There is also no backpropagation or gradient adjustment during model training. Compared to DNNs, it does not require large amounts of data (Yao et al., 2023).

This study focuses on employee attrition prediction using the Deep Forest algorithm. The results obtained using the ML methods, which are referred to as classical ML in the literature, are compared with the results obtained using the Deep forest algorithm. To the best of our knowledge, this is the first study to predict employee attrition using the Deep Forest algorithm. The results demonstrate the success of the proposed DFCEA framework. Accordingly, the proposed DFCEA framework can be used to effectively manage employee attrition, which has become a critical concern for organizational leaders and human resource professionals.

Literature Review

Employee attrition is a common problem faced by organizations, and ML methods have been used to predict and prevent it. Studies have been conducted using the IBM HR Analytics Employee Attrition & Performance (IBM HR) dataset or other datasets to develop ML models for predicting employee attrition.

Jain et al. (2020) used an ML approach to enhance different retention strategies for targeted employees. The authors analyzed 1000 samples from a dataset comprising 14000 records and 10 attributes obtained from Kaggle. They evaluated the performance of binary classifiers such as SVM, DT, and RF, using various features to predict employee attrition in different departments. They obtained the most successful metrics with the proposed RF algorithm compared to other methods.

Mendes and Jesus (2021) explored various ML algorithms for example NB, RF, SVM, K-Nearest Neighbor (KNN), LR, and linear discriminant analysis (LDA) etc., to predict employee attrition. They used the IBM HR dataset and the ADASYN algorithm to eliminate data imbalance, and the best performance was achieved with the RF classifier.

Qutub et al. (2021) presented the development of multiple ML models to predict employee attrition. The IBM HR dataset was used to train and evaluate these models, including the DT, RF, LR, Adaboost Model, and Gradient Boosting Classifier (GBC). After training and evaluating the five basic models, the authors created three ensemble models using multiple combinations of the five basic models. Finally, they compared the results obtained from the considered models. They demonstrated that the AdaBoost algorithm achieved the highest accuracy among the evaluated models.

Usha and Balaji (2021) conducted a comparative study on ML algorithms for employee attrition prediction. To conduct the study, the researchers created a dataset using a questionnaire that includes variables that affect an employee's decision to leave an organization. They collected data from an organization operating in the IT sector. The dataset was used to train and evaluate the NB, DT, J48, RF, and K-means scores. NB performed better than the other algorithms in terms of accuracy and time, whereas J48 was better in terms of building time. They noted that clustering performance achieved lower accuracy than classification algorithms.

Mansor et al. (2021) conducted a comparative analysis of three ML classifiers: DT, SVM, and ANN, using the IBM HR dataset. The evaluation was based on multiple metrics, including accuracy rate, error rate, root mean square error (RMSE), receiver operating characteristic (ROC) curve, and model building time. Their findings

indicate that the optimized SVM model exhibited superior performance in predicting employee attrition, outperforming both ANN and DT classifiers in terms of accuracy.

Al-Darraji et al. (2021) used DNN to predict employee attrition. The authors used the IBM HR and ADASYN dataset resampling techniques. The authors conducted experiments on both balanced and imbalanced datasets to assess the robustness and effectiveness of the classifiers across different data distributions.

Guerranti and Dimitri (2022) employed various ML methods to analyze employee attrition, including LR, Classification Trees, RF, NB, and ANN. The authors used the IBM HR dataset for prediction. They also used the ADASYN algorithm to eliminate data imbalance. Among the proposed methods, LR demonstrated the best.

Wardhani and Lhaksmana (2022) used an LR method to predict employee attrition in a company and evaluated the performance of three feature selection methods, namely, information gain, selecting the best, and recursive feature elimination (RFE). They used the IBM HR dataset. The three data preprocessing steps are data encoding using a label encoder, data scaling using a min-max scaler, and data sampling using the SMOTE-ENN method to handle unbalanced class values. The goal of this study was to demonstrate the performance of feature selection methods. The proposed RFE feature selection method outperformed the other feature selection methods. However, they stated that the performance of the LR classifier was better without feature selection.

Raza et al. (2022) used four ML techniques—ET, SVM, LR, and DTC—to predict employee attrition. The authors used the IBM HR dataset for their research. They examined the employee attrition dataset and the factors causing employee attrition using Employee Exploratory Data Analysis (EEDA). Because of their study, they achieved the highest accuracy score with the ET Classifier.

Shaik et al. (2023) discussed the problem of employee attrition in businesses and the use of ML to predict employee behaviors. The authors proposed the use of the IBM HR dataset to train their ML model and compared three different gradient boosting algorithms, namely XGBoost, CatBoost, and LightGBM.

Kakulapati and Subhani (2023) focused on the predictive analytics of employee attrition using K-Fold methodologies. The authors used the IBM HR dataset for their predictions. They used ML approaches, such as CatBoost, LightGBM, and XGBoost, to compare and identify which employees were more likely to leave their organizations. They achieved the best accuracy with the XGBoost algorithm.

Chung et al. (2023) introduced a predictive employee attrition model that utilizes stacking ensemble learning. Employing the IBM HR dataset, they constructed and assessed eight predictive models, encompassing LR, RF, XGBoost, SVM, ANN, and ensemble models. Their primary objective was to enhance the accuracy of employee attrition prediction using a novel ensemble-based approach. The results of the model development and analysis indicated that the stacking-based ensemble model exhibited superior performance. Furthermore, their study highlighted environmental satisfaction, overtime, and relationship satisfaction as primary contributing factors to employee attrition.

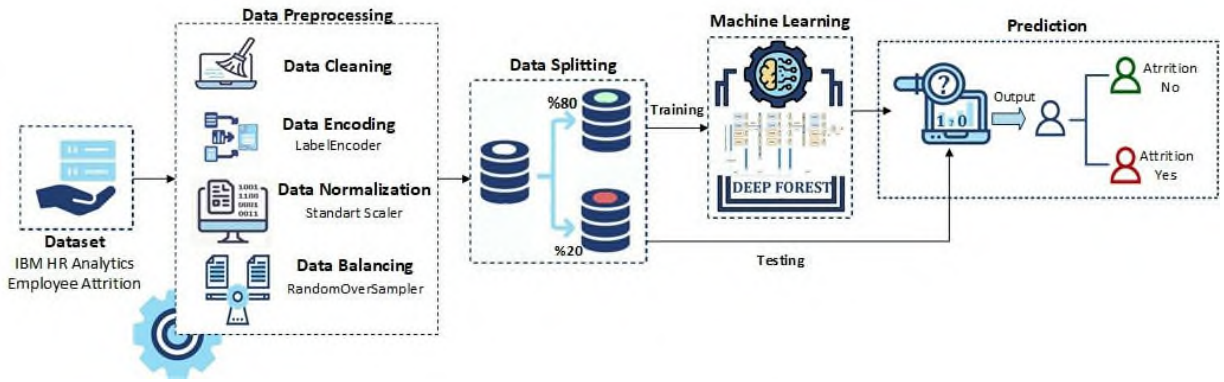
Chaurasia et al. (2023) proposed an employee attrition system that utilizes an ANN to predict the probability of an employee leaving an organization and develop targeted retention strategies. They used their own dataset in their study. Their proposed system demonstrated significant improvement, achieving an accuracy of 96%, which surpasses the performance of the existing system that uses DNN with an accuracy of 89.11%.

Material and Method

In this section, the dataset used in the study, preprocessing steps, methods, and evaluation metrics are explained. A flow chart of the study is presented in Figure 1.

Figure 1

The proposed DFCEA model



Dataset

In the realm of human resources (HR), understanding the factors that influence employee attrition and performance is of paramount importance for organizations seeking to optimize their workforce. One valuable resource in this pursuit is the IBM HR dataset provided by Kaggle, which is a comprehensive collection of employee data that provides insights into the factors contributing to attrition and performance. The IBM HR dataset is a curated collection of employee information that encompasses several attributes relevant to HR analytics. The proposed model comprises anonymized data derived from a fictional company and provides a realistic representation of the challenges faced by HR professionals when managing employee attrition and performance. The proposed dataset includes a diverse set of variables that capture essential aspects of employees' personal information, employment details, performance metrics, compensation, leaves, and training. In the dataset, 9 attributes are categorical and 26 attributes are numeric. The main attributes of the dataset are presented in Table 1.

Table 1

IBM HR dataset attributes

Attribute	Type	Attribute	Type
Age	numeric	MonthlyIncome	numeric
Attrition	categorical	MonthlyRate	numeric
BusinessTravel	categorical	NumCompaniesWorked	numeric
DailyRate	numeric	Over18	categorical
Department	categorical	OverTime	categorical
DistanceFromHome	numeric	PercentSalaryHike	numeric
Education	numeric	PerformanceRating	numeric
EducationField	categorical	RelationshipSatisfaction	numeric
EmployeeCount	numeric	StandardHours	numeric
EmployeeNumber	numeric	StockOptionLevel	numeric
EnvironmentSatisfaction	numeric	TotalWorkingYears	numeric

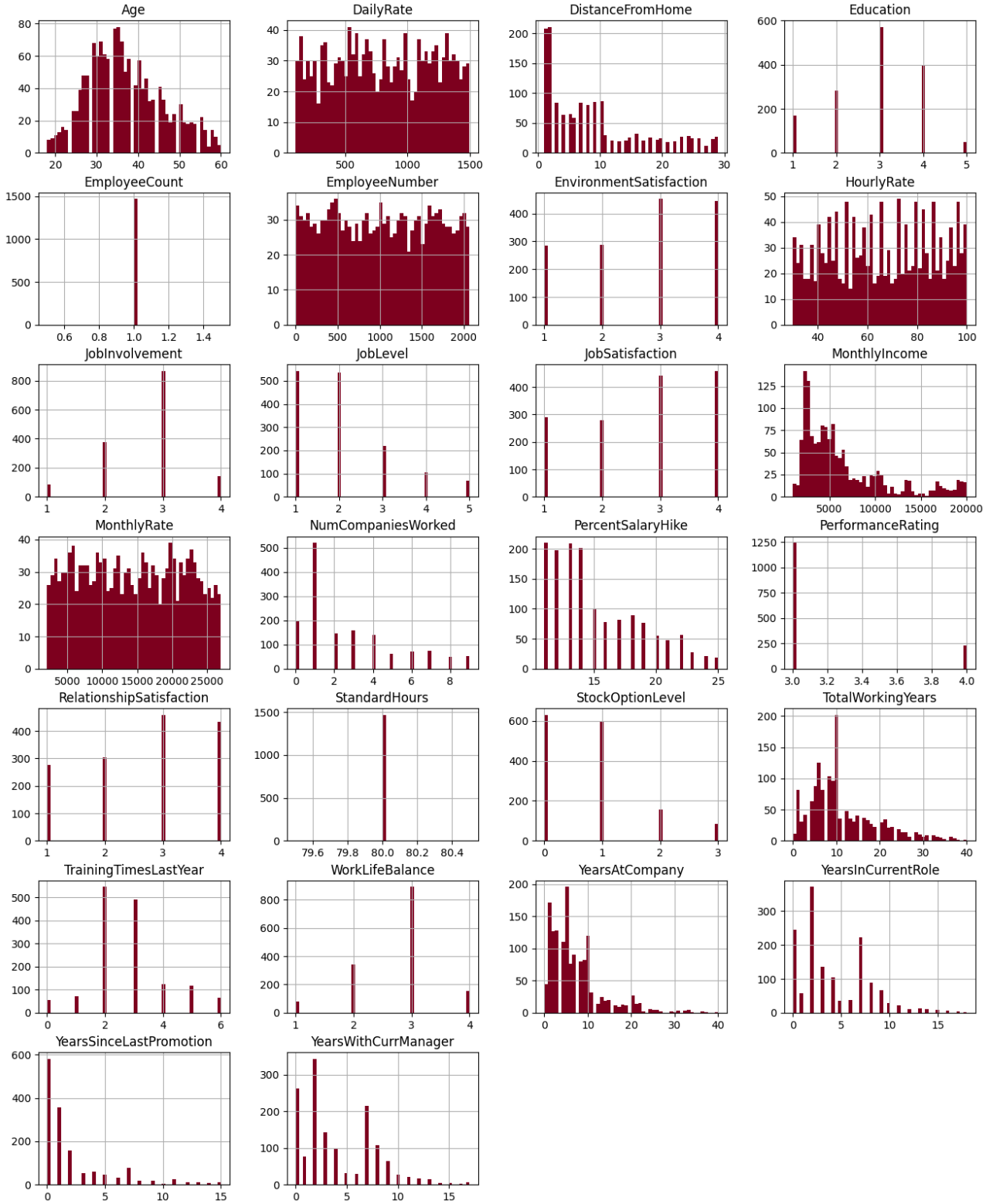
Attribute	Type	Attribute	Type
Gender	categorical	TrainingTimesLastYear	numeric
HourlyRate	numeric	WorkLifeBalance	numeric
JobInvolvement	numeric	YearsAtCompany	numeric
JobLevel	numeric	YearsInCurrentRole	numeric
JobRole	categorical	YearsSinceLastPromotion	numeric
JobSatisfaction	numeric	YearsWithCurrManager	numeric
MaritalStatus	categorical		

Data Preprocessing

Preprocessing began by deleting the EmployeeCount, StandardHours, and Over18 attributes, where all fields in the dataset have the same data (Figure 2). For example, the EmployeeCount attribute has a value of 1 in all rows, the StandardHours attribute has a value of 80 in all rows, and the Over18 attribute is labeled as Y in all rows. Then, the EmployeeNumber attribute, which comprises the numbers given to identify each employee, was deleted. The categorical variables were then transformed into a form suitable for ML algorithms to be used in predictions.

Figure 2

IBM HR dataset attribute frequency.

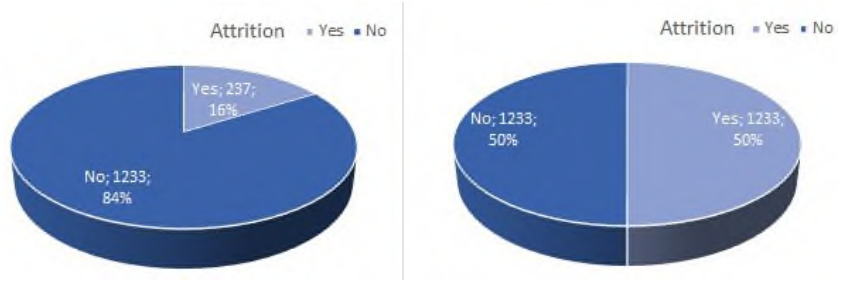


The raw data were then transformed by applying the StandardScaler normalization process. Normalization is a method that rescales or transforms raw data such that its features make uniform contributions. This method is used to eliminate dominant feature and outlier problems that negatively affect the learning process of ML algorithms (Singh & Singh, 2020).

When the distribution of the target class was analyzed, data imbalance was identified. The RandomOverSampler method was used to eliminate data imbalance. The numerical distribution of the target class and balanced version is shown in Figure 3.

Figure 3

Original data and balanced data.



The RandomOverSampler data resampling method attempts to balance the class distribution by randomly duplicating instances from the minority class. The number of instances in the minority class is increased until it reaches the same number as the majority class or a user-defined ratio. This process is performed randomly, which means the duplicated instances are selected randomly from existing minority class instances (Kamalov, Leung, & Cherukuri, 2023; Ustyannie & Suprapto, 2020). The proposed method reduces the risk of model overfitting. Another advantage of this method is that it combines the generation of synthetic data with community ideas (Menardi & Torelli, 2014).

The dataset comprising a total of 2466 samples obtained after data preprocessing was divided into 80% training and 20% testing. The dataset was split for training and testing using the train_test_split method in the scikit-learn library. The train_test_split method randomly splits the training and testing data. The random_state parameter recommended by the library was set to 42 to guarantee randomness. Various experiments were conducted to maximize the performance of the ML models used. According to the characteristics of the models, the following parameters were set, and the models were run. The models and parameter settings are listed in Table 2.

Table 2

Model Hyperparameters

Model Name	Parameters
DT	max_depth=4
GBC	n_estimators = 20, learning_rate = 1.0, max_depth = 3, random_state = 0
SVM	kernel=poly, degree=3, gamma="auto", cache_size=30, C=1
RF	max_depth=6, n_estimators=100, max_features=1
XGBoost	default
ET	n_estimators=100
DeepForest	default

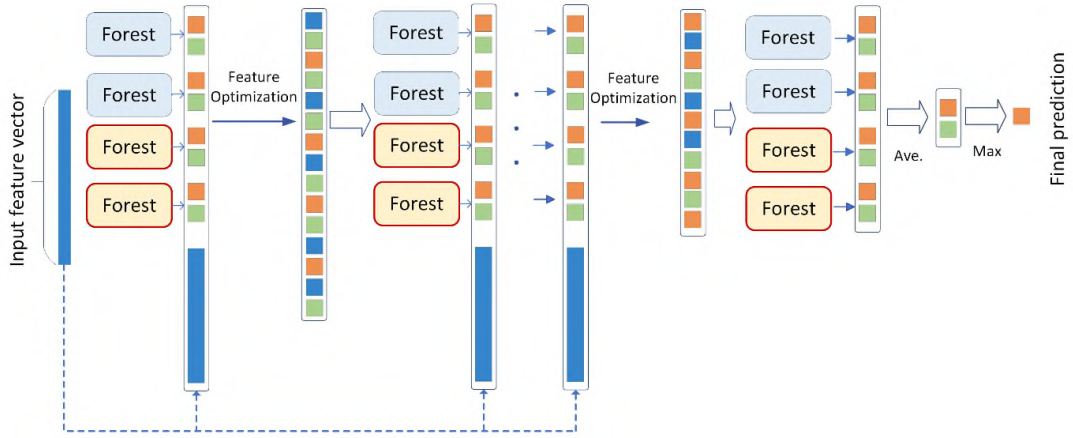
Deep Forest

Deep Forest is a layered architecture proposed by Zhou and Feng that offers a new representation learning approach. Zhou and Feng introduced the concept of a layered architecture called Deep Forest, which offers a novel approach to representative learning. In Deep Forest, features relevant to the target problem are learned at each layer of the cascading structure. The learned features serve as the input for the next layer, allowing information to flow throughout the architecture. To ensure diversity, the Deep Forest includes different types of forests. Diversity plays a crucial role in fostering robust and effective communities (Zhou & Feng, 2019). The overall structure of Deep Forest (Figure 4) comprises multiple levels in the cascade. Each level consists of two RF in gray and two completely RF in red.

Suppose there are two classes to be classified. Each forest in the Deep Forest architecture produces a two-dimensional vector representing the respective classes. These vectors are then combined to represent the original input (Guo, Liu, Li, & Shang, 2018). This unique approach enhances the Deep Forest learning capability, making it a promising technique for various ML tasks.

Figure 4

Deep Forest architecture.



Performance evaluation

The evaluation was performed by calculating accuracy, precision, recall, and F1-scores for the training and test performance of the established models. At the same time, the results obtained in the target class prediction were expressed both numerically and visually with the Confusion Matrix, which contains the numerical data of the real and predicted classes used to evaluate the performance of the classifiers in classification problems (Çavuşoğlu & Kaçar, 2019; Raza et al., 2022). The performance evaluation metrics used in this study are explained below:

True Positive (TP) denotes the number of correctly identified employee attrition. True negative (TN) is the number of correctly detected employee attritions. False Positive (FP) is the number of incorrectly detected employee attrition, which is actually no employee attrition. False Negative (FN) is the number of incorrectly detected no employee attrition, which is actually employee attrition.

Accuracy: Accuracy is calculated as the ratio of the correctly predicted areas in the model to the total dataset. Equation 1 shows the accuracy formula.

$$\text{Accuracy} = \frac{TP + TN}{TP + FN + FP + TN} \tag{1}$$

Precision: The accuracy of positive predictions is determined using the following precision formula. Equation 2 gives the precision formula.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

Sensitivity (Recall): The proportion of positive samples correctly detected by the model. Equation 3 gives the sensitivity formula.

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{3}$$

F1-Score: While calculating the F1 score value, a new value is calculated using the precision and sensitivity values together. Equation 4 gives the F1-score formula.

$$\text{F1-Score} = 2^{*}\frac{precision^{*}\ sensitivity}{precision + sensitivity} \tag{4}$$

Results

The training and testing of the classification models proposed in this study were performed using the Python programming language on the Google Colab Notebook platform. in This section, we present the results of the study. The performance metrics of the proposed DFCEA are summarized in Table 3. Additionally, to demonstrate the performance of the proposed DFCEA, the performance of the other ML methods used in this study is presented in Table 4.

Table 3
DFCEA performance metrics

	Precision	Recall	F1-score	Support
0 (Employee Attrition No)	0.981	0.996	0.988	257
1 (Employee Attrition Yes)	0.996	0.979	0.987	237
accuracy			0.988	494
Macro avg	0.988	0.988	0.988	494

Table 4
Comparison of Deep Forest and other ML algorithm performance metrics

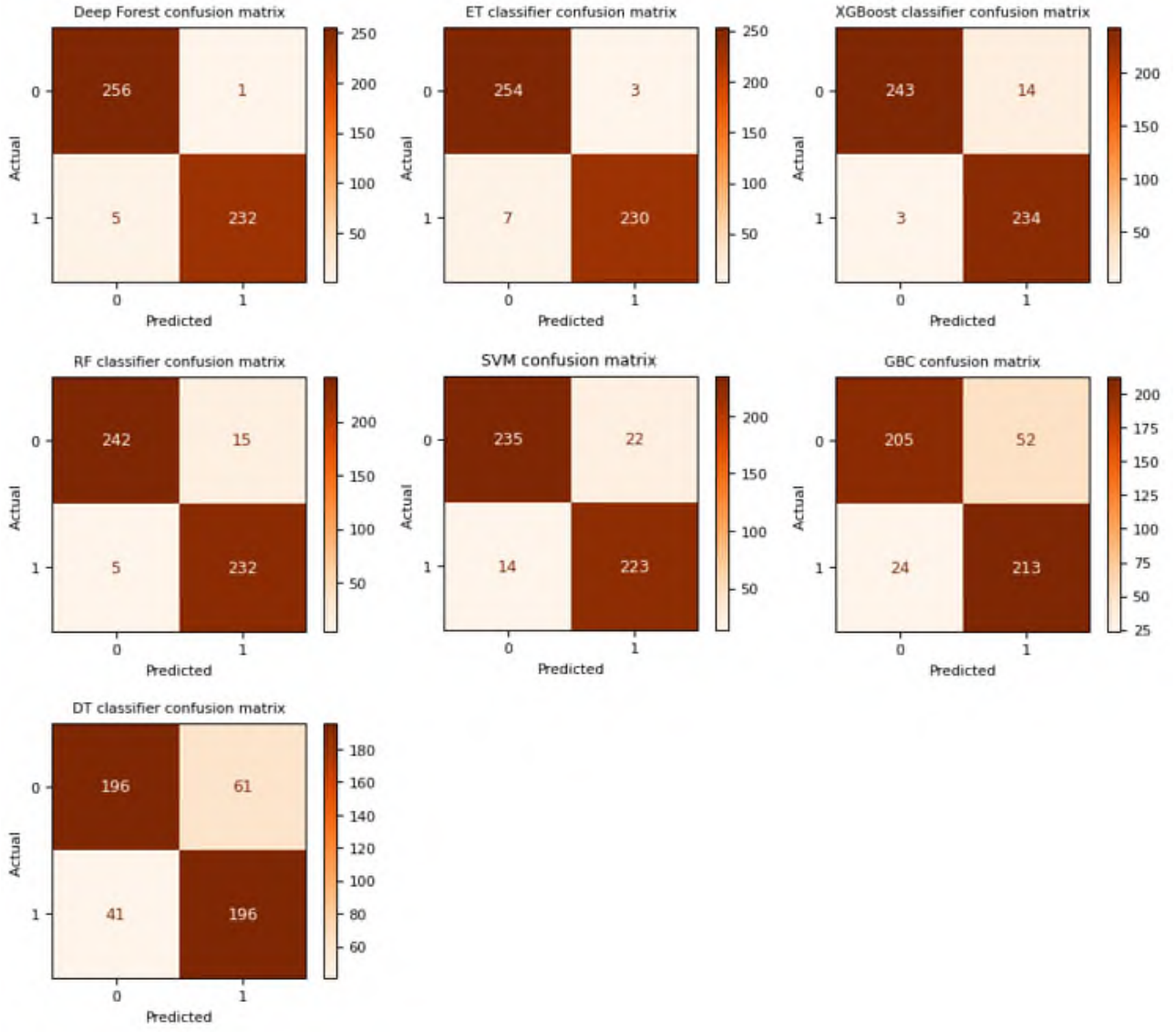
	Accuracy	Precision	Recall	F1-Score	Support
DT	0.794	0.795	0.795	0.794	494
Gboost	0.846	0.849	0.848	0.846	494
SVM	0.927	0.927	0.928	0.927	494
RF	0.960	0.960	0.960	0.959	494
XGBoost	0.966	0.966	0.966	0.966	494
ET	0.980	0.980	0.979	0.980	494
DFCEA	0.988	0.988	0.988	0.988	494

Figure 5 shows the confusion matrix of the all classifiers in this study. The best performer, DFCEA, correctly classified 488 of 494 samples. The 1 sample were identified as a false positive, and 5 samples were identified

as false negatives. The confusion matrix-validated accuracy score was 98.8%. The DT algorithm achieved the worst performance. The other methods used in this study achieved over 84% accuracy and another metrics.

Figure 5

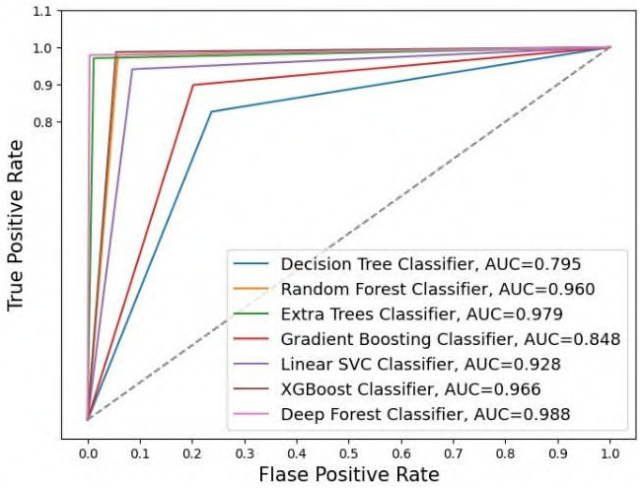
Confusion matrices of the considered methods.



The ROC curve analysis for the proposed technique is presented in Figure 6. A larger area under the curve indicates superior model validation performance in classification tasks. The results of the ROC analysis indicate that the proposed model achieved an accuracy score of 98.8%. These findings validate the effectiveness of my approach in predicting employee attrition.

Figure 6

ROC curve.



Discussion and Conclusion

Employee attrition is a multifaceted phenomenon. Factors such as career development opportunities, work-life balance, corporate culture, job security, job satisfaction, recognition and rewards, interpersonal relationships, and alignment of personal values with the company's mission and values influence employees' decisions to stay or leave a company. Organizations must address these various factors to create an environment that fosters long-term employee engagement. In addition to considering the factors affecting attrition in a broad context, predicting attrition in the early stages is important for an organization to take the necessary steps to retain employees. In this way, companies can take steps to prevent employee turnover. For this purpose, ML methods are used to predict employee attrition.

In this study, the HR data of IBM Analytics were used to construct an employee attrition prediction model. The preprocessing techniques of attribute deletion, data encoding, data normalization, and data balancing were applied to the dataset. The DT, GBC, SVM, RF XGBoost, ET, and DFCEA models were constructed.

The effectiveness of layer-by-layer information processing in DNN models has been demonstrated in previous studies. As models deepen, abstract and informative features emerge (Jiang, Nazarpour, & Dai, 2023). Therefore, this study focused on demonstrating the performance of the Deep Forest algorithm, which is inspired by the layered structure in DNNs, and compared it to other ML methods.

This study is the first to apply a Deep Forest model to predict employee attrition. The importance of this study lies in the high success rate of predicting employee attrition by proposing a Deep Forest model that includes different forest types. After model construction and analysis, the DFCEA model demonstrated the highest performance in terms of all metrics, such as accuracy 0.988, F1-score 0.988, precision 0.988, and recall, 0.988. These metrics demonstrate the performance of the Deep Forest cascade model. These metrics demonstrate the performance of the Deep Forest cascade model. In addition, the importance of data preprocessing methods is also highlighted. Using the same dataset used in this study, Mendes and Jesus (2021) achieved 93.6% accuracy with the RF classifier, whereas the proposed study achieved 96.2% accuracy with the RF classifier. Similarly, Raza et al. (2022) achieved 93% accuracy with the ET classifier, whereas this study achieved 98.4% accuracy with the ET classifier. This demonstrates that the data preprocessing techniques improved the prediction performance of the proposed framework.

Previous studies have estimated employee attrition using ML methods. Table 5 lists the classification accuracy scores of the models examined in the literature and the proposed model. In addition, the ML and balancing methods used in these studies are shown in the table. However, in some studies, performance metrics were found to be low due to the characteristics and limitations of the methods used. Studies have shown that when data balancing algorithms are used for the IBM HR dataset, such algorithms exhibit higher prediction performance. As shown in Table 4, meeting the distribution of models presented by previous studies predicting attrition using the same data usage as in this study, these features confirm that the performance of the proposed preprocessing methods and deep forest model is better from the perspective of accuracy, precision recall, and f1-score.

Table 5

Comparison of classification accuracy of existing models

Literature	ML Tecnique	Dataset	Dataset Balanced	Accuracy Score %
Mendes and Jesus (2021)	NB, RF, SVM, KNN, LR, LDA	IBM HR	ADASYN	93.6
Qutub et al. (2021)	DT, RF, Adaboost, Gradient Boosting Ensemble Model (DT+LR)	IBM HR		86.7
Usha and Balaji (2021)	NB, DT; J48, RF; KMeans	Own Dataset		85.98
Mansor et al. (2021)	DT, SVM, and ANN	IBM HR		88.87
Al-Darraji et al. (2021)	DNN	IBM HR	ADASYN	94.16
Guerranti and Dimitri (2022)	LR, CT, RF, NB, and ANN	IBM HR		87.96
Wardhani and Lhaksmana (2022)	LR	IBM HR	SMOTE-ENN	86.5
Raza et al. (2022)	ET, SVM, LR-DT,	IBM HR	SMOTE	93
Shaik et al. (2023)	CatBoost, LightGBM, and XGBoost	IBM HR		90.47
Chung et al. (2023)	Ensemble Model(Base: RF, ANN; Meta: LR)	IBM HR	SMOTE	97.4
Kakulapati and Subhani (2023)	CatBoost, LightGBM, and XGBoost	IBM HR		90.47
Chaurasia et al. (2023)	ANN, DNN	Own Dataset		96
DFCEA	Deep Forest	IBM HR	Random Over Sampler	98.88

In future, explainable artificial intelligence techniques will be developed. The effects of attributes on the model can be revealed using SHAP or LIME values to evaluate the prediction results in terms of interpretability, explainability, and comprehensibility.

Peer Review Externally peer-reviewed.
Conflict of Interest The author has no conflict of interest to declare
Grant Support The author declared that this study has received no financial support.

Author Details Mahmut Tokmak
[1] Burdur Mehmet Akif Ersoy University, Bucak Zeliha Tolunay School of Applied Technology and Management, Department of Management Information Systems, Burdur, Türkiye
0000-0003-0632-4308 mahmuttokmak@mehmetakif.edu.tr

References

Al-Darraji, S., Honi, D. G., Fallucchi, F., Abdulsada, A., Giuliano, R., and Abdulmalik, H. A. (2021). Employee attrition prediction using deep neural networks. *Computers*. doi: 10.3390/computers10110141

Alharbi, H., Alshammari, O., Jerbi, H., Simos, T. E., Katsikis, V. N., Mourtas, S. D., & Sahas, R. D. (2023). A fresnel cosine integral WASD neural network for the classification of employee attrition. *Mathematics*, *11*(6), 1506.

Alshiddy, M. S., and Aljaber, B. N. (2023). Employee attrition prediction using nested ensemble learning techniques. *International Journal of Advanced Computer Science and Applications*, *14*(7).

Çavuşoğlu, Ü., and Kaçar, S. (2019). Anormal trafik tespiti için veri madenciliği algoritmalarının performans analizi. *Academic Platform-Journal of Engineering and Science*, *7*(2), 205–216.

Chaurasia, A., Kadam, S., Bhagat, K., Gauda, S., and Shingane, P. (2023). Employee attrition prediction using artificial neural networks. *2023 4th International Conference for Emerging Technology (INCET)*, 1–6. IEEE.

Chung, D., Yun, J., Lee, J., & Jeon, Y. (2023). Predictive model of employee attrition based on stacking ensemble learning. *Expert Systems with Applications*, *215*, 119364.

da Silva Mendes, R. F., & de Jesus, J. V. R. (2021). *Exploração de modelos de aprendizado de máquina e seleção de atributos para employee attrition*. doi: 10.14210/cotb.v12.p267-272

Frye, A., Boomhower, C., Smith, M., Vitovsky, L., & Fabricant, S. (2018). Employee attrition: What makes an employee quit? *SMU Data Science Review*, *1*(1), 9.

Fukui, S., Wu, W., Greenfield, J., Salyers, M. P., Morse, G., Garabrant, J., Bass, E., Kyere, E., Dell, N. (2023). Machine learning with human resources data: Predicting turnover among community mental health center employees. *The Journal of Mental Health Policy and Economics*, *26*(2), 63–76.

Guerranti, F., & Dimitri, G. M. (2022). A comparison of machine learning approaches for predicting employee attrition. *Applied Sciences*. doi: 10.3390/app13010267

Guo, Y., Liu, S., Li, Z., & Shang, X. (2018). Bcdforest: A boosting cascade deep forest model towards the classification of cancer subtypes based on gene expression data. *BMC Bioinformatics*, *19*(5), 1–13.

Jain, P. K.; Jain, M.; Pamula, R. (2020). Explaining and predicting employees' attrition: A machine learning approach. *Sn Applied Sciences*. doi: 10.1007/s42452-020-2519-4

Jiang, X., Nazarpour, K., & Dai, C. (2023). Explainable and robust deep forests for EMG-Force modeling. *IEEE Journal of Biomedical and Health Informatics*.

Kakulapati, V., and Subhani, S. (2023). Predictive analytics of employee attrition using k-fold methodologies. *IJ Math. Sci. Comput.*, *1*, 23–36.

Kamalov, F., H.-H. Leung, and A. K. Cherukuri. (2023). Keep it simple: Random oversampling for imbalanced data. *2023 Advances in Science and Engineering Technology International Conferences (ASET)*, 1–4. IEEE.

Mansor, N., Sani, N. F. M., & Aliff, M. (2021). Machine learning for predicting employee attrition. *International Journal of Advanced Computer Science and Applications*. doi: 10.14569/ijacsa.2021.0121149

Menardi, G., & Torelli, N. (2014). Training and assessing classification rules with imbalanced data. *Data Mining and Knowledge Discovery*, *28*, 92–122.

Meraliyev, B., Karabayeva, A., Altynbekova, T., and Nematov, Y. (2023). Attrition rate measuring in human resource analytics using machine learning. *2023 17th International Conference on Electronics Computer and Computation (ICECCO)*, 1–6. IEEE.

Metlek, S. (2021). Disease Detection from Cassava Leaf Images with Deep Learning Methods in Web Environment. *International Journal of 3D Printing Technologies and Digital Industry*, *5*(3), 625–644.

Qutub, A., Al-Mehmadi, A. R., Al-Hssan, M., Aljohani, R., and Alghamdi, H. M. (2021). Predict employee attrition using machine learning and ensemble methods. *Int. J. Machine Learning Comput*. doi: 10.18178/ijmlc.2021.11.2.1022

Raza, A., Munir, K. M., Almutairi, M., Younas, F., & Fareed, M. M. S. (2022). Predict employee attrition using machine learning approaches. *Applied Sciences*. doi: 10.3390/app12136424

Shaik, S. (2023). Machine learning-based employee attrition predicting. *Asian Journal of Research in Computer Science*. doi: 10.9734/ajrcos/2023/v15i3323

Singh, D., & Singh, B. (2020). We investigate the impact of data normalization on classification performance. *Applied Soft Computing*, *97*, 105524.

Usha, P. M., and Balaji, N. V. (2021). A comparative study of machine learning algorithms for employee attrition prediction. *Iop Conference Series Materials Science and Engineering*. doi: 10.1088/1757-899x/1085/1/012029

Ustyannie, W., and Suprapto, S. (2020). Oversampling method to handling imbalanced datasets problem in binary logistic regression algorithm. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, *14*(1), 1–10.

Wardhani, F. H., and K. M. (2022). Predict employee attrition using logistic regression with feature selection. *Sinkron*. doi: 10.33395/sinkron.v7i4.11783

Yao, L., Li, W., Zhang, Y., Deng, J., Pang, Y., Huang, Y., Chung, C., Yu, J., Chiang, Y., Lee, T.-Y. (2023). Accelerating the Discovery of Anticancer Peptides through Deep Forest Architecture with Deep Graphical Representation. *International Journal of Molecular Sciences*, *24*(5), 4328. doi: 10.3390/ijms24054328

Zhou, Z. H., & Feng, J. (2019). Deep forest. *National Science Review*, *6*(1), 74–86.

İSTANBUL UNIVERSITY PRESS

Acta Infologica

Acta Infologica,
ACIN 2025, 9 (1): 34–54

https://doi.org/10.26650/acin.1474115
Submitted: 26.04.2024
Revision Requested: 04.11.2024
Last Revision Received: 23.11.2024
Accepted: 24.01.2025
Published Online 21.02.2025

Research Article

Open Access

Classification of Ventricular Septal Defect Disease Using Deep Learning

Kadir Barut [1], İhsan Pençe [2], Özlem Çetinkaya Bozkurt [3] & Melike Şişeci Çeşmeli [2]

[1] Burdur Mehmet Akif Ersoy University, Department of Management Information Systems, Burdur, Türkiye
[2] Burdur Mehmet Akif Ersoy University, Department of Software Engineering, Burdur, Türkiye
[3] Burdur Mehmet Akif Ersoy University, Department of Business Administration, Burdur, Türkiye

Abstract Ventricular Septal Defect (VSD) disease is the most prevalent type of congenital heart disease. VSD is a hole between the left and right ventricles of the heart structure. VSD disease accounts for approximately one-fifth of all congenital heart disease types. Therefore, accurate disease diagnosis is paramount in determining the most appropriate treatment methods. This study aims to classify VSD disease using the deep learning algorithms VGG16, ResNET50, and Inceptionv3 on Computed Tomography (CT) images and compare the pre-trained algorithms used. One of the reasons why imaging methods such as echocardiography are generally used to detect congenital heart diseases is that there are almost no CT datasets related to this disease. The dataset used in this study is the ImageCHD dataset, which comprises 3D CT scans encompassing 16 distinct types of congenital heart defects. Hyperparameter optimization was performed using the grid search method to enhance the model performance, identifying the VGG16 model as the most effective. The model demonstrated a very high classification accuracy of 99.99% in the training dataset and 99.94% in the test dataset. Gradient-weighted Class Activation Mapping was employed to enhance model explainability, providing visualizations of the regions most critical for the classification, thereby enabling medical professionals to validate AI-driven predictions. An optimized model that successfully classifies VSD using 3D CT image data has been introduced to the literature for the first time. Therefore, this study assumes greater significance in the existing literature and sets a benchmark for future studies.

Keywords Computed tomography · deep learning · medical image classification · transfer learning · Ventricular Septal Defect disease

Citation: Barut, K., Pençe, İ., Bozkurt, Ö. Ç. & Çeşmeli, M. Ş. (2025). Classification of ventricular septal defect disease using deep learning. *Acta Infologica*, *9*(1), 34-54. https://doi.org/10.26650/acin.1474115

This work is licensed under Creative Commons Attribution-NonCommercial 4.0 International License.
© 2025. Barut, K., Pençe, İ., Bozkurt, Ö. Ç. & Çeşmeli, M. Ş.

Corresponding author: İhsan Pençe ihsanpence@mehmetakif.edu.tr

Introduction

With the developing technology in the world, significant developments and innovations are also experienced in the health field. Mainly since the beginning of the 2010s, artificial intelligence technology has been aimed to be used in different fields, following the significant developments and successful results obtained in the field of artificial intelligence, and with the innovations brought by this technology, much more advanced systems have started to be used compared to the existing system. Health is one of these fields, but considering the importance of human health and treatment processes, it has become one of the leading fields. Especially regarding the correct diagnosis of the disease and applying the proper treatment to the patient, the correct detection of some diseases is a process that challenges healthcare professionals. When human health is taken into consideration, a wrong diagnosis or a wrong decision is of such severe importance that it can even lead to the death of a person.

Ventricular Septal Defect (VSD) disease is the most prevalent type of congenital heart disease. VSD is a hole between the left and right ventricles of the heart structure. VSD disease accounts for approximately one-fifth of all congenital heart disease types. Therefore, accurate disease diagnosis is paramount in determining the most appropriate treatment methods. This study aims to classify VSD disease using the deep learning algorithms VGG16, ResNET50, and Inceptionv3 on Computed Tomography (CT) images and compare the pre-trained algorithms used. One of the reasons why imaging methods such as echocardiography are generally used to detect congenital heart diseases is that there are almost no CT datasets related to this disease. The dataset used in this study is the ImageCHD dataset, which comprises 3D CT scans encompassing 16 distinct types of congenital heart defects. Hyperparameter optimization was performed using the grid search method to enhance the model performance, identifying the VGG16 model as the most effective. The model demonstrated an outstanding classification accuracy of 99.99% in the training dataset and 99.94% in the test dataset. Gradient-weighted Class Activation Mapping was employed to enhance model explainability, providing visualizations of the regions most critical for the classification, thereby enabling medical professionals to validate AI-driven predictions. An optimized model that successfully classifies VSD using 3D CT image data has been introduced to the literature for the first time. Therefore, this study assumes greater significance in the existing literature and sets a benchmark for future studies.

Congenital heart disease is a type of congenital heart disease seen in 8 out of every 1000 babies worldwide (Bernier et al., 2010). Heart diseases, one of the most common types of congenital diseases, can cause serious health problems and deaths (Hallıoğlu et al., 2018). Ventricular Septal Defect (VSD) and Atrial Septal Defect (ASD) are the two most common types of congenital heart defects and account for the majority of the most common congenital heart diseases (Geva et al., 2014). It has been a challenging and complex process for doctors to detect and accurately differentiate these diseases, which show severe differences in the heart structure. Specialized radiologists can detect the types of congenital heart diseases by observing the structural changes in some parts of the anatomical structure of the heart (Left Ventricle, Right Ventricle, Left Atrium, Right Atrium, Aorta). There are techniques generally used to diagnose heart diseases. The echocardiography (EC) imaging technique plays an important role. It is widely used, especially in some cases, for diagnosing whether the heart performs its routine functions properly or for diagnosing structural heart diseases (Tay Lik Wui et al., 2011). Thanks to EC, the work of the heart can be monitored live. Thanks to this technique, which transmits sound waves to the heart, the heart valves, vessels, and muscle movements can be monitored, and heart murmurs and structural status can be observed from different angles. It has been observed that some defects can be missed when only the listening method is used to diagnose congenital

heart diseases, and it is emphasized that patients should also be evaluated using the EC imaging method (Yıldız et al., 2015). However, even though EC and similar imaging modalities are more commonly used in diagnosing heart diseases, Computed Tomography (CT) imaging in diagnosing congenital heart diseases has advantages over commonly used modalities such as EC. It can provide more successful results (Dillman & Hernandez, 2009).

Especially since 2010, the rapid rise of artificial intelligence applications worldwide has begun to impact the medical field. When the literature is examined, Sun et al. (2012) used heart sounds to classify VSD disease using a support vector machines (SVM) algorithm and achieved a 98.4% success rate in detecting the heart with VSD disease. Hassani et al. (2017) developed an artificial neural network (ANN) based method to classify heart sounds using discrete wavelet transform to detect the size of VSDs and obtained 96.6% accuracy for the classification of VSDs with small structural size and 93.3% accuracy for VSDs with large structural size. Aziz et al. (2020) used signal-processing techniques to classify ASD and VSD, two types of congenital heart diseases, and achieved an accuracy of 95.24%. They used the SVM algorithm to create the classification algorithm. Wang et al. (2020) obtained sensitivity scores of 96.0% and specificity scores of 96.7% using the automatic murmur recognition system (TAP-CRNN) model in their study using a dataset containing two classes of heart murmurs, normal and VSD. In the heart's second aortic and tricuspid regions, they observed 100% classification success of heart murmurs. They concluded that the automatic murmur recognition system has a promising potential for detecting VSD and other structural heart diseases. Sapitri et al. (2021) reported that they achieved 99.79% pixel accuracy and 97.82% overall accuracy by applying the semantic segmentation method with U-NET architecture on 2D ultrasound videos in their disease detection studies, including three disease classes: VSD, ASD, and AVSD, which is a combination of these two diseases . Yang et al. (2023) compared the classification success of VSD disease using several different models. They reported that the YOLOV5 model was the most successful in classifying VSD disease, with a success rate of 92.59%. In a study by Koundinya et al. (2023), they trained three different congenital heart diseases, including VSD disease, using the CNN model they created by converting heart sounds into spectrograms and achieving 97.12% classification success. Abbas et al. (2024) compared machine learning and deep learning techniques to detect heart diseases by analyzing noisy audio signals and found that the multilayer perceptron model performed the best with an accuracy of 95.65%. Tüzün and Özdemir (2023) demonstrated the efficacy of deep learning models in classifying brain tumors using MRI images, making a significant contribution to accelerating early diagnosis processes and minimizing human errors. This study particularly emphasizes that the EfficientNet-b0 model achieved the highest performance with an accuracy rate of 99.54%. Arslan and Ozdemir (2024) published the DEMxNET deep learning model, which demonstrates success in classifying the stages of Alzheimer's disease and possesses the potential to enhance the reliability of AI-based diagnostic tools in clinical settings through the explainability provided by the LIME method. Dörterler et al. (2024) presented an innovative approach to improve the accuracy of medical datasets by hybridizing meta-heuristic algorithms with K-Means.

Xu et al. (2021) published the world's first 3-dimensional (3D) ImageCHD dataset of CT images of congenital heart diseases. They achieved an overall classification success of 82% in the classification of 16 different congenital heart diseases, including VSD disease, in their congenital heart disease classification study with this dataset. The same study emphasizes that theirs is the first disease classification study developed using CT images on this subject. Their success rates are very suitable for improvement, as they emphasize the classification success they have achieved. When the literature is examined, it is stated that there are very few studies on the classification of congenital heart diseases and the importance of the ImageCHD dataset in

terms of its contribution to new studies in this field. In their study, Dillman and Hernandez (2009) stated that although EC and Cardiac Angiography are the most common methods used in the diagnosis of congenital heart diseases, CT and Magnetic Resonance methods are also essential imaging techniques that can be used in the detection of this disease and that CT imaging method is more advantageous to be used in the diagnosis of congenital heart diseases when evaluated in terms of structural examination of the heart due to some technical limitations of EC and other imaging methods, especially when compared with CT scans. When the EC method is evaluated from this point of view, the fact that this dataset is the world's first dataset containing CT images of 16 different types of congenital heart diseases reveals the importance of the dataset. In particular, the medical images that make up this dataset are in 3D, providing the opportunity to examine the heart's structure in more detail and to examine the details of the physical structure of the heart significantly better in diagnosing congenital heart diseases.

VSD has become the most common type of congenital heart disease in newborns as a structural defect due to the formation of a hole between the right and left ventricle of the heart. While the EC method is widely known to be successful in diagnosing VSD disease by analyzing heart murmurs, the ImageCHD dataset used in this study plays a significant role in better structural understanding and detection of VSD disease with its 44 3D 512x512 resolution medical images in the 3D CT dataset. This can be better understood by examining other studies in which VSD disease has been detected using medical images.

Due to the wide variety of congenital heart diseases, the automatic diagnosis and classification of these diseases has been a field of study that has been felt lacking in the literature due to the scarcity of medical image data sets related to this subject and the difficulty of accessing these data. This study is one of the few studies using image datasets to detect and classify VSD disease on 3D medical images. This study categorizes 3D CT scans in the ImageCHD database into "VSD" and "Non-VSD". A grid search method was used to conduct hyperparameter optimization to ascertain the influence of the hyperparameters on the model efficacy. This approach facilitated the identification of the optimal configuration. The original images with a size of 512x512 were separated into 2D slices and resized to 128x128 after separation. The input layers of the VGG16, ResNET50, and InceptionV3 pre-trained deep learning models were modified to have 128x128x3 input layers and two classes in the output layers. In this study, Gradient-weighted Class Activation Mapping (Grad-CAM) was also employed to provide a better visual representation of the classified medical images and to highlight the regions that play a more significant role in the classification process. Grad-CAM improves the explainability of artificial intelligence models in healthcare, providing reliability and transparency. The novelty and contribution of the study are as follows:

- An optimized model that successfully classifies VSD using 3D CT image data has been introduced to the literature for the first time.
- The optimized VGG16 model achieved a high 99.94% classification accuracy on the test set by hyperparameter optimization and evaluating state-of-the-art pre-trained deep learning models.
- This research sets a benchmark for future studies by demonstrating how artificial intelligence models can support and improve the diagnostic accuracy in congenital heart disease, specifically in VSD.
- Integrating the Grad-CAM method enhances the interpretability of the results, allowing medical professionals to validate AI-driven predictions by visualizing the model's focus areas and pinpointing regions critical to diagnosis.

This article is organized as follows: In the materials and methods section, the ImageCHD dataset is explained, and the pre-processing and CNN architectures are given. The experimental studies of VGG16, ResNET50, and InceptionV3 are presented in Section 3. Finally, Section 4 presents the conclusions.

Materials And Methods

This study aimed to classify 3D CT scans obtained from the ImageCHD database into “VSDs” and “Non-VSDs.” The original 512x512 images were sliced into 2D slices of 128x128 pixels during the image resizing. The deep learning models VGG16, ResNET50, and InceptionV3 were modified to accept 128 x 128 x 3 inputs and predict two output classes.

ImageCHD dataset

The data used in this study were obtained from the ImageCHD dataset. The ImageCHD dataset, which contains 110 3D CT images, is the first medical image dataset in the field to be published to detect congenital heart disease (Xu et al., 2021). This dataset contains 16 different congenital heart disease data, including 44 congenital heart diseases with VSDs and 110 congenital heart diseases with 512x512x3-dimensional medical images. In addition, some of the 3D CT scan images in the ImageCHD dataset contained more than one congenital heart disease. In this case, it is essential to note that any CT image in the dataset may contain more than one disease.

Table 1 shows the number of medical images of VSD and other diseases in the dataset. Some diseases can be a combination of two or more different diseases.

Table 1

Distribution of diseases in the ImageCHD dataset

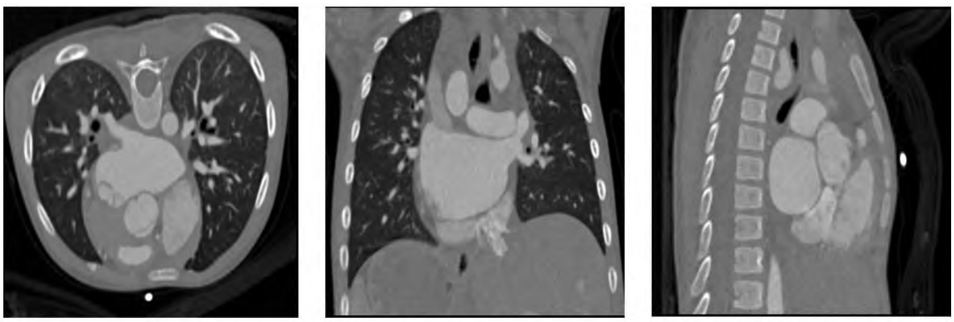
Common Congenital Heart Diseases	**ASD**	**AVSD**	**VSD**	**TOF**	**PDA**	**TGA**	**CA**	**PuA**
	32	18	44	12	14	5	6	16
Rare Congenital Heart Diseases	**PAS**	**DORV**	**CAT**	**DAA**	**APVC**	**AAH**	**IAA**	**DSVC**
	3	8	4	5	6	3	3	8
Normal	6							

Because some diseases are more common than other types of diseases, the table is divided into "Common Diseases" and "Rare Diseases". In addition to 16 different disease types, the dataset includes six disease-free healthy heart CT scans.

Figure 1 shows the structural shapes of one of the 3D CT images in the ImageCHD dataset in different axes.

Figure 1

Illustration of the width (X), height (Y), and depth (Z) layers of a 3D image



Since the CT image is 3D, it is seen that training a 3D CNN model with much more detailed data is a more precise and accurate classification approach than a 2D classification model in terms of revealing structural defects in the classification of this disease. In addition, given that each 3D image is of the float64 type, each image has a higher file size and much more detail.

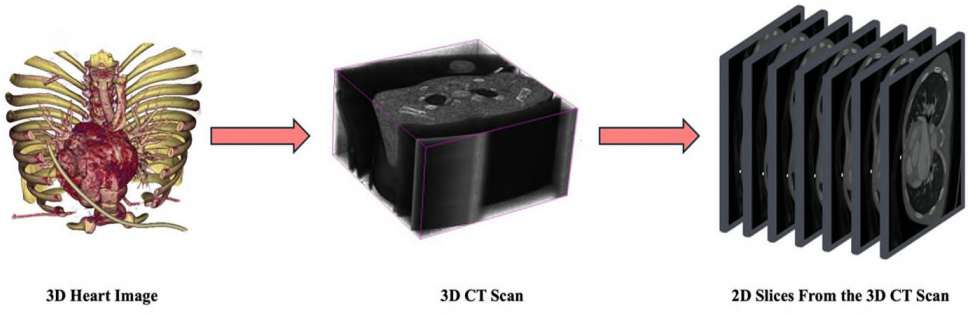
Transforming 3D images into 2D slices

Different methods exist for training a convolutional neural network (CNN) to classify 3D images. In this study, 3D medical images were reduced to 2 dimensions before training the CNN model. As a result, new 2-dimensional (2D) images are obtained from the segmented 3D image from the number of slices that make up the 3D image.

Figure 2 shows how one of the 3D tomography images in the dataset looks when converted to 2D.

Figure 2

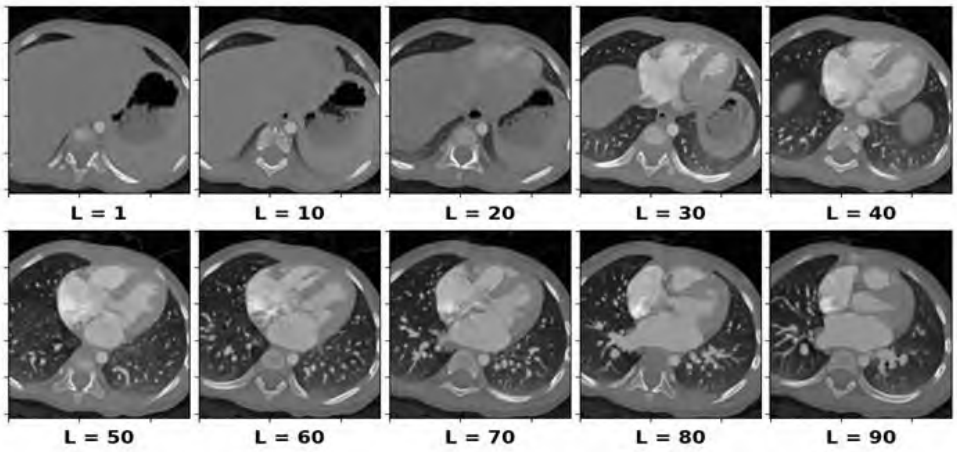
Representing a 3D image as 2D slices



This image, which has a size of (275x512x512), has become 275 (512x512) medical images after being reduced to 2 dimensions. Each of the 110 3D images was separated into two dimensions in this way, and then all the 2D images were resized to (128x128x3). In this way, it was suitable for pre-trained 2D CNN models for model training. Figure 3 shows the specific layers of a 3D image in the ImageCHD dataset in 2D.

Figure 3

Some 2D slices of a 3D image



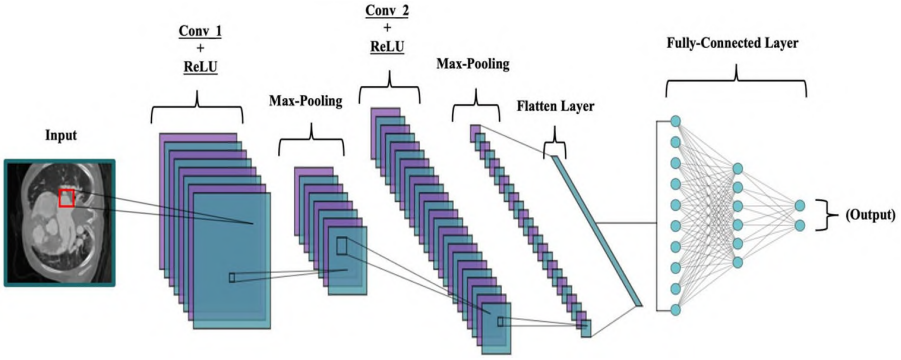
L denotes the number of layers on the 3D image in Figure 3. There are ten 2D CT images between each image. As the layers change, the structural differences in the image can be seen more clearly.

Convolutional neural networks

This neural network model, first introduced by Yann LeCun et al. in 1998, has seen a rapid increase in usage in recent years and has become the most popular neural network model used in image classification problems. It can give outstanding results, especially in classification problems. Today, CNN models are used in many fields, and different neural network models can be created in a specific field by developing the model according to the field needed. It is frequently used in image classification, natural language processing, style transfer, and text mining. Figure 4 shows the general structure of the CNN model.

Figure 4

The general structure of a CNN



The CNN structure, which is highly preferred in the medical field due to its successful results in disease classification in recent years, is an improved version of ANN. The CNN model, which works on image data, attempts to obtain the attributes of the object to be detected in the image by processing the images in layers consisting of some basic pre-processing before entering the ANN. Before these attributes enter the ANN, the image is reduced to a much smaller size than its original size, and the extracted attributes are reduced to a single dimension. This process is called flattening. After the flattening process, the data becomes one-dimensional and thus enters the ANN (Yamashita et al., 2018). The CNN model has become increasingly popular, especially after its highly successful results in high-impact image classification competitions like ImageNET (Russakovsky et al., 2014). Using the CNN model, binary or multiclass classification can be performed on the image.

Unlike the structure of the regular ANN model, the CNN model includes some essential layers, such as convolution and pooling. In this structure, pre-processing is performed to reduce the data size before the data enters the neural network and is trained. After obtaining the features that separate the attributes specific to these data instead of the entire data, they are included in the neural network.

ImageNet data set

When the ImageNet dataset was first published in 2010 under ImageNet LSVRC-2010, it contained 1.2 million high-resolution images in 1000 categories (Deng et al., 2009). As of 2024, it has become one of the most popular image datasets in the world, containing 14 million images. Many pre-trained deep learning models have been trained using this dataset. In addition, an image classification competition called "The ImageNet Large Scale Visual Recognition Challenge (ILSVRC)" has been organized annually since 2010. This competition determines the best algorithm for object detection using the ImageNet dataset (Russakovsky et al., 2014). InceptionV3, ResNET50, and VGG16 pre-trained deep learning models used in this study were also trained using the ImageNet dataset.

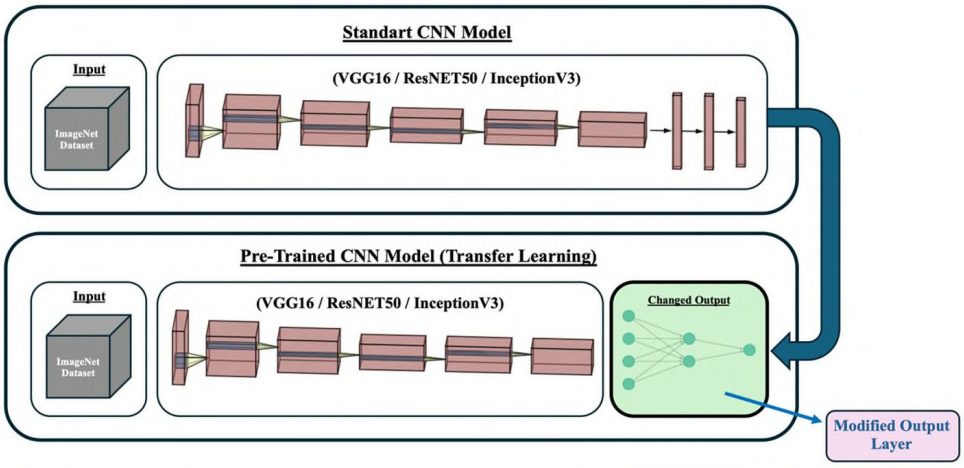
Transfer learning

Although CNN produces excellent results in image classification problems, it is a more challenging model for dealing with some problems. Some reasons are the lack of data sets for model training, the data size being too high, or the need for more costly hardware and time, as the CNN model consists of too many layers. Such problems that stand in the way of better model training and high classification success are significant challenges for scientists and researchers to overcome in their work. Transfer learning is often used to deal with these problems, and good results can be obtained. Transfer learning is the reuse of a previously trained artificial intelligence model by modifying it for different problems (Tan et al., 2018). Modifying and using some pre-trained neural networks with data that is too large to be obtained personally under normal conditions increase the success rate (Accuracy) obtained from the model. Some transfer learning models that gave successful results in most cases were used in this study. These models are called pre-trained deep learning models because they are trained with extensive data and have many layers in their structure.

Figure 5 shows a structural comparison between a typical CNN model and the modified version of the pre-trained deep learning models.

Figure 5

Comparison of the CNN and the pre-trained model



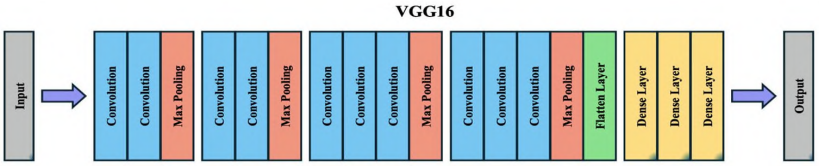
In this study, ResNET50, VGG16, and InceptionV3 pre-trained deep learning algorithms were utilized, and the original output layers were modified to have two neurons.

VGG16 model

In 2014, Karen Simonyan and Andrew Zisserman from the University of Oxford introduced the VGG16 and VGG19 models at the ILSVRC competition organized by ImageNet to investigate the effects of neural network depth on the success rate of large-scale data classification problems (Simonyan & Zisserman, 2015). The most important differences between these models and the previous deep CNN models are the smaller filter sizes and more convolutional layers. Figure 6 shows the structure of the VGG16 model. While developing these deeper CNN models, Krizhevsky et al. (2017) referenced their classification study using the ImageNet dataset. VGG models have been widely used in recent studies because of their high performance. This technique offers high performance primarily because it is already optimized for training with multidimensional data. Consequently, the model's feature map is highly developed (Kara, 2023).

Figure 6

Internal structure and layers of the VGG16 model



The structure of the VGG models consists of convolution, pooling, and fully connected layers. The input layer accepts 224x224 fixed-size inputs. The only pre-processing is to extract the average RGB (Red, Green, Blue) value from each pixel value. In the convolution layers, 3x3-dimensional filters are used. The five convolution layers have a 2x2 maximum pooling layer. The successive convolution and pooling layers are connected to fully connected layers of 1000 classes in the output layer (Simonyan & Zisserman, 2015).

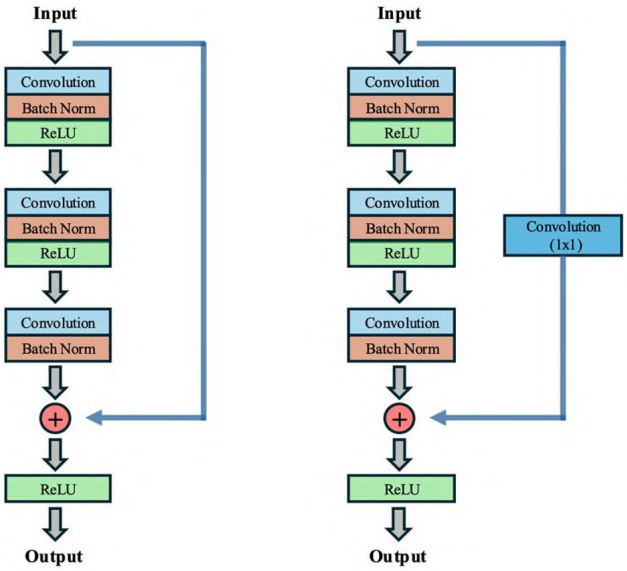
ResNET50 model

The ResNET50 model has less complexity than other popular deep learning models and simplifies the difficulty of training high-dimensional deep learning models with significantly more layers. Although it is much deeper than the VGG models, it has less complexity (He et al., 2016). The ResNET50 model trained with the ImageNet dataset won first place in classification in the ILSVRC competition held in 2015. In the Microsoft Common Object in Context (COCO) competition, they won first place in detection and segmentation (He et al., 2016).

The ResNET50 model accepts 224x224 color image data at the input layer. Instead of the standard convolution layer, a specially modified structure called "Residual Block" is used, and the model contains 16 Residual Blocks. Each of these so-called Residual Blocks contains two convolution layers. The convolution layers contain different filter sizes, such as 7x7 and 3x3, and there are connections between each Residual Block structure. Thanks to these connections, the gradient vanishing problem is minimized. The Residual Block structure of the ResNET50 Model is shown in Figure 7.

Figure 7

The residual block structure of the ResNET50 model



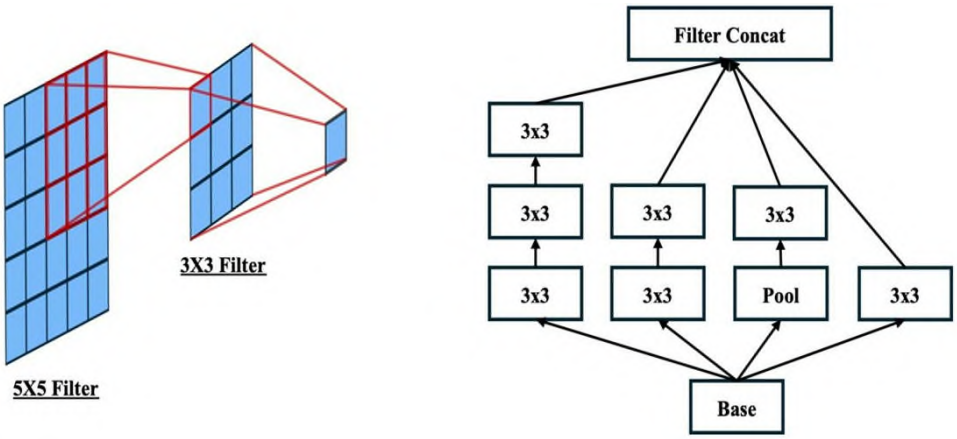
Since the ResNET50 model is trained with the ImageNet dataset, it consists of fully connected networks with 1000 classes in the output layer (He et al., 2016).

InceptionV3 model

Inception, also known as GoogLeNet, has been developed to outperform other large deep-learning models with respect to high computational cost and memory usage (Szegedy et al., 2016). InceptionV3 is a fast, powerful, and popular deep learning model with a high success rate in classification, even though it uses only 5 million parameters in training and is trained with a much smaller number of parameters compared to a popular deep learning algorithm such as VGG (Szegedy et al., 2016). The InceptionV3 model, such as the VGG16 and ResNET50 models, was one of the high-performing deep-learning models in the ILSVRC competition. Figure 8 shows the Asymmetric Convolution structure of the InceptionV3 model.

Figure 8

Asymmetric convolution structure of the InceptionV3 model

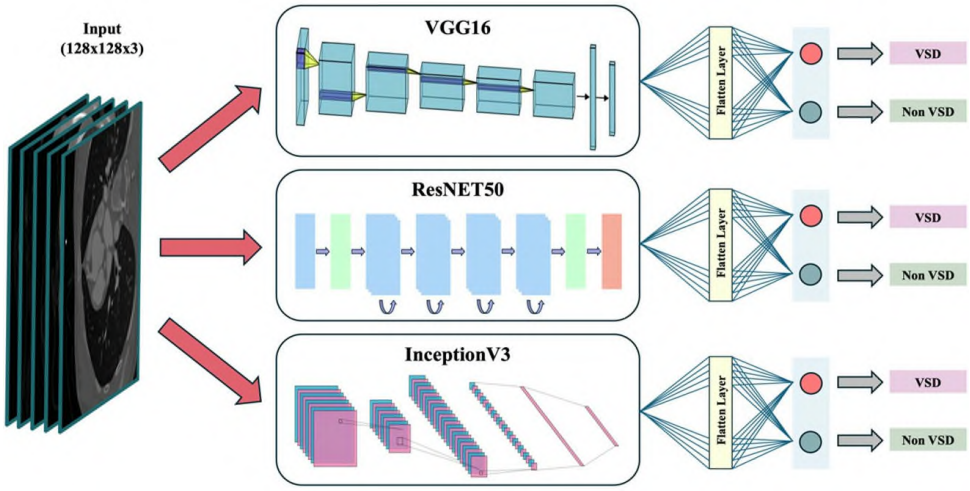


The factorized convolution process used in the model's structure is one of the most important reasons for the low hardware cost, speed, and high performance of the InceptionV3 model trained with few parameters. Filter sizes considered relatively large, such as 5x5 or 7x7, incur a disproportionate computational cost. For example, instead of a 5x5 filter, using two fully connected 3x3 filters is preferable. In the InceptionV3 model, 1x1 and 3x3 filters are generally preferred. In this way, the computational cost remains lower than that of large-size filters, and a significant performance increase is achieved.

Proposed method

For VSD disease classification, the ImageCHD dataset containing 3D CT images was converted into 2D CT images. VGG16, ResNET50, and InceptionV3 deep learning models were used in the training. These three models using the transfer learning method were modified, as shown in Figure 9, with input layers of 128x128x3 and output layers modified to be directly 2-class. The models were built with an input layer of 128x128x3 2D CT and an output layer of a categorical binary classifier as "with VSD" or "without VSD". To utilize the pre-trained weights, it was necessary to turn off the trainable feature of the layers before training the models. Furthermore, pre-trained with the ImageNet dataset, the default model weights were employed to avoid training all models from scratch.

Figure 9
Visualization of the classification method used



A 10-fold cross-validation technique and 25 training cycles (epoch) were preferred for the model training. The batch size was 16, and the activation function was "Softmax". For model optimization, the Adam optimization algorithm was preferred.

Cross-validation and performance metrics

The k-fold cross-validation method aims to evaluate the statistical success of an artificial intelligence model more objectively and to reveal the actual success rate more accurately. In this study, a 10-fold cross-validation was applied for model training.

The performance metrics used to evaluate the results of the VGG16, ResNET50, and Inceptionv3 models are analyzed in this section. Accuracy, Recall, Precision, F1-Score, Confusion Matrix, and Roc Curve performance metrics frequently used in image classification problems are used to analyze the trained models (Sevli, 2023).

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{1}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

$$\text{F1-Score} = \frac{2^{*}\ \text{Presicion} * \text{Recall}}{\text{Presicion} + \text{Recall}} \tag{4}$$

Results

This study aimed to classify VSD disease using deep learning models. VGG16, ResNET50, and InceptionV3 pre-trained deep learning models were trained using 3D CT images from the ImageCHD dataset. The 3D CT scans were converted to 2D and resized to 128x128 before model training. The models were modified to have 2 class labels, "VSD" and "Not VSD", and the output layers were set as binary classifiers. To determine the effects of the hyperparameters on the model success, hyperparameter optimization was performed

using the grid search method. In addition, to achieve statistically more accurate results, the "K-Fold Cross Validation" method was applied and set to work in 10 folds, and the entire data set was entered into the model for training as training and test data in pieces for each model.

Comparative analysis of the effects of the hyperparameters on the models

A comparative analysis of the effects of the hyperparameters on the models' classification metrics results was evaluated under varying batch sizes and learning rates over 10-fold cross-validation, as shown in Table 2. Nvidia P100 GPU and 16 GB VRAM were used as the computational resources.

Table 2

Demonstrating the effects of different hyperparameter values on the performance of the models through performance metrics

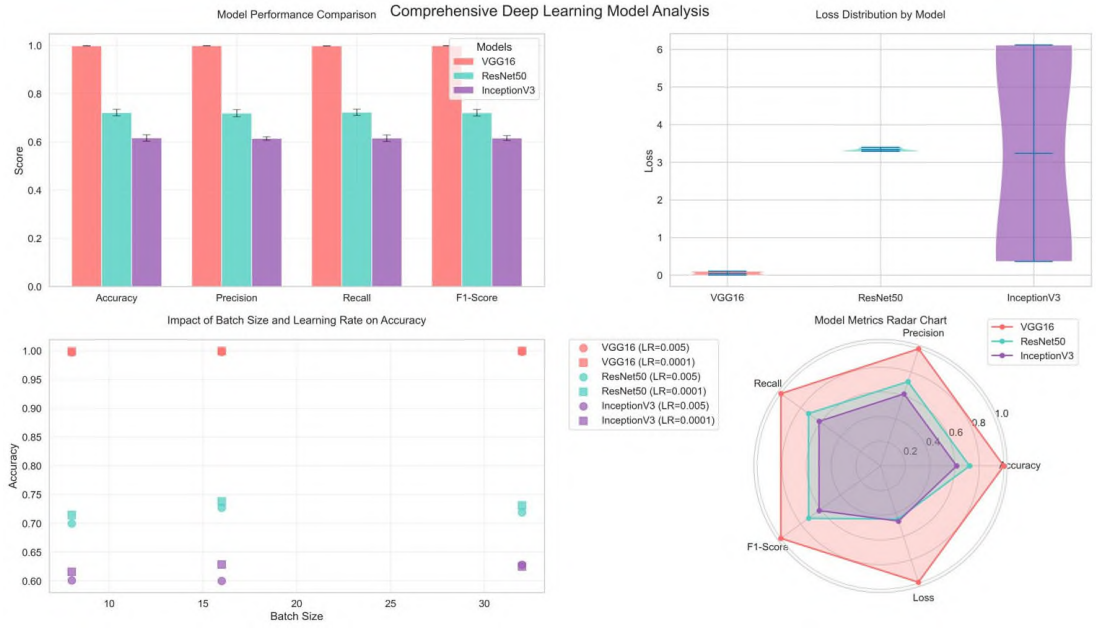
Model	Learning Rate	Batch Size	Train					Test				
			Accuracy	Loss	Recall	Precision	F1-Score	Accuracy	Loss	Recall	Precision	F1-Score
VGG16	0.005	8	0.9987	0.126	0.999	0.999	0.999	0.9967	0.123	0.997	0.997	0.997
	0.005	16	0.9981	0.204	0.998	0.998	0.998	0.9959	0.352	0.996	0.996	0.996
	0.005	32	0.9986	0.096	0.999	0.999	0.999	0.9973	0.104	0.997	0.997	0.997
	0.0001	8	0.9989	0.286	0.999	0.999	0.999	0.9977	0.342	0.998	0.998	0.998
	0.0001	16	0.9998	0.061	1.000	1.000	1.000	0.9993	0.110	0.999	0.999	0.999
	0.0001	**32**	**0.9999**	**0.004**	**1.000**	**1.000**	**1.000**	**0.9994**	**0.041**	**0.999**	**0.999**	**0.999**
ResNet50	0.005	8	0.7232	3392	0.723	0.754	0.720	0.7250	3305	0.725	0.755	0.723
	0.005	16	0.7272	4.225	0.727	0.753	0.720	0.7262	4.294	0.726	0.751	0.718
	0.005	32	0.6959	5.072	0.696	0.742	0.690	0.6956	5.132	0.696	0.744	0.690
	0.0001	8	0.6837	6.428	0.684	0.758	0.679	0.6829	6.267	0.683	0.756	0.678
	0.0001	16	0.6687	8.594	0.669	0.747	0.654	0.6705	8.755	0.671	0.747	0.656
	0.0001	32	0.7381	3.289	0.738	0.747	0.736	0.7346	3.124	0.735	0.745	0.733
InceptionV3	0.005	8	0.5627	6.117	0.563	0.550	0.512	0.5648	5.967	0.565	0.552	0.513
	0.005	16	0.6000	6.938	0.600	0.607	0.519	0.5972	7.168	0.597	0.573	0.516
	0.005	32	0.5786	6.448	0.579	0.569	0.495	0.5788	6.164	0.579	0.572	0.495
	0.0001	8	0.6248	0.900	0.625	0.612	0.545	0.6272	1.512	0.627	0.609	0.548
	0.0001	16	0.6260	0.527	0.626	0.614	0.558	0.6277	1.030	0.628	0.625	0.560
	0.0001	32	0.6286	0.368	0.629	0.613	0.561	0.6277	1. 217	0.628	0.610	0.559

Table 2 presents the classification metrics of these models, namely VGG16, ResNet50, and InceptionV3, that were trained using various combinations of batch size (8, 16, and 32) and learning rate (0.005, 0.0001). The variability in the VGG16 graphical network unfolds a clear advantage over other approaches in terms of illustrative performance and tenability, where the model performs a variety of fundamental classification techniques through vital metrics such as Accuracy, Loss, Recall, Precision, and F1-Score. These outcomes imply that VGG16 is the best and most stable model from the group for conducting the classification tasks that were tested. Specifically, the combination of a batch size of 32 and a learning rate of 0.0001 resulted in the highest classification performance, achieving an accuracy of 99.94%. Even though the ResNet50 model, being better than the InceptionV3 one, shows good classification performance, it has limitations with a

loss value of 3.124, which means it cannot efficiently classify some samples. The network architecture is characterized by the integration of "residual blocks" specifically created to solve the vanishing gradient problem and give a structural privilege over the InceptionV3 algorithm in this case. However, the model's relatively high loss value indicates the need for further optimization or refinement in hyperparameter tuning to achieve greater reliability across diverse datasets. The InceptionV3 model demonstrated the lowest performance across all metrics in this study. With an accuracy of 62.77%, the model is significantly less effective than the VGG16 and ResNet50 models regarding the classification accuracy on this dataset. Even though it is an advanced model, the InceptionV3 model does not appear to be very successful in classification. One of the major drawbacks of InceptionV3 is that it uses multiple parallel convolutional layers to extract details at various feature scales, increasing the complexity. Although such an architecture is advantageous for particular datasets, it might introduce redundancy in all computing. Thus, the performance might be under productivity when other datasets with a simple or even less varied configuration are applied. In Figure 10, the classification success of the VGG16, ResNet50, and InceptionV3 models in all metrics is very detailed as comparative graphs.

Figure 10

Comparative visualizations presenting the results achieved by the VGG16, ResNet50, and InceptionV3 models



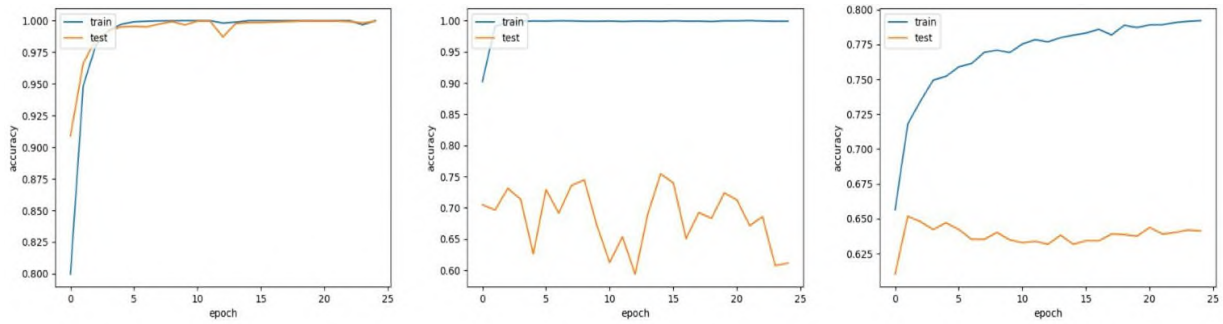
In Figure 10, the inconsistent results obtained from the InceptionV3 model are especially noticeable in the loss values graph. It is observed that the VGG16 and ResNet50 models obtain more consistent loss values. When the accuracy graph is examined, the accuracy results obtained from the models at different learning rate values (0.005, 0.0001) are seen. It is seen that the VGG16 model is again the most consistent and successful model with the results it obtains. The results of the ResNet50 model show that similar results were obtained at two different learning values, but the model trained with the 0.0001 learning rate performed slightly better. When the results of the InceptionV3 model are examined, it is more clearly seen that the model trained with the learning rate value of 0.0001 obtains more successful results than the model trained with the value of 0.005. When the radar chart is examined, the models are compared according to the metric results obtained. It can be observed very clearly that the VGG16 model is the most successful model that gives the best results by far. In addition, the superiority of VGG16 over the ResNe50 and InceptionV3

models, especially in the Loss value, is very clearly visible in the radar chart. It is seen that the Loss values of the other two models are close to each other. The charts reveal the superiority of the VGG16 model in classification success over the other models.

Figure 11 shows the accuracy curve graphs for the first fold of the VGG16, ResNET50, and InceptionV3 models with the specified learning rate of 0.0001 and batch size of 32.

Figure 11

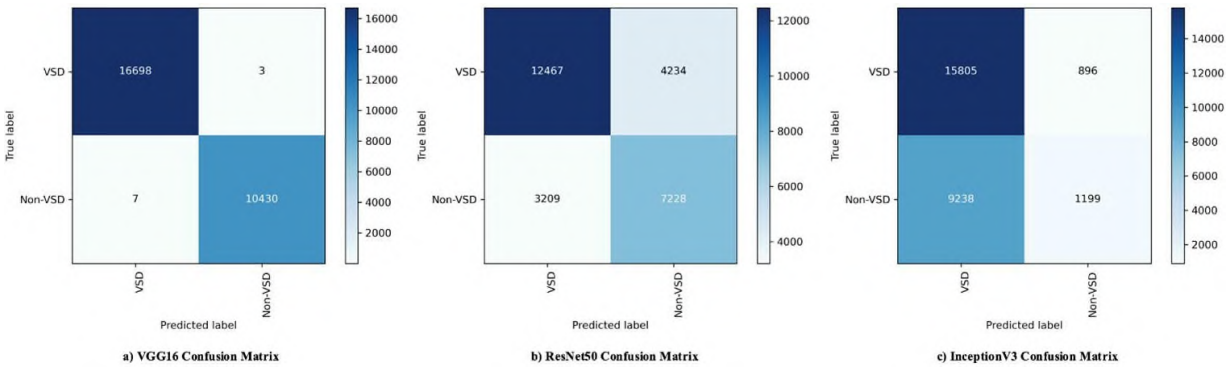
Accuracy metric results for the first fold of the optimized VGG16, ResNET50, and InceptionV3 models



The relationship between the actual and predicted data of the optimized models is illustrated in Figure 12, which depicts the confusion matrices. By looking at this table showing the TP, TN, FP, and FN values, inferences can be made about the performance of the models. When the results are compared with accuracy graphs and confusion matrix tables, inferences can be made more easily about the success of the models.

Figure 12

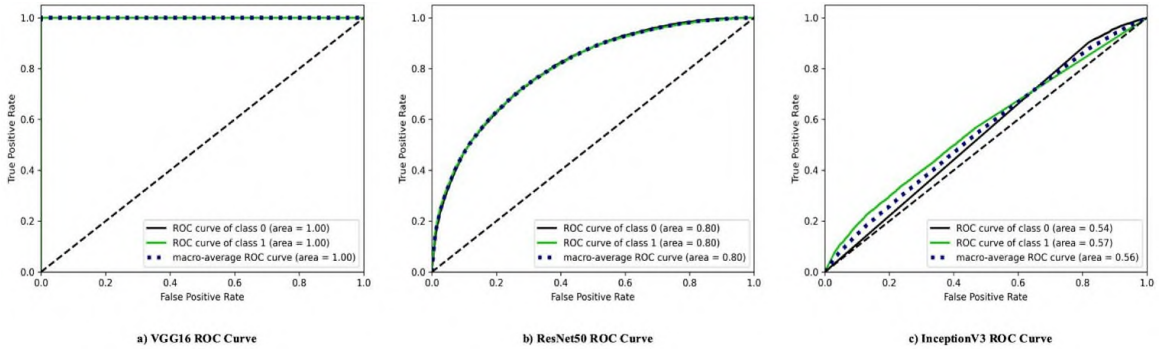
The confusion matrices of the optimized VGG16, ResNET50, and InceptionV3 models



When the confusion matrices are examined, the classification success of the VGG16 model attracts attention. The performance superiority of the VGG16 model, which stands out with its high accuracy rate and classification success, is seen compared to the ResNET50 and InceptionV3 models. The ResNET50 and InceptionV3 models achieve average classification success. When looking at the accuracy graphs, it is seen that the training and test curves are consistent and have a high success rate in the VGG16 model. At the same time, it is noteworthy that the training and test curves are inconsistent and have low success rates in the ResNET50 and InceptionV3 models. Figure 13 shows the Roc Curve graphs of the VGG16, ResNET50, and InceptionV3 models.

Figure 13

Roc curve of the optimized VGG16, ResNET50, and InceptionV3 models



When the Roc Curve graphs, another statistical metric, are examined, the success of the VGG16 model stands out compared to the ResNET50 and InceptionV3 models. The results of the Roc Curve graphs obtained from the VGG16, ResNET50, and InceptionV3 models support the results of all other performance metrics. When the graph of the VGG16 model is examined, it has achieved a very high success. When the graphics of the ResNET50 model are examined, it is seen that it gives better results than the InceptionV3 model, but it is still not good enough and has a slightly above-average performance. When the Roc Curve chart of the InceptionV3 model is examined, it is seen to give bad results. According to these results, the InceptionV3 model can be interpreted as the model with the most unreliable results. It can be clearly stated that the most reliable and successful model is the VGG16. As a result, all models' performance metrics and performances were analyzed statistically. According to the results obtained, the performance metrics confirmed each other.

Comparative analysis of the effects of hyperparameters on the training times of the models

A comparative analysis of the effects of the hyperparameters on the models' training times under varying batch sizes and learning rates over 10-fold cross-validation is presented in Table 3. Nvidia P100 GPU and 16 GB VRAM were used as the computational resources.

Table 3

Comparative analysis of training durations for the VGG16, ResNet50, and InceptionV3 models evaluated under varying batch sizes and learning rates over a 10-fold cross-validation

Model	Number of Parameters	Learning Rate	Batch Size	Total Training Time (minutes)	Average Training Time per Fold (minutes)
VGG16	138 million	0.005	8	130.51	13.05
		0.005	16	116.17	11.61
		0.005	32	91.14	9.11
		0.0001	8	130.31	13.03
		0.0001	16	114.92	11.49
		0.0001	32	93.73	9.37
ResNet50	25.6 million	0.005	8	207.28	20.72
		0.005	16	193.72	19.37
		0.005	32	127.79	12.77
		0.0001	8	205.88	20.58

Model	Number of Parameters	Learning Rate	Batch Size	Total Training Time (minutes)	Average Training Time per Fold (minutes)
		0.0001	16	209.03	20.90
		0.0001	32	130.50	13.05
InceptionV3	23.8 million	0.005	8	254.98	25.49
		0.005	16	212.24	21.22
		0.005	32	133.80	13.38
		0.0001	8	253.97	25.39
		0.0001	16	210.16	21.01
		0.0001	32	137.92	13.79

Table 3 compares the VGG16, ResNet50, and InceptionV3 models regarding training times under different learning rates and batch size values. Each model's architectural features and number of parameters significantly affected the training times. Although the VGG16 model, with 138 million parameters, is significantly larger than the ResNet50 and InceptionV3 models, its training time is shorter than the other two. This shorter training time indicates that the architectural complexity of the models has a more significant impact on the training time than the number of parameters they contain. The results indicate that the training time significantly decreased with larger batch sizes, while the learning rate (0.005 and 0.0001) had minimal impact on the training duration. Although the InceptionV3 model has a lighter architecture than the VGG16 and ResNet50 models regarding the number of parameters, its training time is longer than that of the other models. The results demonstrate that with larger batch sizes, the training duration significantly decreased while the impact of the learning rates (0.005 and 0.0001) on the training time was minimal. The visualization hyperparameter analysis for the VGG16, ResNET50, and InceptionV3 models is shown in Figure 14.

Figure 14

Visualization hyperparameter analysis for VGG16, ResNET50, and InceptionV3 models

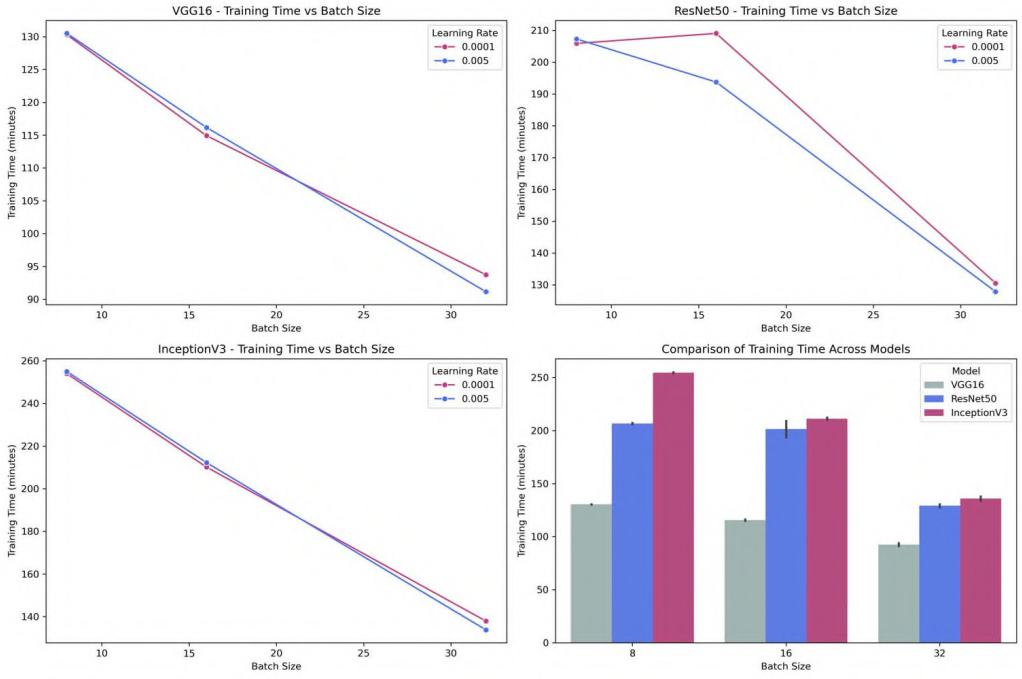


Figure 14 shows a comparative analysis of the results obtained by the VGG16, ResNet50, and InceptionV3 models when different hyperparameters (learning rate and batch size) were used. VGG16 is the model that exhibits the most efficient training time when testing different hyperparameter configurations. The lowest training time indicates that, among the models tested, VGG16 demonstrates the best training efficiency. It is the optimal choice for reducing the training time in this experimental context. These results demonstrate the necessity of grasping the influence of various deep learning models and hyperparameter settings on the training time. This model selection process should be treated as one of the most crucial factors in the decision-making process.

Gradient-weighted Class Activation Mapping Analysis

Grad-CAM is an eminent interpretable artificial intelligence (XAI) technique that seeks to elucidate the inner workings of artificial intelligence models, including their decision-making processes and outputs, for better human understanding. The approach thus allows the connection between the magnitude of AI programs and human thought to be bridled by providing visualizations of the rationale behind a model's predictions.

In this experiment, the Grad-CAM algorithm (Selvaraju et al., 2017), one of the famous XAI techniques, was implemented as a method of localization of regions in CT scans believed by the model to be affected by VSD disease. Grad-CAM creates heatmaps by distributing larger activation values to the pixels influencing the model's prediction. This heatmap leads to a color-coded diagram according to different diseased parts of the patient, henceforth providing doctors with a map indicating where the model is looking.

The Grad-CAM algorithm was employed to assess the output of a deep learning model trained for VSD disease diagnosis. By partially covering the CT images with the overlayed heatmaps, the regions of the disease were visually separated from the rest of the image. Therefore, the zones with more intensive red hues are where the model the presence of disease, and the blue zones are where no disease was found. This approach furnishes a comprehensible rationale for the model's thinking to healthcare experts, thus making it easily understood. The Grad-CAM output for a CT scan image with and without VSD Disease is presented in Figure 15.

Figure 15

View of CT scans with and without VSD Disease using Grad-CAM

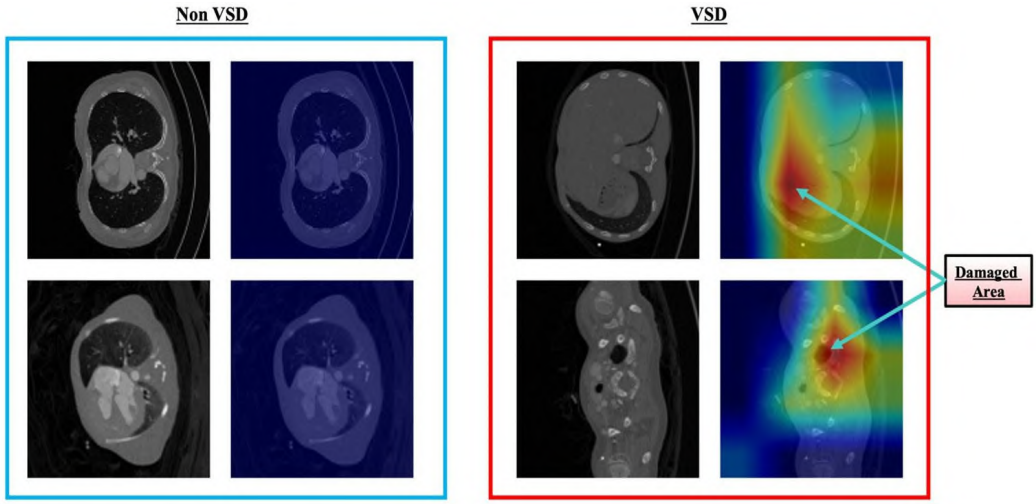
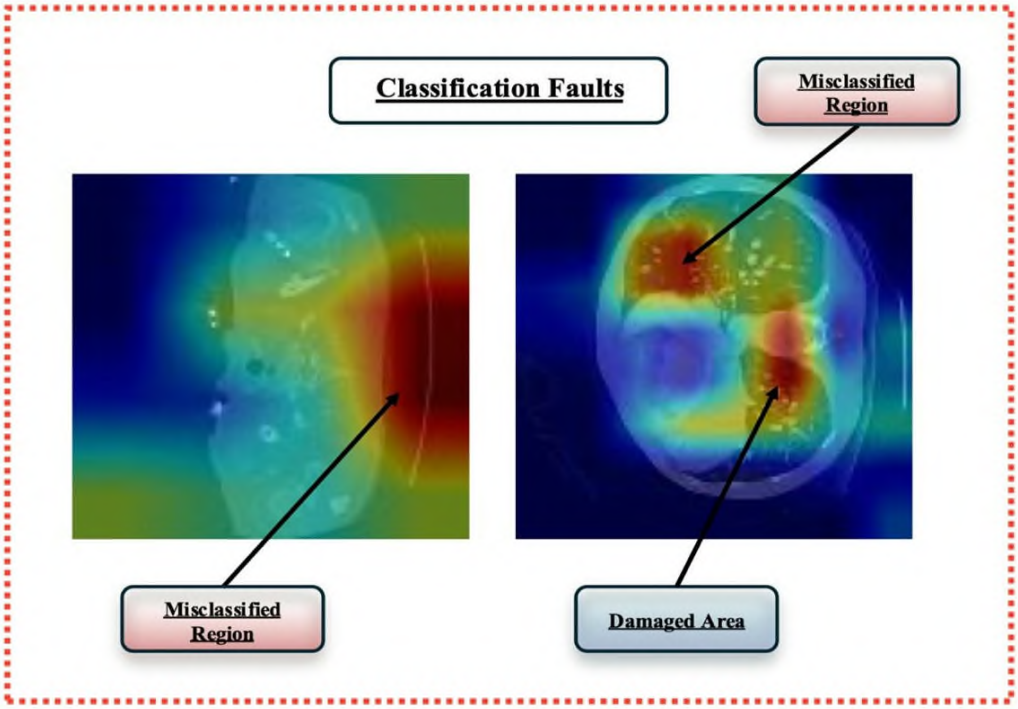


Figure 15 presents an example of the Grad-CAM output for a CT scan image. The red-highlighted areas represent regions where the model detects VSD disease, whereas the blue-highlighted areas correspond to regions where the disease is absent. This visual representation enables a straightforward and interpretable understanding of the model's predictions. Visualization of the classification faults can be seen in Figure 16.

Figure 16

Visualization of the classification faults



The Grad-CAM method, as shown in Figure 16, mistakenly marked a part of a CT scan image as a VSD case. This picture shows a problem that the deep learning model applied in this study has since it wrongly identified the normal area as the one with the disease. Therefore, the Grad-CAM heatmap was incorrectly focused on this image region indicated by the red spot, which could confuse the medics interpreting the results. Noise in the input data: Artifacts or inconsistencies in CT scans are the main reasons for wrongful feature extraction by the model. The misclassification in Figure 16 underscores the importance of not relying solely on AI-based models for clinical decision-making, especially in high-stakes fields like healthcare.

In addition to verifying the predicted results of the deep learning model, these visual outputs help clinicians interpret the data. Thus, it acts as a vital item. Using this method, medical professionals can act as supervisors and thus verify or disprove the model's reasoning. Thus, physicians will have additional confidence in their already great expertise in diagnosis. In conclusion, Grad-CAM is a useful tool for understanding artificial intelligence models. However, it must be applied with other validation methods and domain expertise to ensure reliable results.

Conclusion

The fact that VSD disease is the most common type of congenital heart disease reveals the importance of new treatment methods to be developed for this disease. To develop methods, it is of great importance to diagnose the disease correctly. Deep learning algorithms can achieve successful results in disease detection.

However, the fact that there are very few available data sets for VSD disease, especially the almost complete absence of medical image data sets, has caused the number of studies in this field to remain limited.

ImageCHD, the first 3D CT medical image dataset published on congenital heart diseases, was used in this study. Because of the research, when the literature is examined, this study is the first study conducted on the classification of VSD disease on CT image data. In order to ascertain the optimal model and parameters for classifying the VSD disease, hyperparameter optimization was conducted using the grid search method. By using optimized VGG16, InceptionV3, and ResNET-50 pre-trained deep learning models for classification, it was determined that the model with the highest degree of accuracy was the VGG16 model with a value of 99.99% in the training set and 99.94% in the test set. The high classification success achieved shows that CT image data are suitable for detecting or classifying this disease.

Moreover, the activation map generated by the Grad-CAM method, considering the weights in the deep learning model's last convolutional layer, showed that the model focused on the exact regions and made successful predictions. The Grad-CAM method allows for the visualization of the regions of an image that are the focus of the model's attention during the decision-making process. This visualization enables medical professionals to gain insight into and validate the decisions made by artificial intelligence systems. Additionally, the Grad-CAM model can be employed to analyze instances of inaccurate predictions, which can help identify the model's limitations and facilitate improvements in the development process.

This study will facilitate new research endeavors in this domain in the coming years. This is the first study to use CT images to classify VSD disease, and it has yielded the most optimal outcomes regarding general classification efficacy. In future work, classification studies will be conducted for the disease categories in the ImageCHD dataset.

Peer Review Externally peer-reviewed.
Author Contributions Conception/Design of Study- Ö.Ç.B.; Data Acquisition- K.B.; Data Analysis/Interpretation- İ.P., M.Ş.Ç.; Drafting Manuscript- K.B., İ.P.; Critical Revision of Manuscript- Ö.Ç.B., M.Ş.Ç.; Final Approval and Accountability- K.B., İ.P., Ö.Ç.B., M.Ş.Ç.; Supervision- İ.P.
Acknowledgments The numerical calculations reported in this paper were fully/partially performed at TUBITAK ULAKBIM, High Performance and Grid Computing Center (TRUBA resources).
Conflict of Interest The authors have no conflict of interest to declare.
Grant Support The authors declared that this study has received no financial support.

Author Details
Kadir Barut
[1] Burdur Mehmet Akif Ersoy University, Department of Management Information Systems, Burdur, Türkiye
0000-0002-6288-5954

İhsan Pençe
[2] Burdur Mehmet Akif Ersoy University, Department of Software Engineering, Burdur, Türkiye
0000-0003-0734-3869 ihsanpence@mehmetakif.edu.tr

Özlem Çetinkaya Bozkurt
[3] Burdur Mehmet Akif Ersoy University, Department of Business Administration, Burdur, Türkiye
0000-0002-6218-2570

Melike Şişeci Çeşmeli
[2] Burdur Mehmet Akif Ersoy University, Department of Software Engineering, Burdur, Türkiye
0000-0001-9541-2590

References

Abbas, S., Ojo, S., Al Hejaili, A., Sampedro, G. A., Almadhor, A., Zaidi, M. M., & Kryvinska, N. (2024). Artificial intelligence framework for heart disease classification from audio signals. *Scientific Reports*, *14*(1), 3123. doi:10.1038/s41598-024-53778-7

Arslan, N. N., & Ozdemir, D. (2024). Analysis of CNN models in classifying Alzheimer's stages: comparison and explainability examination of the proposed separable convolution-based neural network and transfer learning models. *Signal, Image and Video Processing, 18*(S1), 447–461. DOI: 10.1007/s11760-024-03166-5

Aziz, S., Khan, M. U., Alhaisoni, M., Akram, T., & Altaf, M. (2020). Phonocardiogram signal processing for the automatic diagnosis of congenital heart disorders through the fusion of temporal and cepstral features. *Sensors*, *20*(13), 3790. doi:10.3390/s20133790

Bernier, P. L., Stefanescu, A., Samoukovic, G., & Tchervenkov, C. I. (2010). The challenge of congenital heart disease worldwide: epidemiologic and demographic facts. *Seminars in Thoracic and Cardiovascular Surgery: Pediatric Cardiac Surgery Annual*, *13*(1), 26–34. https://doi.org/10.1053/j.pcsu.2010.02.005

Deng, J., Dong, W., Socher, R., Li, L.-J., Kai Li, & Li Fei-Fei. (2009). ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 248–255). IEEE. doi: 10.1109/CVPR.2009.5206848

Dillman, J. R., & Hernandez, R. J. (2009). Role of CT in the evaluation of congenital cardiovascular disease in children. *American Journal of Roentgenology*, *192*(5), 1219–1231. doi: 10.2214/AJR.09.2382

Dörterler, S., Dumlu, H., Özdemir, D., & Temurtaş, H. (2024). Hybridization of meta-heuristic algorithms with k-means for clustering analysis: Case of medical datasets. *Gazi Journal of Engineering Sciences, 10*(1), 1–11. https://doi.org/10.30855/gmbd.0705N01

Geva, T., Martins, J. D., & Wald, R. M. (2014). Atrial septal defects. *The Lancet*, *383*(9932), 1921–1932. https://doi.org/10.1016/S0140-6736(13)62145-5

Hallıoğlu, O., Karpuz, D., Giray, D., Demetgül, H., & Öztaş, A. (2018). Frequency of the congenital heart diseases according to the risk groups: Fetal echocardiographic screening. *The Journal of Gynecology-Obstetrics and Neonatology*, *15*(1), 1–4.

Hassani, K., Jafarian, K., & Doyle, D. J. (2017). Heart sounds feature used for the classification of ventricular septal defect size in children. In *IFMBE Proceedings* (Vol. 61, pp. 28–31). Springer Verlag. doi:10.1007/978-981-10-4220-1_6

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 770–778). IEEE. doi: 10.1109/CVPR.2016.90

Kara, İ. (2023). Query by image examination: Classification of digital image-based forensics using deep learning methods. *Acta Infologica*, *7*(2), 348-359. doi:10.26650/acin.1282567

Koundinya, T. S., Sudhanva, S., Shashank, C. S., Chandrasekar, S., & Munavalli, J. R. (2023). Congenital heart disease detection using spectral analysis and CNN. In *2023, 4th International Conference for Emerging Technology, INCET 2023*. Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/INCET57972.2023.10170429

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84–90. https://doi.org/10.1145/3065386

Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324. doi: 10.1109/5.726791

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... Fei-Fei, L. (2015). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, *115*(3), 211–252. doi:10.1007/s11263-015-0816-y

Sapitri, A. I., Nurmaini, S., Sukemi, Rachmatullah, M. N., & Darmawahyuni, A. (2021). Segmentation atrioventricular septal defect by using convolutional neural networks based on U-NET architecture. *IAES International Journal of Artificial Intelligence*, *10*(3), 553–562. doi: 10.11591/ijai.v10.i3.pp553-562

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)* (pp. 618–626). IEEE. doi: 10.1109/ICCV.2017.74

Sevli, O. (2023). A deep learning-based classification study for diagnosing corneal ulcers on ocular staining images. *Acta Infologica*, 7(2), 281-292. https://doi.org/10.26650/acin.1173465

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *3rd International Conference on Learning Representations (ICLR 2015)*, 1–14. Retrieved from http://arxiv.org/abs/1409.1556

Sun, S., Jiang, Z., Wang, H., Fang, Y., & Tao, T. (2012). Heart sound feature parameter distribution and support vector machine-based classification boundary determination method for ventricular septal defect auscultation. *Journal of Computational Science and Technology*, *6*(3), 198–206. doi:10.1299/jcst.6.198

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the Inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2818–2826). IEEE. https://doi.org/10.1109/CVPR.2016.308

Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., & Liu, C. (2018). A survey on deep transfer learning. *The 27th International Conference on Artificial Neural Networks (ICANN 2018)*. Retrieved from http://arxiv.org/abs/1808.01974

Tay Lik Wui, E., Yip, J. W. L., & Li, W. (2011). Echocardiography. In *Diagnosis and Management of Adult Congenital Heart Disease* (pp. 28–43). Elsevier. https://doi.org/10.1016/B978-0-7020-3426-8.00005-8

Tüzün, B. N., & Özdemir, D. (2023). Classification of brain tumors using deep learning models. *Journal of Scientific Reports-A*, 054, 296–306. doi: 10.59313/jsr-a.1293119

Wang, J.-K., Chang, Y.-F., Tsai, K.-H., Wang, W.-C., Tsai, C.-Y., Cheng, C.-H., & Tsao, Y. (2020). Automatic recognition of ventricular septal defect murmurs using convolutional recurrent neural networks with temporal attentive pooling. *Scientific Reports*, *10*(1), 21797. doi:10.1038/s41598-020-77994-z

Xu, X., Wang, T., Zhuang, J., Yuan, H., Huang, M., Cen, J., ... Shi, Y. (2021). ImageCHD: A 3D computed tomography image dataset for classification of congenital heart disease. *Segmentation, Classification, and Registration of Multi-modality Medical Imaging Data*, *4*, 77–87. Retrieved from http://arxiv.org/abs/2101.10799

Yamashita, R., Nishio, M., Do, R. K. G., & Togashi, K. (2018). Convolutional neural networks: an overview and application in radiology. *Insights into Imaging*, *9*(4), 611–629. doi:10.1007/s13244-018-0639-9

Yang, Y., Wu, B., Wu, H., Xu, W., Lyu, G., Liu, P., & He, S. (2023). Classification of normal and abnormal fetal heart ultrasound images and identification of ventricular septal defects based on deep learning. *Journal of Perinatal Medicine*, *51*(8), 1052–1058. doi:10.1515/jpm-2023-0041

Yıldız, J., Çetin, İ. İ., Aktaş, D., Arı, M. E., Kocabaş, A., Ekici, F., & Şaylı, T. R. (2015). Is an echocardiographic evaluation necessary for all children with a cardiac murmur? *Turkish Journal of Pediatric Disease*. doi:10.12956/tjpd.2015.158

İSTANBUL UNIVERSITY PRESS

Acta Infologica,
ACIN 2025, 9 (1): 55–73

https://doi.org/10.26650/acin.1584030
Submitted: 29.11.2024
Revision Requested: 31.12.2024
Last Revision Received: 07.01.2025
Accepted: 21.02.2025
Published Online 27.02.2025

Acta Infologica

Research Article

Open Access

A Novel and Robust LSTM Model for Customer Churn Analysis Using Deep, Machine Learning, and Ensemble Learning: A Telecommunications Case

Muhammet Sinan Başarslan [1], Aslıhan Ünal [2] & Fatih Kayaalp [3]

[1] İstanbul Medeniyet University, Faculty of Engineering and Natural Sciences, Computer Engineering, İstanbul, Türkiye
[2] İstanbul Gelişim University, Faculty of Economics, Administrative and Social Sciences, Department of Business Administration, İstanbul, Türkiye
[3] Düzce University, Faculty of Engineering, Computer Engineering, Düzce, Türkiye

Abstract Customer churn is an important issue in increasing both the long- and short-term revenues. If companies identify customers' churn behavior, they can prevent churn, ensure customer loyalty, and, in turn, gain better financial returns. The telecommunications sector is a customer-oriented sector that requires customer retention to survive in the market. In this sector, customer churn is observed at a high level. In recent years, artificial intelligence-based customer churn analysis has been widely used to predict customer churn behavior. In this study, a customer churn analysis was conducted using publicly shared Telco telecommunications data. Predictive models were constructed using machine learning (LR, KNN, SVM, DT, RF, ANN), ensemble learning (XGBoost, Majority Voting), and deep learning (LSTM) methods. In addition, a 3-layered LSTM model was proposed. Accuracy (Acc), F1-score (F1), Precision (Prec), and Recall (Rec) rates were used to evaluate the models. As a result, the novel 3-layered LSTM model achieved 91.90% Acc, 91.49% Prec, 92.31% Rec, and 91.90% F1 values. The proposed model is competitive with the existing models.

Keywords Customer Churn Analysis · Ensemble Learning · Machine Learning · Deep Learning · Telecommunication

Citation: Başarslan, M. S., Ünal, A. & Kayaalp, F. (2025). A novel and robust LSTM model for customer churn analysis using deep, machine learning, and ensemble learning: A telecommunications case. *Acta Infologica*, *9*(1), 55-73. https://doi.org/10.26650/acin.1584030

This work is licensed under Creative Commons Attribution-NonCommercial 4.0 International License.
© 2025. Başarslan, M. S., Ünal, A. & Kayaalp, F.

Corresponding author: Muhammet Sinan Başarslan muhammet.basarslan@medeniyet.edu.tr

Introduction

Incremental developments in Information Technologies (IT) and the Internet have led to radical changes in consumer behaviors, organizational structures, and business processes. The new age is based on a knowledge-based view (KBV), and in this age, transforming data into information for decision-making is a vital strategic tool. In data-based decision-making, executives use the outcomes of predictive data analysis to make predictions and provide recommendations for current and future activities (Kozak et al., 2021).

The rise of e-commerce in the mid-1990s triggered fierce competition. In such an environment, retaining customers has become even more difficult for companies. The emergence of online businesses has led to new revenue models. Web catalogs, digital content, advertising-supported, advertising-subscription mixed, and fee-based models are new sources of income for online businesses (Gary Schneider, 2016). E-commerce companies require appropriate methods for predicting consumer behavior and decreasing customer churn rates to survive or gain a competitive advantage in the market. Customer churn analysis is a widely employed method in the sectors that adopt subscription-based income models as telecommunication, insurance, and banking (Osmanoglu, 2019).

In today's world, speed, easiness, user-friendly applications, and websites have a positive impact on consumer experience by providing access to information in the right place at the right time. These positive user experiences play an important role in consumers' buying decisions (PwC consumer Experience Series, n. d.). Customers can react in three dimensions when their expectations are not satisfied or they cannot experience the user experience they desire: Exit, voice, and loyalty. Customer churn is the behavior a customer engages in to end his/her relationship with the company (e.g., unsubscribe, transfer, or suspend the subscription), and when the customer leaves the company and starts to buy another company's product, it is called customer churn (Fu, 2022). In the case of customer churn, the customer can be convinced to stay with the company by implementing customized marketing strategies (Cao et al., 2019).

The saturation of markets due to globalization gave rise to a highly competitive environment in the business world. As a result, the conversion costs of customers decreased, and the customer acquisition costs of companies increased (Ahn et al., 2020). Because customer loyalty directly affects company revenues, it is one of the primary goals of customer relationship management (CRM). As acquiring new customers is more costly than retaining existing customers, companies conduct customer churn prediction analysis (Karvana et al., 2019).

Customer churn analysis is used widely in various sectors to increase the level of predictability (Karvana et al., 2019). and plays an important role in the sectors that adopt a subscription-based revenue model as telecommunication, banking, and insurance (Osmanoglu, 2019). The telecommunications sector witnesses high churn rates in its highly competitive market environment (Su et al., 2022), and identifying churn risk is required for retaining existing customers and gaining a competitive advantage (Edwine et al., 2022).

Customer churn in the telecommunications sector can be defined as customers who tend to unsubscribe from a company's service. Telecommunication companies need to know the reasons for churns, and this information can be extracted from the data of the customers (Gaur & Dubey, 2018). Customer churn behavior in the telecommunication sector came into prominence in the 2000s in the literature, and "churn prediction and modeling" have become one of the most examined areas (Bhattacharyya & Dash, 2022). Research in this field includes building classification models using traditional machine learning methods (eg. Decision Tree-DT, Random Forest- RF, Naïve Bayes-NB, Logistic Regression-LR, Support Vector Machine-SVM) and/or deep

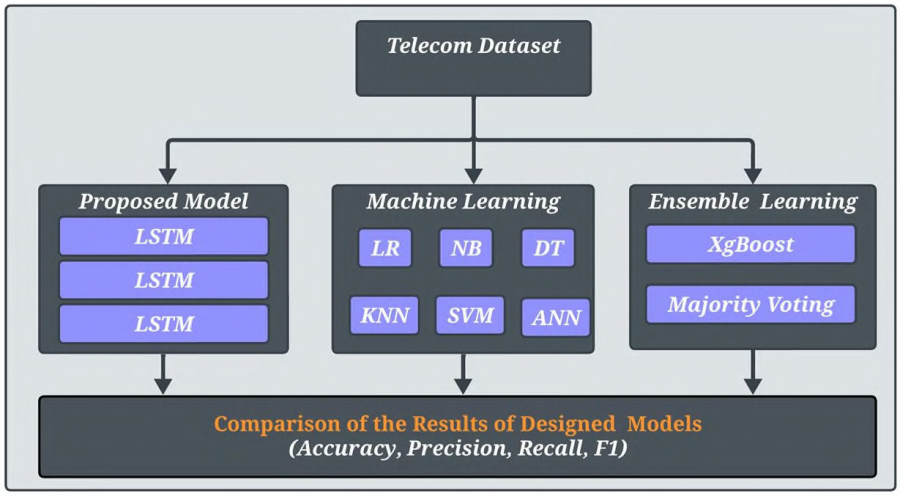
learning algorithms (e.g.. Artificial Neural Network-ANN, Convolution Neural Network (CNN), and determining the Acc levels of the models using various evaluation metrics (Mishra & Reddy, 2017), (Indulkar & Patil, 2021). Because of developments in big data technology, models based on deep learning algorithms increase prediction efficiency by reducing processing time and providing higher accuracy (Mishra & Reddy, 2017).

The performance of predictive models based on traditional machine learning algorithms, deep learning algorithms, and hybrid algorithms has played an important role in the literature. For example, Gaur and Dubey compared the performance of several classification models (LR, SVM, RF, and gradient-boosted tree) (Gaur & Dubey, 2018). Findings show that the Gradient boosting model is the best model with a performance of 84.57% AUC. In addition, in Zatonatska et al.'s research, all four models were defined as "high quality" with Acc rates of over 80%. have Acc rates (Zatonatska et al., 2023). Jeyaprakaash and Sashirekha compared predictive models based on Adaboost and K-Nearest Neighbor (KNN) algorithms and as a result, the Adaboost model with an accuracy level of 90% was defined as the best predictive model. (Jeyaprakaash & Sashirekha, 2022). Bayrak et al. compared two deep learning models based on CNN and ANN algorithms, and as a result, the CNN model with a 97.62% accuracy rate was defined as the best predictive model. Bayrak et al. compared a hybrid model based on CNN and long short-term memory (LSTM) algorithms to an RNN model and found that the hybrid model demonstrated the best performance model (Bayrak et al., 2021).

If churn customers can be predicted as much as possible as earlier and accurately, it will be possible to plan new marketing offers, for whom customers who are considering leaving can be predicted in advance and accurately, making it possible to plan and implement new marketing campaigns and offers to increase the loyalty of these customers. Therefore, the classification performance of churn customer analysis is a crucial issue. Various factors affect model success. The types of attributes in the dataset, missing/incorrect data, conformity of data types to the algorithms used, and parameter optimizations of the algorithms are some of these factors. Hence, the processes related to these factors should be examined through various experiments, and appropriate models should be constructed. To objectively evaluate the success of the models, it would be useful to compare them with the findings of similar research in the literature.

In this study, experiments were conducted on an open-source Telco dataset to compare the performances of the widely used machine learning, deep learning, and ensemble learning methods for predicting customer churn. The experimental results were evaluated according to the Acc, Prec, and F1 metrics. As a result, we observed that the LSTM deep learning method, which is an emerging technology, provides better results than traditional machine learning methods. The flow diagram of the study is shown in Figure 1.

Figure 1
Flow diagram of the study



Customer Churn Management

Customer churn management (CCM) aims to predict the customers who most likely intend to transfer to competing service providers and convince them to stay with the company by providing customized offers to these customers. (Hadden et al., 2007). There are two kinds of churn customers for companies (Wanchai, 2017): voluntary and involuntary. Involuntary churns include cases in which customers are unwilling to quit. Death, relocation, or termination of subscription due to nonpayment of invoices are examples of involuntary churns. Voluntary churn is the churn of customers because they voluntarily give up the company's goods and services (Ahn et al., 2020). Voluntary customers are also called "unhappy customers" (Kurt et al., 2019).

Research shows that the cost of acquiring a new customer is higher than retaining existing ones. Therefore, customer churn management focuses on convincing voluntary customers to continue to use the company's services. Involuntary churns are excluded from consideration because they are out of the company's control (Wanchai, 2017). Customer churn prediction is performed in various businesses to increase the revenue of the company by increasing the retention rate of customers (Ahn et al., 2020). Telecommunication, banking, insurance, gaming, music streaming, newspaper subscriptions, social media-based services, public relations, and energy are examples of industries where customer churn analysis is utilized (Ahn et al., 2020). Particularly, customer churn prediction has a vital role in industries where value is directly related to active customer amount and requires customer retention in telecommunication, banking, and insurance (Sadi Evren Şeker, 2016). Hence, companies struggle to prevent customer churn by using data mining methods when developing strategies to increase customer loyalty (Guo & Qin, 2017).

The success of CCM is related to predicting voluntary churns and understanding the underlying reasons (Routh et al., 2021). To achieve this goal, companies use data science and machine learning methods to define customer satisfaction, reveal the reasons for dissatisfaction, and predict customer churn (Zatonatska et al., 2023). Predictive models play a key role in developing customer churn strategies and sustaining customer loyalty (Lalwani et al., 2022).

Customer Churn Analysis

Customer churn analysis (CCA), a subfield of customer analytics (Seymen et al., 2023), aims to predict customer churn based on historical data and behavioral patterns and to convince customers by providing customized offerings (Karvana et al., 2019).

Data mining methods and machine learning algorithms are traditionally used in CCA. Developments in big data technology have made it possible to build predictive models based on deep learning and obtain real-time warnings (S. Cao et al., 2019). Data mining is the process of exploring relationships in the data without relying on assumptions or premises. This feature distinguishes it from traditional data analysis (Guo & Qin, 2017). Data mining can be used to classify data into series, patterns, and trends (descriptive mining) and to predict by processing data with various algorithms (predictive mining) (Wanchai, 2017). The behavior of churn customers can be extracted from the data analyzed using data mining methods (Ullah et al., 2019).

A predictive model for churn analysis follows five steps (Wanchai, 2017). The first step involves collecting appropriate data for analysis. Examples of data used in CCA include demographic data, customer account information, call data in the telecommunications sector (Wanchai, 2017), and sales data in the retail sector (Bayrak et al., 2021). In the second step, data are organized and analyzed to explore patterns and relationships. First, a threshold value should be defined to distinguish “churn customer” from “ordinary customer” (Bayrak et al., 2021). The time frame when several customers are inactive is also referred to as the “time window”. Customers who exceed this threshold are identified as “churn customers”. The time window period can change due to the type of services provided by the company (Ahn et al., 2020). Then, training data are prepared for data mining. (Bayrak et al., 2021). The training data can be used as input to a predictive model (Karvana et al., 2019). In the third step, an appropriate data mining method is selected for the given data. Several machine learning methods frequently used in this process are LR, SVM, RF, gradient boosted tree (GBT), KNN algorithms, DT, ANN, rough set approach, linear discriminant analysis, market basket analysis, and sequential pattern mining, NB (Gaur & Dubey, 2018) (Kurt et al., 2019). Data mining can be performed in two ways: supervised and unsupervised. In CCA, the supervised learning method is preferred to classify churning and nonchurning customers (Wanchai, 2017). In the final step, a predictive model is built based on the preferred data mining method, and the results are evaluated and verified in terms of usability and accuracy.

Literature

Essentially, customer churn analysis is a classification problem. Hence, there are many classification-methods-based CCA research in the literature. in the literature. Ullah et al. analyzed two samples of telecom customer data using the Bayes method and found that the test results ranged from 51% to 82.11% depending on the type of sample and iteration level (Amin et al., 2019). Arifin and Samopa conducted customer churn classification research in Indonesia using the SVM method. As a result, the predictive model achieved an 80% accuracy rate. (Arifin & Samopa, 2018). In addition, De Caigny et al. combined LR and DT methods and proposed a hybrid model. Acc rates for individual classifiers were defined as 81.6% for LR and 79.6% and 79.6% for RF, respectively, and with the hybrid model, the result increased to 86% (De Caigny et al., 2018). Coussement et al. built classification models based on various machine learning algorithms such as RF, SVM, DT, SGB, and NB, and they analyzed the customer data of a large telecom service provider in Europe. The Acc levels of the models were found as follows. NB-62%, RF = 63%, SVM = 53%, and SGB = 61%. (Coussement et al., 2017). Zhang et al. The LR, DT, and ANN methods were analyzed to classify churn customers in the

South Asia telecom sector, and as a result, the ANN model achieved a 75.53% Acc level (Zhang et al., 2010). In Azeem et al.'s study, a model based on fuzzy classifiers was applied to a dataset of 600.000 subscribers with 84 attributes, and as a result, the model provided a 57% Acc level (Azeem et al., 2017). Qureshi et al analyzed telecom customer data with LR-model prediction and predicted customer churn with a 78% Acc level (Qureshi et al., 2013). Varan and Bahara's NB model achieved an acc of 68%.

Wanchai analyzed the customer data of a telecommunication company in Thailand using the ANN model and obtained an acc level of 86.13% (Wanchai, 2017). Zatonatska et al. built classification models with LR, RF, SVM, and Extreme Gradient Boosting (XGBoost) and analyzed Telco Telecom customer data using a Python program. As a result, the performance of the models for Acc rates was defined as follows: SVM-82%, RF-80.9%, LR-80.8%, and XGBoost-83% (Zatonatska et al., 2023). Lalwani et al. proposed a six-layered customer churn model to analyze telecommunications customer data. The authors performed preprocessing and attribute selection in the first three layers and selected attributes using a gravitational search algorithm in the third layer. They separated the data into training and test data using the data holdout method (80% and 20%) in the fourth layer. In the last layer, the models were evaluated using the AdaBoost and XGBoost methods. As a result, the AdaBoost model performed at 81.71% acc rate, and the XGBoost model at 80.8% acc rate (Lalwani et al., 2022).

Gaur and Dubey built LR, SVM, RF, and GBT classification models for churn prediction. The data holdout method was applied to the data with rates of 75% and 25%, and as a result, the Acc rates of the models were defined as follows: LR-82.86%; SVM-79.75%; RF-81.26%, and GBT-84.59% (Gaur & Dubey, 2018). Seymen et al. We performed customer churn analysis in the telecommunications industry using an RF model and achieved an Acc rate of 88.63% (Ullah et al., 2019).

The RNN and LSTM deep learning architectures are widely used in many fields due to their high performance and compatibility. For example; energy forecasting of wind turbines (Çelebi & Fidan, 2024), sales forecasting in e-commerce (Ecevit, et.al, 2023), and energy forecasting generated by wind power in SCADA systems (Çelebi & Karaman, 2023). Although LSTM is widely used in classification problems, it has not been applied in the abovementioned churn analysis studies. In order to investigate the performance of LSTM in this domain, we build an LSTM-based model using the proposed method.

Material and Method

This section provides information about the dataset used in this study and the machine learning, deep learning, and ensemble learning methods applied to the dataset.

Dataset

The experiments were conducted on the Telco dataset (Zhuang, 2022), which is an open-source dataset of customer data from a telecommunications company in California in the second quarter of 2022. The Telco customer churn dataset contains data about a fictional telecommunications company operating in California during the third quarter. This dataset comprises 7043 customer records, including customer retention, departure, and sign-ups for services. This dataset contains 21 attributes. Table 1 lists the attributes and their descriptions (Zhuang, 2022).

Table 1
Attributes and their descriptions

Attribute Name	Description
Customer Id	Customer ID
Gender	Gender of the customer
SeniorCitizen	Customers' senior citizenship status
Partner	Does the client have a partner?
Addicts	The presence of dependents
Mission time	The month that he stayed with the company
PhoneService	Telephone service subscription
MultipleLines	Multiple lines
Internet Service	Internet service provider
OnlineSecurity	Online safety status
OnlineBackup	Online backup status
Device Protection	Device protection ownership status
Tech Support	Availability of technical support
StreamingTV	TV subscription status
Streaming Movies	Movie subscription status
Agreement	Contract duration
Paperless Billing	E-billing status
payment method	Customers' payment methods
MonthlyFees	Amount to be paid by the customer per month
TotalCharges	Total fees paid by the customer
Churn	Customer churn status

The dataset was preprocessed for missing data and outliers. Missing data are only present in the Total Charges attribute. This missing data was completed by assigning the median value of the relevant attribute, as is commonly done in the literature (Hanif et al 2017, Mokhairi et al 2016).

Machine Learning

Machine learning is a branch of artificial intelligence that enables computer systems to learn from data, recognize patterns, and make decisions. Machine learning is used in many fields today due to its ability to understand and utilize various data. Machine learning is a revolutionary approach to improve the data analysis and learning capabilities of computers. Machine learning has made solving complex problems more accessible and effective (Öztürk et al., 2022). The machine learning methods used in this study are described below.

Random Forest

RF was developed by Leo Breiman and combines the decisions of multiple trees trained on different training datasets rather than building a single DT (Breiman, 2001).. By creating multiple DTs, each trained on a different dataset, the decisions of these trees are combined, resulting in a more powerful and stable classifier. RF is a popular algorithm that provides successful results for various machine learning problems, such as classification and regression. It is also favored because of its ability to work effectively on high-

dimensional and large datasets. RF is widely used in data analysis because it has various applications and is generally resistant to overfitting (Breiman, 2001).

Support vector machine

SVM can be defined as a vector space-based machine learning method that finds a decision boundary between two classes and places this decision boundary at the furthest point from any point in the training data. SVM is a structural risk minimization approach in statistical learning theory. It is noteworthy that SVM is a structural risk minimization approach in statistical learning theory. One of the basic assumptions of the SVM is that all samples in the training set are independent of each other and are distributed similarly (Ahmed et al., 2023a). This property makes SVM an effective solution for linear and nonlinear data classification and regression problems. Due to its high performance and generalizability on low-dimensional and high-dimensional data sets, SVM has a wide range of applications and is a powerful tool for machine learning and data analysis (Ahmed et al., 2023a).

Decision tree

DT is an important tool in data analysis and machine learning. Used in several data mining processes, decision trees facilitate information extraction with the ability to transform complex data structures into simple decision rules. DTs provide an effective way to make data-driven decisions by analyzing features in a dataset (Uzun et al., 2012). The DT building process is a fundamental part of data mining processes. In the first step, the data are preprocessed, and feature selection is performed. Then, the tree-building process begins, where nodes are created to identify the best discriminative decisions based on the features in the dataset. However, the tree was pruned to prevent overfitting and ensure generalizability. Dts have yielded successful results in many fields, such as disease diagnosis in medicine, risk assessment in finance, and customer segmentation in marketing. As a result, DTs are an important tool that provides valuable information in data-driven decision-making processes and constitute a cornerstone of data analysis (Uzun et al., 2012).

Artificial Neural Networks

ANNs are artificial model systems inspired by the biological nervous system and provide impressive results in the fields of machine learning and data analysis. These artificial models can be used in various tasks due to their data processing and pattern recognition capabilities (Ustebay et al., 2019).

LR is a classification technique that is widely used in statistical analysis and machine learning. The proposed method, which is generally used in binary classification problems, attempts to estimate the probability of a data point belonging to a particular class. This estimation is realized by determining a threshold value of the probability values and classifying them. LR derives its name from the "logit" function. This function maps probabilities to an unlimited range of real numbers. The primary purpose of the model is to estimate the probability of a given event occurring based on the given input characteristics. Therefore, the output value is a probability value and lies between 0 and 1 (Baltsou et al., 2022).

At the heart of LR is the idea of finding a line or hyperplane (this line or plane is called the "decision boundary") that best separates the dataset. This line or plane attempts to best capture the separation between classes. The training process involves adapting the parameters (weights) of the model to the data. At the end of training, the model can be used to predict new data points (LaValley, 2008).

With a subfield called deep learning, deep structured networks can produce outstanding results by extracting features from large and complex datasets. ANNs play an important role in processing large amounts of data and understanding complex relationships (Öztürk et al., 2022).

Logistic Regression

LR is a classification technique that is widely used in statistical analysis and machine learning. The proposed method, which is generally used in binary classification problems, attempts to estimate the probability of a data point belonging to a particular class. This estimation is realized by determining a threshold value of the probability values and classifying them. LR derives its name from the "logit" function. This function maps probabilities to an unlimited range of real numbers. The primary purpose of the model is to estimate the probability of a given event occurring based on the given input characteristics. Therefore, the output value is a probability value and lies between 0 and 1 (Baltsou et al., 2022).

At the heart of LR is the idea of finding a line or hyperplane (this line or plane is called the "decision boundary") that best separates the dataset. This line or plane attempts to best capture the separation between classes. The training process involves adapting the parameters (weights) of the model to the data. At the end of training, the model can be used to predict new data points (LaValley, 2008).

K-Nearest Neighbor

The KNN algorithm is widely used for pattern recognition, classification, and regression. The proposed KNN algorithm has a simple structure. When a data point needs to be classified or predicted, the k nearest neighbors around it are determined. The class labels or values of these neighbors are examined. In the classification task, the class to which the largest number of neighbors belongs is selected. In the case of regression, the values of the neighbors are averaged, and prediction is performed (Ahmed et al., 2023b).

Deep Learning

Deep learning is a branch of machine learning that aims to automatically learn data analysis, pattern recognition, and predictive capabilities using complex model structures, such as artificial neural networks. By processing large datasets, the proposed method can identify more complex and high-level features. This can lead to more accurate results and better understanding (Zavrak & Iskefiyeli, 2023). Deep learning is characterized by making the layers of artificial neural networks wider and deeper. These neural networks use many intermediate layers to process the data, allowing more complex features to be discovered. The learning process involves automatically adjusting weights and features using large amounts of data. Deep learning attempts to extract features directly from the data. This makes it possible to learn more information and achieve better results. The application areas of deep learning are wide-ranging. It has been successfully used in many fields, such as image recognition, audio processing, natural language processing, medical diagnosis, and autonomous vehicles. For example, deep learning models can identify objects in images, understand text, and diagnose diseases (Zavrak & Yilmaz, 2023). The LSTM deep learning method is described in this section.

Long Short-Term Memory

LSTM is a powerful recurrent neural network developed for deep learning, especially for processing sequential data (Hochreiter & Schmidhuber,1997). LSTM has achieved great success in areas such as natural language processing, time series analysis, and video processing. The proposed model is designed to

overcome the traditional methods for learning and understanding sequential data (Tan et al., 2022). The key feature of LSTM is its ability to understand long-term dependencies in sequential data. This determines the superiority of LSTM over regular Recurrent Neural Network (RNN) (Malik et al., 2022).

The LSTM cell has three components; Forget gate, Input gate, and Output gate. The forget gate controls the length of the forgotten cell state. The input gate sets the insertion of new information. The exit gate sets the time at which the updated cell state is used.

Ensemble Learning

Ensemble learning is used in many fields such as social networks, communication networks, and biological networks. With this method, groups within a network and their relationships with each other can be better understood. Ensemble learning algorithms identify communities by identifying similar nodes or grouping nodes according to the strength of their connections to each other. The ensemble learning methods used in this study are described below (Başarslan & Kayaalp, 2023).

Xgboost

XGBoost is an ensemble learning algorithm used in machine learning. It demonstrates high performance, especially in classification and regression problems. XGBoost offers a more powerful and faster solution than previous Gradient Boosting methods.

XGBoost works as a tree-based algorithm. The model combines multiple weak learners (usually DTs) to construct a robust prediction model. The algorithm performs the learning process with an approach that focuses on the sequential addition of trees and corrects the errors of previous trees (Chen & Guestrin, 2016).

Voting

Voting is a machine learning and statistical forecasting method. In this method, the predictions of different prediction models or classifiers are combined, and the result is the prediction of the majority vote. Voting was used to compensate for the weaknesses of the models used alone and to obtain more reliable predictions. There are two types: majority voting and probabilistic voting (J. Cao et al. (2012).

Majority voting (Hard Voting): This is a method in which most predictions produced by different models are taken. For example, we consider three different classifiers: A, B, and C. If models A and B predict an instance as Class 1, while model C predicts it as Class 2, the Majority Voting method predicts it as Class 1 (Mohammadifar et al., 2022).

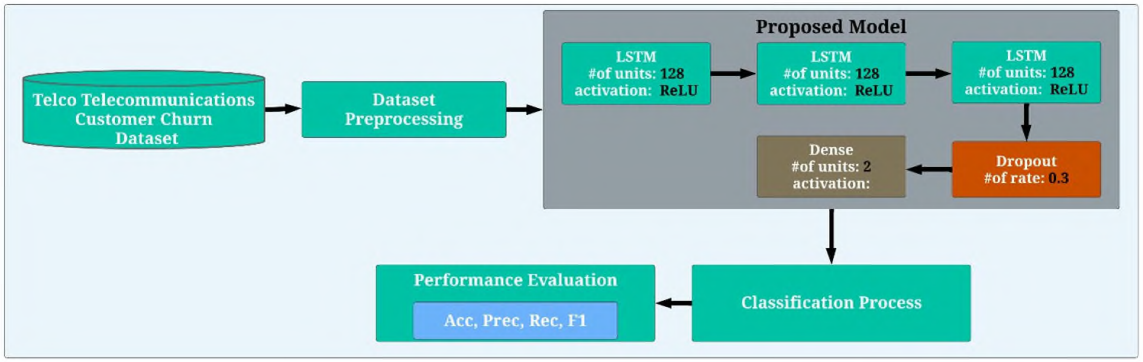
Probabilistic Voting (Soft Voting): This method takes the average or weighted average of the forecast probabilities of different models. The proposed method can obtain more accurate predictions by considering the reliability levels of the models. The prediction probabilities of each model are combined, and the class with the highest probability is predicted (J. Cao et al., 2012; Mohammadifar et al. (2022)

Voting ensemble learning is used to combine the predictions of different prediction models or classifiers to obtain more robust and reliable results. In this study, all machine learning methods were used with majority voting.

Proposed Model

In the proposed model, three LSTM neural networks were used. A graphical representation of this model is shown in Figure 2.

Figure 2
Proposed model



As shown in Figure 2, the Telco dataset was preprocessed for missing and outlier data, and then scaled and prepared for the classification process. The prepared dataset was then modeled with a 3-layer LSTM, dropout layer, and dense layer. The model is then evaluated using performance metrics. Experiments on the model were conducted on Google Colaboratory (Colab) using the TensorFlow 2.9.0 and Keras 2.9 libraries and Python version 3.9.13. The Colab Pro version was used for faster and uninterrupted results of the experiments. In addition, various callbacks were used to create a model with better results. These are ReduceLROnPlateau and Early stopping algorithms. In this study, hyperparameter optimization was performed using KerasTuner. Parameters such as epoch, batch size, activation function, optimization function, number of neural network nodes, and dropout rate are given in Table 2.

Table 2
Hyperparameters of proposed model

Parameters	Values
Batch Size	32, **64**,128
Epoch	**10**, 20, and 30
Activation Functions	**ReLU**, Sigmoid, Tanh,
Optimizaton Functions	**Adam**, RMSprop, and SGD
Number of Units	64, **128**, and 256
Learning Rate Starting Value	**0.001** for all optimizers
Loss Function	Binary cross-entropy
Dropout rate	**0.3**; 0.4; 0.5; 0.6

The parameters shown in bold in Table 2 represent the most optimum parameters with the KerasTuner model. The results of the proposed model with these parameters are given under the experimental results section. The comparison of this model with the results obtained using other machine and ensemble learning methods is also given under the same heading.

Experimental Results and Evaluation

In this section, we describe the metrics used to compare model performance and obtained results.

Performance Evaluation

The performance evaluation of the constituent models was performed using the complexity matrix. The matrix from which Acc and other metrics are obtained is given in Table 3 (Şentürk & Şentürk, 2016).

Table 3
Confusion matrix

		Actual value	
		Positive	**Negative**
Predicted Value	Positive	TP	FP
	Negative	FN	TN

The Acc of the model is expressed as the ratio of correctly classified samples to the total number of samples. High Acc indicates that the model has a high ability to predict correctly, whereas low Acc indicates that improvements in the model's performance are needed. The Acc value is given in Equation (1) (Şentürk & Şentürk, 2016).

$$\mathrm{Acc} = \frac{T_P + T_N}{T_P + F_P + F_N + T_N} \tag{1}$$

Prec denotes the proportion of samples predicted by the model as positive to true positives. A high Prec indicates that the number of false positive predictions is low, and most of the samples classified as positive are positive. Prec is calculated as in Equation (2)(Sinan Basarslan & Kayaalp, 2021).

$$\mathrm{Prec} = \frac{T_P}{T_P + F_P} \tag{2}$$

The Rec value is a performance measure used in classification problems and refers to the rate at which all true positive examples are predicted as positive. Prec is an important measure to reduce the number of false negatives and minimize the number of missing true positive examples. Rec is calculated using Equation (3) (Sinan Basarslan & Kayaalp, 2021)

$$\mathrm{Rec} = \frac{T_P}{T_P + F_N} \tag{3}$$

Mathematically, the harmonic mean of Rec and Prec defines F1. Equation (4) provides the formula for F1 (Şentürk & Şentürk, 2016).

$$\mathrm{F1} = \frac{2^* \ \mathrm{Recall} \ * \ \mathrm{Precision}}{\mathrm{Recall} + \mathrm{Precision}} \tag{4}$$

In this study, the Acc performance criterion was used in the classification models created for the customer churn. A comparison of the literature on similar studies on the same dataset using the Acc criterion is provided in the conclusion and discussion section.

Experimental Results

Experimental results from a 5-fold cross-validation decomposition on the Telco dataset are presented in this section. Within the scope of this research, the results of experiments on models created using machine learning, deep learning, and ensemble learning in the Google Colaboratory environment are given in Table 4.

Table 4

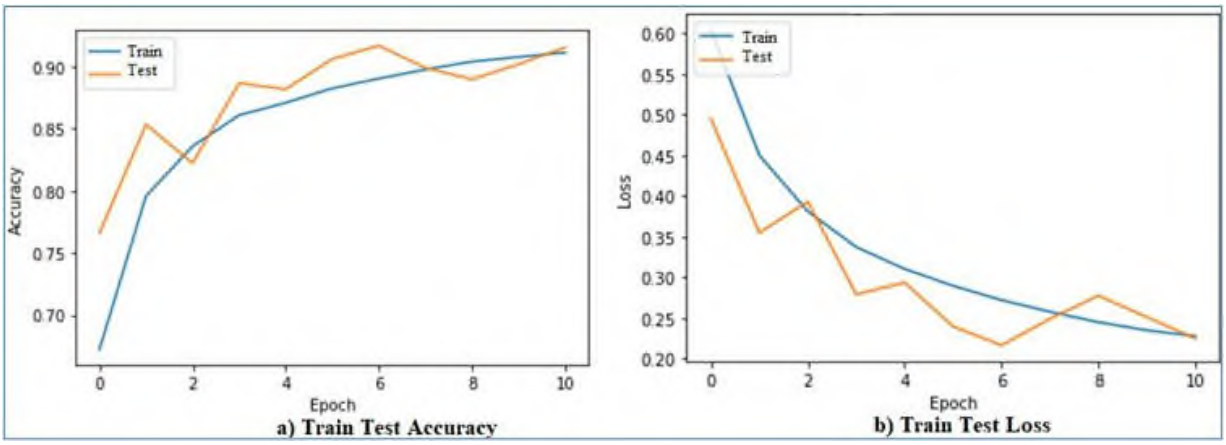
Experimental results of models on Telco Telecom dataset

Category	Classifiers	Acc	Prec	Rec	F1
Machine Learning	LR	80.48	88.48	78.55	83.20
	KNN	77.71	85.86	76.36	80.83
	SVM	80.62	88.70	78.60	83.34
	DT	80.70	88.89	78.43	83.33
	RF	80.84	89.15	78.48	83.48
	ANN	84.95	90.33	74.24	81.50
Ensemble Learning	XGBoost	82.97	86.77	72.97	79.27
	Majority Voting	80.98	81.52	72.41	76.70
Deep Learning	LSTM	89.03	94.32	81.91	87.68
	Proposed Model	**91.90**	**91.49**	**92.31**	**91.90**

According to Table 4, the Proposed Model achieved the highest accuracy (91.90%) among all methods applied on the Telco Telecom dataset in training-test separation with 5-fold cross-validation. In the deep learning category, the LSTM model achieved an accuracy of 89.03%, which is lower than the Proposed Model but higher than all other methods. Among the machine learning methods, RF obtained the highest accuracy (80.84%), followed closely by DT (80.70%), SVM (80.62%), and LR (80.48%). However, the proposed KNN method lagged other methods with an accuracy of 77.71%. Ensemble learning methods demonstrated better performance than machine learning methods, with XGBoost achieving an accuracy of 82.97% and the Majority Voting classifier achieving 80.98%. The training-validation accuracy and loss graph for the proposed model are shown in Figure 3.

Figure 3

Acc and loss graphics during training and testing.



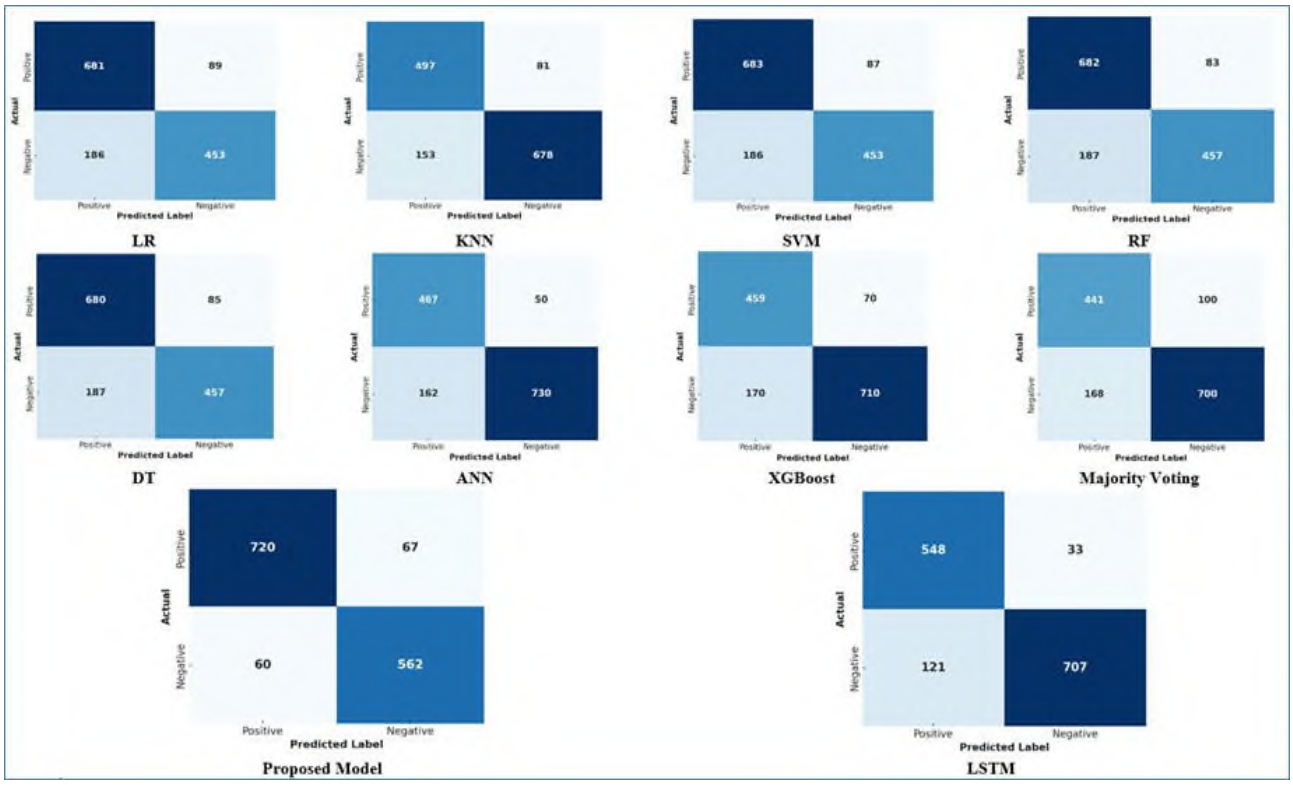
In Figure 3, we present the training and test accuracy and loss graphs over 10 epochs to illustrate the learning dynamics of the proposed model. In the training and test accuracy graph (Figure 3 a), both curves show a steady increase, with the test accuracy exceeding the training accuracy in the early epochs (e.g., 2-4). This demonstrates that the model generalizes well from the outset, possibly due to effective regularization techniques By epoch 10, the training and test accuracy converged, demonstrating minimal overfitting and strong generalization to the test data. In the training and test loss plots (Figure 3 b), both losses decreased

significantly during the first epochs, indicating that the model effectively optimized the error function. The test loss stabilized after epoch 6, indicating the consistency of the model in handling unseen data. More importantly, there was no discernible deviation between the training and test loss, which further supported the robustness and stability of the model across the datasets.

Taken together, the two plots in Figure 3 (3a and 3b) confirm that the proposed model learns effectively, avoids overfitting, and generalizes well, making it reliable for predictions on unseen data. The confusion matrices of the models that were created in the study are shown in Figure 4.

Figure 4

Results of the confusion matrix of the models



Limitations of The Study

The dataset itself is a limitation because the performance of the model may vary depending on the characteristics of the dataset. Another limitation is that the domain of the dataset is defined as communications because customer churn is also important for other sectors, such as banking or subscription-based domains. At the same time, the selected deep learning and machine learning methods can also be considered limitations.

Conclusion and Discussion

In today's market conditions, the value of existing customers for companies is quite high. Therefore, companies attach great importance to customer disruption analysis. There are many studies in this field in the literature. In this study, churn analysis experiments were conducted on the Telco Telecommunication open dataset with models based on some of the abovementioned machine learning, deep learning, and

ensemble learning methods. The Acc, Prec, Rec, and F1 performance metrics were used to evaluate the experimental results.

The results of the experiments performed on the Telco dataset using the above methods are presented in Table 4. When Table 4 is analyzed, it can be seen that the method that gives the best Acc result is the three LSTM models, which consist of a hybrid of deep learning methods with 91.90%. The comparison of the best results obtained from the models developed in this study with those from other studies in the literature on the same dataset is presented in Table 5.

Table 5
Comparison with literature

Model	Acc (%)
(Azeem et al., 2017)	57.00
(Coussement et al., 2017)	63.00
(Varan & Behara, 2014)	68.00
(Zhang et al., 2010)	75.53
(Qureshi et al., 2013)	78.00
(Arifin & Samopa, 2018)	80.00
(De Caigny et al., 2018)	81.60
(Lalwani et al., 2022)	81.71
(Amin et al., 2019)	82.11
(Zatonatska et al., 2023)	83.00
(Gaur & Dubey, 2018)	84.59
(Wanchai, 2017)	86.13
(Ullah et al., 2019)	88.63
Proposed Model	**91.90**

It can be seen that this study competes with the literature by examining the values given in Table 5. Future studies are planned to evaluate the performance of other deep learning methods, which are cutting-edge technology, in customer churn analysis studies.

Although churn analysis is focused on the telecom sector, which is a subscription-based industry, it is also relevant to industries such as banking, insurance, and healthcare. In addition, it is expected that different deep learning methods such as LSTM can be applied as single or ensemble learning. The proposed model stands out for its high accuracy and generalizability. Tested with 5-fold cross-validation, the model demonstrated consistent performance for different data partitions and reduced error rates for critical business problems, such as customer churn prediction. This demonstrates both the analytical accuracy and operational flexibility of the proposed model. As a result, companies will be able to predict churns with greater accuracy and reduce losses as much as possible. These studies provide companies with an advantage in all types of actions, including marketing strategies.

Investigating the effects of Bidirectional Long Short-Term Memory Networks (BiLSTM) and transfer learning methods on enhancing model performance are planned as future work.

Peer Review Externally peer-reviewed.
Author Contributions Conception/Design of Study- M.S.B., A.Ü., F.K.; Data Analysis/Interpretation- A.Ü., M.S.B., F.K.; Drafting Manuscript- A.Ü., M.S.B., F.K.; Critical Revision of Manuscript- A.Ü., F.K., M.S.B.; Final Approval and Accountability- M.S.B., A.Ü., F.K.
Conflict of Interest The authors have no conflict of interest to declare.
Grant Support The authors declared that this study has received no financial support.

Author Details

Muhammet Sinan Başarslan
[1] İstanbul Medeniyet University, Faculty of Engineering and Natural Sciences, Computer Engineering, İstanbul, Türkiye
0000-0002-7996-9169 muhammet.basarslan@medeniyet.edu.tr

Aslıhan Ünal
[2] İstanbul Gelişim University, Faculty of Economics, Administrative and Social Sciences, Department of Business Administration, İstanbul, Türkiye
0000-0001-5896-8880

Fatih Kayaalp
[3] Düzce University, Faculty of Engineering, Computer Engineering, Düzce, Türkiye
0000-0002-8752-3335

References

Ahmed, R., Bibi, M., & Syed, S. (2023a). Improving Heart Disease Prediction Accuracy Using a Hybrid Machine Learning Approach: A Comparative study of SVM and KNN Algorithms. *International Journal of Computations, Information and Manufacturing (IJCIM)*, *3*(1), 49–54. https://doi.org/10.54489/ijcim.v3i1.223

Ahmed, R., Bibi, M., & Syed, S. (2023b). Improving Heart Disease Prediction Accuracy Using a Hybrid Machine Learning Approach: A Comparative study of SVM and KNN Algorithms. *International Journal of Computations, Information and Manufacturing (IJCIM)*, *3*(1), 49–54. https://doi.org/10.54489/ijcim.v3i1.223

Ahn, J., Hwang, J., Kim, D., Choi, H., & Kang, S. (2020). A Survey on Churn Analysis in Various Business Domains. *IEEE Access*, *8*, 220816–220839. https://doi.org/10.1109/ACCESS.2020.3042657

Amin, A., Al-Obeidat, F., Shah, B., Adnan, A., Loo, J., & Anwar, S. (2019). Customer churn prediction in telecommunication industry using data certainty. *Journal of Business Research*, *94*, 290–301. https://doi.org/10.1016/j.jbusres.2018.03.003

Arifin, S., & Samopa, F. (2018). Analysis of Churn Rate Significantly Factors in Telecommunication Industry Using Support Vector Machines Method. *Journal of Physics: Conference Series*, *1108*, 012018. https://doi.org/10.1088/1742-6596/1108/1/012018

Azeem, M., Usman, M., & Fong, A. C. M. (2017). A churn prediction model for prepaid customers in telecom using fuzzy classifiers. *Telecommunication Systems*, *66*(4), 603–614. https://doi.org/10.1007/s11235-017-0310-7

Baltsou, G., Tsichlas, K., & Vakali, A. (2022). Local community detection with hints. *Applied Intelligence*, *52*(9), 9599–9620. https://doi.org/10.1007/s10489-021-02946-7

Başarslan, M. S., & Kayaalp, F. (2023). Sentiment analysis with ensemble and machine learning methods in multi-domain datasets. *Turkish Journal of Engineering*, *7*(2), 141–148. https://doi.org/10.31127/tuje.1079698

Bayrak, A. T., Aktas, A. A., Tunali, O., Susuz, O., & Abbak, N. (2021). *Personalized Customer Churn Analysis with Long Short-Term Memory*. 2021 IEEE International Conference on Big Data and Smart Computing (BigComp), 79–82. https://doi.org/10.1109/BigComp51126.2021.00024

Bhattacharyya, J., & Dash, M. K. (2022). What Do We Know About Customer Churn Behaviour in the Telecommunication Industry? A Bibliometric Analysis of Research Trends, 1985–2019. *FIIB Business Review*, *11*(3), 280–302. https://doi.org/10.1177/23197145211062687

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Cao, J., Lin, Z., Huang, G.-B., & Liu, N. (2012). Voting based extreme learning machine. *Information Sciences*, *185*(1), 66–77. https://doi.org/10.1016/j.ins.2011.09.015

Cao, S., Liu, W., Chen, Y., & Zhu, X. (2019). *Deep Learning Based Customer Churn Analysis*. 2019 11th International Conference on Wireless Communications and Signal Processing (WCSP), 1–6. https://doi.org/10.1109/WCSP.2019.8927877

Chen, T., & Guestrin, C. (2016). *XGBoost*. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. https://doi.org/10.1145/2939672.2939785

Coussement, K., Lessmann, S., & Verstraeten, G. (2017). A comparative analysis of data preparation algorithms for customer churn prediction: A case study in the telecommunication industry. *Decision Support Systems*, *95*, 27–36. https://doi.org/10.1016/j.dss.2016.11.007

Çelebi, S. B., & Fidan, Ş. (2024). RNN-Based Time Series Analysis for Wind Turbine Energy Forecasting. *International Journal of Engineering and Innovative Research*, *6*(1), 15-28. https://doi.org/10.47933/ijeir.1387314

Çelebi, S. B., & Karaman, Ö. A. (2023). Multilayer LSTM Model for Wind Power Estimation in the Scada System. *European Journal of Technique (EJT)*, *13*(2), 116-122. https://doi.org/10.36222/ejt.1382837

De Caigny, A., Coussement, K., & De Bock, K. W. (2018). A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees. *European Journal of Operational Research*, *269*(2), 760–772. https://doi.org/10.1016/j.ejor.2018.02.009

Ecevit, A., Öztürk, İ., Dağ, M., & Özcan, T. (2023). Short-Term Sales Forecasting Using LSTM and Prophet Based Models in E-Commerce. *Acta Infologica*, *7*(1), 59-70, 2023. https://doi.org/10.26650/acin.1259067

Edwine, N., Wang, W., Song, W., & Ssebuggwawo, D. (2022). *Detecting the Risk of Customer Churn in Telecom Sector: A Comparative Study*. Mathematical Problems in Engineering, 2022, 1–16. https://doi.org/10.1155/2022/8534739

Fu, W. (2022). Research on the construction of early warning model of customer churn on an e-commerce platform. *Applied Mathematics and Nonlinear Sciences*, 0(0). https://doi.org/10.2478/amns.2022.1.00016

Gary Schneider. (2016). *Electronic Commerce (12th Edition)*. Cengage Learning.

Gaur, A., & Dubey, R. (2018). *Predicting Customer Churn Prediction In Telecom Sector Using Various Machine Learning Techniques*. 2018 International Conference on Advanced Computation and Telecommunication (ICACAT), 1–5. https://doi.org/10.1109/ICACAT.2018.8933783

Guo, F., & Qin, H. (2017). Data Mining Techniques for Customer Relationship Management. *Journal of Physics: Conference Series*, *910*, 012021. https://doi.org/10.1088/1742-6596/910/1/012021

Hadden, J., Tiwari, A., Roy, R., & Ruta, D. (2007). Computer assisted customer churn management: State-of-the-art and future trends. *Computers and Operations Research*, *34*(10), 2902–2917. https://doi.org/10.1016/j.cor.2005.11.007

Hanif, A., & Azhar, N. (2017, December). *Resolving class imbalance and feature selection in customer churn dataset*. In 2017 International Conference on Frontiers of Information Technology (FIT) (pp. 82-86). IEEE.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Indulkar, Y., & Patil, A. (2021). *Comparative Study of Machine Learning Algorithms for Twitter Sentiment Analysis*. 2021 International Conference on Emerging Smart Computing and Informatics (ESCI), 295–299. https://doi.org/10.1109/ESCI50559.2021.9396925

Jeyaprakaash, P., & Sashirekha, K. (2022). Accuracy Measure of Customer Churn Prediction in Telecom Industry using Adaboost–Decision Tree Algorithm. *Journal of Pharmaceutical Negative Results*, *13*(SO4). https://doi.org/10.47750/pnr.2022.13.S04.179

Karvana, K. G. M., Yazid, S., Syalim, A., & Mursanto, P. (2019). *Customer Churn Analysis and Prediction Using Data Mining Models in Banking Industry*. 2019 International Workshop on Big Data and Information Security (IWBIS), 33–38. https://doi.org/10.1109/IWBIS.2019.8935884

Kozak, J., Kania, K., Juszczuk, P., & Mitręga, M. (2021). Swarm intelligence goal-oriented approach to data-driven innovation for customer churn management. *International Journal of Information Management*, *60*, 102357. https://doi.org/10.1016/j.ijinfomgt.2021.102357

Kurt, K. K., Atay, H. T., Cicek, M. A., Karaca, S. B., & Turkeli, S. (2019). *The Importance of Multidisciplinary in Data Science: Application of the Method in Health Sector to Telecommunication Sector*. 2019 International Artificial Intelligence and Data Processing Symposium (IDAP), 1–4. https://doi.org/10.1109/IDAP.2019.8875917

Lalwani, P., Mishra, M. K., Chadha, J. S., & Sethi, P. (2022). Customer churn prediction system: a machine learning approach. *Computing*, *104*(2), 271–294. https://doi.org/10.1007/s00607-021-00908-y

LaValley, M. P. (2008). Logistic Regression. *Circulation*, *117*(18), 2395–2399, doi:10.1161/CIRCULATIONAHA.106.682658

Malik, S., Bansal, P., Sharma, P., Jain, R., & Vashisht, A. (2022). *Image Retrieval Using Multilayer Bi-LSTM* (pp. 745–755). https://doi.org/10.1007/978-981-16-2597-8_64

Mishra, A., & Reddy, U. S. (2017). *A Novel Approach for Churn Prediction Using Deep Learning*. 2017 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), 1–4. https://doi.org/10.1109/ICCIC.2017.8524551

Mohammadifar, A., Gholami, H., & Golzari, S. (2022). Stacking- and voting-based ensemble deep learning models (SEDL and VEDL) and active learning (AL) for mapping land subsidence. *Environmental Science and Pollution Research*, *30*(10), 26580–26595. https://doi.org/10.1007/s11356-022-24065-7

Mokhairi, M., Nordin, A.R. & Khalid, A. (2016). Predictive modeling for telco customer churn using rough set theory. *ARPN Journal of Engineering and Applied Sciences*, *11*(5), 3203-3207.

Osmanoglu, O. O. (2019). *Comparing to Techniques Used in Customer Churn Analysis.* In Journal of Multidisciplinary Developments (Vol. 4, Issue 1).

PwC consumer Experience Series. (n.d.). https://www.pwc.de/de/consulting/pwc-consumer-intelligence-series-customer-experience.pdf. https://www.pwc.de/de/consulting/pwc-consumer-intelligence-series-customer-experience.pdf

Qureshi, S. A., A. S. Rehman, A. M., & Qamar, A. (2013). *Telecommunication subscribers' churn prediction model using machine learning*. Eighth International Conference on Digital Information Management (ICDIM 2013), 131–136. https://doi.org/10.1109/ICDIM.2013.6693977

Routh, P., Roy, A., & Meyer, J. (2021). Estimating customer churn under competing risks. *Journal of the Operational Research Society*, *72*(5), 1138–1155. https://doi.org/10.1080/01605682.2020.1776166

Seymen, O. F., Ölmez, E., Doğan, O., Er, O., & Hızıroğlu, K. (2023). Customer Churn Prediction Using Ordinary Artificial Neural Network and Convolutional Neural Network Algorithms: A Comparative Performance Assessment. *Gazi University Journal of Science*, *36*(2), 720–733. https://doi.org/10.35378/gujs.992738

Shi Long Zhuang. (n.d.). *Telecom Churn Dataset*. 2022. https://www.kaggle.com/datasets/shilongzhuang/telecom-customer-churn-by-maven-analytics

Sinan Basarslan, M., & Kayaalp, F. (2021). *Performance Evaluation of Classification Algorithms on Diagnosis of Breast Cancer and Skin Disease* (pp. 27–35). https://doi.org/10.1007/978-981-15-6321-8_2

Su, C., Wei, L., & Xie, X. (2022). *Churn Prediction in Telecommunications Industry Based on Conditional Wasserstein GAN*. 2022 IEEE 29th International Conference on High Performance Computing, Data, and Analytics (HiPC), 186–191. https://doi.org/10.1109/HiPC56025.2022.00034

Şeker, S. E. (2016). Müşteri Kayıp Analizi (Customer Churn Analysis). *YBS Ansiklopedi*, *3*(1).

Şentürk, A., & Şentürk, Z. K. (2016). Yapay Sinir Ağları İle Göğüs Kanseri Tahmini. *El-Cezeri Fen ve Mühendislik Dergisi*, *3*(2). https://doi.org/10.31202/ecjse.264199

Tan, K. L., Lee, C. P., Anbananthen, K. S. M., & Lim, K. M. (2022). *RoBERTa-LSTM: A Hybrid Model for Sentiment Analysis With Transformer and Recurrent Neural Network*. IEEE Access, 10, 21517–21525. https://doi.org/10.1109/ACCESS.2022.3152828

Ullah, I., Raza, B., Malik, A. K., Imran, M., Islam, S. U., & Kim, S. W. (2019). A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector. *IEEE Access*, *7*, 60134–60149. https://doi.org/10.1109/ACCESS.2019.2914999

Ustebay, S., Turgut, Z., & Aydin, M. A. (2019). *Cyber Attack Detection by Using Neural Network Approaches: Shallow Neural Network, Deep Neural Network and AutoEncoder* (pp. 144–155). https://doi.org/10.1007/978-3-030-21952-9_11

Uzun, E., Agun, H. V., & Yerlikaya, T. (2012). *Web content extraction by using decision tree learning*. 2012 20th Signal Processing and Communications Applications Conference (SIU), 1–4. https://doi.org/10.1109/SIU.2012.6204476

Varan, S., & Behara, R. S. (2014). *Customer churn analysis in the wireless industry: A data mining approach*. https://www.researchgate.net/publication/228411974

Wanchai, P. (2017). *Customer churn analysis: A case study on the telecommunication industry of Thailand*. 2017 12th International Conference for Internet Technology and Secured Transactions (ICITST), 325–331. https://doi.org/10.23919/ICITST.2017.8356410

Zatonatska, T., Fareniuk, Y., & Shpyrko, V. (2023). Churn Rate Modeling for Telecommunication Operators Using Data Science Methods. *Marketing and Management of Innovations*, *14*(2), 163–173. https://doi.org/10.21272/mmi.2023.2-15

Zavrak, S., & Iskefiyeli, M. (2023). Flow-based intrusion detection on software-defined networks: a multivariate time series anomaly detection approach. *Neural Computing and Applications*, *35*(16), 12175–12193. https://doi.org/10.1007/s00521-023-08376-5

Zavrak, S., & Yilmaz, S. (2023). Email spam detection using hierarchical attention hybrid deep learning method. *Expert Systems with Applications*, *233*, 120977. https://doi.org/10.1016/j.eswa.2023.120977

Zhang, X., Liu, Z., Yang, X., Shi, W., & Wang, Q. (2010). *Predicting customer churn by integrating the effect of the customer contact network*. Proceedings of 2010 IEEE International Conference on Service Operations and Logistics, and Informatics, 392–397. https://doi.org/10.1109/SOLI.2010.5551545

İSTANBUL UNIVERSITY PRESS

Acta Infologica,
ACIN 2025, 9 (1): 74–89

https://doi.org/10.26650/acin.1604272
Submitted: 19.12.2024
Revision Requested: 30.01.2025
Last Revision Received: 18.02.2025
Accepted: 24.02.2025
Published Online 26.02.2025

Acta Infologica

Research Article

Open Access

Optimizing the Sentiment Recognition in Spotify Playlists Through Ensemble-Based Approaches

M. Erdem İsenkul [1]

[1] İstanbul University-Cerrahpaşa, Faculty of Engineering, Department of Computer Engineering, İstanbul, Türkiye

Abstract Spotify, with over 320 million monthly active users reported in 2020, offers a unique platform for data science and machine learning applications. This study leverages Spotify's extensive music library of over 50 million songs to analyze the emotional tone of user-created playlists using machine learning algorithms. By employing advanced classification methods, including Random Forest, Decision Tree, and Support Vector Machines (SVM), the research compares their effectiveness in sentiment classification tasks. The Random Forest model achieved the highest test accuracy of 87%, closely followed by the Decision Tree model at 86%. These results highlight the potential of sentiment-informed data to enhance music recommendation systems by tailoring suggestions to users' emotional preferences. This work not only contributes to the evolving domain of sentiment-aware recommendation models but also demonstrates the technical challenges and practical implications of applying machine learning in music streaming platforms. The study's findings underscore the value of integrating emotional intelligence into recommendation algorithms to improve user engagement and satisfaction in digital music services.

Keywords Sentiment Analysis · Spotify API · Playlist · Machine Learning · Music Recommendation Systems

Citation: İsenkul, M. E. (2025). Optimizing the sentiment recognition in spotify playlists through ensemble-based approaches. *Acta Infologica*, *9*(1), 74-89. https://doi.org/10.26650/acin.1604272

This work is licensed under Creative Commons Attribution-NonCommercial 4.0 International License.

© 2025. İsenkul, M. E.

Corresponding author: M. Erdem İsenkul eisenkul@iuc.edu.tr

Introduction

The proliferation of digital music platforms has transformed how individual access, curate, and experience music. Among these platforms, Spotify, with its expansive user base and extensive music catalog, stands out as a dynamic ecosystem for exploring user engagement, musical preferences, and emotional expression through playlist creation. As playlists are inherently personalized, they often reflect the moods and emotional states of their creators, offering an opportunity to delve into the sentiment embedded within them.

Sentiment analysis, a branch of natural language processing (NLP), traditionally used for understanding opinions in textual data, now finds meaningful applications in music streaming. Recognizing the emotional tone of playlists not only provides insights into user behavior but also enhances the quality of music recommendations by enabling streaming platforms to deliver more personalized content. Emotionally aware recommendations can improve user engagement, satisfaction, and overall streaming experience. Spotify, with its API offering vast data on playlists, songs, and user activity, provides an ideal environment to study sentiment classification in a musical context.

The primary aim of this study was to classify the sentiment of Spotify playlists using machine learning algorithms. By leveraging a dataset constructed from the Spotify API, we apply a range of classifiers, including DT, RF, and Support Vector Machines (SVM), to explore the sentiment within playlists and assess model performance. Through this research, we seek to contribute to the evolving field of sentiment-aware recommendation systems by addressing both technical challenges and practical applications in music streaming.

An important feature of this research is its exploration of the unique challenges of sentiment analysis on the playlist level of music streaming sites. The extraction of sentiment from the titles of songs is very much unlike sentiment analysis on well-structured text, where the titles of songs are generally brief, ornamental, and contextually reliant. Furthermore, the reasons behind people using playlists are multi-formed, making the overall sentiment determination of a playlist on its titles only complex. The use of sentiment classification by using diverse machine learning models is explored by this research, including its focus on sentiment classification on the playlist level, where very little is seen in the sentiment classification literature up to date. Contrasting previous sentiment classification literature that has mostly been on sentiment classification concerning tracks or lyrics, this contribution presents a new dimension by considering sentiment on the playlist level, hence bringing new ideas on sentiment-based recommendation of music. The findings of this research contribute to making emotionally aware recommendation systems better, hence enhancing the content personalization of streams of music.

This paper is structured as follows: Section 2 reviews the existing literature on sentiment analysis in music and user-generated content. Section 3 outlines the methodology, including data collection, pre-processing, feature extraction, and classifier implementation. Section 4 presents and discusses the experimental results, followed by Section 5, which explores the practical implications and potential for future research. The study concludes by summarizing the findings and their contributions to sentiment analysis and music recommendation systems.

Related Work

The intersection of music recommendation and sentiment analysis has spurred significant interest, leading to the development of various machine learning approaches that enhance user engagement by

personalizing the emotional tone of playlists. This section reviews the foundational work in sentiment analysis, its application in user-generated content such as playlists, and specific methodologies in music recommendation systems.

Sentiment analysis as a field has grown from the initial approaches in opinion mining and review classification. Hu and Liu's (2004) pioneering work established a framework for extracting sentiment from customer reviews, laying a foundation for text-based sentiment analysis in various domains. Pang and Lee (2008) further advanced this area with a comprehensive review of opinion mining methodologies, underscoring the evolution of sentiment classification techniques. Their work highlights the shift toward machine learning models and deep learning methods for handling the intricacies of sentiment within social media and user-generated content.

In the context of music, Chi et al. (2010) explored emotion-based playlist generation using reinforcement learning to create playlists that match specific emotional states. This early attempt to integrate sentiment into playlist curation demonstrated the potential for algorithmic approaches to enhance music personalization. Additionally, Jiang et al. (2011) examined sentiment classification on Twitter, expanding sentiment analysis to social platforms with user-generated data, further bridging the gap between NLP and user engagement applications. Abel et al. (2011) contributed to this domain by applying sentiment-aware models to Twitter for personalized news recommendations, presenting sentiment as a valuable feature in content recommendation systems.

Recent advancements have focused on incorporating sophisticated machine learning models into sentiment analysis for playlist recommendation. Delbouys et al. (2018) applied deep neural networks for mood detection in music based on audio and lyrical features, setting a benchmark for the integration of text and audio data in music mood analysis. Similarly, Schedl et al. (2021) provided a comprehensive overview of music recommendation systems, discussing the challenges and potential of emotion-aware recommendation models to capture the nuanced emotional preferences of users.

In addition to these foundational studies, Ferraro et al. (2018) explored hybrid recommendation systems that combine text and audio features, demonstrating the benefits of using multimodal data for playlist personalization. Their work, along with that of Irene et al. (2019), who used neural networks for automatic playlist generation, has informed modern approaches to sentiment-informed playlist creation. Both studies emphasize the role of feature extraction in accurately capturing the emotional context in music data.

To enrich user modeling, Pappas and Popescu-Belis (2016) developed sentiment-aware collaborative filtering, a method that personalizes recommendations based on user preferences for certain moods or sentiments. Their work, combined with the analysis of common-sense reasoning frameworks like ConceptNet by Liu and Singh (2004), supports the notion that incorporating external knowledge and user-centric features enhances recommendation accuracy in sentiment-driven applications.

Schedl's (2016) introduction of the LFM-1b dataset provided researchers with a valuable resource for evaluating sentiment analysis methods in large music datasets, contributing to benchmarking efforts in the field. The LFM-1b dataset continues to be widely used in music recommendation and sentiment analysis research, enabling the evaluation of models on real-world user behavior.

Within the current research landscape, our study stands out by applying an ensemble of machine learning models, including Random Forest, Decision Tree, and Support Vector Machines (SVM), to the sentiment classification of Spotify playlists. Leveraging Spotify's API and the Spotipy library, we constructed a dataset

of 2000 songs and applied both textual (VADER and TextBlob) and metadata-based feature extraction. By using machine learning classifiers with cross-validation techniques, including Leave-One-Subject-Out (LOSO) validation, our approach not only addresses playlist sentiment classification but also evaluates model robustness across user preferences, contributing to sentiment-aware recommendation systems in music streaming platforms.

In summary, this work builds upon existing sentiment analysis methodologies (Hu & Liu, 2004; Pang & Lee, 2008) and incorporates recent advances in music recommendation systems (Ferraro et al., 2018; Delbouys et al., 2018). By expanding sentiment analysis to the realm of playlist curation, we aim to enhance user engagement in music streaming applications, providing a foundation for further research in emotion-driven recommendation systems.

Material and Methods

This section details the methodology for extracting, preprocessing, and analyzing playlist data to classify sentiment in Spotify playlists. The approach integrates various natural language processing (NLP) techniques, feature extraction strategies, and machine learning models to ensure robust sentiment classification.

Dataset description

The dataset used in this study was curated by extracting songs from publicly accessible playlists on Spotify, leveraging the Spotify Web API to obtain a comprehensive set of features. The dataset comprises 2000 songs, carefully selected to include an equal number of tracks representing positive and negative emotions. This dataset forms the basis for exploring sentiment classification within Spotify playlists, with each song labeled according to its emotional tone, as supported by previous research on sentiment analysis of user-generated content (Abel *et al*., 2011; Chi *et al*., 2010).

To construct the dataset, the Spotipy library, a Python client for the Spotify Web API, was employed. Spotipy enables access to Spotify's extensive music catalog and metadata, including song titles, artist information, album details, and playlist characteristics. Through the Spotify API's RESTful architecture, JSON metadata about each song was retrieved, which included not only textual information but also user-defined playlist attributes, artist popularity, and genre distribution, enriching the dataset with contextual features (). Studies have shown that this type of metadata can enhance sentiment classification by providing a context that aligns with the emotional tone of user-generated playlists (Ferraro *et al*., 2018; Howells & Ertugan, 2017).

To achieve diversity along with the representativeness of the dataset, there is a mix of musical genres within the playlists. Particularly, there is around 30% of pop, 25% of rock, 15% of electronic, 10% of jazz, 10% of classical, and 10% of others, hence allowing diverse musical content within diverse categories. The diverse musical content is important to ensure that there is no bias that might result from domination by one kind of musical genre, while also promoting the overall generalizability of the sentiment classification schemes.

Additionally, the playlists were taken from users with diverse engagement levels ranging from occasional listeners to actively engaged editors of music constantly changing and adjusting respective playlists. The practice ensures that the dataset represents the diverse listening habits of people, hence making the findings applicable to real-world situations of recommending music.

Sentiment labeling

The labeling process was essential for categorizing songs as having either positive or negative sentiment. Two sentiment analysis libraries, VADER (Valence Aware Dictionary and sEntiment Reasoner) and TextBlob, were used to assess and label the emotional tone of each song title. VADER, specifically tailored for sentiment analysis in social media, was selected due to its ability to capture the nuances of informal language, making it well-suited for analyzing user-generated playlists on Spotify (Hu & Liu, 2004; Pang & Lee, 2008). TextBlob, another NLP tool, provided a secondary reference for sentiment labeling, offering valuable cross-validation insights. Ultimately, the VADER-labeled dataset was chosen for further analysis because of its higher alignment with user-perceived emotional tones in music.

Data preprocessing

Before applying the sentiment classification algorithms, the dataset underwent several preprocessing steps to optimize it for machine learning. This included standard NLP techniques such as tokenization, punctuation removal, and conversion of all text to lowercase. These steps were crucial for maintaining consistency across the dataset, enabling the machine learning models to accurately interpret and analyze song titles without unnecessary noise. To streamline this process, the NLTK (Natural Language Toolkit) library in Python was used, along with the string manipulation functions for cleaning symbols and extraneous characters. Such preprocessing is essential for sentiment classification, as demonstrated by Pang & Lee (2008) and Schedl, Knees, McFee & Bogdanov (2021), who highlight the importance of clean text inputs in improving classifier performance.

Feature extraction

Feature extraction focused on both textual and metadata attributes, ensuring a holistic approach to sentiment classification. Textual features were derived from song titles and artist names, using techniques such as bag-of-words and TF-IDF (Term Frequency-Inverse Document Frequency) to quantify word importance within the dataset (Turney, 2002). Additionally, word embeddings such as Word2Vec were used to capture semantic relationships, allowing the model to recognize and leverage similarities in sentimentally relevant vocabulary. Metadata features, including playlist creation dates, number of followers, and genre distribution, were incorporated to provide context that may influence playlist sentiment. Studies suggest that metadata features can add contextual value in sentiment classification, especially in user-generated music data (Delbouys *et al*., 2018; Liu & Singh, 2004).

Data distribution and balance

To ensure model reliability, the dataset was evenly divided, with 1000 songs labeled as positive and 1000 as negative. This balance was intentionally maintained to prevent bias toward either class, enhancing the model's ability to generalize across different emotional tones. The dataset spans various genres, including pop, rock, jazz, and electronic, reflecting several user preferences and increasing the robustness of the sentiment classification model (Schedl, 2016). Ensuring a balanced dataset aligns with best practices in machine learning, as imbalance can significantly affect classifier performance by introducing bias (Pappas & Popescu-Belis, 2016).

In summary, the curated dataset captures a diverse spectrum of musical content, enriched with both textual and contextual metadata. The inclusion of balanced sentiment labels, extensive preprocessing, and

feature extraction processes contribute to a comprehensive dataset that serves as a robust foundation for the sentiment classification of Spotify playlists.

Methodology

All experiments were conducted in Python using Jupyter Notebook as the primary development environment. The classification approach relied on implementing multiple machine learning models, each chosen for its strengths in handling text data and binary classification tasks, with a focus on extracting informative features and optimizing model performance through cross-validation and hyperparameter tuning.

The Random Forest classifier was selected for its robustness and ensemble nature, which improves accuracy by reducing overfitting. This model is particularly effective in handling complex datasets by combining multiple decision trees, yielding a more generalized model that performs well with high-dimensional data (Breiman, 2001; Delbouys *et al*., 2018). Random Forests have been shown to enhance performance in sentiment classification tasks by leveraging ensemble learning techniques, which combine predictions from multiple models to improve robustness and accuracy (Pang & Lee, 2008).

Support Vector Machines (SVM) were also applied, given their strength in high-dimensional spaces, which suits sentiment classification tasks where distinct boundaries between classes need to be defined. SVM's performance in text analysis tasks has been well-documented, with studies indicating its effectiveness in distinguishing between sentiment classes (Ferraro *et al*., 2018; Pappas & Popescu-Belis, 2016). SVMs are especially effective when the dataset involves clear class separations, as they seek to maximize the margin between classes, which has been beneficial in both text-based and multimedia sentiment analysis (Schedl, 2016).

Additionally, the Decision Tree classifier was used as a benchmark due to its interpretability and hierarchical structure, providing insights into feature importance and classification pathways (Pang & Lee, 2008; Turney, 2002). Decision Trees are often favored in sentiment analysis for their transparency, as they allow researchers to track classification decisions across different feature splits, contributing to explainable AI practices in recommendation systems (Jiang *et al*., 2011).

XGBoost, a gradient boosting algorithm, was included for its ability to enhance classification accuracy by iteratively correcting errors, which is advantageous for datasets with complex patterns. XGBoost's capacity to handle imbalanced data is well-documented, making it a valuable asset in sentiment analysis where certain emotions may be more prevalent (Chen & Guestrin, 2016; Schedl *et al*., 2021). XGBoost's iterative approach has proven to be highly effective in various domains of sentiment analysis, including music data, where its boosting mechanism addresses potential biases in training data (Howells & Ertugan, 2017).

Logistic Regression was applied as a straightforward linear model, valuable for its interpretability and the insights it provides regarding feature weights. Logistic Regression's scalability and fast training times of logistic regression make it a practical choice for initial classification testing in sentiment analysis, as its linear nature simplifies the interpretability of feature impact on class prediction (Abel *et al*., 2011; Hu & Liu, 2004).

Finally, Multinomial Naive Bayes was employed due to its computational efficiency and effectiveness with high-dimensional data, especially in text classification scenarios. Naive Bayes models have been widely used in sentiment classification, particularly when datasets include sparse text-based features, as they calculate class probabilities to handle high-dimensional spaces (Mumcuoğlu, 2007; Tarone *et al*., 2011).

This model's probabilistic approach serves as a reliable baseline for comparing the performance with more complex algorithms in sentiment analysis (Wang *et al*., 2020).

To ensure reliable evaluation, hyperparameter tuning was conducted for each model using grid search and cross-validation techniques. Leave-One-Subject-Out (LOSO) cross-validation was implemented to assess the generalizability of the models across diverse user preferences, thereby simulating real-world variability in playlist sentiment. This approach helped avoid model bias toward specific user data, a consideration that has proven crucial in similar sentiment classification research (Ferraro et al., 2018; Pang & Lee, 2008; Schedl, 2016). The comparative table for those methods is provided in Table 1.

Table 1

Comparative Analysis of the Classifiers

Classifier	Advantages	Disadvantages
Random Forest	• High accuracy and robustness • Handles large datasets with high dimensionality well • Reduces overfitting by combining multiple trees	• Computationally expensive and slow training • Lack of interpretability • Requires more memory
Logistic Regression	• Simple and easy to implement • Fast training and prediction • Provides probabilities for predictions	• Assumes a linear relationship between the features and output • Limited to linear decision boundaries • Sensitive to outliers
Decision Tree Classifier	• Intuitive and easy to understand • No need for feature scaling • Handles both numerical and categorical data well	• Prone to overfitting, especially on small datasets • Instability, small changes in data can result in a different tree • Lack of smoothness in the decision boundaries
Naive Bayes	• Simple and computationally efficient • Performs well with high-dimensional data • Fast training and prediction	• Assumes independence between features • Sensitivity to irrelevant features
Support Vector Machines (SVM)	• Effective in high-dimensional spaces • Versatile, different kernel functions for various data types • Effective in cases where the number of dimensions is greater than the number of samples	• Computationally expensive, especially with large datasets • Sensitivity to parameter tuning • Lack of transparency and interpretability

The training of these models involved a through exploration of hyperparameter tuning and optimization to enhance their performance. The ensemble of machine learning methods employed in this study aimed to provide a comprehensive understanding of sentiment dynamics within playlists, allowing for nuanced and accurate classification of emotional tones. This methodology lays the foundation for a detailed examination of sentiment patterns, contributing valuable insights to the intersection of machine learning and music sentiment analysis.

Classification

To evaluate the robustness and generalizability of our sentiment classification models, we employed the Leave-One-Subject-Out (LOSO) cross-validation strategy. In this approach, the dataset was divided into folds, each representing the playlists of a specific user. During each iteration, the data from one user was held out for testing, while the playlists from all other users were used to train the models.

This method is particularly valuable for playlist sentiment analysis because it closely resembles real-world conditions where sentiment patterns can vary significantly across individual user preferences. By systematically omitting one user's playlists in each iteration, the models were exposed to a range of user behaviors and musical tastes, enhancing their ability to learn generalizable sentiment dynamics within playlists. LOSO cross-validation enabled us to examine the models' capability to perform consistently across different users, ensuring that the sentiment classification was not overly influenced by specific user profiles. The aggregated results from each iteration offered a thorough assessment of the models' effectiveness in recognizing sentiment patterns across a diverse user base, thereby reinforcing the reliability and applicability of our sentiment analysis framework.

To comprehensively evaluate the performance of our sentiment classification models, we employed several key metrics: accuracy, precision, recall, F1 score, and Matthews Correlation Coefficient (MCC). These metrics provided insights into the models' ability to accurately classify sentiment across varying user playlists, while highlighting specific strengths and potential areas for improvement.

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \tag{1}$$

where TP is the number of true positives, TN true negatives, FP false positives, and FN false negatives. Sensitivity and specificity are statistical measures that assess the accuracy of correctly classified positive and negative instances:

$$\text{sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

$$\text{specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{3}$$

The Matthews Correlation Coefficient (MCC) serves as a quality measure for binary classification in machine learning. It remains stable even when there are significant differences in class densities. MCC functions as a correlation coefficient between the predicted and observed binary classifications, yielding a value between 1 and +1. The formulation of the MCC metric is provided as follows:

$$\text{MCC} = \frac{(\text{TP} \times \text{TN}) - (\text{FP} \times \text{FN})}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}} \tag{4}$$

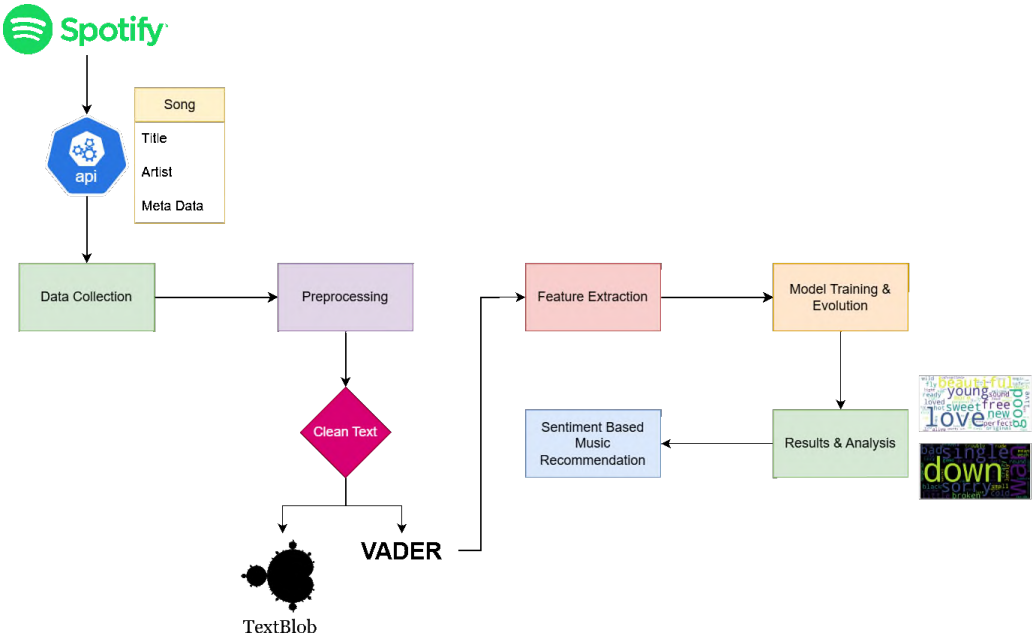
The Matthews Correlation Coefficient (MCC) attains a value of +1 when the classifier makes perfect predictions, –1 when the predictions and actual values are in complete disagreement, and 0 when the classification is no better than a random prediction. By using these metrics, we gained a holistic view of each model's performance, enabling us to identify the most reliable classifiers for sentiment classification in the playlist data. This multifaceted evaluation helped ensure that our models were both accurate and consistent in handling varied sentiment patterns across diverse user playlists.

To present an overall view of the proposed methodology, Figure 1 displays a flowchart that summarizes the key steps of the sentiment classification process. It incorporates acquisition of data through the Spotify

API, preprocessing measures like sanitizing of text and sentiment tagging, feature extraction processes using both textual features along with metadata features, and model training using diverse machine learning classifiers. The evaluation stage includes performance measures on diverse metrics like MCC to determine the most skilled classifier. Finally, the sentiment classification outcomes contribute to enhancing context-based music recommendations. The structured framework ensures that the resilience of the proposed approach is guaranteed along with its applicability.

Figure 1

Detailed Flowchart of the Proposed Method



Experimental Results

In this study, we compiled a dataset of 2000 songs and artist names using the Spotify API. To facilitate the dataset creation, a function was developed to identify and retrieve songs from user playlists by matching the playlist names. Specifically, we focused on a playlist titled *The Longest Playlist Ever*, created by user "Willis Orr" and set to public. This playlist was selected due to its extensive song collection, which provided a rich source of data. Additionally, another playlist containing primarily positive songs was created to serve as the test data.

Once the dataset was constructed, the data were labeled as positive or negative for sentiment classification. The sentiment labeling process employed two libraries, VADER and TextBlob, each of which provided sentiment polarity values for each song title. Before applying these libraries, the dataset underwent data cleaning procedures, including the removal of symbols and punctuation marks and the conversion of all text to lowercase using the NLTK library in Python.

After data cleaning, the sentiment polarity scores were calculated for both VADER and TextBlob, and the songs were labeled accordingly. The initial analysis indicated a predominance of positive sentiment within the data. However, VADER showed a higher frequency of negative labels compared to TextBlob, as shown in Figure 2 and Figure 3. Based on this comparison, we opted to proceed with the VADER-labeled dataset for further analysis.

Figure 2

Sentiment analysis results using VADER

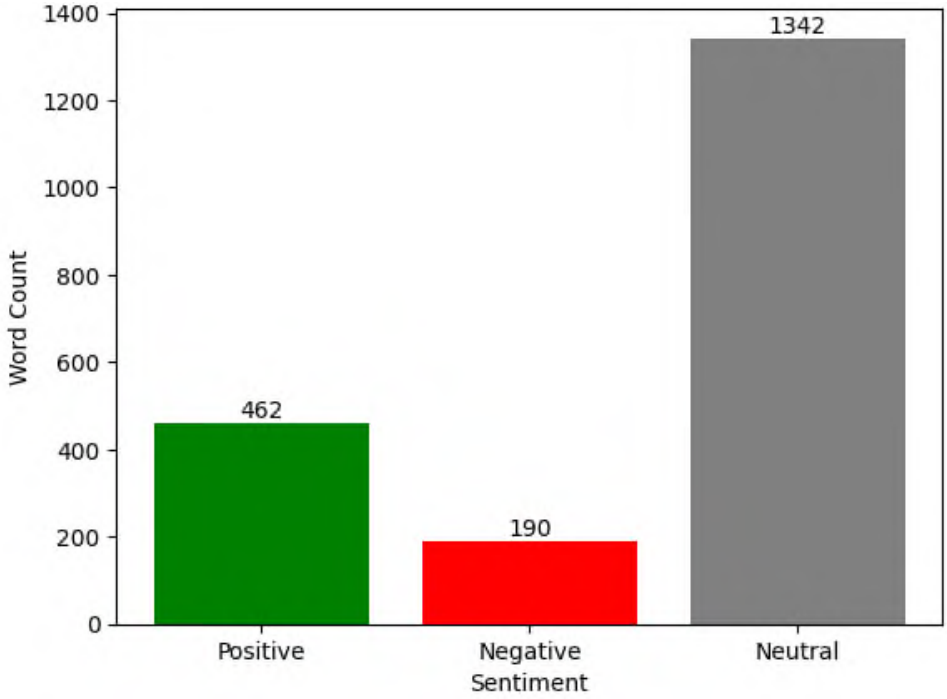
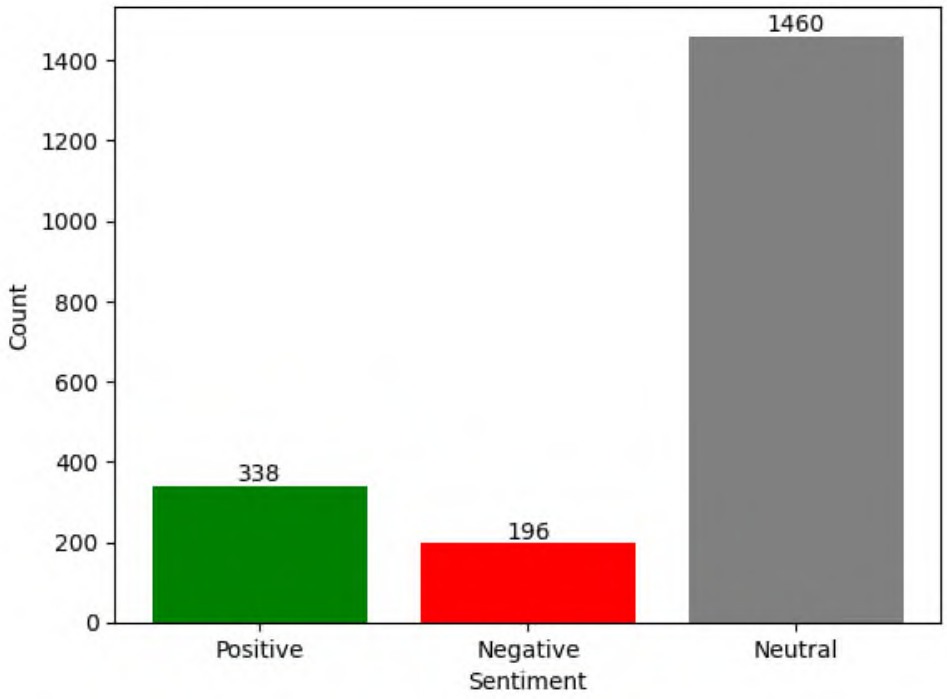


Figure 3

Sentiment analysis results using TextBlob



Following the labeling process, we calculated the frequency of the most commonly used words using count vectorization. The most frequently occurring word was "love," reflecting the common theme of happiness in music. Further analysis of word usage revealed distinct patterns in sentiment; for instance,

negatively labeled songs predominantly featured words such as "down," "sorry," "bad," "broken," and certain profanities (Figure 4 and Figure 5).

Figure 4

Word cloud of negative words

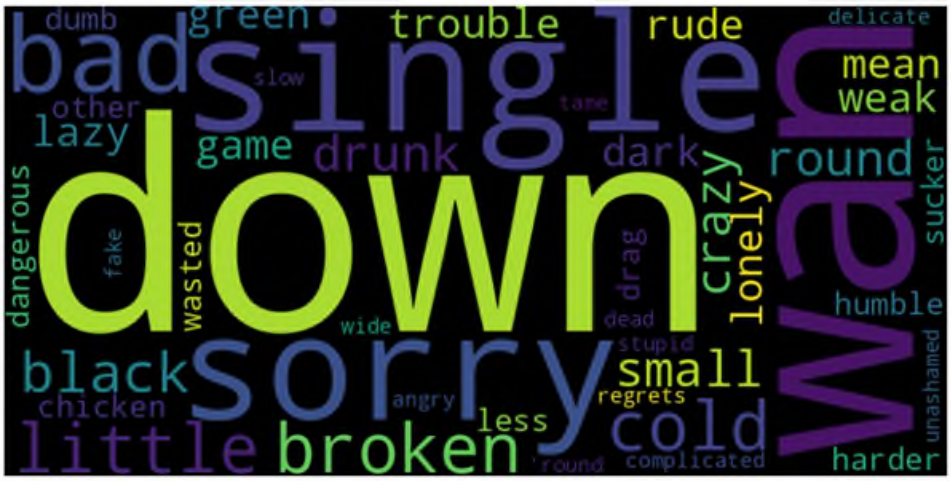
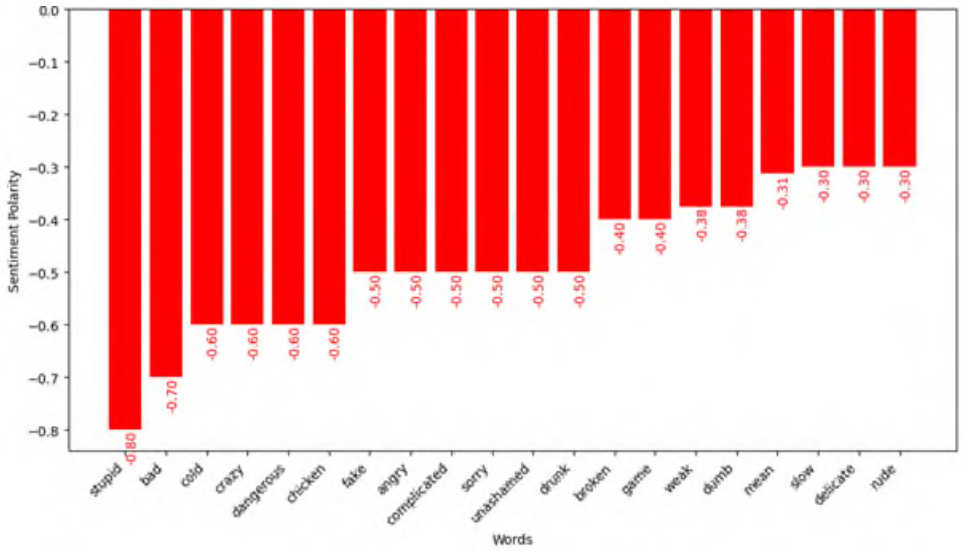


Figure 5

Top 20 words for highest negative sentiment polarities



Conversely, positively labeled songs frequently contained words like "beautiful," "love", "sweet", "young", and "good" (Figure 6 and Figure 7). Based on these word frequency patterns, the dataset was transformed into a bag-of-words model, a representation commonly used in text classification for machine learning applications.

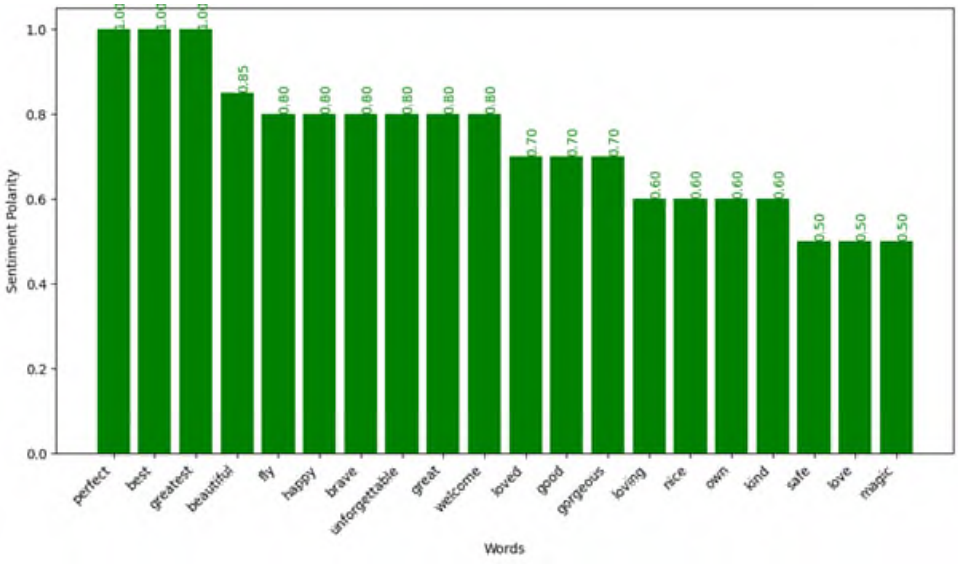
Figure 6

Word cloud of negative words



Figure 7

Top 20 words for highest negative sentiment polarities



For sentiment classification, six machine learning models were employed: Logistic Regression, Random Forest, Decision Tree, Support Vector Machine (SVM), XGBoost, and Multinomial Naive Bayes. Performance metrics, including accuracy, F1 score, recall, and precision, were calculated for each classifier to evaluate its effectiveness in sentiment classification (Table 2 and Table 3).

Table 2

Performance Metrics of Various Classifiers

Classifier	Training Accuracy	Validation Accuracy	Test Accuracy
Support Vector Machine	1.000	0.803	0.774
Logistic Regression	0.9311	0.757	0.741
Random Forest	1.000	0.840	0.866
Decision Tree	1.000	0.868	0.878

Classifier	Training Accuracy	Validation Accuracy	Test Accuracy
XGBClassifier	0.815	0.730	0.692
MultinomialNB	0.981	0.852	0.792

The training, validation, and test accuracies for each classifier are presented in Table 2. The SVM model achieved perfect training accuracy (1.000) but showed slightly lower validation (0.803) and test accuracies (0.774), suggesting potential overfitting. Logistic Regression attained a high training accuracy of 0.931, with validation and test accuracies of 0.757 and 0.741, respectively. The Random Forest classifier also achieved perfect training accuracy (1.000) and demonstrated strong validation (0.840) and test (0.866) accuracies. Similarly, the Decision Tree classifier achieved perfect training accuracy (1.000) and performed well on both the validation (0.868) and test (0.878) datasets. In contrast, the XGBClassifier showed a moderate training accuracy of 0.815, with lower validation (0.730) and test accuracies (0.692). Multinomial Naive Bayes reached a high training accuracy of 0.981, with validation and test accuracies of 0.852 and 0.792, respectively. Overall, the Decision Tree and Random Forest classifiers exhibited the highest test accuracies, indicating their robustness in generalizing to new data.

Table 3

Classifier Evaluation Metrics

Classifier	Precision	Recall	F1 Score	MCC
Support Vector Machine	0.711	1.000	0.826	0.577
Logistic Regression	0.683	1.000	0.807	0.509
Random Forest	0.833	0.991	0.903	0.722
Decision Tree	0.862	0.968	0.910	0.736
XGBClassifier	0.659	0.946	0.771	0.368
MultinomialNB	0.736	0.980	0.834	0.598

In terms of specific performance metrics, the Random Forest and Decision Tree classifiers demonstrated superior precision, recall, and F1 scores (Table 3). The Random Forest classifier achieved a precision of 0.833, recall of 0.991, and an F1 score of 0.903, resulting in a Matthews Correlation Coefficient (MCC) of 0.722. Similarly, the Decision Tree classifier displayed a precision of 0.862, recall of 0.968, and an F1 score of 0.910, with an MCC of 0.736. These results that the Random Forest and Decision Tree classifiers excel in achieving a balanced trade-off between precision and recall, making them particularly suitable for sentiment classification tasks in playlist data.

In addition to the F1-score, precision, recall, and accuracy, we have also added the Matthews Correlation Coefficient (MCC) as another metric of evaluation, considering its effectiveness to measure classification performance, especially where there is an imbalanced class situation. Unlike accuracy, where its result can have a lopsided view where one of the classes is dominant within the set of instances, MCC makes its result closer to fairness by considering false negatives, false positives, true negatives, and true positives. This characteristic makes it especially useful when there is a sentiment classification task under the imbalanced conditions of the class. With the use of MCC within our framework of evaluation, we make our classification performance evaluation richer and more reliable.

Discussion

The findings of this study highlight the varying effectiveness of machine learning models in classifying sentiment within Spotify playlists, with the random forest and decision tree classifiers emerging as the top-performing models. The overall success of these classifiers aligns with prior research in text-based sentiment classification, where ensemble methods such as Random Forest are often favored for their robustness in handling complex, multi-feature datasets (Breiman, 2001; Delbouys *et al*., 2018). The Random Forest model's ability to combine multiple decision trees allows it to capture subtle sentiment patterns that simpler models, such as Logistic Regression, may miss. The decision tree classifiers also demonstrated strong performance, achieving high test accuracy and an impressive balance between precision and recall. While individual Decision Trees are known to be prone to overfitting, the structured data preprocessing in this study likely minimized noise, allowing the model to generalize effectively across sentiment categories (Pang & Lee, 2008).

Support Vector Machines (SVM) also performed well, though with signs of overfitting, as evidenced by its high training accuracy and slightly lower validation and test accuracies. This phenomenon is common in SVM applications for text classification, where the high-dimensional nature of the data can challenge model generalizability (Pappas & Popescu-Belis, 2016, Sancar, 2024). SVM's reliance on well-defined margins works effectively for sentiment classes with clear separations; however, the ensemble nature of Random Forest allowed it to handle the varied sentiment expressions across playlists more flexibly. This observation aligns with that of Ferraro et al. (2018), who noted that ensemble approaches often outperform SVM in tasks involving diverse user data.

Logistic Regression achieved moderate accuracy, providing insights into feature weights yet potentially limiting its capacity to capture complex sentiment shifts within playlist data. Its linear nature made it less capable of handling nuanced patterns than non-linear models such as XGBoost and Random Forest. Studies on sentiment analysis have similarly shown that linear classifiers may struggle with data requiring deeper contextual interpretation (Hu & Liu, 2004).

XGBoost displayed mixed performance, with moderate training accuracy and lower validation and test accuracies compared to the other classifiers. While XGBoost is known for its strength in handling imbalanced datasets through iterative boosting, its performance here may indicate that the playlist sentiment data contained more complex variations than it could optimize for. Recent studies by Schedl et al. (2021) suggest that XGBoost performs well in mood-based music classification but may encounter limitations with polarity-focused sentiment analysis due to the subtleties involved in positive versus negative classification.

Multinomial Naive Bayes provided useful baseline insights but underperformed relative to other models. The Naive Bayes model assumes feature independence, an oversimplification that can limit its ability to capture sentiment nuances. This tendency to rely on individual word probabilities may explain why Naive Bayes struggles with sentiment classification in multi-dimensional data like playlists, as noted in prior sentiment analysis studies (Pang & Lee, 2008).

Recently, deep learning models have become increasingly prominent in sentiment analysis, particularly in multimedia contexts. Models like Recurrent Neural Networks (RNN) and their variants, such as Long Short-Term Memory (LSTM) networks, have shown notable success in capturing sentiment by considering sequential dependencies in text data (Zhang *et al*., 2022). In music sentiment analysis, Convolutional Neural Networks (CNNs) combined with audio features have further enhanced model performance, offering multi-

modal capabilities that integrate both text and audio for a richer sentiment profile (Birdal, 2024; Li *et al.*, 2023). While deep learning models require more computational power and extensive data for training, they hold promise for future applications in playlist sentiment classification, particularly when combined with ensemble approaches for increased accuracy and generalizability.

This study's use of the random forest and decision tree classifiers emphasizes the value of interpretable, high-performing models in sentiment classification. However, integrating deep learning models could advance this approach by capturing sequential patterns within playlists, enhancing both sentiment analysis and recommendation accuracy. Future work could explore hybrid models that blend traditional machine learning with deep learning, creating a multi-modal system that leverages both textual and acoustic features. Such an approach could significantly improve user engagement in music recommendation systems, addressing a key challenge in personalized content delivery.

In summary, while the random forest and decision tree models proved effective for playlist sentiment analysis, the potential of deep learning in this field remains considerable. This study contributes to the ongoing research in music recommendation by underscoring the value of sentiment-aware models, suggesting that future work may benefit from integrating ensemble and deep learning approaches to capture the complexities of user sentiment across diverse musical content.

Peer Review Externally peer-reviewed.
Conflict of Interest The author have no conflict of interest to declare.
Grant Support The author declared that this study has received no financial support.

Author Details M. Erdem İsenkul
[1] İstanbul University-Cerrahpaşa, Faculty of Engineering, Department of Computer Engineering, İstanbul, Türkiye
0000-0003-0856-2174 eisenkul@iuc.edu.tr

References

Abel, F., Gao, Q., Houben, G. J. & Tao, K. (2011). *Analyzing user modeling on Twitter for personalized news recommendations*. User Modeling, Adaption, and Personalization, 1–12. Springer.

Birdal, R. G. (2024). The Influence of Air Pollution Concentrations on Solar Irradiance Forecasting Using CNN-LSTM-mRMR Feature Extraction. *CMC-Computers Materials and Continua*, *78*(3), 4015-4028.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Chen, T. & Guestrin, C. (2016). *XGBoost: A scalable tree boosting system*. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794.

Delbouys, R., Hennequin, R., Piccoli, F., Royo-Letelier, J. & Moussallam, M. (2018). *Music mood detection based on audio and lyrics with deep neural net*. Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR), 688–695.

Ferraro, S., Bogdanov, D., Yoon, S., Kim, Y., & Serra, X. (2018). *Cross-cultural analysis of user behavior in music streaming services*. Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR), 1–6.

Hu, M. & Liu, B. (2004). *Mining and summarizing customer reviews*. Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 168–177.

Hutto, C. J., & Gilbert, E. (2014). *VADER: A parsimonious rule-based model for sentiment analysis of social media text*. Proceedings of the 8th International Conference on Weblogs and Social Media (ICWSM), 216–225.

Jiang, L., Yu, M., Zhou, M., Liu, X., & Zhao, T. (2011). *Target-dependent Twitter sentiment classification*. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 151–160.

Li, Y., Wang, X., Chen, J., & Xu, Z. (2023). Sentiment analysis on music recommendation: A convolutional neural network approach. *IEEE Transactions on Multimedia*, *25*(3), 689–700.

Liu, H., & Singh, P. (2004). ConceptNet: A practical commonsense reasoning toolkit. *BT Technology Journal*, *22*(4), 211–226.

Mumcuoğlu, K. Y. (2007). *Biotherapy laboratory protocol department of parasitology*. Israel: Hebrew University-Hadassah Medical School Jerusalem.

Pang, B. & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, *2*(1–2), 1–135.

Pappas, N. & Popescu-Belis, A. (2016). *Sentiment analysis of user comments for one-class collaborative filtering over text-based recommendations*. Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, 773–776.

Sancar, Y. (2024). Enhanced Classification of Skin Lesions Using Fine-Tuned MobileNet and DenseNet121 Models with Ensemble Learning. *Erzincan University Journal of Science and Technology*, *17*(3), 870-883.

Schedl, M. (2016). *The LFM-1b dataset for music recommendation and user modeling*. Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval, 103–110.

Schedl, M., Knees, P., McFee, B., & Bogdanov, D. (2021). Music recommendation: State of the art and challenges. *IEEE Signal Processing Magazine*, *38*(3), 29–44.

Turney, P. D. (2002). *Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews*. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 417–424.

Zhang, J., Li, H., Zhao, R., & Huang, Y. (2022). Comparative study of LSTM and GRU for music sentiment analysis. *IEEE Transactions on Neural Networks and Learning Systems*, *33*(5), 2345–2356.

İSTANBUL UNIVERSITY PRESS

Acta Infologica,
ACIN 2025, 9 (1): 90–111

https://doi.org/10.26650/acin.1538357
Submitted: 26.08.2024
Revision Requested: 06.02.2025
Last Revision Received: 17.02.2025
Accepted: 17.03.2025
Published Online 30.04.2025

Acta Infologica

Research Article

Open Access

Evaluating the Accessibility, Usability, and Security of Corporate Investor Relations Web Pages: A Case Study of the Turkish Stock Market

Merve Garip [1] & Serhat Peker [1]

[1] İzmir Bakırçay University, Department of Management Information Systems, İzmir, Türkiye

Abstract This study investigates the accessibility, usability, and security of investor relations (IR) web pages of 53 companies listed on the Istanbul Stock Exchange (BIST), focusing on both BIST-30, representing the most liquid and highly capitalized companies, and BIST-2022, companies that had their initial public offerings (IPOs) in 2022. Using automated testing tools, the analysis examines adherence to the WCAG 2.1 guidelines, mobile responsiveness, and security measures. The findings reveal that accessibility issues, particularly related to the robust and perceivable principles, are prevalent across the assessed web pages, with BIST-30 companies showing a higher rate of errors. While BIST-2022 companies generally performed better in usability, with faster loading times and fewer broken links, security assessments indicated low to medium risks across both categories. This study contributes to the limited literature on the quality assessment of corporate IR web pages and highlights the necessity for enhancements in accessibility, usability and security, providing valuable insights for web administrators and developers to improve the quality of IR web pages.

Keywords Human Computer Interaction · Website Performance Evaluation · Automated Testing Tools · Corporate Websites · Investor Relations

Citation: Garip, M. & Peker, S. (2025). Evaluating the accessibility, usability, and security of corporate investor relations web pages: A case study of the Turkish stock market. *Acta Infologica*, *9*(1), 90-111. https://doi.org/10.26650/acin.1538357

This work is licensed under Creative Commons Attribution-NonCommercial 4.0 International License.

© 2025. Garip, M. & Peker, S.

Corresponding author: Serhat Peker serhat.peker@bakircay.edu.tr

Introduction

Websites have evolved into powerful universal communication tools, serving as platforms for information sharing, interaction, and engagement. In this digital age, it is crucial to recognize the importance of designing, usable, and secure. Accessibility ensures that individuals of all abilities and disabilities can access and navigate the website, regardless of their physical, sensory, or cognitive limitations (Sevilla, Herrera, Martínez, & Alcantud, 2007). Incorporating accessible design elements, like adding alternative text for images, using straightforward language, and enabling keyboard navigation, makes websites inclusive, allowing all users to fully engage with the online experience. Usability focuses on creating intuitive and user-friendly interfaces that enable seamless navigation and interaction (Kumar et al., 2023). By understanding user behaviors and preferences, designers can optimize the layout, structure, and functionality of websites, thereby enhancing user satisfaction and engagement. Usable websites are efficient, well-organized, and provide clear calls to action, enabling visitors to easily find the information they need and accomplish their goals. In today's digital landscape, where online threats and data breaches are common, ensuring website security is also crucial. This requires implementing strong security measures like encryption, secure authentication, and consistent software updates (Chatterjee, Gupta, & De, 2013; Sicari, Rizzardi, Grieco, & Coen-Porisini, 2015) . By prioritizing security, websites protect user data, maintain trust, and safeguard sensitive information.

In today's competitive business landscape, corporate websites serve as powerful instruments for facilitating communication between companies and their stakeholders (Brennan & Merkl-Davies, 2018). Corporate websites also serve as the primary platform for disseminating Investor Relations (IR) information to a wide range of stakeholders, including current and potential investors (Halim, Basiruddin, & Ali, 2015). The purpose of Investor Relations (IR) is to attract and retain investors by effectively communicating the company's financial performance, strategic direction, and future prospects. By offering detailed financial reports, annual statements, investor presentations, and other relevant materials, companies demonstrate transparency and accountability, instilling confidence and trust among investors. The availability of IR information on corporate websites is crucial as it allows investors to make informed decisions about their investment choices. Moreover, it enables companies to showcase their strengths, competitive advantages, and growth strategies, attracting new investors and retaining existing ones. By prioritizing Investor Relations on their corporate websites, companies can establish strong investor relationships, enhance their market credibility, and foster long-term investment partnerships that contribute to their sustained success.

Recognizing the pivotal role of corporate websites in today's competitive business landscape, this research aims to assess the usability, accessibility, and security of investor relations web pages of companies listed on the Istanbul Stock Exchange in Turkey. Our comprehensive analysis covers 53 web pages, using automated testing tools to scrutinize various aspects such as adherence to accessibility guidelines, loading speed, broken links, and security measures. The results of these analyses offer valuable insights into the overall performance of corporations in their online presence for effective Investor Relations (IR) communication. This research offers a nuanced understanding of how these corporate websites facilitate communication with investors, contributing valuable insights to the limited literature available on the quality assessment of corporate IR web pages.

Related Work

Many studies have been undertaken to assess corporate websites. This section examines studies of corporate websites, and each study provides valuable insights into the distinct aspects of web quality

and accessibility. For instance, Fogli and Guida (2013) presented an innovative quality model for corporate websites, highlighting the essential concepts of ultimate quality, external quality, and internal quality. The emphasis on these three key dimensions provides a nuanced perspective on the overall quality of corporate websites, considering both user satisfaction and organizational objectives. While exploring the quality of other websites, Fogli and Guida (2015) introduced a viewpoint on website quality, emphasizing stakeholder relationships. The research aligns the concept of quality with stakeholder perspectives, distinguishing final quality for site owners from quality in use, basic quality, and internal quality for users, service managers, and designers.

Other studies have also delved into the analysis of business websites. Huang and Cappel (2012) evaluate and contrast the website usability of INC. 500 companies using 11 distinct measures of usability. The study aims to determine the extent to which companies follow selected website usability guidelines and to identify any differences between the Fortune 500 and INC. 500 companies in this regard. In the initial assessment, it is evident that Fortune 100 companies have enhanced the accessibility of their websites throughout the study period (Loiacono, Romano, & McCoy, 2009). In another study, Loiacono and McCoy (2004) investigated and compared the accessibility of websites in the federal government, nonprofit, and corporate sectors for people with disabilities. The study reveals that federal government websites are significantly more accessible than those of non-profits and corporations.

Examining web accessibility in the corporate context, Singh and Singh (2020) explore the web accessibility of corporate information disclosed by 100 major Indian companies listed on the BSE. Their objective was to identify the factors that impact the accessibility of this information. Their results reveal that larger, more established Indian firms with lower risk tend to have websites with better accessibility, around 55%, emphasizing the significance of web accessibility for effective corporate communication. Additionally, Leitner et al. (2016) aimed to study an exploratory case study across three industry sectors to examine organizations' motivations, consequences, and determinants regarding the implementation of Web accessibility standards, aiming to illuminate managerial rationales and support better decision-making in organizations. Consequently, only 12% of the assessed websites successfully met the accessibility criteria.

The web accessibility practices of 311 Dow Jones Sustainability World Index-listed firms were evaluated through a content analysis of adherence to World Wide Web Consortium (W3C) guidelines by Conte et al. (2022). This study identifies variations in corporate accessibility practices for disabled users, offering insights to enhance alignment with sustainability objectives. Advancing in the investigation of web accessibility, Gonçalves et al. (2013) assessed web accessibility among the top 250 Forbes enterprises, revealing significant shortcomings and emphasizing the ethical responsibilities of leaders. It reveals a significant need for improvement in web accessibility levels among the Forbes 250 largest enterprises, indicating widespread errors and low compliance with established standards.

Embarking on an in-depth analysis, Ashbaugh et al. (1999) investigated the practice of Internet Financial Reporting (IFR) among firms, assessing the extent to which they utilize the internet to enhance the relevance of their financial reporting. The study examines firms' IFR practices, finding substantial variation in quality and timeliness, with 70% of the sampled firms engaging in IFR, and discusses implications for stakeholders. Xiao et al. (2004) examined the factors that impact the voluntary adoption of Internet-based financial reporting by Chinese listed companies, along with the extent of their disclosures. The results, derived from an analysis of the top 300 listed Chinese companies, indicate that choices in Internet-based disclosure are influenced by particular attributes of their environment.

The field studies of web accessibility in the literature are mentioned sequentially. Numerous studies investigate website accessibility across different sectors, with a particular focus on industries such as tourism, e-commerce, airline, banking, and fintech. Domínguez Vila et al. (2018) evaluated the accessibility of 210 official national tourism organization websites globally, utilizing a website accessibility evaluation tool to identify patterns and rankings, intending to provide an indicator for inclusive organizational practices in online accessibility for individuals with disabilities in the realm of accessible tourism. Gonçalves et al. (2018) evaluated Portuguese e-commerce websites for accessibility and usability, revealing issues through automated tools, manual checks, heuristic evaluations, and user tests, and proposed seven recommendations for improvement. Agrawal et al. (2019) analyzed the quality of Indian airline websites in terms of accessibility, usability, and readability, emphasizing their adherence to web accessibility guidelines, identification of usability issues, and assessment of mobile friendliness. Additionally, Fatima et al. (2020) assessed the usability, accessibility, and vulnerability of Pakistani and International Banking Websites, employing manual evaluations, automated tools, and user perspectives, with a focus on identifying issues and comparing the performance of these websites. Lastly, Nour (2022) focuses on 32 Saudi FinTech websites and reveals notable accessibility and usability issues, emphasizing the urgent need for improvements to better serve people with disabilities in accessing crucial services like banking and healthcare.

Table 1 summarizes previous studies assessing corporate web pages in various domains worldwide in terms of different dimensions such as accessibility and usability in chronological order.

Table 1

A summary of previous research work

Article	Year	Sector	Dimensions	Sample	Tools and techniques Used
Gonçalves et al. (2018)	2018	E-commerce	Accessibility and Usability	Portuguese e-commerce websites	SortSite and manual assessments, heuristic evaluation for usability
Agrawal et al. (2019)	2019	Airline	Accessibility, Usability, and Readability	Home pages of 9 websites of airlines based in India	TAW, Pingdom, Websitepulse, WAVE, Webpagefx, Google mobile-friendly test
Fatima et al. (2020)	2020	Banking	Accessibility and Usability	Homepage on 10 bank websites in Pakistan, Canada, Australia, United States of America, and United Kingdom	Heuristic evaluation of usability Wave, TAW, AChecker
Nour (2022)	2022	Fintech	Accessibility and Usability	The Arabic version of the 32 Saudi FinTech websites' homepages	MAUVE++, WAVE
Başeğmez et al. (2023)	2023	Technology and Informatics	Usability	Websites of 24 Turkish companies in the BIST Technology and Informatics index	GTmetrix, gray relational analysis,

As shown in Table 1, there are several studies conducted in various sectors, including technology, fintech, e-commerce, banking, tourism, and airline, assessing dimensions such as accessibility and usability. These studies used tools like MAUVE++, WAVE, SortSite, and manual assessments, among others, to evaluate website performance. While these studies have contributed to the understanding of website accessibility and usability, a noticeable gap exists in the assessment of accessibility, usability, and security specifically for corporate investor relations web pages. Our study addresses this gap by conducting a comprehensive

evaluation of investor relations web pages in the context of the Turkish stock market, considering accessibility with WCAG 2.1 standards, usability, and security.

Methodology

Data

This study focuses on conducting an assessment of investor relations web pages associated with companies listed on the BIST-30, which represents the performance of the 30 most liquid and highly capitalized companies traded in the Istanbul stock market and the companies with initial public offerings (IPOs) in the Istanbul Stock Exchange (ISE) during the year 2022. The data were collected between March 28 and April 1, 2023. It was thoroughly processed, tabulated, and visualized by April 27, 2023. Throughout the article, we used "BIST-30" to refer to the first group of companies and "BIST-2022" for the second group.

In this study, we initially identified 68 investor relations (IR) web pages associated with companies listed on the BIST-30, which represents the most liquid and highly capitalized companies on the Istanbul Stock Exchange, and companies that had their initial public offerings (IPOs) in 2022. However, 53 of these web pages were included in the final analysis based on the following criteria:

- **Inclusion Criteria:**
  - Web pages that are explicitly dedicated to investor relations and provide financial reports, presentations, investor announcements, and other related content.
- **Exclusion Criteria**:
  - Two company web pages were excluded because they presented financial statements, reports, and investor-related content directly on their corporate websites instead of having a dedicated IR page.
  - Thirteen companies were excluded because they do not maintain a dedicated IR web page, meaning their investor relations information is either absent or merged within general corporate sections without a clear structure.

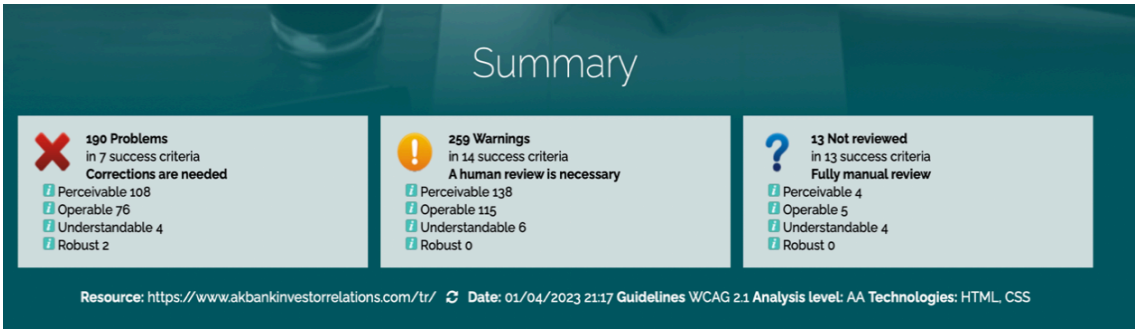
A detailed list of companies and their analyzed IR web pages can be found in the Appendix.

Tools and techniques

For accessibility testing, the TAW tool was employed. The TAW is one of the most frequently used automated testing tools for website evaluation and has been widely used in academic research (Macakoğlu & Peker, 2022). This tool evaluates investor relations websites based on the WCAG 2.1 standards, providing an online accessibility analysis and generating a comprehensive report for a specific URL (TAWdis, 2001). The generated TAW report encompassed an overview of the issues and warnings, as well as a detailed breakdown of the perceivable, operable, understandable, and robust aspects, as illustrated in Figure 1. The generated reports were used to identify accessibility shortcomings and categorize them based on WCAG conformance levels (A, AA, AAA). The data collection process using TAW was conducted manually, with each web page being individually tested by entering its URL the tool.

Figure 1

Taw's interface



Usability testing was conducted using Pingdom and GTmetrix, both of which are commonly used in performance evaluation studies. Pingdom was employed to analyze website load speed and optimize performance, evaluating key usability aspects such as page size, number of requests, and load time (Pingdom, 2005). This tool is widely used in research, such as Akgül (2021), who applied it to assess the usability of state and private university websites. GTmetrix is an online tool that evaluates website speed and performance by measuring metrics such as fully loaded time, first contentful paint, speed index, and total blocking time (GTmetrix, 2009). The tool was applied manually by inputting each website URL, following a methodology similar to that of Csontos and Heckl (2021), who assessed usability across various domains.

For security evaluation, the Sucuri tool was used. This tool scans for viruses, checks the blacklisting status, identifies website errors, assesses the risk level of the website, and examines the TLS and server configurations (Sucuri, 2007). It is widely employed in security assessments, as demonstrated by Macakoğlu et al. (2022) in their study on website domain scanning. The security assessment was carried out by inputting each URL into Sucuri. The Dead Link Checker tool (Dead Link Checker, 2013.) was used to identify broken links, which can have adverse effects on user experience and SEO. Each website was manually checked using this tool, which analyzed and reported any dead or broken links. This tool has also been widely employed in academic research, such as Gopinath et al. (2016), who used it to examine broken links on Sri Lankan government websites.

Furthermore, the Google Mobile-Friendly tool has been widely used to evaluate mobile responsiveness. (Mobile-Friendly Test, 2014). This online tool evaluates how well a website performs on mobile devices by assessing key mobile usability factors, such as text readability, ease of navigation, and the responsiveness of images and other elements when viewed on smaller screens. The tool provides a straightforward "pass" or "fail" result, indicating whether the site is optimized for mobile use. Studies such as Al-Sakran and Alsudairi (2021) have used this tool to evaluate mobile e-government websites. Each web page was tested manually by entering its URL the tool, and the results were recorded. Lastly, all evaluations were conducted manually, with URLs being individually entered into the respective tools. The data collection process lasted five days, and all results were systematically recorded for further analysis.

Results

Accessibility analysis

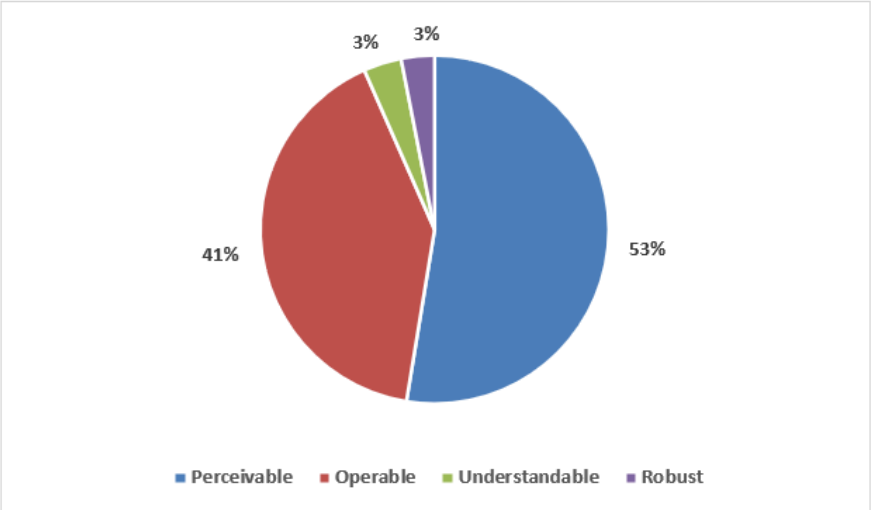
Accessible web pages play a crucial role in providing content that is easily accessible to all individuals. The WCAG 2.1 guidelines specifically emphasize the importance of web content accessibility for people with disabilities (Web Content Accessibility Guidelines (WCAG) 2.1, 2018). In this study, a total of 51 web pages

belonging to companies listed on the BIST-30 firms and BIST-2022 firms indices in Turkey were evaluated using the TAW accessibility tool.

Three web pages were excluded from the analysis due to analysis errors, resulting in a final sample size of 51 web pages. This exclusion was specifically due to errors encountered during the evaluation process with the TAW accessibility tool, which failed to properly load the results for these pages. The principles of accessibility—namely, perceivable, operable, understandable, and robust, which are essential for ensuring that web content can be accessed and used by everyone—are illustrated in Figure 2.

Figure 2

Accessibility error distribution based on the four principles



In the "Perceivable" category, the recorded accessibility error rate was 53%. This suggests a notable presence of challenges in making system elements or content perceivable to users. These challenges may arise from various factors, including insufficient provision of alternative text for images, absence of captions for videos, or inadequate color contrast for text.

Within the "Operable" category, the accessibility error rate was recorded at 41%. This indicates notable challenges in providing the system's functionality and operability for all users. These challenges may stem from various issues, including difficulties with navigation, limitations in keyboard-only controls, or inconsistencies in the provision of clear and cohesive interactive elements.

The accessibility error rate of 3% in the "Understandable" category indicates that the system generally excels in terms of user comprehension and understanding. However, there may still exist areas that warrant improvement, such as the provision of clear instructions, error messages, or the avoidance of jargon or complex language. Similarly, the "Robust" category also demonstrated a commendable accessibility error rate of 3%, signifying the system's compatibility and adaptability across various platforms and technologies.

Figure 3 illustrates the distribution of accessibility errors based on basic principles, with the results separately presented to two BIST categories. The accessibility error rates for the BIST-2022 companies demonstrate noteworthy differences.

Figure 3

Distribution of errors according to principles by website page type



In the "Perceivable" category, BIST-2022 companies exhibit a lower error rate of 30.40%, signifying relatively fewer challenges compared to BIST-30 companies. Similarly, in the "Operable" category, BIST-2022 shows a moderate error rate of 43.40%, suggesting moderate difficulties in ensuring system functionality and operability, which is lower than the error rate of BIST-30 companies. Notably, the "Understandable" category in BIST-2022 presents a relatively stronger performance with an error rate of 33.18%. Lastly, the "Robust" category of BIST-2022 companies showcases the lowest error rate of 26.84%.

The analysis reveals that a significant majority of the accessibility errors in the "Perceivable" category were found on web pages belonging to companies listed on BIST-30. Likewise, the "Understandable" category also exhibited a substantial concentration of errors originating from BIST-30 company web pages, surpassing two-thirds of the total errors. Moreover, it is noteworthy that the BIST-30 companies consistently demonstrated the highest number of errors across all evaluated principles.

Table 2 presents an overview of the rate of accessibility errors by success criteria. The Web Content Accessibility Guidelines (WCAG) define three conformance levels to assess website accessibility: Level A, Level AA, and Level AAA. Level A establishes the fundamental accessibility criteria, ensuring that the key needs of individuals with impairments are met. If a webpage does not meet the Level A criteria, it may lack essential functionality, such as alternative text for images or keyboard navigation, making it difficult for users with disabilities to access the content. Level AAA, on the other hand, represents the highest level of accessibility, designed to provide maximum usability for all users, particularly those with impairments. However, achieving full compliance at this level is a challenging task and is often regarded as an ideal goal rather than a mandatory requirement for all web content.

Table 2

Distribution of the success criterion errors

Success Criteria	Level	BIST Category					
		Overall		BIST-30 (n=24)		BIST-2022 (n=27)	
		%	Total	%	Avg	%	Avg
1.1.1 - Non-text Content	A	84.31	2284	87.5	77.04	81.48	16.11
1.3.1 - Information and Relationships	A	90.19	767	87.5	14.37	92.59	15.62
2.1.3 - Keyboard (No Exception)	AAA	9.80	37	12.5	1.37	7.41	0.14
2.4.2 - Page Titled	A	3.92	2	0	0	7.41	0.07
2.4.4 - Link Purpose (In Context)	A	86.27	1249	87.5	31.95	85.19	17.85
2.4.9 - Link Purpose (Link Only)	AAA	76.47	600	79.16	16	74.07	8
2.4.10 - Section Headings	AAA	80.39	528	79.16	5.91	81.48	14.29
3.1.1 - Language of the Page	A	35.29	18	45.83	0.45	25.93	0.25
3.2.2 - On Input	A	33.33	22	33.33	0.45	33.33	0.40
3.3.2 - Labels or Instructions	A	52.94	145	62.5	4.83	44.44	1.07
4.1.2 - Name, Role, Value	A	60.78	179	75	5.5	48.15	1.74

The analysis of the BIST-30 and BIST-2022 companies, based on the data provided in Table 2, highlights key similarities and differences in their adherence to accessibility standards. The results indicate that the highest number of errors were found at Conformance Level A, indicating that many investor relation web pages do not meet basic accessibility requirements. The most common issues at Level A were related to Non-Text Content (1.1.1), Information and Relationships (1.3.1), and Link Purpose (2.4.4). Additionally, some errors were identified at Level AAA, particularly in the Link Purpose (2.4.9) and Section Headings (2.4.10). These findings suggest that while some accessibility measures are in place, further improvements are necessary, especially in addressing Level A errors to ensure a more inclusive user experience.

Both BIST-30 companies and BIST-2022 companies encounter challenges in meeting the accessibility requirements for non-text content (criterion 1.1.1) and information and relationships (criterion 1.3.1). The relatively high error percentages in these areas indicate a significant proportion of non-compliant web pages within both groups, highlighting the need for substantial improvements in ensuring accessibility. The high error percentages suggest a significant number of non-compliant web pages in both groups. Additionally, keyboard accessibility without exceptions (criterion 2.1.3) is relatively better although there is room for improvement. The error percentages indicate the need for enhancing accessibility in terms of non-text content and information structures.

In terms of specific criteria, BIST-30 companies demonstrate better compliance with page titles (criterion 2.4.2), while BIST-2022 companies achieve perfect compliance in this regard. However, both categories struggle with providing clear and meaningful link purposes within the context (criterion 2.4.4), indicating a need for improvement. The comparison also highlights challenges in fulfilling link purposes when links are presented in isolation (criterion 2.4.9). BIST-30 companies and BIST-2022 companies exhibit relatively high error percentages, suggesting the necessity to enhance link accessibility.

Broken link analysis

Table 3 presents the results of the broken link analysis for both BIST-30 and BIST-2022 companies separately. The analysis was conducted using the Deadlink Checker tool. However, this tool has the limitation

of checking a maximum of 2000 pages per website. As a result, for websites exceeding this limit, only the first 2000 pages were analyzed. Table 3 provides an overview of the average number of URLs checked per website, along with the minimum, average, and maximum number of broken links identified.

Table 3

Number of broken links

BIST Category	Average of URLs checked	Broken Links		
		Min.	Avg.	Max.
BIST-30	1577.07	0	270	2000
BIST-2022	1095.76	0	52.51	2000

For the BIST-30 companies, an average of 1577.07 URLs were evaluated, revealing a heterogeneous distribution of broken links. The range indicates that some web pages within this category had a significant number of broken links, while others had none. In the case of BIST-2022, an average of 1095.76 URLs were checked. Overall, the table demonstrates that both BIST-30 companies and BIST-2022 companies had varying degrees of broken links, with BIST-30 companies showing a wider range in the number of broken links. The results highlight the importance of regularly checking and maintaining the integrity of links on web pages to ensure seamless user experience and to avoid issues such as broken navigation and inaccessible content.

Mobile-friendly analysis

Mobile-friendly analysis conducted using the Google Mobile-Friendly tool enabled a comprehensive examination of 52 web pages to evaluate their mobile responsiveness. Figure 4 illustrates the results of this evaluation, indicating the rate of web pages that successfully passed the mobile-friendly standards. Only 10% of all BIST company web pages failed.

Figure 4

Result of the mobile-friendly test

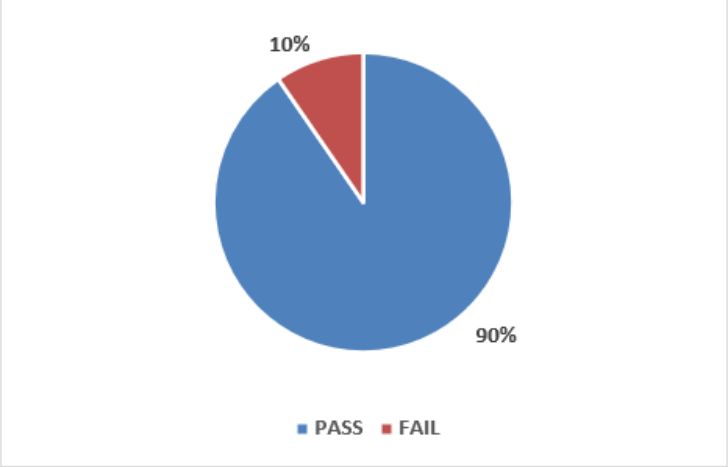


Table 4 demonstrates that company IR web pages failed or passed the mobile-friendly evaluation. 88% of the web pages successfully passed the mobile-friendliness evaluation, indicating a significant proportion of compliant web pages in terms of mobile design within the BIST-30 category. Conversely, 12% of the web pages in this category did not meet the requirements for mobile friendliness, indicating areas that need improvement.

Table 4
Mobile-friendly results by the BIST category

BIST Category	Pass	Fail
BIST-30 (n=25)	22 (%88)	3 (12%)
BIST-2022 (n=27)	25 (%92.59)	2 (7.41%)

In the BIST-2022 category, 92.59% of the web pages achieved a pass result, denoting a slightly higher percentage of mobile-friendly web pages compared to BIST-30 companies. A mere 7.40% of the web pages failed to fulfill the mobile-friendliness criteria. The findings underscore the overall positive outcomes in both BIST categories, as a substantial majority of the web pages in these categories are mobile-friendly. Nonetheless, there is still a small subset of web pages in each category that necessitates enhancements to ensure optimal mobile compatibility.

Usability analysis

In the analysis of 53 web pages, page sizes and loading times were evaluated using Pingdom and GTmetrix. Pingdom's grading system assigns scores from A (excellent) to F (very bad), based on factors such as page load time, number of requests, and overall optimization. Table 5 presents an overview of the performance grade distribution based on the Pingdom scores.

Table 5
Overview of the Pingdom grades

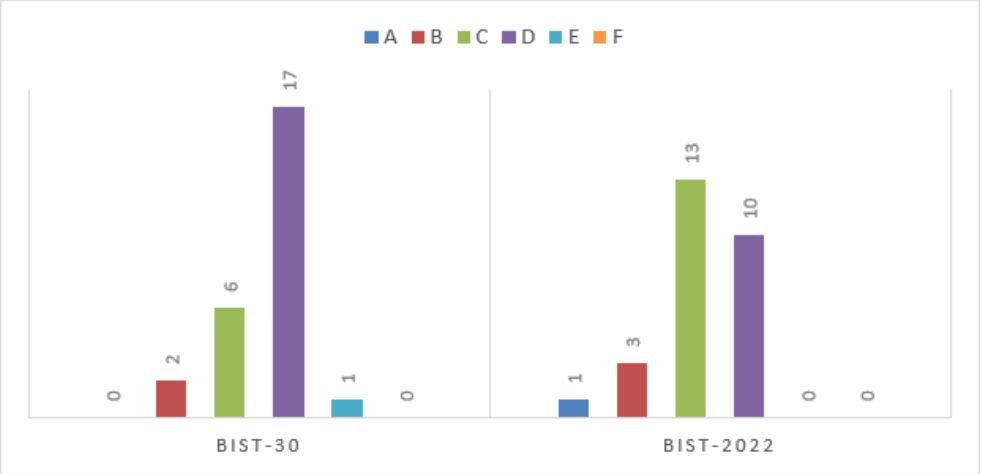
Pingdom Grade (A-F)	Number of websites
A (excellent)	1 (2%)
B (good)	5 (9%)
C (average)	19 (36%)
D (below average)	27 (51%)
E (bad)	1 (2%)
F (very bad)	0 (0%)

The results indicate that most web pages fall within the “Average” (C) and “Below Average” (D) categories, highlighting opportunities for optimization. A considerable 51% of the web pages received a “D” grade, indicating performance below the average while the F category excluded any web pages. This shows that a significant portion of the web pages examined need improvements in their performance metrics.

Figure 5 illustrates a detailed breakdown of the performance grades, shown as percentages, for both the BIST-30 and BIST-2022 companies.

Figure 5

Percentage of performance grades by BIST categories



For BIST-30 companies, it is noteworthy that no web pages were categorized under the F grade. Instead, most web pages gravitated toward the higher end of the performance spectrum, particularly in the D and C grades. For BIST-2022 companies, a similar pattern was observed, with no web pages having received an F grade. All web pages managed to avoid the lowest performance grade, and most of them obtained performance grades ranging from A to D. This figure indicates that web pages in both BIST categories tended to achieve better performance scores, highlighting a notable proficiency in web performance among the assessed web pages. In conclusion, it is discerned that the webpages of BIST-30 companies are predominantly categorized within Grade D, whereas those of BIST-2022 exhibit a preponderance in Grade C.

The website performance analysis for the BIST-30 and BIST-2022 companies, as summarized in Table 6, examines key GTmetrix performance metrics to assess loading efficiency and responsiveness. The evaluated metrics include:

- **Fully Loaded Time (s):** The total time required for a web page to load completely.
- **First Contentful Paint (FCP) (s):** The time it takes for the first visible element to appear on the screen.
- **Speed Index (s):** A measure of how quickly the content is visually displayed during loading.
- **Total Blocking Time (TBT) (ms):** The amount of time a webpage remains unresponsive due to background processes.

Each metric is presented with its minimum, mean, and maximum values to provide a comprehensive performance comparison between the two groups.

Table 6

Quality performance summary of websites by BIST categories

Performance Metrics	BIST categories	
	BIST-30 (n=26)	**BIST-2022** (n=27)
Fully loaded time (s)		
Min	0.57	0.3
Mean	2.73	1.21
Max	8.13	2.77

Performance Metrics	BIST categories	
	BIST-30 (n=26)	**BIST-2022** (n=27)
First contentful point (s)		
Min	1.1	0.60
Mean	3.15	1.71
Max	12.9	3.9
Speed index (s)		
Min	1.4	0.86
Mean	4.55	2.44
Max	13.6	5.2
Total blocking time (ms)		
Min	0	0
Mean	110.21	33.66
Max	781	228

An analysis of the fully loaded time metric reveals that the BIST-30 companies exhibit a wider range, varying from 0.57 s to 8.13 s, with an average of 2.73 s. In contrast, the BIST-2022 companies demonstrated a more limited range, spanning from 0.3 seconds to 2.77 seconds, with an average of 1.21 seconds. For the first contentful paint metric, BIST-30 companies show a broader range from 1.1 s to 12.9 s, averaging 3.15 s. Conversely, the BIST-2022 companies have a narrower range, from 0.60 to 3.9 seconds, with an average of 1.71 seconds. The speed index metric follows a similar trend, with BIST-30 companies showing higher values and a broader range from 1.4 seconds to 13.6 seconds, with an average of 4.55 seconds, compared to BIST-2022 companies, which have a smaller range from 0.86 seconds to 5.2 seconds and an average of 2.44 seconds.

The total blocking time, measured in milliseconds, highlights a significant difference between the two categories. BIST-30 companies exhibited higher values, ranging from 0 ms to 781 ms, with an average of 110.21 ms. In contrast, BIST-2022 companies show lower values, ranging from 0 ms to 228 ms, with an average of 33.66 ms. A thorough examination of the results reveals that the BIST-30 companies consistently display higher values across various performance metrics compared to the BIST-2022 companies. This discrepancy is particularly evident in the fully loaded time, first contentful paint, and speed index metrics, where BIST-30 companies demonstrate both wider ranges and higher average values.

GTmetrix also reports the total page size in megabytes (MB), which directly impacts the loading speed and overall performance. Table 7 presents a detailed distribution of the total page sizes for websites categorized under BIST-30 and BIST-2022. The data are classified into different size ranges (MB) to provide insights into how page weight affects display efficiency and user experience.

Table 7
Distributions of the total sizes of websites

Total page size (MB)	Number of websites	
	BIST-30 (n=26)	**BIST-2022** (n=27)
0-1.4 (very good)	4 (%15)	12 (44%)
1.5-2.4 (good)	8 (%31)	11 (41%)
2.5-3.4 (average)	3 (%11,5)	3 (11%)
3.5-4.4 (bad)	3 (%11,5)	0 (0%)

Total page size (MB)	Number of websites	
	BIST-30 (n=26)	**BIST-2022** (n=27)
4.5 and above (very bad)	8 (%31)	1 (4%)

For the "4.5 and above (very bad)" range, BIST-30 companies encompass 31% with 8 web pages, while BIST-2022 holds 4% with 1 web page. This analysis provides insights into the distribution of web pages across different total page size categories, revealing performance variations within the BIST-30 and BIST-2022 categories.

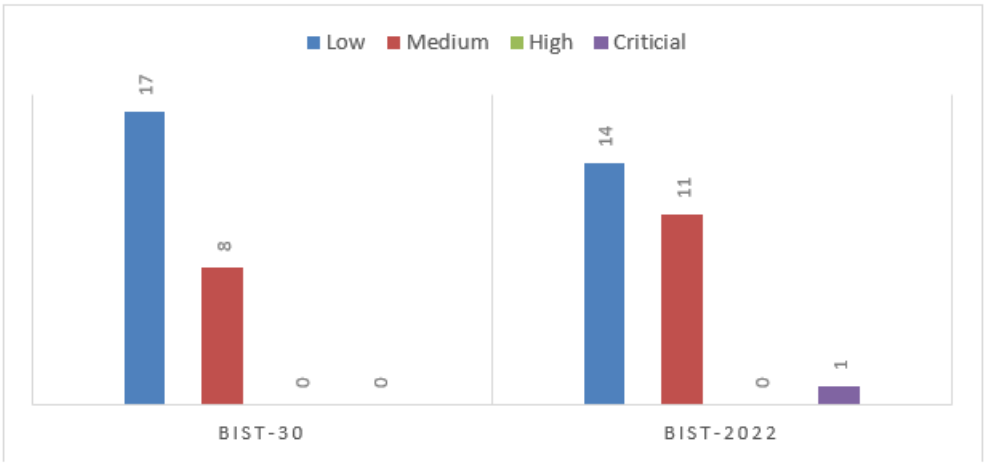
Security analysis

In the security analysis, the evaluation focused on assessing the risk levels of the web pages using the Sucuri tool. The risk levels were categorized as low, medium, high, and critical, providing a comprehensive understanding of the security posture of the analyzed web pages. Sucuri examines various security vulnerabilities, including malware presence, blacklist status, security anomalies, outdated software, SQL injection vulnerabilities, and cross-site scripting (XSS) risks. This detailed analysis helps identify potential threats that could compromise the integrity and safety of the websites.

Figure 6 presents an overview of the security risk levels across the BIST-30 and BIST-2022 web pages. In the "Low" risk category, there are 17 instances for BIST-30 companies and 14 for BIST-2022 companies, totaling 31. In the "Medium" risk category, BIST-30 companies have 8 instances, while BIST-2022 companies have 11, resulting in a combined 19.

Figure 6

Percentage of risk levels of the sites by BIST type



Notably, neither BIST-30 nor BIST-2022 companies have web pages classified under the "High" risk category. In the "Critical" risk category, BIST-30 companies show no instances, whereas BIST-2022 has 1, resulting in a cumulative 1.

Additionally, Table 8 highlights the distribution of the web server technologies used by these companies.

Table 8

Distribution of the web servers

Type of web server	Number of websites
Apache	5 (9%)

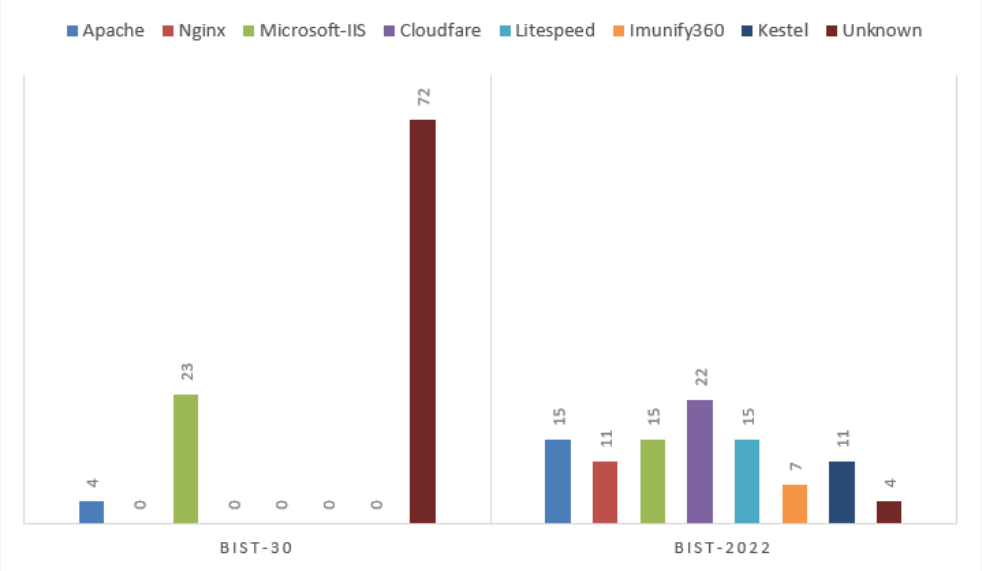
Type of web server	Number of websites
Nginx	3 (6%)
Microsoft-IIS	10 (19%)
Cloudfare	6 (11%)
Linseed	4 (8%)
Imunify360	2 (4%)
Kestrel	1 (2%)
Unknown	22 (41%)

The results indicate a diverse selection of web servers, although a significant portion (41%) is categorized as "unknown." This classification arises from instances where server details are either not publicly disclosed by the company or cannot be retrieved due to security configurations or privacy settings. Among the identified web servers, Microsoft-IIS was the most prevalent, constituting 19% of the total, followed by Cloudfare (11%), Apache (9%), Litespeed (8%), Nginx (6%), Imunify360 (4%), and Kestrel (2%).

The percentage of web servers by BIST type is given in Figure 7. For BIST-30 companies, Unknown is the dominant choice, constituting 72% of the total, while Apache is present at 4%. In contrast, BIST-2022 displays a more diverse selection, with Cloudfare leading at 22%, followed closely by Microsoft-IIS (15%), Litespeed (15%), Apache (15%), and Nginx (11%).

Figure 7

Percentage of web servers by BIST type



This analysis shows the nuanced distribution of web server technologies across the two BIST categories, indicating distinct preferences in server infrastructure. In the final stage of the security analysis, the examination focused on the use of TLS (SSL) encryption on web pages. The findings indicated that among the 54 web pages scrutinized, only 1 does not implement SSL encryption on BIST-2022 companies.

Discussion and Conclusion

This study examined the accessibility, usability, and security of investor relation web pages of corporate companies in Turkey by employing automated testing tools. A total of 53 web pages were subjected to analysis, and the assessment encompassed the utilization of online automated testing tools, namely TAW for the identification of accessibility errors, Deadlink Checker for the evaluation of broken links, and Google Mobile-Friendly Test for the assessment of mobile responsiveness. The study followed the WCAG 2.1 guideline, and the findings were subject to distinct analysis, systematically documented, and followed by subsequently analyzing the results.

The results of the study highlighted a significant number of accessibility problems, indicating that the most frequently violated standards were related to robust and perceivable principles. Additionally, both groups encountered challenges related to non-text content, indicating the need for improvement in adhering to accessibility guidelines. When comparing BIST categories, BIST-30 company web pages exhibited a higher rate of accessibility errors. Fixing these errors for each page type is crucial to improving the accessibility of web pages, providing a more user-friendly digital environment to individuals with disabilities. The analysis of broken links showed that the BIST-2022 company web pages had better performance in this regard. However, nearly each site contains at least one broken link. Mobile-friendliness assessments indicated positive results for both categories, with BIST-2022 companies demonstrating a slight advantage.

In terms of usability, BIST-2022 companies emerged as the leader, consistently displaying faster loading times and smaller page sizes. These metrics contribute to an enhanced user experience, positively influencing factors such as user engagement and satisfaction. The security evaluation for both BIST-30 and BIST-2022 companies indicated low to medium risks, with no web pages raising alarms in the "High" risk category. Additionally, web server types varied, with unknown or unspecified web server being the most used in the BIST categories. Except for one web page in the BIST-2022 group, all web pages implemented SSL encryption, demonstrating a strong commitment to secure communication. In summary, BIST-2022 companies demonstrated superior performance in accessibility, usability, and security compared to BIST-30. Nevertheless, both BIST-30 and BIST-2022 companies have specific areas that require attention and improvement for overall enhancement.

The study's significance is based on recognizing company web pages as crucial information sources for stakeholders and researchers, filling a notable research gap by evaluating BIST-30 and BIST-2022 company web pages. Similarly, the outcomes have practical implications for equitable access to information on investor relation pages, providing valuable guidance with raising public awareness. The findings reveal the frequency of errors, indicating insufficient compliance to WCAG 2.1 standards for site administrators and developers.

Despite the comprehensive analysis conducted in this study, certain limitations should be acknowledged. First, the study focused solely on automated testing tools to evaluate accessibility, usability, and security, which, while effective for identifying technical issues, may not fully capture user experience and subjective usability aspects. Future research could complement automated analyses with user-based evaluations, incorporating real user feedback and usability testing to gain deeper insights. Second, this study utilized data corresponding to the year 2022, providing a snapshot of the accessibility, usability, and security of investor relations web pages at that time. Given that corporate web pages are subject to change over time due to evolving technologies and company strategies, this study could be repeated periodically (e.g., yearly) in the future. By applying the same methodology to data from different years, researchers can assess

changes in website performance and identify trends in accessibility, usability, and security enhancements. Lastly, expanding the scope to include a broader range of companies or comparing results across different global stock markets could provide a more comprehensive understanding of IR web page quality.

Peer Review Externally peer-reviewed.
Author Contributions Conception/Design of Study- M.G., S.P.; Data Acquisition- M.G.; Data Analysis/Interpretation- M.G., S.P.; Drafting Manuscript- M.G., S.P.; Critical Revision of Manuscript- M.G., S.P.; Final Approval and Accountability- M.G., S.P.
Conflict of Interest The authors have no conflict of interest to declare.
Grant Support The authors declared that this study has received no financial support.

Author Details
Merve Garip
[1] İzmir Bakırçay University, Department of Management Information Systems, İzmir, Türkiye
0009-0007-6441-8288

Serhat Peker
[1] İzmir Bakırçay University, Department of Management Information Systems, İzmir, Türkiye
0000-0002-6876-3982 serhat.peker@bakircay.edu.tr

References

Agrawal, G., Kumar, D., Singh, M., & Dani, D. (2019). Evaluating accessibility and usability of airline websites. *Communications in Computer and Information Science*, *1045*, 392–402. Springer Verlag. doi:10.1007/978-981-13-9939-8_35

Akgül, Y. (2021). Accessibility, usability, quality performance, and readability evaluation of university websites of Turkey: a comparative study of state and private universities. *Universal Access in the Information Society*, *20*(1), 157–170. doi:10.1007/s10209-020-00715-w

Al-Sakran, H. O., & Alsudairi, M. A. (2021). Usability and Accessibility Assessment of Saudi Arabia Mobile E-Government Websites. *IEEE Access*, *9*, 48254–48275. doi: 10.1109/ACCESS.2021.3068917

Ashbaugh, H., Johnstone, K. M., & Warfield, T. D. (1999). Corporate reporting on the internet. *Accounting Horizons*, *13*(3), 241–257. doi:10.2308/acch.1999.13.3.241

Başeğmez, H., & Özdemir, K. (2023). Website Performance Evaluation by Gray Relational Analysis: A Research on Companies in BIST Technology and Informatics Index. *Acta Infologica*, *0*(0), 0–0. doi:10.26650/acin.1165100

Brennan, N. M. and Merkl-Davies, D. M. (2018). Do firms effectively communicate with financial stakeholders? A conceptual model of corporate communication in a capital market context. *Accounting and Business Research*, *48*(5), 553–577. doi:10.1080/00014788.2018.1470143

Chatterjee, K., Gupta, D. and De, A. (2013). "A framework for development of secure software." *CSI Transactions on ICT*, *1*(2), 143–157. doi:10.1007/s40012-013-0010-8

Conte, F., Coppola, C., Sardanelli, D., Vollero, A., & Siano, A. (2022). *Accessibility and social inclusion: an empirical investigation on the adoption of the World Wide Web Consortium guidelines on corporate websites*.

Csontos, B., & Heckl, I. (2021). Accessibility, usability, and security evaluation of Hungarian government websites. *Universal Access in the Information Society*, *20*(1), 139–156. doi:10.1007/s10209-020-00716-9

Domínguez Vila, T., Alén González, E. and Darcy, S. (2018). Website accessibility in the tourism industry: an analysis of official national tourism organization websites around the world. *Disability and Rehabilitation*, *40*(24), 2895–2906. doi:10.1080/09638288.2017.1362709

Fatima, K., Bawany, N. Z., & Bukhari, M. (2020). Usability and accessibility evaluation of banking websites. *2020 International Conference on Advanced Computer Science and Information Systems, ICACSIS 2020*, 247–256. Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/ICACSIS51025.2020.9263083

Fogli, D., & Guida, G. (2013). LNCS 8295-Assessing corporate websites: Quality Model and Methodology. In *LNCS* (Vol. 8295).

Fogli, D. and Guida, G. (2015). A practical approach to the assessment of quality in use of corporate web sites. *Journal of Systems and Software*, *99*, 52–65. https://doi.org/10.1016/j.jss.2014.09.006

"Free Broken Link Checking Tool-Dead Link Checker." (2013). Accessed: Dec. 13, 2023. [Online]. Available: https://www.deadlinkchecker.com/.

Gonçalves, R., Martins, J., Pereira, J., Oliveira, M. A. Y., & Ferreira, J. J. P. (2013). Enterprise Web Accessibility Levels Amongst the Forbes 250: Where Art Thou O Virtuous Leader? *Journal of Business Ethics*, *113*(2), 363–375. doi:10.1007/s10551-012-1309-3

Gonçalves, R., Rocha, T., Martins, J., Branco, F. and Au-Yong-Oliveira, M. (2018). Evaluation of e-commerce websites' accessibility and usability: an e-commerce platform analysis with the inclusion of blind users. *Universal Access in the Information Society*, *17*(3), 567–583. DOI: 10.1007/s10209-017-0557-5

Gopinath, S., Senthooran, V., Lojenaa, N. and Kartheeswaran, T. (2016). Usability and accessibility analysis of selected government websites in Sri Lanka. *IEEE Region 10 Symposium (TENSYMP)*, 394–398.

GTmetrix | Website Performance Testing and Monitoring." (2009). Accessed: Dec. 13, 2023. [Online]. Available: https://gtmetrix.com/

Halim, N. A., Basiruddin, R. and Ali, N. R. M. (2015). Measuring the Extent of Investor Relations on the Web: A Multi-dimensional Approach. *Procedia Economics and Finance*, *31*, 714–721. https://doi.org/10.1016/s2212-5671(15)01160-0

Huang, Z. and Cappel, J. J. (2012). A Comparative Study of Website Usability Practices of Fortune 500 Versus INC. 500 Companies. *Information Systems Management*, *29*(2), 112–122. doi:10.1080/10580530.2012.661633

Kumar Gupta, A., Singh, V., Kulkarni, S., Khatri, V., Sumit,], & Sonkar, S. (2023). Human-Computer Interaction: Designing Intuitive User Experiences [1]. In *Tuijin Jishu/Journal of Propulsion Technology* (Vol. 44).

Leitner, M. L., Strauss, C., & Stummer, C. (2016). Web accessibility implementation in private sector organizations: motivations and business impact. *Universal Access in the Information Society*, *15*(2), 249–260. doi:10.1007/s10209-014-0380-1

Loiacono, E. T., & McCoy, S. (2004). Website accessibility: A cross-sector comparison. *Universal Access in the Information Society*, *4*(4), 393–399. DOI: 10.1007/s10209-005-0003-y

Loiacono, E. T., Romano, N. C. and McCoy, S. (2009). The state of corporate website accessibility. *Communications of the ACM*, *52*(9). https://doi.org/10.1145/1562164.1562197

Macakoglu, S. S., & Peker, S. (2022). Web accessibility performance analysis using web content accessibility guidelines and automated tools: a systematic literature review. *HORA 2022-4th International Congress on Human-Computer Interaction, Optimization and Robotic Applications, Proceedings*. Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/HORA55278.2022.9799981

Macakoğlu, Ş. S., Peker, S., & Medeni, İ. T. (2022). Accessibility, usability, and security evaluation of universities' prospective student web pages: a comparative study of Europe, North America, and Oceania. *Universal Access in the Information Society*. doi:10.1007/s10209-022-00869-9

"Mobil Accessibility Testing-Google Search Console." (2014). Accessed: Dec. 13, 2023. [Online]. Available: https://search.google.com/test/mobile-friendly

Nour, R. (2022). An Assessment of Accessibility and Usability of Saudi Online FinTech Services for People with Disabilities. *Computational and Mathematical Methods in Medicine*, *2022*. doi: 10.1155/2022/8610844

"Pingdom: Website Performance and Availability Monitoring." (2005). Accessed: Dec. 13, 2023. [Online]. Available: https://www.pingdom.com

Sevilla, J., Herrera, G., Martínez, B. and Alcantud, F. (2007, September 1). Web accessibility for individuals with cognitive deficits: A comparative study between an existing commercial Web and its cognitively accessible equivalent. *ACM Transactions on Computer-Human Interaction*, Vol. 14. https://doi.org/10.1145/1279700.1279702

Sicari, S., Rizzardi, A., Grieco, L. A., & Coen-Porisini, A. (2015, January 15). Security, privacy and trust in Internet of things: The road ahead. *Computer Networks*, Vol. 76, pp. 146–164. Elsevier B.V. https://doi.org/10.1016/j.comnet.2014.11.008

Singh, H., & Singh, A. (2020). Factors influencing web accessibility of corporate information: Indian evidence. *International Journal of E-Business Research*, *16*(3), 1–19. https://doi.org/10.4018/IJEBR.2020070101

"Sucuri SiteCheck: Website Security Checker | Malware Scan.." (2007). Accessed: Dec. 13, 2023. [Online]. Available: https://sitecheck.sucuri.net

TAW | Web accessibility and W3C standardization services." (2001). Accessed: Dec. 13, 2023. [Online]. Available: https://www.tawdis.net/index

Web Content Accessibility Guidelines (WCAG) 2.1. (2018, June 5). Retrieved May 19, 2023, from W3C website: https://www.w3.org/TR/WCAG21/#background-on-wcag-2

Xiao, J. Z., Yang, H. and Chow, C. W. (2004). The determinants and characteristics of voluntary Internet-based disclosures by listed Chinese companies. *Journal of Accounting and Public Policy*, *23*(3), 191–225. https://doi.org/10.1016/j.jaccpubpol.2004.04.002

Appendix

See Table 9.

Table 9

List of companies and their IR web pages analyzed

Company	Type	IR Web page URL
Akbank	BIST-30	https://www.akbankinvestorrelations.com/tr/
Aksa Enerji	BIST-30	https://www.aksainvestorrelations.com/tr/
Alarko	BIST-30	https://www.alarko.com.tr/tr/yatirimci-iliskileri/
Arçelik	BIST-30	https://www.arcelikglobal.com/tr/sirket/yatirimci-iliskileri/yatirimci-iliskileri/
Aselsan	BIST-30	https://www.aselsan.com/tr/yatirimci-iliskileri
Bim	BIST-30	https://www.bim.com.tr/Categories/105/yatirimci-iliskileri.aspx
Emlak Konut	BIST-30	https://www.emlakkonut.com.tr/tr-tr/yatirimci-iliskileri
Erdemir	BIST-30	https://www.erdemir.com.tr/yatirimci-iliskileri/
Ford	BIST-30	https://www.fordotosan.com.tr/tr/yatirimcilar/yatirimci-iliskileri-anasayfasi
Garanti BBVA	BIST-30	https://www.garantibbva.com.tr/yatirimci-iliskileri
Gübretaş	BIST-30	https://www.gubretas.com.tr/yatirimci-iliskileri/
Hektaş	BIST-30	https://hektas.com.tr/yatirimci-iliskileri
İş Bankası	BIST-30	https://www.isbank.com.tr/bankamizi-taniyin/yatirimci-iliskileri
Koç Holding	BIST-30	https://www.koc.com.tr/yatirimci-iliskileri
Odaş Enerji	BIST-30	https://www.odasenerji.com.tr/tr/yatirimci-iliskileri/
Petkim	BIST-30	https://www.petkim.com.tr/yatirimci-iliskileri
Pegasus	BIST-30	https://www.pegasusyatirimciiliskileri.com/tr
Sabancı Holding	BIST-30	https://yatirimciiliskileri.sabanci.com/tr/
Sasa Polyester Sanayi	BIST-30	https://www.sasa.com.tr/yatirimci-iliskileri
Şişecam	BIST-30	https://www.sisecam.com.tr/tr/Sayfalar/Inverstor%20Relations/Investor-Relations.aspx
Tav Havalimanları	BIST-30	https://ir.tav.aero/tr-TR/

Company	Type	IR Web page URL
Turkcell	BIST-30	https://m.turkcell.com.tr/tr/hakkimizda/yatirimci-iliskileri
Türk Hava Yolları	BIST-30	https://investor.turkishairlines.com/tr
Tekfen Holding	BIST-30	https://www.tekfen.com.tr/yatirimci-iliskileri-birimi-4-15
Tofaş	BIST-30	https://www.tofas.com.tr/YatirimciIliskileri/GenelBakis#
Tüpraş	BIST-30	https://www.tupras.com.tr/genel-bakis
Tera Yatırım	BIST-2022	https://www.terayatirim.com/kurumsal/yatirimci-iliskileri
Özsu Balık	BIST-2022	https://www.ozsubalik.com.tr/yatirimci/kategori/yatirimci-iliskileri
Sanica Isı	BIST-2022	https://www.sanicaisi.com.tr/yatirimci-iliskileri/
Hidropar	BIST-2022	https://www.hktm.com.tr/yatirimci-iliskileri
Aztek Teknoloji	BIST-2022	https://www.aztekteknoloji.com/yatirimci/
Barem Ambalaj	BIST-2022	https://www.baremambalaj.com/yatirimci-iliskileri.aspx
Obase Bilgisayar	BIST-2022	https://obase.com/yatirimci-iliskileri-ana-sayfa/?lang=tr
Koroplast	BIST-2022	https://www.koroplast.com/tr/yatirimci-iliskileri
Kiler Holding	BIST-2022	https://www.kilerholding.com.tr/yatirimciliskileri.php
Rubenis Tekstil	BIST-2022	https://www.rubenis.com/yatirimci-iliskileri/
Kocaer Çelik	BIST-2022	https://www.kocaersteel.com/tr/investor-relations
Europen Endüstri	BIST-2022	https://yatirimciiliskileri.europen.com.tr
Sun Tekstil	BIST-2022	https://www.suntekstil.com.tr/yatirimci-iliskileri
Kimpur	BIST-2022	https://kimpur.com/tr/yatirimci-iliskileri/
BMS Tel	BIST-2022	https://www.bmstel.com.tr/yatirimci-iliskileri/
Consus Enerji	BIST-2022	https://consusenerji.com.tr/yatirimci-iliskileri/
Smart Güneş	BIST-2022	https://www.smartsolar.com.tr/yatirimci-iliskileri.html
Gür-Sel Turizm	BIST-2022	https://www.gurseltur.com.tr/yatirimci-iliskileri
Gezinomi	BIST-2022	https://www.gezinomi.com/yatirimci-iliskileri

Company	Type	IR Web page URL
Koleksiyon Mobilya	BIST-2022	https://koleksiyon.com.tr/yatirimci-iliskileri/
Hitit Bilgisayar	BIST-2022	https://hitit.com/tr/yatirimci-iliskileri
InvestCo Holding	BIST-2022	https://investco.com.tr/tr/yatirimciiliskileri
DAP Gayrimenkul	BIST-2022	https://dapgayrimenkulgelistirme.com.tr/yatirimci-iliskileri/
Hun Yenilebilir	BIST-2022	https://www.hunyenilenebilirenerji.com/yatirimci-iliskileri
Panelsan	BIST-2022	https://www.panelsan.com/tr/page/35/yatirimci-iliskileri
Oncosem	BIST-2022	https://oncosem.com/yatirimci-iliskileri/
SDT	BIST-2022	https://www.sdt.com.tr/tr/yatirimci-iliskileri

İSTANBUL UNIVERSITY PRESS

Acta Infologica,
ACIN 2025, 9 (1): 112–132

https://doi.org/10.26650/acin.1604516
Submitted: 20.12.2024
Revision Requested: 02.03.2025
Last Revision Received: 12.04.2025
Accepted: 13.04.2025
Published Online 20.05.2025

Acta Infologica

Research Article

Open Access

YOLOv8 for Road DamageRecognition: Deep Learning-Based Segmentation and Detection System

Mehmet Tevfik Ağdaş [1] & Kaan Arık [2]

[1] Computer Technologies, Çemişgezek Vocational School, Munzur University, Tunceli, Türkiye
[2] Multidimensional Modeling and Animation, Information Tech. Vocational School Vocational School, Sakarya Applied Sciences University, Sakarya, Türkiye

Abstract Road surface detection is critical for improving traffic safety and reducing road maintenance costs. Because traditional methods are time-consuming and costly, deep learning-based image processing techniques offer an important alternative in this field. This study aims to develop a model that automates the detection and segmentation of road surface defects such as potholes, manhole covers, and culverts using deep-learning-based image processing techniques. In this study, a dataset previously used in the literature was preferred. It was observed that object detection was performed using a dataset from the literature. In this study, both object detection and object segmentation were performed using different parameters. To prove the success of object segmentation, both object detection and segmentation were performed using the YOLOv8 algorithm, which has previously obtained successful results. AdamW optimization and Auto Batch parameters were selected for this study. With these parameters, object detection was first performed with the YOLOv8s model, which is one of the variances of the YOLOv8 algorithm with the most successful results in the literature, and a successful 92.8% mAP@50 performance value was obtained according to the sources in the literature. In this study, the YOLOv8n, YOLOv8s, YOLOv8m, and YOLOv8l variance models of the YOLOv8 algorithm were used with the preferred parameters, and segmentation was performed. In object segmentation, a map@50 performance value of 90.9% in all classes and 99.1% in culverts was obtained using the YOLOv8l model. A map@50 performance value of 89.1% for pothole segmentation and 88% for manhole cover segmentation was obtained using the YOLOv8s model. The results of the analyses showed consistency in precision and recall values. These findings contribute significantly to improving road safety, reducing maintenance costs, and supporting sustainable urban infrastructure. Future research could explore integrating multiple data sources and adapt these models to more complex road conditions.

Keywords YOLOv8 · pothole detection · road damage segmentation · image preprocessing · surface cracks

Citation: Ağdaş, M. T. & Arık, K. (2025). YOLOv8 for road damageRecognition: Deep learning-based segmentation and detection system. *Acta Infologica*, *9*(1), 112-132. https://doi.org/10.26650/acin.1604516

This work is licensed under Creative Commons Attribution-NonCommercial 4.0 International License.
© 2025. Ağdaş, M. T. & Arık, K.

Corresponding author: Mehmet Tevfik Ağdaş mtagdas@munzur.edu.tr

Introduction

In the last few years, research into artificial intelligence (AI), particularly deep learning, has been very active. It has been widely applied in image processing and object detection technologies. Deep learning is a subset of general AI. It has enabled computers to use large amounts of data with exceptional accuracy, inspired by the neural networks of the human brain. It has applications in agriculture, automotive, commerce, economics, finance, agriculture, gaming, entertainment, and marketing (Han et al., 2008). Going beyond the traditional pattern recognition algorithms used in vision systems, deep learning has helped transform autonomous driving, surveillance, and retail. With this revolutionary approach to image processing, they are now equipped for real-time perception and decision making (Chen et al., 2023; Li et al., 2024). The combination of AI, deep learning, and image processing has a significant impact on various industries worldwide. This is an opportunity that business people cannot afford to miss, helping everyone work more efficiently with information. Rapid developments in this field will influence the way in which we live and the entire global economy in the future.

Image processing is an important area in the transition to deep learning. These networks, extended on Neural Networks (NN) such as Convolutional Neural Networks (CNN) and Recurrent neural networks (RNN), offer successful performance from object detection to tabular data. CNN, as well as RNN, holds a significant position in the area of deep learning. RNNs may be used to process sequential data, making them suitable for solving a variety of data analysis problems (Kumar et al., 2023). CNNs are commonly preferred deep learning architecture for image processing and object recognition tasks. The key benefit of CNNs is that they can learn image data features automatically. According to Kilic et al. (2019) CNN can achieve high success rates in image classification problems. This characteristic has made it a popular choice worldwide, such as in medical imaging, agriculture, and biomedicine. Another common approach for sequential data and time-series forecasting is RNN. Due to their ability to remember information about past events, they are often used in tasks like natural language processing and time series forecasting.

Future studies can expand visualization and analysis methods to handle more complex datasets. In this context, You Only Look Once (YOLO) is a powerful real-time object detection tool. The efficiency of the proposed method stems from requiring only one forward pass for object detection, which makes it significantly faster and more accurate than traditional methods. Chang and Kim (2023) outline the process of creating a real-time object detection system using YOLO, focusing on its impact on both object detection and segmentation. The broad coverage and faster processing rates further highlight the wide application of YOLO in various fields. One of the areas of work mentioned above is to include applications that will increase the level of life of the society and the potential of the service by applying computer vision (Diwan et al., 2022). For example, the automatic detection and segmentation of potholes, bumps, and drainage systems is crucial for maintaining modern city infrastructure. Such applications are essential for improving traffic safety and reducing infrastructure maintenance costs. These algorithms, which are based on advances in real-time object detection, are highly effective in specialized applications, such as pothole detection. In this context, Babulal (2022) investigated the effect of deep learning-based object detection on image segmentation and used the YOLO algorithm to identify bounding boxes and segment objects such as wires and shaded trees. It has been stated that acceptable error rates can be achieved in object detection with appropriate parameter start.

YOLO is also very efficient in object detection, especially in noisy environments. Ghanem and Bakour (2023) examined the object detection ability in noisy images using the YOLO v5 algorithm and found that

the performance in noisy images significantly decreased when the added noise ratio changed. In addition, it stated that image quality is important for complex tasks. In conclusion, in image processing and object detection, in addition to the benefits of the YOLO algorithm, the effects of noise and other ambient conditions should also be considered in terms of accuracy.In the literature, Faster R-CNN, YOLO, and U-Net models are the most widely used deep learning-based models. These models enable easy detection of potholes and road area segmentation, which is vital for monitoring and managing urban infrastructure. The following sections introduce various mean-variance neural network technologies, datasets, and pothole detection applications. The methods developed for the detection of road surface defects range from deep learning-based models to hybrid 2D/3D approaches (Rastogi et al., 2020; Arjapure and Kalbande, 2021; Ragab, 2021; Jakubec et al., 2023).

- **Deep learning-based models** like YOLO, Faster R-CNN, and U-Net are mostly used for pothole detection. In some studies, special network architectures were developed to improve efficiency, i.e., MVGG16-based Faster R-CNN (Safyari et al., 2024; Ghanam and Khaled, 2023)
- **2D/3D Hybrid Approaches** enable classical image processing techniques along with 3D point cloud data and deep learning algorithms. (Safyari et al., 2024)
- **Special Dataset (Data and Metrics):** Very few pothole datasets are available, and even fewer give in to the creation of high-quality diverse datasets. To address this issue, several datasets have been created based on stereo camera images, drones, and crowdsourcing techniques (Bello et al., 2020).
- **Application Areas:** These areas accelerate road surface maintenance and repair activities while lowering costs. Applications for mobile phones and drone-assisted systems can be used to gather data on the status of potholes, bumps, and drainage systems.

DifferentTypes of Road Damage

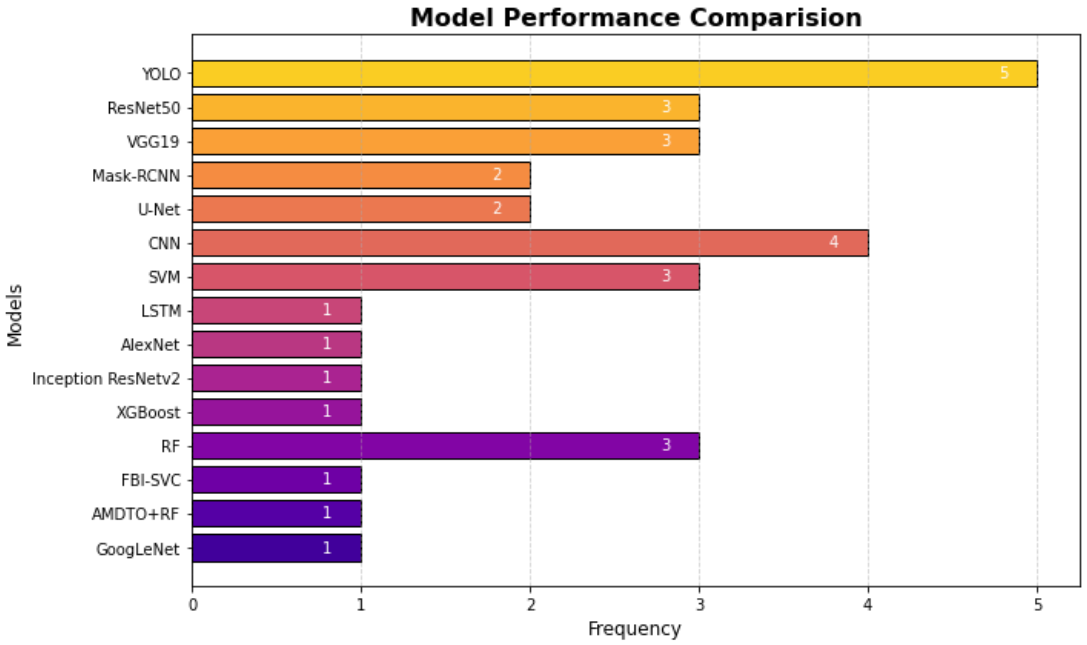
Weather conditions, traffic density, the type of vehicles in use, and the conditions of the road pavement can all lead to road damage. There are many different types of road damage: cracks, potholes, sinkholes, wiped-out lanes, and pedestrian crossings. Potholes become more dangerous the deeper (or bigger) they become. Least of all, small potholes that are often ignored can be fatal, while large potholes cause vehicle tires to burst and get stuck, as well as causing serious accidents. Thus, the features of the pothole need to be extracted and addressed. Detection was performed along with the edge of the pothole, and its size was estimated for this purpose. An edge is a line that separates two different light regions. There are four main types of edges: loop, line, step, and ramp edges. A loop edge is a place in an image where the brightness changes gradually and then returns to its normal value at a certain point. A line edge is an area where the brightness transition occurs abruptly and fades over time. The step edge is an area where the brightness change is abrupt. Here, an edge is a region where the brightness transition is rapid. Edge detection is a technique used to obtain the pixel that corresponds to the edge; this process gets the gradient through a partial differentiation process. This method uses the first differential to detect edges in images by identifying the presence of edges. The sign of the graph gradient indicates the location of bright and dark pixels in the edge pixels, as determined by the second differential. Potholes in the road damage image data are darker than the adjacent environment. Thus, edge detection can extract features in real-time.

Literature Review

It has been mentioned that the detection of various damages on roads is usually performed using deep learning algorithms. Automatic detection of damage such as potholes, bumps, and culverts is based on studies in which image processing and computer vision techniques are at the forefront. Deep learning algorithms work effectively to classify and segment road surface damage and become even more efficient when real-time data processing is performed.

Figure 1

Frequency Distribution of Models in Literature



According to the data in Figure 1, the YOLO model stands out as the most frequently preferred algorithm, being used 5 times, highlighting its strong performance and widespread use in computer vision problems such as object detection. The CNN model ranked second with 4 uses, indicating that traditional CNNs are still popular in the literature. ResNet50, VGG19, SVM, and RF (Random Forest) models can be counted among the widely preferred methods with 3 uses. In particular, ResNet50 and VGG19 are deep learning-based architectures, which suggests that these methods are preferred for complex image classification tasks. The Mask-RCNN and U-Net models are used in detailed operations, such as segmentation with 2 uses. In contrast, algorithms such as LSTM, AlexNet, Inception ResNetv2, XGBoost, FBI-SVC, AMDTO+RF, and GoogleNet were used only once. This suggests that there are areas of use for more niche or specific problems. In general, it has been observed that fast and effective object detection models such as YOLO have a dominant usage trend, and classical CNN and deep learning methods continue to have a strong presence in the literature.

Research conducted by Kim and Ryu (2014) revealed that various pothole detection methods have been developed in recent years to improve road safety and maintenance processes. The first study aimed to develop a decision support system to prevent pothole-related accidents, which are increasing because of climate change. This system aims to create pothole classification guides from the data collected from video recordings, integrate these guides into a 2D database, and accelerate the repair process. Using the YOLO algorithm, Baek and Chung (2020) achieved good accuracy with edge detection algorithms, having previously

converted RGB images to grayscale. In 3D laser data, the highest accuracy, precision, and recall rate on the "watershed" method were achieved by Tsai and Chatterjee (2018), which resulted in 94.97%, 90.80%, and 98.75%, respectively. Hoang et al. (2021) obtained a classification accuracy of 94.833% on publicly available roads using texture-based descriptors to automatically detect patched and unpatched potholes on asphalt road surfaces. With a deep learning-based approach, Saisree and Kumaran (2023) found the VGG19 model to be 97% and 98% accurate for the detection of potholes on highways and muddy roads, respectively.

Image-based, vibration-based, and three-dimensional reconstruction methods for automatic pothole detection were studied by Zhang et al. (2022), who compared their advantages and disadvantages. The road conditions were classified with 93% accuracy using a mobile application developed by Thiruppathiraj (2020). Li et al. (2017) achieved an accuracy of 80% by applying machine learning techniques to pothole detection using smartphone sensors and camera images. These studies propose novel approaches for enhancing road infrastructure and improving traffic safety. Chen et al. proposed a sublimate approach to the road pothole detection problem and proposed an algorithm for object detection in road images based on neural network Net CIFAR 10 classification boundary area identifications; it outperformed another existing detection method. Alhussan (2022) suggested that road potholes should be detected in real time to improve the safety of autonomous cars, namely, an adaptive mutation and dip-throated Optimization (AMDTO)-based Random Forest (RF) classifier as a new feature selection algorithm with an accuracy rate over 99%. Agrawal (2021) proposed a CNN-based model for pothole detection and classified potholes with 88% accuracy and road types with 96% accuracy by evaluating a subset of road images in two groups: paved and unpaved. Kim and Ryu (2014) highlighted the need for deep learning-based studies, emphasized that manual methods are expensive and time-consuming, and outlined future directions to improve the efficiency and accuracy of automated systems.

Ma et. al. (2022) investigated classical image processing, 3D point cloud modeling, and deep learning algorithms for pothole detection using 2D and 3D data collection systems. The study predicts that classical methods are insufficient, CNN-based approaches provide successful results, and multimodal semantic segmentation with self-supervised learning will develop in the future. Rehana and Remya (2022) proposed automatic detection and classification of road damages using the YOLOv5 algorithm and reported that the model achieved 92% accuracy. Al Haqi and Hidayat (2022) developed a system with supervised learning methods to support users' visual assessments and showed that these results depend on human factors.

Considering the studies conducted in 2024 and 2025, Overall, these studies highlight the effectiveness of deep learning-based YOLO models, particularly YOLOv7 and YOLOv8, in detecting and classifying potholes under various conditions, improving road safety, and aiding autonomous navigation and accessibility solutions. Bhavana et al. (2024) introduced a YOLOv8-based pothole detection framework with image preprocessing using Contrast Stretching Adaptive Gaussian Star Filter and Sobel edge detection, achieving 99.10% accuracy, outperforming Faster R-CNN, SSD, and Mask R-CNN. Omar and Kumar (2024) compared YOLOv5, YOLOv6, and YOLOv7 for pothole detection in Intelligent Transportation Systems (ITS), concluding that YOLOv7 achieved the highest precision of 93%. Khan et al. (2024) proposed a YOLOv8-based system (2024) for autonomous vehicles and demonstrated superior pothole detection performance, showing promise for real-time applications in self-driving technology.

Saranya et al. (2024) evaluated real-time pothole detection. They found that it was more efficient than CNN models, with a successful deployment potential in edge devices. Reddy et al. (2024) focused on real-time pothole detection to mitigate traffic congestion and accidents and reported improved detection rates

using YOLO. Paramarthalingam et al. (2024) applied YOLO for visually impaired individuals by integrating an auditory feedback system for safer navigation and achieved 82.7% accuracy. Zanevych et al. (2024) incorporated YOLOv11, Grad-CAM++, and Feature Pyramid Networks and demonstrated significant advancements in nighttime pothole detection, achieving 0.88 mAP50, improving detection under low-light conditions.

Pothole detection research in 2024-2025 has mainly focused on the benchmarking and optimization of advanced deep learning models such as YOLOv7 and YOLOv8. The studies aimed to improve accuracy through image pre-processing techniques, improve detection capabilities at night and in low light conditions, implement real-time applications in autonomous vehicles, provide accessibility solutions for visually impaired individuals, and the applicability of the systems on edge devices. This research has led to significant advances in road safety and intelligent transportation systems.

Therefore, road damage prediction is an important tool in the detection and management of road damages. The integration of deep learning and machine learning methods increases the effectiveness of the proposed technology and accelerates road maintenance processes. In the future, further research in this area will be critical to ensure the sustainability and safety of road infrastructure.

Table 1

Pothole Detection Methods and Their Performance Metrics: A Comprehensive Overview

Year	Authors	Method	Success rate/advantages
2011	Yu and Salari,	CCD image sensors and laser imaging	High precision and efficiency
2017	Li et al.	Machine learning	%80.1 accuracy
2018	Tsai and Chatterje	Watershed method	%94.97 accuracy
2020	Baek and Chung	YOLO, edge detection	%80 above accuracy
2020	Thiruppathiraj	RF, XGBoost, and ANN	%93 accuracy
2020	Dhoundiyal	İmage processing	Fast and economical solutions
2021	Hoang et al.	Texture-based descriptors	94.833% accuracy
2021	Agrawal	CNN	%96 road-type accuracy
2022	Alhussan	AMDTO+RF classifier	%99.795 accuracy
2022	Rehana and Remya	YOLOv5	%92 accuracy
2022	Lin et al.	Field-compatible methods	Data compatibility in different countries
2022	Saha et al.	YOLO	
2023	Saisree and Kumaran	VGG19	97% on motorways and 98% ACC on muddy roads
2024	Bhavana et al.	YOLOv8 with Contrast-Stretching Adaptive Gaussian Star Filter	99.10% accuracy
2024	Omar and Kumar	YOLOv5, YOLOv6, and YOLOv7 comparative analysis	%93.00 accuracy with YOLOv7
2024	Khan et al.	YOLOv8-based system	YOLOv8 significantly outperformed
2024	Saranya et al.	Evaluation of pothole detection by YOLOv7	YOLOv7 is more efficient than the CNN model
2024	Reddy et al.	Real-time pothole detection	YOLOv7 outperformed
2024	Paramarthalingam et al.	Detect potholes using YOLOv5	82.7% accuracy
2024	Zanevych et al.	YOLOv11, Grad-CAM++, and Feature Pyramid Networks	0.88 mAP50

Research breakthroughs have been made in detecting damage to road surfaces, especially potholes, bumps, and culverts, and other automatic detection of depressions is given in Table 1. When the researches in Table 1 are examined, deep learning based algorithms have been applied especially in road surface analysis. In fact, object detection and segmentation algorithms can be particularly beneficial for road maintenance and driver safety by rapidly and accurately identifying these damages. For the classification of road surface deteriorations like potholes, algorithms like YOLO and Mask-RCNN can be used, while models such as CNN and U-Net are also very effective for performing surface segmentation. Using image processing and computer vision methods, these technologies improve traditional manual detection processes with increased efficiency and speed up maintenance processes. At the same time, such applications enable real-time road analysis by processing images collected from vehicles or drones, thus being a part of the growing trend of automation of maintenance and repair work. Although popular algorithms, such as YOLO, ResNet50, SVM, and CNN, are used more frequently, studies have intensively preferred deep learning-based methods, segmentation, and machine learning methods. Machine learning methods (SVM) and deep learning methods (CNN) are superior to these data with various criteria such as accuracy, segmentation, etc.

Dataset

This study aims to automatically detect and segment potholes, culverts, and manholes on roads. Deep learning-based image processing technologies were used in this study. To perform experimental analyses in the image processing process, a dataset containing potholes, culverts, and manhole covers was required. For this study, previously conducted datasets were investigated in this context to obtain the dataset. Khare et al. (2023) used a dataset comprising 3770 images. This dataset was used in the study by Khare et al. as 3033 training data, 491 validation data, and 246 test data. To measure the performance of the dataset accurately, the same separation values were used in this study. Figure 2 shows examples of the classes contained in the dataset.

Figure 2

Sample Images from Different Categories



Data Collection andPre-Processing

Image pre-processing is critical in the first step of our study because it can improve the data quality and relevance to object detection. This project used a dataset that contained road surface anomalies (potholes, bumps, and drainage covers) in high-resolution images. These images were taken from real-life environments and cover a variety of lighting, weather, and road surface conditions. The bounding boxes indicating the locations of the anomalies in the dataset are also fed into this dataset, allowing them to be labelled, which is used as the ground truth by the object detection model.

For the dataset to be pre-processed for object detection, the following methods were used:

- **Resize Process:** Images are resized to the same resolution to facilitate their passing as input to the detection model.
- **Normalization:** All pixel values were normalized to be between 0 and 255, adjusted according to the chosen model architecture, which helped the model converge faster during training.
- **Data Augmentation:** Depending on the variety of road conditions on which the model will be implemented, data augmentation in terms of rotation, flipping, and color jittering will be performed. Therefore, this approach expands the dataset and helps prevent overfitting.
- **Noise Reduction:** Some images in the dataset are noisy, which can result from environmental conditions. In the data preparation process, noise reduction techniques (e.g., Gaussian blurring, median filtering) are employed to maintain the salient features while removing the noise.
- **Bounding Box Adjustment:** Bounding boxes are adjusted as necessary. This can be helpful in situations where initial annotations may have minor deviations due to human error.

All these pre-processing measures are crucial to enhance the model's capacity to precisely identify and categorize potholes, bumps, and drainage covers during the final detection stage, resulting in more reliable outcomes in practical environments.

In the study where the dataset was obtained, various pre-processing operations were performed on the existing data;

- **Image Flip:** Horizontal flipping of images
- **Image Scaling:** The scaling process of images was performed and 640x640 image size.
- **Motion Blurring:** The motions in the images are blurred.
- **Color Manipulation:** RGB color adjustment to adapt images to low ambient conditions
- **Fog addition:** This step adds fog to images.

In this study, unlike the current study, not only horizontal orientation but also Auto-Orient, that is, automatic orientation, was applied. Because of this pre-processing process, the dataset size was increased by 43. Finally, the dataset included 3076 training, 491 validation, and 246 test data.

Implementation and Evaluation of Model

This work poses the performance of road damage with a mixed dataset on an NVIDIA P100 on Google Colab Plus. In this dataset, the input image sizes contain different image sizes. But, 640x640 image size was applied to the dataset for the YOLO algorithm. The dataset was split into 70% training, 20% validation, and 10% test sets to prevent overfitting. Edge detection is performed for pothole classification, and the performance of the model is evaluated based on image data corruption and repair rates, classification results obtained during the final test stage, and the efficiency of the suggested approaches. Results were reported by considering the mAP, precision, and recall metrics. Those metrics; mAP (mean average precision), precision, and recall are used to evaluate the model's accuracy and performance. Precision measures the ratio of correctly predicted objects or segmentations to all predictions, which indicates the accuracy of the model. Recall calculates the ratio of correctly detected objects or segmentations to the total number of real objects, indicating the comprehensiveness of the model. mAP summarizes the model's performance at different thresholds by combining the precision and recall values; it is a metric that evaluates the overall

performance of the model. These metrics are critical for understanding the model's ability to perform accurate detection and segmentation while balancing false positives and missing detections.

Method

Object detection is a popular computer vision task, and one of the most successful solutions is YOLOv8 (You Only Look Once) model. This model is notable because it handles images in one network while simultaneously completing the classification and location estimation (bounding box) tasks. YOLOv6 and YOLOv7 increased the object detection speed and improved the model accuracy. The model splits the input image into a grid of cells (in fact using k-means clustering) and, for each cell, returns whether an object is present in that cell according to the confidence/threshold. Instead of using regression to classify and locate objects, the proposed YOLOv8 uses direct linear classification methods. This eliminates the need for a model to predict accurately in advance, which makes it a much faster model.

Lidar-based models have been used in the literature. The Lidar-based model uses the laser method. Laser beams are sent parallel to the target area, and gaps are detected. In this study, it is aimed to detect road defects using image processing technology. Road defects can be detected using these two methods. When road defects are to be detected by a moving highway vehicle, the use of image processing technologies is expected to provide more efficient operation and results, considering that the process of sending laser beams parallel to the road will cause various problems. For this reason, image processing was preferred in this study.

In the proposed YOLOv8, CNN (Convolutional Neural Network) is the primary network. The new version of the network is deeper but has fewer parameters, which decreases the training time and computational burden. YOLOv8 can work in real-time, which is one of the most significant innovations of the proposed model. Then, for each cell, it is calculated whether there are objects in a certain area of the image. One of the four parameters is estimated: class probability, center coordinates, width, and height. Yolov8 also refines regression strategies to improve localization results. The neural network learns the estimated bounding boxes and class probabilities using this model.

In the Yolov8 model, an input image is split into an S×S grid, and B bounding boxes are predicted for each cell, with each class being a C class predicted for every class. This can be represented as the estimated outputs for each cell in Formula 1:

$$p_{cell} = (x, y, h, p_1, p_2, ..., p_c) \tag{1}$$

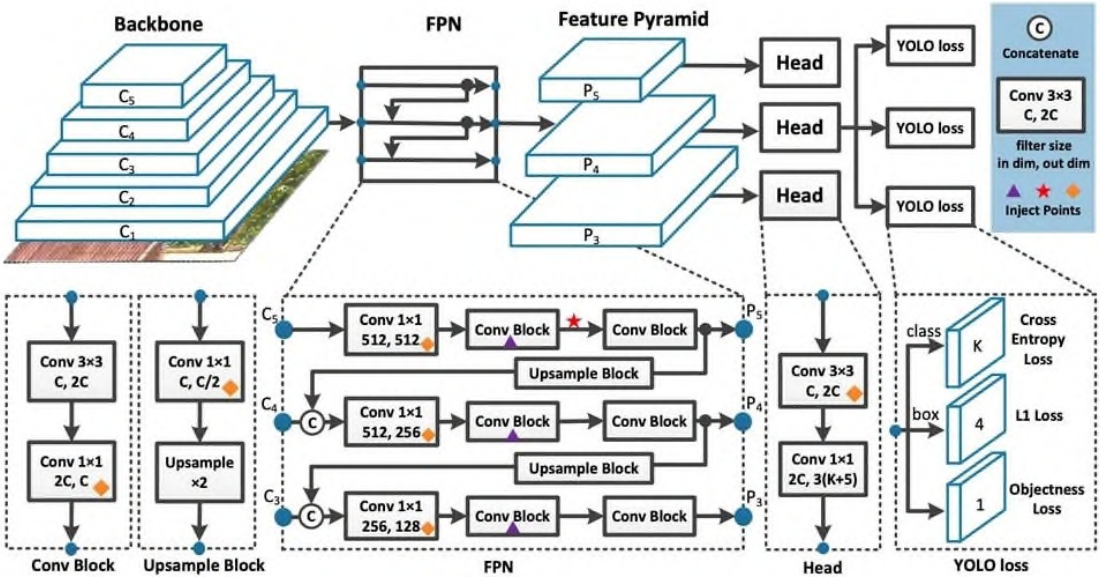
where x and y are the center coordinates of the object; w and h are the width and height of the object; p1, p2, p(c) are the probability values of each class. For this purpose, it employs a certain loss function, focusing on the predictions. The loss function is responsible for optimizing the class and location accuracy in YOLOv8. Therefore, the loss can be expressed as follows in Formula 2:

$$\mathcal{L} = \lambda_{coord}\Sigma_i\mathcal{L}_{coord}(i) + \lambda_{obj}\Sigma_i\mathcal{L}_{obj}(i) + \lambda_{coord}\Sigma_i\mathcal{L}_{noobj}(i) + \lambda_{cls}\Sigma_i\mathcal{L}_{cls}(i) \tag{2}$$

This combination of components helps improve the model's accuracy. There are many optimizations in Yolov8, which are now also used to obtain better and quicker results. The success of this model stems from the new generation of structures that offer context awareness and optimized training strategies. Yolov8 is one of the most powerful steps influencing the future of deep learning object detection and segmentation (Figure 3).

Figure 3

Algorithmic Schema of Yolov8



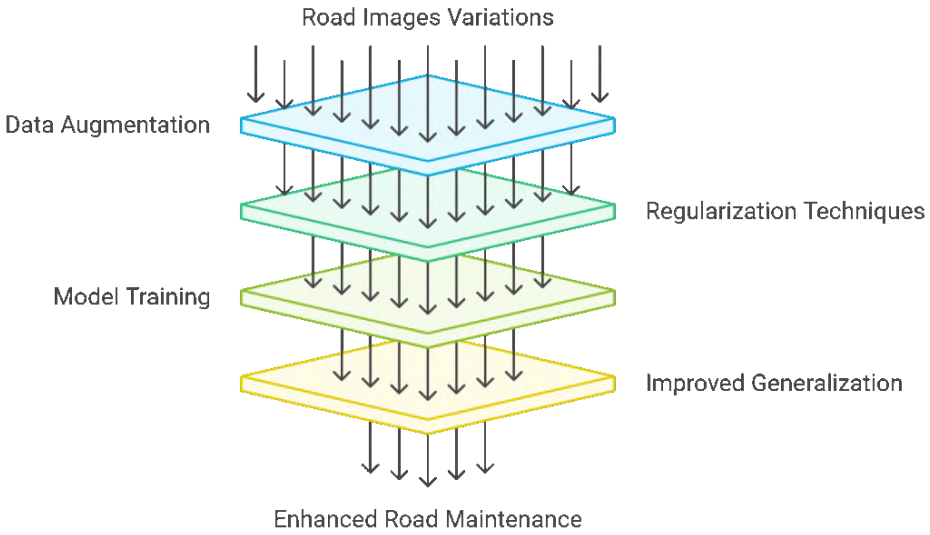
Detecting road damages is an important task because it helps improve road safety by avoiding accidents caused by potholes and cracks (Fan et al., 2021; Nawale et al., 2023). However, conventional classification models that primarily determine the presence of potholes, segmentation, and detection are designed to find and classify the exact location and contours of potholes in road images (Atluri&Bandi, 2022; Fan, Wang, Wang, Liu, & Pitas, 2021). In fact, this step is very important because it involves not only identifying the existence of a pothole but also determining its exact position and size, which are necessary for maintenance and repair planning (Fan, Wang, Bocus, & Liu, 2020; Fan et al., 2021). Segmentation involves determining which pixels of the image belong to the pothole (i.e., that represent the pothole in the image), while detection describes only the localization of the pothole, that is, predicting bounding boxes or masks around the damaged area.

Segmenting and detecting roads also have an additional challenge in background noise analysis like cracks or other road defects, which are not of immediate concern (A Review of Vision-Based Pothole Detection Methods Using Deep Learning, 2023; Fan, Wang, Bocus, & Liu, 2020). Also, we had to filter out these non-pothole objects, as we wanted the model to only pay attention to the pothole regions (Atluri&Bandi, 2022; Nawale et al., 2023). This is done by pre-processing the images by detecting only non-pothole objects using the YOLO algorithm and edge detection to detect the pothole boundaries to ensure that only the right boundaries of the potholes are preserved (Khare et al., 2023, Fan et al., 2021). As the background can be standardized to a white background, it increases the visibility of segmentation and detection, thus helping the model to separate potholes even better from other features.

Finally, data augmentation and regularization are often used to increase the robustness and accuracy of segmentation and detection models. This approach of feeding new variations of road images will help the model generalize through multiple types of potholes and road conditions. Techniques such as dropout and batch normalization may be applied at the training stage to avoid overfitting and ensure that the model is not biased toward the examples used for training, which may not necessarily be representative. In summary, in relation to road maintenance, pothole segmentation and detection using edge detection and modern object detection algorithms like YOLO serve as a crucial methodology for detecting and localizing road deterioration, enhancing road maintenance operations, and ultimately bolstering road safety.

Figure 4

Road Damage Image Analysis Process for Our Work



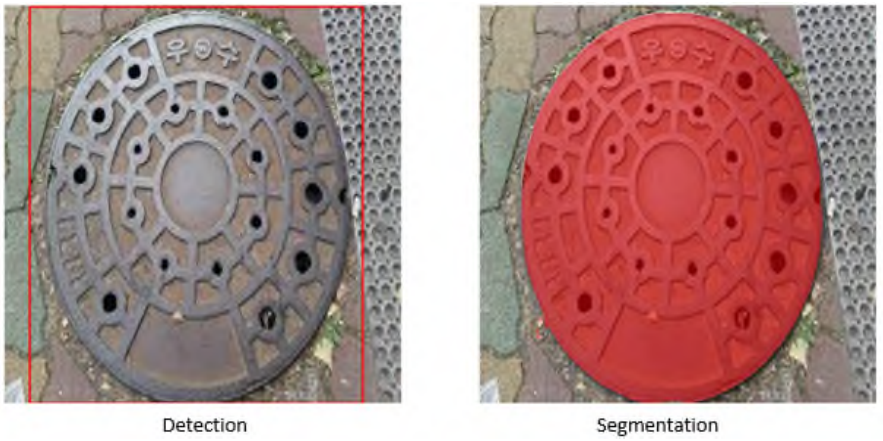
According to the model shown in Figure 4, in the first step, Data Augmentation, allows the model to learn different situations by creating variations in the road images. In the second step, the risk of overfitting is reduced using regularization techniques. In the third step, Model Training is performed, and the generalizability of the model is increased by the techniques developed in this process. In the last step, this improved model improves the efficiency of field operations by providing Enhanced Road Maintenance. This structure shows a systematic progression from data to real-world applications. The dataset used in this study consisted of images taken from wet and dry ground. In particular, for the pothole data, images with dry and water-filled potholes were selected.

Findings and Performance Evaluation

In this study, the detection and segmentation of potholes, culverts, and manhole covers on roads were carried out. In the detection of images using image processing technologies, the target area is enclosed in a rectangle. However, if segmented, the images enclosed in a rectangle are limited to two dimensions. In this way, the target is revealed more clearly. Figure 5 shows the sample detection and segmentation outputs.

Figure 5

Samples of Detection and Segmentation



Training processes were performed on the experimental YOLOv8s, YOLOv8s, and segmentation datasets and YOLOv8m datasets. Some hyper-parameters were used in the experimental training process. The values of these parameters are given in the Table 2 compared with those of the study from which the dataset was taken.

Table 2

Hyper-parameters and Values on Our Implementation

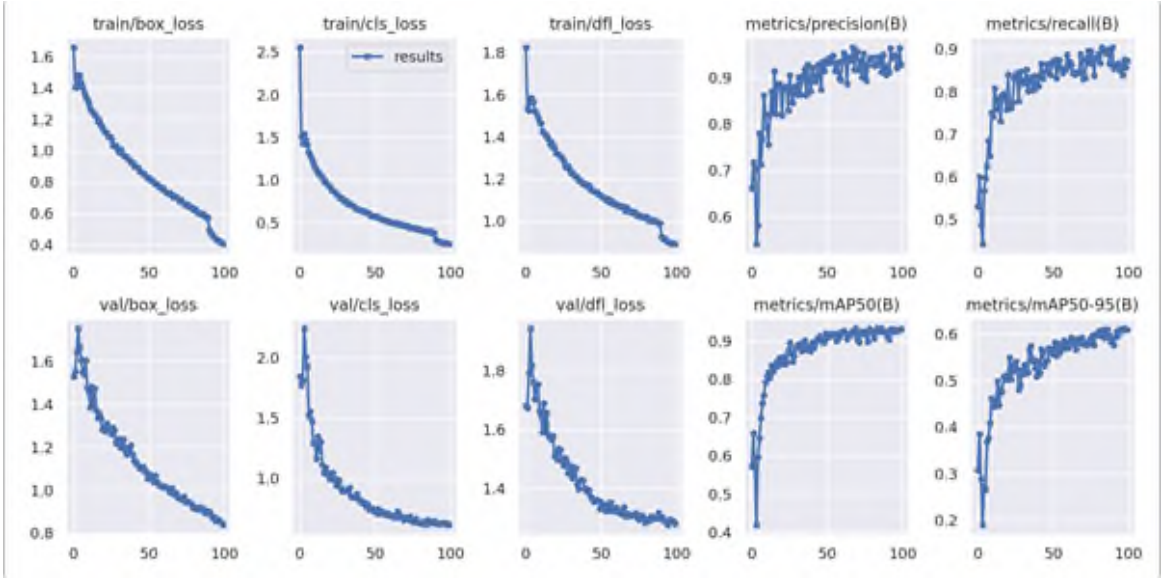
Hyperparameter	Kharevd. (2023) Value	Our Value
Epoch	250	**100**
Batch Size	16	**Auto**
Learning Rate	0.01	0.01
Weight Decay	0.0005	0.0005
Optimizer	Adam	**AdamW**
Momentum	0.937	0.937

Unlike the study of Khare et al. (2023), the epoch value was 100, and the batch size value (-1), that is, automatic, and AdamW was used for optimization. Because of the experimental training process, the performance values obtained from the models are given in Table 2.

YOLOv8s Detection: YOLOv8s and the parameters given in the hyperparameter table, 87.6% recall, 92.7% precision, and 92.8% mAP50 performance values were obtained in the detection process of images. In the present study, the detection process was performed only at this stage. The results of this detection process were compared with those of an experimental study using the dataset. Thus, the success of the hyperparameters used in this study was tested. The detection and segmentation processes were continued according to their success. In addition, the performance outputs of this model are presented in the visual below.

Figure 6

Graphs from Training Process of Our YOLO Model



Graphics from the traning process of our YOLO model are shown in Figure 6. The graphs obtained during the training process of the proposed YOLO model demonstrate that the model's performance improved steadily. The bounding box loss (box_loss), classification loss (cls_loss), and distribution focal loss (dfl_loss)

values in the training and validation sets have decreased regularly; this proves that the model is successful in correctly detecting object coordinates and class prediction. In particular, the similar decrease in training and validation losses indicates that the model avoids the overfitting problem and exhibits strong generalizability.

In terms of performance metrics, precision increased to 90% and recall increased to 85%. In addition, the mAP@50 and mAP@50-95 values have shown a steady increase during the training process, revealing that the model made high-accuracy predictions even at different IoU thresholds. These results demonstrate that the proposed YOLO model increases the true positive prediction rate and reduces the number of missed objects in the object detection task, indicating that the model is trained successfully and reliably. The performances obtained in the study of Khare et al. (2023) are comparatively given in the Table 3.

Table 3

mAP@50, processing time, and model size in different works

Authors	Model	mAP@50 (%)	Processing Time	Size of Model
Kharevd (2023)	YOLOv5n	0.84	28 ms	14.8 mb
	YOLOv5s	0.86	38 ms	15.1 mb
	YOLOv7	0.90	35 ms	74.8 mb
	YOLOv8n	0.911	8.8 ms	6.3 mb
	YOLOv8s	0.92	11 ms	21.5 mb
Our Model	YOLOv8s	**0.928**	5.9 ms	21.5 mb

In this study, the dataset of Khare et al. (2023) was used. Table 3 compares this study with the previous study. In our study, the best performance and success values were obtained using the YOLOv8s model. The primary purpose of our study was to perform segmentation. However, to observe the performance of the most appropriate parameters during segmentation, the appropriate parameters were first selected, and the most successful model in Khare et al. (2023) was tested. When the selected parameters and the YOLOv8s model were used, the most successful result was obtained, as shown in Table 3. This indicates that the proposed parameters yield successful results. These parameters were used in the partitioning process, which is the main objective of this study. The reason why the object detection Map@50 values are different in Table 3 and Table 4 is that Table 3 shows the one-time object detection result, and Table 4 shows the simultaneous object detection value of this segmentation process.

Instance Segmentation (Detection and Segmentation): At this stage of the study, dataset training was performed using image processing technologies to detect and segment potholes, culverts, and manhole covers from road defects. The YOLOv8s and YOLOv8m models were used from the image processing models. The performance values of the YOLOv8s and YOLOv8m models are given in Table 4.

Table 4

Performance metrics on YOLOv8s, YOLOv8m, YOLOv8n, and YOLOv8l

		Detection (%)				Segmentation (%)			
Model	Class	Precision	Recall	mAP@50	mAP@50-95	Precision	Recall	mAP@50	mAP@50-95
YOLOv8n	All	88.7	85.6	**91.1**	59	87.8	84.8	90.3	56.8
	Manhole	93.5	90.7	98.4	63	93.5	90.7	98.4	61.7
	Sewer Cover	92.1	85.1	**89.3**	54.2	89.4	82.7	86.8	52.6

		Detection (%)				Segmentation (%)			
	Pothole	80.6	81	85.7	59.7	80.6	81	85.8	56.1
YOLOv8s	All	89.4	84	88.7	57.3	88.5	83.2	87.9	54.7
	Manhole	86.6	80.8	86.5	48	86.6	80.8	86.5	49.4
	Sewer Cover	91.2	87.8	90.3	55.2	88.7	85.4	88	49.8
	Pothole	90.3	83.3	**89.5**	68.7	90.3	83.3	89.1	64.7
YOLOv8m	All	91.1	84.3	90.8	61.3	90.2	83.5	89.3	58.8
	Manhole	93.3	87.4	93.9	59.7	93.3	87.4	93.9	59.6
	Sewer Cover	89.5	83.5	89.4	56.1	87	81.3	85.8	51.8
	Pothole	90.4	81.9	88.9	68	90.3	81.8	88.3	63.9
YOLOv8l	All	93.9	78.7	91	59.8	93.9	78.7	90.9	58.2
	Manhole	1	83.8	**99.1**	64.5	1	83.9	99.1	64.3
	Sewer Cover	88.8	82.9	86.8	52.6	88.6	82.9	86.8	51.3
	Pothole	92.8	69.3	87	62.4	92.9	69.4	86.7	59

According to the results, when the four models are compared, it can be seen that the YOLOv8l model produces more successful results in the All and Manhole classes than the thermal models. The YOLOv8s model produces more successful results in the Sewer Cover and Pothole classes than the thermal models. In general, when all classes are considered, it is concluded that the YOLOv8l model produces more successful results. When looking outside the YOLOv8l model, it can be seen that the YOLOv8n model produces more successful results than the thermal models in the All and Manhole classes. According to Table 4, when interpreting the results of the analysis, on average, the most successful models in all classes are the YOLOv8l and YOLOv8n models.

In the study by Khare et al. (2023), where the dataset was obtained, detection was performed. In this study, detection and simultaneous detection and segmentation were performed. The difference between our work and that of Khare et al. (2023) is that we used AdamW optimization, automatic batch (-1) and epoch 100 parameters, as shown in Table 2. With these parameters, we applied detection to the same dataset as the YoloV8s model, which achieved the most successful 92% mAP@50 performance in the study by Khare et al. With the parameters applied in this study, it has been shown that successful results can be obtained by obtaining 92.8% mAP@50 performance according to the data presented in Table 3. In this study, in addition, segmentation was performed with the existing data set and the parameters that we have pre-tuned, which we think will contribute to the literature. In the segmentation process, different variances of the Yolov8 model, which has the most successful results in Khare et al. (2023), were used. Table 4 presents the results of these variance models. Detection is also performed by the system during the simultaneous training process for the experimental training of segmentation. The detection process given in Table 4 is performed within the segmentation training; thus, it differs from the detection training results shown in Table 3. The most successful segmentation experimental result was obtained with the Yolov8l model, which extracts the Yolov8 variances.

Figure 7

YOLOv8n performance metrics

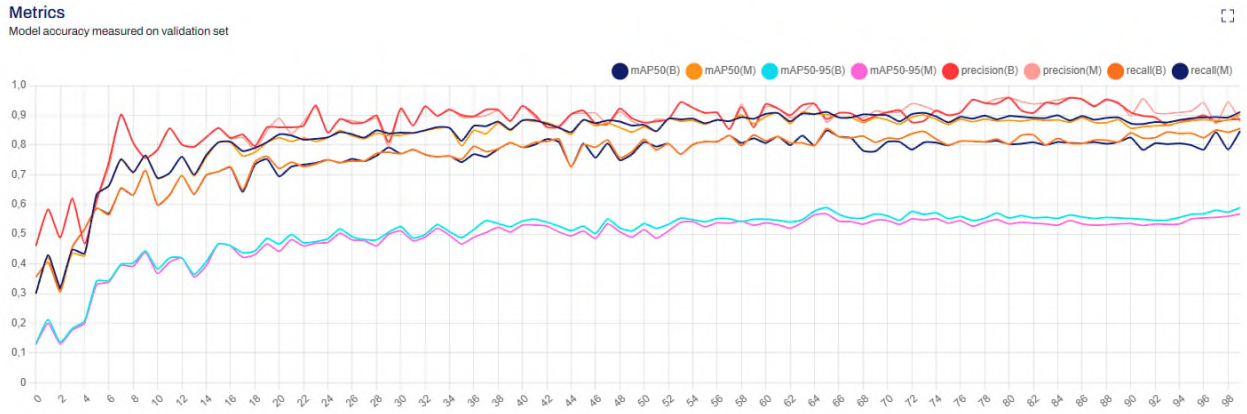
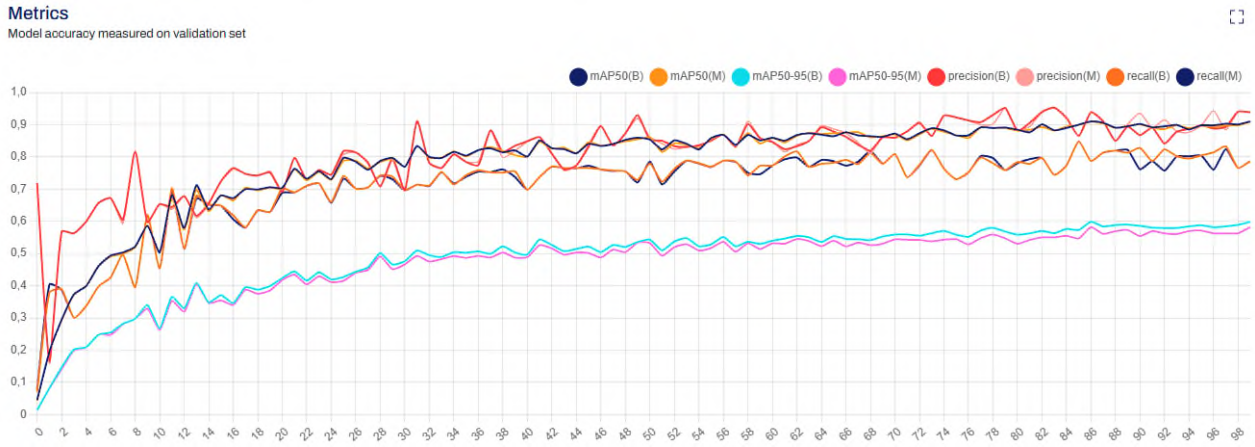


Figure 8

YOLOv8l performance metrics



Looking at the visualization of the performance metrics of YOLOv8n and YOLOv8l models, it can be seen that the YOLOv8n model has a parabolic distribution similar to YOLOv8l model. It can be seen that there is no memorization error in either model in Figure 7 and Figure 8.

Figure 9

YOLOv8n performance parameters

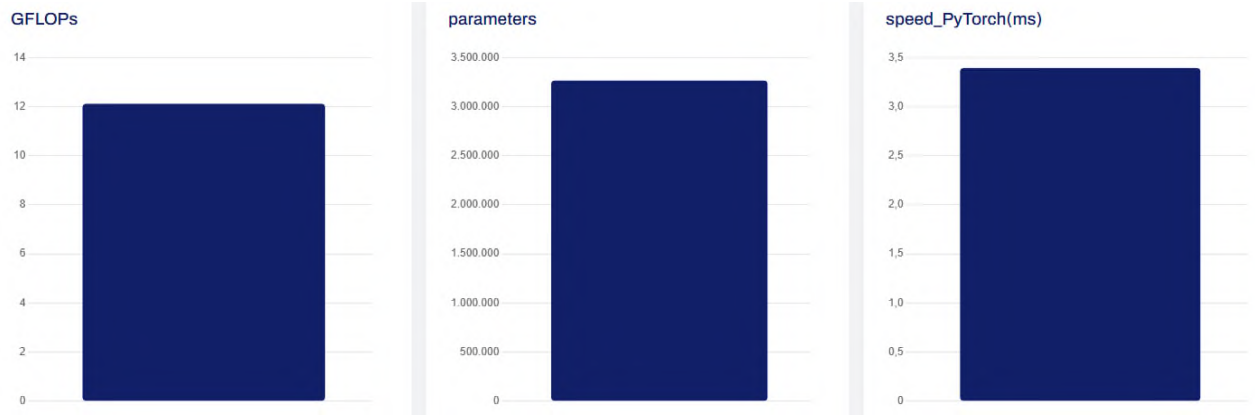
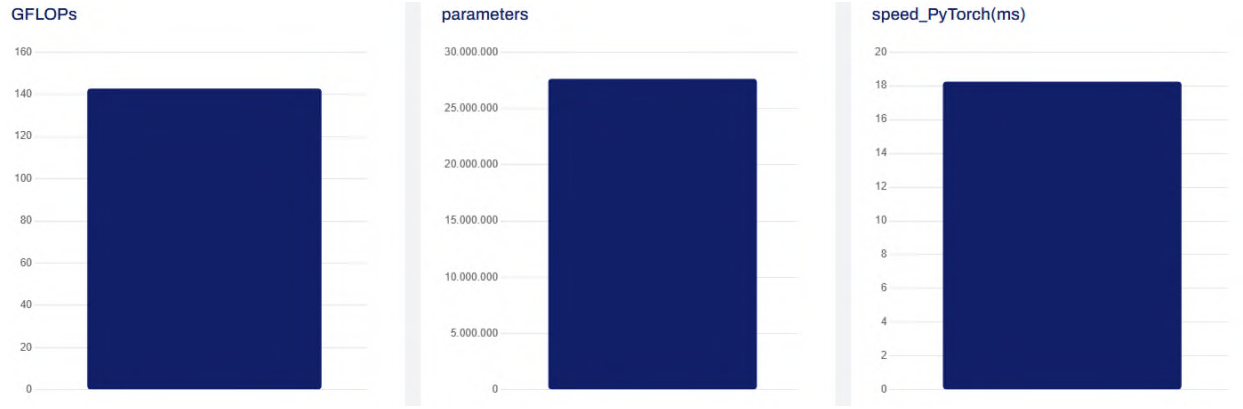


Figure 10

YOLOv8l performance parameters

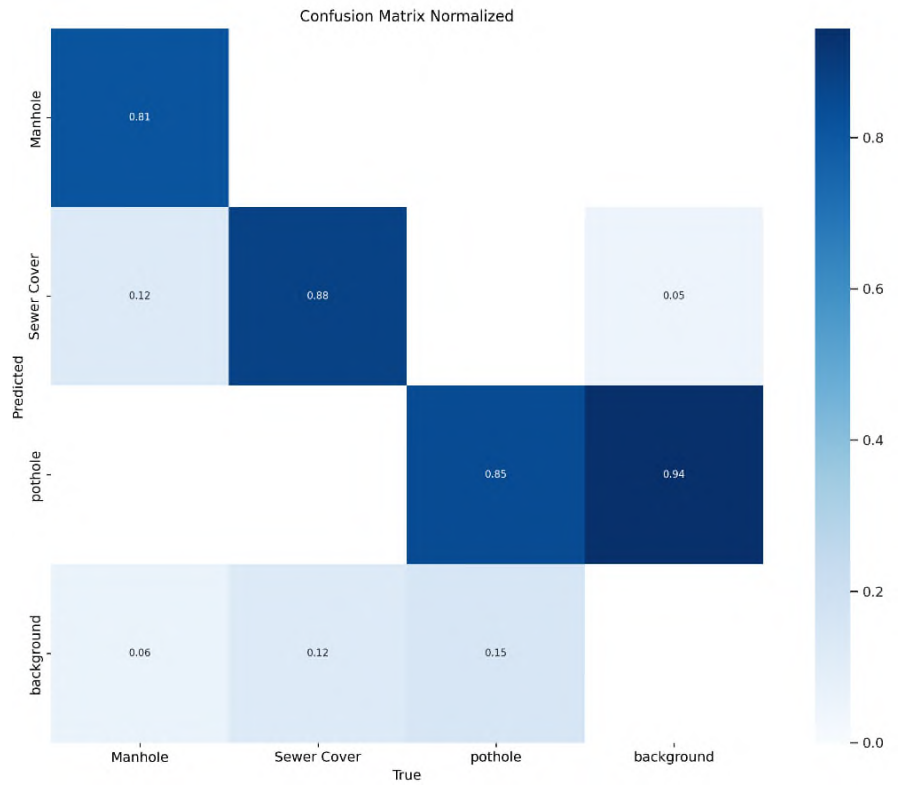


According to the graphs shown in Figure 9 and Figure 10, When the parametric performances of the YOLOv8n and YOLOv8l models are examined, it is seen that YOLOv8l model works with higher hyperparameters and is faster. The YOLOv8n model also works with a processing time of 3.7 ms and a size of 6.48 mb/s. It is observed that YOLOv8l model works with 17.2 ms processing time and 53.3 mb/s size of 53.3 mb/s. According to these data, the YOLOv8l model was more successful with its data read speed per second.

An analysis of the hyperparameter and Yolov8 variance models used in this study and the data obtained because of training shows that the Yolov8l model was the most successful. The confusion matrix values obtained from training the dataset using the Yolov8l model are shown in Figure 9 and Figure 10.

Figure 11

Yolov8l Confusion Matrix Chart



As shown in Figure 11, the confusion matrix elucidates the classification efficacy of the road defect detection model. The analysis of the normalized values reveals exceptional performance across the primary categories. The model demonstrated high accuracy in identifying manholes with a classification rate of 0.89, while sewer covers and potholes exhibited robust detection rates of 0.88 and 0.85, respectively.

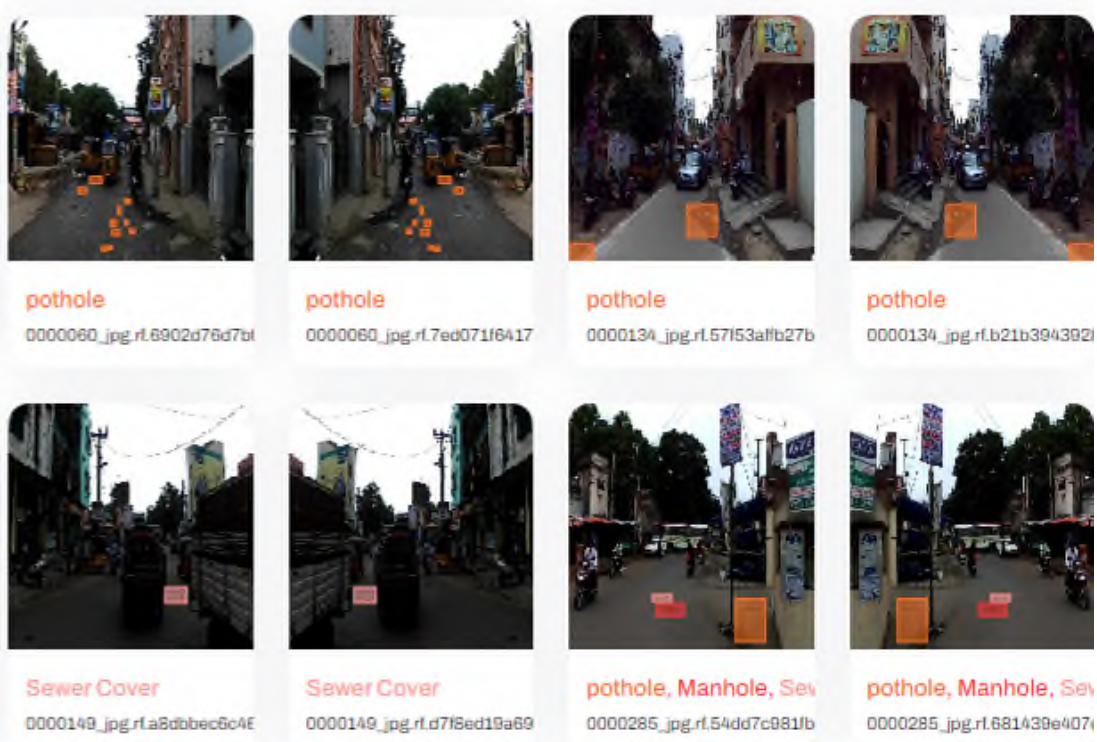
Examination of the error distribution indicates minor classification challenges between categories. Specifically, the model occasionally misclassifies sewer covers as background elements (0.12 misclassification rate) and confuses potholes with sewer covers in approximately 10% of cases. Similarly, manholes are erroneously identified as sewer covers in 8% of the cases. These metrics substantiate the model's high discriminative capability for critical road safety features. The pronounced diagonal values in the confusion matrix affirm the algorithm's reliability in distinguishing various road defects. In infrastructure maintenance applications, accurate identification of potentially hazardous road conditions is paramount.

The minimal cross-category confusion demonstrates the model's refined feature extraction capabilities and suggests effective hyperparameter optimization. The classification outcomes documented in this confusion matrix represent a significant advancement in automated road defect detection systems, with potential applications in preventive maintenance planning and road safety enhancement initiatives.

An experimental study was performed with the YOLOv8m algorithm and sample test data is shown in Figure 12.

Figure 12

Samples of YOLOv8m Test Data



Results and Discussion

According to the results of this study, the detection and segmentation of road surface defects were successfully performed using the YOLOv8-based model. In the detection of road defects, the success of the

parameters was demonstrated by adding hyperparameters to the dataset and model in the study by Khari et al. (2023) and achieving a success value of 92.8% for mAP@50. Segmentation was performed on the same dataset using the parameters that demonstrated successful object detection. In the segmentation process, unlike single-object detection, the variances of the YOLOv8 model were used. The YOLOv8n, YOLOv8s, YOLOv8m, and YOLOv8l variances of the YOLOV8 model were used.

In addition, segmentation analyses were performed with the YOLOv8n model with 98.4% manhole, 86.8% sewer cover, 85.8% pothole, and 90.3% mAP@50 performance value. With the YOLOv8s model, segmentation analyses were performed using 86.5% manhole, 88% sewer cover, 89.1% pothole, and 87.9% mAP@50 performance value. With the YOLOv8m model, manholes were 93.9%, sewer covers were 85.8%, potholes were 88.3%, and all classes were 89.3% with the mAP@50 performance value. With the YOLOv8l model, 99.1% manhole, 86.8% sewer cover, 86.7% pothole, and all classes with a performance value of 90.9% mAP@50 were performed. In the segmentation analysis processes, the most successful Manhole 99.1% mAP@50 performance value was obtained with the YOLOv8l model, Sewer Cover 88% mAP@50 performance value was obtained with the YOLOv8s model, Pothole 89.1% mAP@50 performance value was obtained with the YOLOv8s model, and the average 90.9% mAP@50 performance value in all classes was obtained using the YOLOv8l model.

Furthermore, analyses on test data showed that the YOLOv8n, YOLOv8s, YOLOv8m, and YOLOv8l models showed significant detection and segmentation performance, but the YOLOv8l model had higher data processing capacity (Tang, Li, & Wang, 2023; ultralytics, 2024). These findings support the usability of the developed model under real-life conditions and show that it is an important step to accelerate road maintenance processes. In future, studies can be conducted on the integration of the model with drone-based data collection systems and adaptation to different geographical conditions.

During the study, we observed that image preprocessing techniques significantly affected model performance. The noise reduction and data augmentation methods enabled the model to generalize better in noisy environments; this increased the accuracy of the data obtained, especially in outdoor conditions. In addition, the fast-working capacity of the YOLOv8 algorithm confirmed its suitability for real-time road analysis.

In comparison with previous studies, this study achieved higher mAP@50 values through hyperparameter optimization. While Khare et al. (2023) reported 91.1% mAP using YOLOv5, our implementation with YOLOv8 reached 92.8% mAP, demonstrating the superior detection and segmentation capabilities of YOLOv8's optimized architecture. Our research used the model variants from Khare et al. (2023) with parametric presets applied to the existing dataset, allowing us to evaluate the effectiveness of the parameter adjustments while maintaining a comparable framework.

As a result of the study by Khare et al. (2023), a detection process was performed and contributed to the literature. In this study, the dataset used in Khare et al. (2023) was used, and more successful results were obtained by proposing a different method. It has been shown that more successful results can be obtained using hyperparameters in such studies. In addition to detection, segmentation was performed. The use of segmentation systems in road-defect detection and similar studies can be a successful method for detecting defects and marking their boundaries. It has been shown that different performance values can be obtained by using different variances of the models commonly used in the literature.

Future work may focus on feature fusion techniques and integration with different deep learning architectures to further improve the performance of the proposed YOLOv8 algorithm. In addition, the

generalization capability of the model can be increased by using larger and more diverse datasets; in particular, the success of the model can be tested under harsh weather conditions, such as extreme lighting, shadow, or precipitation. However, the development of models with low hardware requirements will increase the efficiency of real-time applications. Finally, building on the current achievements of YOLOv8, detailed analyses, such as multi-class detection and grading of different road defects, can be added to provide more comprehensive solutions for maintenance prioritization systems.

Peer Review: Externally peer-reviewed.
Author Contributions: Conception/Design of Study- M.T.A., K.A.; Data Acquisition- M.T.A., K.A.; Data Analysis/Interpretation- M.T.A., K.A.; Drafting Manuscript- M.T.A., K.A.; Critical Revision of Manuscript- M.T.A., K.A.; Final Approval and Accountability- M.T.A., K.A.
Conflict of Interest: The authors have no conflict of interest to declare.
Grant Support: The authors declared that this study has received no financial support.

Author Details

Mehmet Tevfik Ağdaş
[1] Computer Technologies, Çemişgezek Vocational School, Munzur University, Tunceli, Türkiye
0000-0002-5608-6240 mtagdas@munzur.edu.tr

Kaan Arık
[2] Multidimensional Modeling and Animation, Information Tech. Vocational School Vocational School, Sakarya Applied Sciences University, Sakarya, Türkiye
0000-0002-0930-8955

References

Agrawal, R., Chhadva, Y., Addagarla, S., and Chaudhari, S. (2021). *Road surface classification and subsequent pothole detection using deep learning.* In 2021 2nd International Conference for Emerging Technology (INCET) (pp. 1-6). IEEE. https://doi.org/10.1109/INCET51464.2021.9456126.

Ahmed, K. R. (2021). Smart pothole detection using deep learning based on dilated convolution. *Sensors*, 21(24), 8406. https://doi.org/10.3390/s21248406.

Alhussan, A., Khafaga, D. S., El-Kenawy, E. S. M., Ibrahim, A., Eid, M. M., and Abdelhamid, A. A. (2022). Pothole and plain road classification using adaptive mutation dipper throated optimization and transfer learning for self driving cars. *IEEE Access*, 10, 84188-84211. https://doi.org/10.1109/ACCESS.2022.3196670.

Al Haqi, N., and Hidayat, F. (2022). *Classification of road damage using supervised learning to assist visual assessment of road damage.* In 2022 International Conference on Information Technology Systems and Innovation (ICITSI). IEEE. https://doi.org/10.1109/icitsi56531.2022.9971082.

Arjapure, S., and Kalbande, D. (2021). *Deep learning model for pothole detection and area computation.* In 2021 International Conference on Communication information and Computing Technology (ICCICT) (pp. 1-6). IEEE. https://doi.org/10.1109/ICCICT50803.2021.9510153.

Atluri, S. and Bandi, S. (2022). Segmentation of potholes from road images. *IEEE Xplore*. doi:10.1109/ICACCS54159.2022.9716468.

Babulal, K. S., and Das, A. (2022). Deep learning-based object detection: An investigation. G. S. Tomar, N. Chaki, L. C. Jain, R. Garg, and T. K. Gandhi (Eds.), *Multimedia Processing, Communication and Computing Applications (pp. 669-681).* Springer. https://doi.org/10.1007/978-981-19-5037-7_50.

Baek, J. W., & Chung, K. (2020). Pothole classification model based on edge detection in road image. *Applied Sciences, 10(19),* 6662. https://doi.org/10.3390/app10196662.

Bello, S. A., Yu, S.; Wang, C. (2020). Review: Deep learning on 3D point clouds. *Remote Sensing, 12(11),* 1729. https://doi.org/10.3390/rs12111729.

Bhatt, U., Mani, S., Xi, E., &Kolter, J. Z. (2017). *Intelligent pothole detection and road condition assessment.* arXiv preprint arXiv:1710.02595. https://arxiv.org/abs/1710.02595.

Bhavana, N., Kodabagi, M. M., Kumar, B. M., Ajay, P., Muthukumaran, N., &Ahilan, A. (2024). POT-YOLO: Real-time road potholes detection using edge segmentation based YOLOv8 network. *IEEE Sensors.* https://doi.org/10.1109/JSEN.2024.3354780.

Chen, H., Yao, M., &Gu, Q. (2020). *Pothole detection using location-aware convolutional neural networks.* Int. J. Machine Learning Cybern. 11, 899–911. https://doi.org/10.1007/s13042-020-01078-7.

Chen, Y., Wang, S., Lin, L., Cui, Z., &Zong, Y. (2024). Computer vision and deep learning transforming image recognition and beyond. *International Journal of Computer Science and Information Technology. 15(1), 153-187.*

Chunmian, L., Tian, D., Duan, X., Zhou, J., Zhao, D., & Cao, D. (2022). DA-RDD: Toward domain adaptive road damage detection across different countries. *IEEE Transactions on Intelligent Transportation Systems.* https://doi.org/10.1109/tits.2022.3221067.

Dhoundiyal, P., Sharma, V., & Vats, S. (2023). *Deep learning framework for automated pothole detection.* In 2023 International Conference on Sustainable Communication Networks and Application (ICSCNA) (pp. 1382-1387). IEEE. https://doi.org/10.1109/ICSCNA58489.2023.10370471.

Diwan, T., Anirudh, G. S., &Tembhurne, J. V. (2022). Object detection using YOLO: Challenges, architectural successors, datasets and applications. *Multimedia Tools and Applications*, 82, 9243-9275. https://doi.org/10.1007/s11042-022-13702-5.

Fan, J., Bocus, M. J., Hosking, B., Wu, R., Liu, Y., Vityazev, S., & Fan, R. (2021). Multi-scale feature fusion: Learning better semantic segmentation for road pothole detection. *arXiv*. https://arxiv.org/abs/2112.13082.

Fan, R., Wang, H., Bocus, M. J., & Liu, M. (2020). We improve road pothole detection: From attention aggregation with adversarial domain adaptation. *arXiv.* https://arxiv.org/abs/2008.06840.

Fan, R., Wang, H., Wang, Y., Liu, M., & Pitas, I. (2021). The graph attention layer evolves semantic segmentation for road pothole detection: A benchmark and algorithms. *arXiv.* https://arxiv.org/abs/2109.02711.

Ghanem, R., and Khaled, B. (2023). The robustness of the YOLO object detection algorithm in terms of detecting objects in noisy environment. *Journal of Scientific Reports-A.* https://doi.org/10.59313/jsr-a.1257361.

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques (3rd ed.).* Morgan Kaufmann Publishers.

Hoang, N. D., Huynh, T. C., and Tran, V. D. (2021). Computer vision-based patched and unpatched pothole classification using machine learning approach optimized by forensic-based investigation metaheuristic. *Complexity,* https://doi.org/10.1155/2021/3511375.

Jakubec, M., Lieskovská, E., Bučko, B., &Zábovská, K. (2023). Comparison of CNN-based models for pothole detection in real-world adverse conditions: Overview and evaluation. *Applied Sciences*, 13(10), 6139. https://doi.org/10.3390/app13106139.

Kang, C. H., & Kim, S. Y. (2023). Real-time object detection and segmentation technology: An analysis of the YOLO algorithm. *JMST Advances, 5*, 69–76. https://doi.org/10.1007/s42791-023-00049-7.

Khan, M., Raza, M. A., Abbas, G., Othmen, S., Yousef, A., &Jumani, T. A. (2024). Pothole detection for autonomous vehicles using deep learning: A robust and efficient solution. *Frontiers in Built Environment, 9*, 1323792. https://doi.org/10.3389/fbuil.2023.1323792.

Khare, O., Gandhi, S., Rahalkar, A., & Mane, S. (2023). *YOLOv8-based visual detection of road hazards: Potholes, sewer covers, and manholes.* In 2023 IEEE Pune Section International Conference (PuneCon) (pp. 1-6). IEEE. https://doi.org/10.1109/PuneCon58716.2023.10149050.

Kilic, E., & Ozturk, S. (2019). A subclass supported convolutional neural network for object detection and localization in remote-sensing images. *International journal of remote sensing, 40(11),* 4193-4212.

Kim, T., &Ryu, S. K. (2014a). A guideline for pothole classification. *International Journal of Engineering and Technology, 4(10),* 618-622.

Kim, T., and Ryu, S. K. (2014b). Review and analysis of pothole detection methods. *Journal of Emerging Trends in Computing and Information Sciences, 5(8),* 603-608.

Kim, Y. M., Kim, Y. G., Son, S. Y., Lim, S. Y., Choi, B. Y., and Choi, D. H. (2022). Review of recent automated pothole-detection methods. *Applied Sciences, 12(11),* 5320. https://doi.org/10.3390/app12115320.

Kumar, A. C., Gaur, N., Chakravarty, S., Alsharif, M. H., Uthansakul, P., &Uthansakul, M. (2023). Analysis of spectrum sensing using deep learning algorithms: CNNs and RNNs. *Ain Shams Engineering Journal, 14(8),* 102164. https://doi.org/10.1016/j.asej.2023.102164.

Li, C., Li, X., Chen, M., & Sun, X. (2023). *Deep learning and image recognition.* In 2023 IEEE 6th International Conference on Electronic Information and Communication Technology (ICEICT) (pp. 557-562). IEEE. https://doi.org/10.1109/ICEICT59239.2023.10373631.

Liao, K. (2022). *Road damage intelligent detection with deep learning techniques.* In 2022 IEEE 5th International Conference on Information Systems and Computer Aided Education (ICISCAE). IEEE. https://doi.org/10.1109/ICISCAE55891.2022.9913326.

Ma, N., Fan, J., Wang, W., Wu, J., Jiang, Y., Xie, L., & Fan, R. (2022). Computer vision for road imaging and pothole detection: A state-of-the-art review of systems and algorithms. *Transportation Safety and Environment, 4(4),* 403-439. https://doi.org/10.1093/tse/tdac019.

Nawale, S., Khut, D., Dave, D., Sawhney, G., Aggrawal, P., &Devadakar, K. (2023). PotholeGuard: A pothole detection approach by point cloud semantic segmentation. *arXiv*. https://arxiv.org/abs/2311.02641

Omar, M., & Kumar, P. (2024). PD-ITS: Pothole detection using YOLO variants for intelligent transport system. *SN Computer Science, 5,* 552. https://doi.org/10.1007/s42979-024-02887-1

Paramarthalingam, A., Sivaraman, J., Theerthagiri, P., Vijayakumar, B., &Baskaran, V. (2024). A deep learning model to assist visually impaired in pothole detection using computer vision. *Decision Analytics Journal, 12.*https://doi.org/10.1016/j.dajour.2024.100507

Ragab, K. (2021). Smart pothole detection using deep learning based on dilated convolution. *Sensors, 21(24),* 8406. https://doi.org/10.3390/s21248406

Rastogi, R., Kumar, U., Kashyap, A., Jindal, S., and Pahwa, S. (2020). *A comparative evaluation of the deep learning algorithms for pothole detection.* In 2020 IEEE 17th India Council International Conference (INDICON). IEEE. doi:10.1109/INDICON49873.2020.9342267

Reddy, J. R., Raghava Deekshith, R., Venu Gopal, B., & Rajwanth, C. (2024). *Real time pothole detection using YOLO algorithm. Available* at SSRN 5088842. https://doi.org/10.2139/ssrn.5088842

Rehana, K. C., &Remya, G. (2022). *Road damage detection and classification using YOLOv5.* In 2022 Third International Conference on Intelligent Computing Instrumentation and Control Technologies (ICICICT). IEEE. https://doi.org/10.1109/icicict54557.2022.9917763

Safyari, Y., Mahdianpari, M., & Shiri, H. (2024). A review of vision-based pothole detection methods using computer vision and machine learning. *Sensors, 24(17),* 5652. https://doi.org/10.3390/s24175652

Saha, P., Arya, D., Kumar, A., Maeda, H., & Sekimoto, Y. (2022). *Road rutting detection using deep learning on images.* In 2022 IEEE International Conference on Big Data (Big Data). IEEE. https://doi.org/10.1109/bigdata55660.2022.10020458

Saisree, C., and Umaran, U. (2023). *Pothole detection using a deep learning classification method.* Procedia Computer Science, 218, 2143-2152. https://doi.org/10.1016/j.procs.2023.01.193

Saranya, E., Nivetha, R., Abirami, S., MohaideenArsath, M., &Dharaneesh, S. (2024). *Revolutionizing road maintenance: YOLO-based pothole detection system.* In 2024 10th International Conference on Advanced Computing and Communication Systems (ICACCS) (Vol. 1, pp. 1991-1997). IEEE. https://doi.org/10.1109/ICACCS59260.2024.10475429

Tang, J., Li, Y., & Wang, S. (2023). *A comparison of road damage detection based on YOLOv8.* ResearchGate. https://www.researchgate.net/publication/376021160_A_Comparison_of_Road_Damage_Detection_Based_on_YOLOv8

Thiruppathiraj, S., Kumar, U. and Bagke, S. (2020). *Automatic pothole classification and segmentation using android smartphone sensors and camera images with machine learning techniques.* In 2020 IEEE Region 10 Conference (TENCON) (pp. 1386-1391). IEEE. https://doi.org/10.1109/TENCON50793.2020.9293756

Tsai, Y. C., & Chatterjee, A. (2018). Pothole detection and classification using 3D technology and watershed method. *Journal of Computing in Civil Engineering, 32(2),* 04017078. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000726

Ultralytics. (2025). *YOLOv8: State-of-the-art computer vision model.* Retrieved from https://yolov8.com/

Varona, B., Monteserin, A., &Teyseyre, A. (2020). A deep-learning approach for automatic road surface monitoring and pothole detection. *Personal and Ubiquitous Computing, 24,* 519–534. https://doi.org/10.1007/s00779-019-01234-z

Yu, X., &Salari, E. (2011). *Sidewalk pothole detection and severity measurements using laser imaging.* In 2011 IEEE International Conference on Electro/Information Technology (pp. 1-5). IEEE. https://doi.org/10.1109/EIT.2011.5978573

Zanevych, Y., Yovbak, V., Basystiuk, O., Shakhovska, N., Fedushko, S., &Argyroudis, S. (2024). Evaluation of pothole detection performance using deep learning models under low-light conditions. *Sustainability. 16(24),*10964. https://doi.org/10.3390/su16241

İSTANBUL UNIVERSITY PRESS

Acta Infologica,
ACIN 2025, 9 (1): 133–146

https://doi.org/10.26650/acin.1631851
Submitted: 02.02.2025
Revision Requested: 10.03.2025
Last Revision Received: 17.04.2025
Accepted: 07.05.2025
Published Online 12.05.2025

Acta Infologica

Research Article

Open Access

Artificial Intelligence and Cost Reduction Strategies for Healthcare Management: Opportunities and Limitations

Canan Bulut [1]

[1] İstanbul Kültür University, Vocational School, Department of Medical Services and Technics, İstanbul, Türkiye

Abstract

This study employs a traditional literature review approach to examine the role of artificial intelligence (AI) in cost-reduction strategies in healthcare management. As healthcare systems face increasing financial pressures, AI has been widely explored for its potential to enhance operational efficiency and optimize resource utilization. By synthesizing recent academic literature and theoretical discussions, this review aims to provide a comprehensive evaluation of AI-driven cost reduction strategies and their broader implications for healthcare management.

This study explores AI applications in automating administrative workflows, enhancing predictive analytics, and optimizing clinical decision-making. AI's contributions to supply chain management and early disease detection are also highlighted as significant factors in achieving operational cost savings. However, despite its transformative potential, AI adoption in healthcare presents notable challenges, including high initial investment costs, data security risks, ethical dilemmas, and regulatory constraints**.** These challenges necessitate a structured and strategic approach to ensure sustainable and effective AI implementation.

Through an extensive review of the existing literature, this study critically analyzes both the opportunities and limitations associated with AI-driven healthcare cost reduction. The findings underscore the need **for** robust regulatory frameworks, ethical AI deployment, and continuous workforce adaptation to facilitate the successful integration of AI technologies into healthcare systems. By offering key strategic insights, this study contributes to the academic discourse on AI's cost-effectiveness in healthcare and provides a foundation for future research and policy development.

Keywords Healthcare Management · Artificial Intelligence · Cost Reduction Strategies · Digitalization in Healthcare · AI-Driven Decision-Making

Citation: Bulut, C. (2025). Artificial intelligence and cost reduction strategies for healthcare management: Opportunities and limitations . *Acta Infologica*, *9*(1), 133-146. https://doi.org/10.26650/acin.1631851

This work is licensed under Creative Commons Attribution-NonCommercial 4.0 International License.

© 2025. Bulut, C.

Corresponding author: Canan Bulut c.bulut@iku.edu.tr

Introduction

The increasing financial burden on healthcare systems, coupled with the demand for more efficient resource utilization, has necessitated the adoption of innovative technologies in healthcare management. In this regard, artificial intelligence (AI) has emerged as a transformative tool, offering significant potential to enhance operational efficiency, optimize clinical workflows, and reduce costs. AI-driven technologies, including machine learning, deep learning, natural language processing, and robotic automation, facilitate the automation of routine administrative tasks, support data-driven decision-making through big data analytics, and enable cost-effective solutions for early diagnosis and preventive care (Tiwari & Etienne, 2024).

Despite these advantages, AI adoption continues to face significant challenges. The substantial initial investment requirements, data security vulnerabilities, and regulatory constraints pose major barriers to widespread implementation (Rajkomar et al., 2018; Topol, 2019). Furthermore, ethical concerns related to AI-driven decision-making demand ongoing regulatory oversight to safeguard patient rights and align with healthcare policies. Effectively addressing these challenges requires regulatory adaptation, ethical AI governance, and workforce training investments to ensure effective and sustainable integration of AI into healthcare systems.

This study employed a traditional literature review approach to analyze the role of AI in cost-reduction strategies in healthcare management. By synthesizing findings from recent academic research, this study examines both the advantages and limitations of AI-driven cost optimization. In addition, it provides strategic recommendations for healthcare managers and policymakers, outlining key considerations for the effective integration of AI technologies into healthcare services. This study aims to contribute to a deeper understanding of AI's role in enhancing healthcare efficiency, financial sustainability, and decision-making processes in an evolving technological landscape.

In line with the characteristics of a traditional literature review, this study does not rely on formal inclusion metrics or statistical synthesis, as found in meta-analyses. Instead, it adopts a narrative and interpretative approach to map the conceptual landscape of AI applications in healthcare cost management. According to Paul and Barari (2022), traditional literature reviews are particularly valuable in emerging or interdisciplinary fields where quantitative data may be limited. However, theoretical development and critical reflection are essential. This method emphasizes thematic coherence, analytical depth, and contextual relevance over statistical generalizability. By drawing upon a wide range of peer-reviewed articles published between 2020 and 2025 and retrieved from academic databases such as Web of Science and ScienceDirect, the review offers a comprehensive understanding of the current discourse while identifying conceptual gaps and opportunities for future research (Snyder, 2019).

Literatur Review

Artificial intelligence (AI) has emerged as a transformative tool in healthcare management, significantly enhancing both clinical decision-making and operational efficiency. AI-driven technologies, including machine learning, deep learning, natural language processing, and robotic automation, are increasingly being integrated into healthcare management processes (Tiwari & Etienne, 2024). These technologies play a crucial role in optimizing hospital workflows, supporting clinical decision-making, and reducing costs in resource management (Jiang et al., 2017).

In recent years, the role of AI in cost-reduction strategies in healthcare has gained significant attention in academic literature. AI-based applications have demonstrated the potential to enhance healthcare service efficiency while reducing operational costs, particularly through big data analytics, robotic process automation, and machine learning-driven decision-support mechanisms (Sahni et al., 2024). Studies have suggested that the integration of these technologies into healthcare management significantly improves both administrative and clinical efficiency, allowing for better resource allocation and streamlined service delivery.

Recent traditional literature reviews have further reinforced the necessity of evaluating the cost-effectiveness of AI applications in healthcare. A literature review conducted by Prabhod (2024) found that AI-supported clinical decision support systems can reduce hospital costs by approximately 15-20%. Similarly, Rossi et al. (2022) highlighted the role of AI-driven predictive analytics in early diagnosis and preventive healthcare, demonstrating substantial long-term economic benefits for healthcare systems.

Moreover, comparative studies by the World Economic Forum (2024), Sahni et al. (2024), and Rajkomar et al. (2018) indicated that AI's effectiveness in reducing healthcare costs varies across sectors and geographical regions. For instance, Topol (2019) argued that while robotic surgery and AI-driven predictive systems have demonstrated significant cost reductions in hospitals across the United States and Europe, their implementation in developing countries remains limited because of infrastructure deficiencies. These variations highlight the importance of tailored AI implementation strategies that consider technological readiness, economic conditions, and healthcare policies at both national and regional levels.

This study adopts a traditional literature review approach with the primary aim of offering a thematic and conceptual synthesis of recent academic research on the use of artificial intelligence in cost reduction strategies in healthcare management. Unlike systematic reviews, which follow rigid protocol-driven inclusion and exclusion processes, traditional reviews allow for greater narrative flexibility and interpretive depth, making them especially appropriate for emerging, complex, or interdisciplinary fields (Snyder, 2019).

This review explores how academic discourse has framed AI's role in improving healthcare efficiency, financial sustainability, and operational decision-making. By engaging with peer-reviewed literature published between 2020 and 2025 and focusing on conceptual patterns, challenges, and opportunities, this study maps the intellectual terrain rather than providing quantitative generalizations. As Paul and Barari (2022) argued, traditional literature reviews are particularly valuable when the goal is to synthesize diverse perspectives, highlight theoretical developments, and guide future research directions without the constraints of rigid methodological filters.

Table 1

Examples of Current Academic Research on the Use of Artificial Intelligence in Healthcare

Author-Year	Publication Title	Topic	Findings
Rossi et. al. (2022)	Cost-effectiveness of artificial intelligence as a decision-support system in healthcare	Cost-effectiveness of AI-based decision-support systems for healthcare	AI-based decision-support systems can reduce hospital costs by 15-20%.
Matheny et al. (2019)	Artificial intelligence in healthcare: The hope, hype, promise, and peril	The promise and challenges of AI applications in healthcare	AI is promising for healthcare applications but may pose significant risks if misused.
Obermeyer et al. (2019)	Dissecting racial bias in an algorithm used to manage population health	Impact of algorithmic bias on healthcare management	Algorithmic biases in healthcare systems have been

Author-Year	Publication Title	Topic	Findings
			identified and need to be addressed.
Prabhod (2024)	The role of artificial intelligence in reducing healthcare costs and improving operational efficiency	AI's role in reducing healthcare costs and improving operational efficiency	AI applications can enhance operational efficiency and lead to cost savings in healthcare.
Rajkomar et al. (2019)	Machine learning in medicine	Application of machine learning to medicine	Machine learning can improve diagnostic processes faster and more effectively than human doctors.
Sahni et al. (2024)	The potential impact of artificial intelligence on healthcare spending	Potential financial impact of AI on healthcare spending	AI-driven solutions can provide significant cost savings in healthcare expenditures.
Tiwari and Etienne (2024)	Artificial intelligence and healthcare: A journey through history, present innovations, and future possibilities	Historical evolution, current innovations, and future possibilities of AI in healthcare	AI is undergoing a transformational process that could revolutionize healthcare services.
Topol (2019)	Deep medicine: How artificial intelligence can make healthcare human again	How AI can enhance human-centric healthcare services	AI can enable faster, personalized, and more efficient healthcare services.
World Economic Forum (2024)	How AI can transform healthcare and treatment	AI's potential to transform healthcare management and treatment processes	AI can improve healthcare treatment efficiency, reduce costs, and enhance patient care quality.

One of the most essential concepts in the use of AI in healthcare management is big data analytics. AI can analyze a wide range of data, from hospital records to public health data, thereby optimizing resource allocation and supporting healthcare workers in their decision-making processes (National Academy of Medicine, 2022). This approach not only improves patient care quality but also contributes to cost reduction.

To fully harness AI's potential, a strategic management approach that considers both technological and human factors should be adopted (Buch et al., 2018).

Cost-Reduction Potentials with Artificial Intelligence

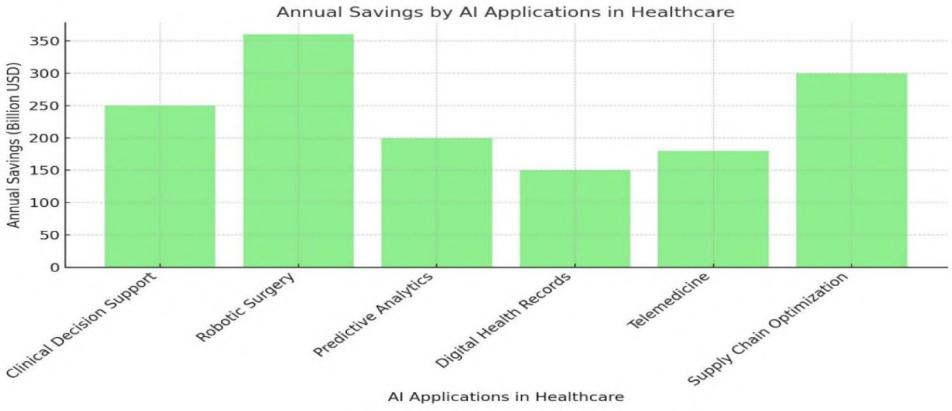
Artificial intelligence (AI) offers several solutions for reducing costs in healthcare management and improving operational processes. AI cost-reduction strategies are used across various areas of healthcare services, driving a comprehensive transformation.

AI-driven solutions enhance clinical process efficiency, streamline resource allocation, and optimize administrative workflows. By automating routine tasks, AI can minimize labor costs and reduce operational inefficiencies, leading to significant savings. Predictive analytics, one of AI's key applications, enables early diagnosis and preventive care, thereby reducing the need for costly treatments at later stages.

AI's ability to analyze large datasets allows for better decision-making, improved resource management, and lower overall healthcare expenditures. These strategies not only contribute to cost efficiency but also enhance the quality of patient care and sustainability of healthcare systems.

Figure 1

Annual Savings by AI Applications in Healthcare (Sahni et al., 2024; Prabhod, 2024; World Economic Forum, 2024)



Annual savings from AI applications in healthcare have been estimated to range between $150-360 billion, primarily driven by advancements in robotic surgery, predictive analytics, and supply chain optimization (Sahni et al., 2024; Prabhod, 2024; World Economic Forum, 2024) (Figure 1). However, to substantiate these figures, recent studies have conducted meta-analyses and systematic reviews examining the direct financial impacts of AI-driven solutions.

For instance, Rossi et al. (2022) investigated the cost-effectiveness of AI-based decision-support systems for melanoma, dental caries, and diabetic retinopathy detection, demonstrating that these AI applications led to hospital cost reductions of approximately 15-20%. Additionally, Prabhod (2024) analyzed AI-driven hospital administration automation and reported a potential 20% decrease in administrative costs through AI-powered scheduling, billing automation, and electronic health record (EHR) management. Furthermore, the World Economic Forum (2024) highlighted the financial impact of predictive analytics, illustrating how AI-driven early diagnosis models can reduce hospital admission rates by 10-15%, leading to significant long-term cost savings.

Comparative analyses by Sahni et al. (2024), Rajkomar et al. (2018), and the World Economic Forum (2024) indicate that while AI solutions are effective in cost reduction, their financial impact varies across different healthcare systems and regional economic conditions. Topol (2019) emphasized that while AI-powered robotic surgeries have demonstrated notable cost savings in high-income countries, their implementation in developing healthcare systems faces infrastructural and economic challenges.

These estimates of AI-driven cost reductions are supported by multiple empirical studies and real-world implementations, reinforcing the argument that AI has a substantial financial impact on healthcare management when effectively integrated with existing systems.

Automation and Workflow Improvement: AI-driven automation eliminates time-consuming and repetitive healthcare tasks, thereby reducing costs and allowing healthcare workers to focus on more complex tasks. The AI algorithms used in scheduling, billing, and medical record management reduce the need for manual labor and minimize errors. Automation-focused AI solutions can lower administrative costs by up to 20%, allowing clinicians to dedicate more time to direct patient care and further cost reduction (Prabhod, 2024).

Predictive Analytics and Early Diagnosis: AI analyzes past patient data to facilitate early diagnosis and preventive treatment. This approach facilitates early disease detection, thereby reducing the need for costly

treatments at later stages. For instance, AI-supported predictive models for early detection of chronic kidney disease have significantly reduced treatment costs. AI's predictive analytics potential can reduce hospital admission rates and associated costs by 10-15% (World Economic Forum, 2024).

Personalized Medicine and Targeted Treatments: Personalized treatment approaches are enabled by AI's analysis of large datasets, such as genetic information and past medical records. AI refines treatment plans more accurately and reduces trial-and-error methods, thus lowering medication and treatment costs. These AI applications minimize drug side effects and ineffective treatments, thereby increasing cost efficiency (Sahni et al., 2024).

Resource Management and Supply Chain Optimization: AI plays a critical role in supply chain management by predicting material demands and making processes more efficient. This approach prevents unnecessary stock accumulation and reduces the costs associated with material shortages. It can cut healthcare facilities' supply costs by 15-20% (Rossi et al., 2022).

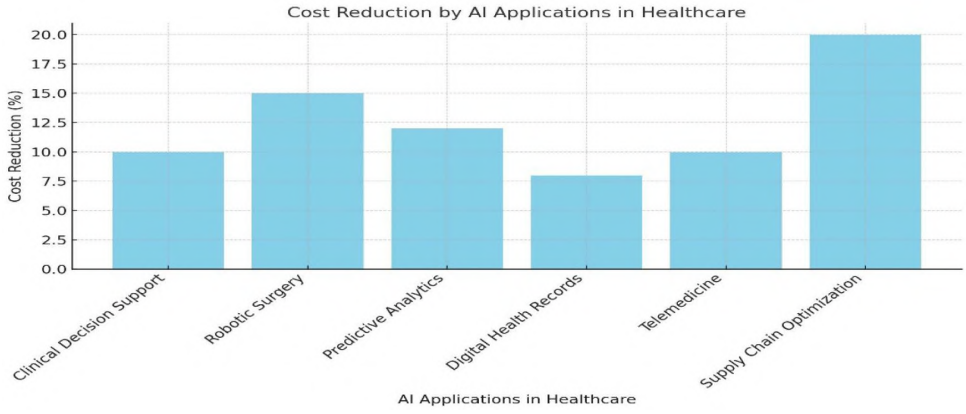
AI's cost-reduction potential in healthcare is substantial; however, realizing this potential requires the adoption of appropriate strategies and the optimization of technological infrastructure. To fully leverage AI's cost advantages, healthcare managers must implement strategic planning, enforce data security measures, and ensure continuous staff training.

Limitations of Cost Reduction through Artificial Intelligence in Healthcare Management

Artificial intelligence (AI) offers potential for cost reduction in healthcare management; however, various limitations are encountered during the implementation and integration process. These limitations can hinder the effective realization of AI and the development of a sustainable cost reduction strategy.

Figure 2

Cost Reduction by AI Applications in Healthcare (Cutler & Sahni, 2020; Wang et al., 2024; Davenport & Kalakota, 2019; Bennett & Hauser, 2013; McGillion et al., 2021)



Artificial intelligence (AI) has demonstrated significant potential in reducing healthcare costs by optimizing various operational and clinical processes. Clinical decision support systems, robotic surgery, and supply chain optimization have been identified as key AI applications that contribute to cost reductions ranging between 10-20% (Cutler & Sahni, 2020; Wang et al., 2024; Davenport & Kalakota, 2019). AI-driven automation enhances resource allocation efficiency, streamlines workflow management, and reduces redundant expenditures, leading to substantial financial savings for healthcare institutions.

High Initial Costs: The integration of AI into healthcare requires substantial investments in technological infrastructure, hardware, software, data management, and workforce training (Cutler & Sahni, 2020). Developing and maintaining AI systems tailored for healthcare applications often necessitate significant financial resources, making adoption challenging—especially in low- and middle-income healthcare systems (Davenport & Kalakota, 2019). The financial burden of initial implementation can be a deterrent for widespread AI adoption, despite its long-term cost-saving potential.

Data Security and Privacy Risks: One of the primary concerns associated with AI in healthcare is its reliance on vast amounts of sensitive patient data, which raises critical privacy and security concerns (Bennett & Hauser, 2013). AI-based systems are vulnerable to cybersecurity threats, data breaches, and unauthorized access, posing legal and ethical risks to healthcare organizations (Wang et al., 2024). Ensuring compliance with data protection regulations and the implementation of robust security measures is essential to safeguard patient confidentiality and maintain institutional trust.

Implementation Challenges: Successful deployment of AI in healthcare requires overcoming technological and organizational barriers. Healthcare institutions must allocate time and resources for AI integration, including staff training, workflow adjustments, and infrastructure upgrades (McGillion et al., 2021). Resistance from healthcare professionals, coupled with a lack of managerial support and inadequate technological infrastructure, can hinder the seamless adoption of AI-driven solutions (Davenport & Kalakota, 2019). Additionally, integrating AI into clinical environments often necessitates workflow restructuring, which may lead to short-term cost increases before realizing financial benefits.

Ethical and Legal Limitations: The ethical implications of AI in healthcare remain a subject of ongoing debate. AI has the potential to enhance clinical decision-making and reduce human error; however, it also introduces ethical dilemmas related to accountability, patient autonomy, and trust in automated systems (Bennett & Hauser, 2013). For instance, AI-based misdiagnoses could raise legal concerns regarding medical malpractice and liability (Cutler & Sahni, 2020). Furthermore, regulatory restrictions on AI usage, particularly regarding data access and patient consent, impose additional constraints on its implementation (McGillion et al., 2021).

Despite its significant cost-saving potential, the adoption of AI in healthcare faces multiple challenges, including high initial costs, data security concerns, implementation barriers, and ethical considerations. For AI to achieve sustainable and widespread adoption, healthcare managers must strategically plan its integration, ensure compliance with regulatory frameworks, address ethical concerns, and provide adequate training to healthcare professionals. Future research should focus on developing AI governance models that balance technological advancement with ethical and financial sustainability (Davenport & Kalakota, 2019; Wang et al., 2024).

Successful AI-Based Healthcare Applications

Artificial intelligence (AI) has been widely recognized for its transformative impact on healthcare management, offering innovative solutions to enhance cost efficiency and patient care quality. This section explores key case studies demonstrating how AI can be successfully integrated into healthcare systems to yield both financial and operational benefits.

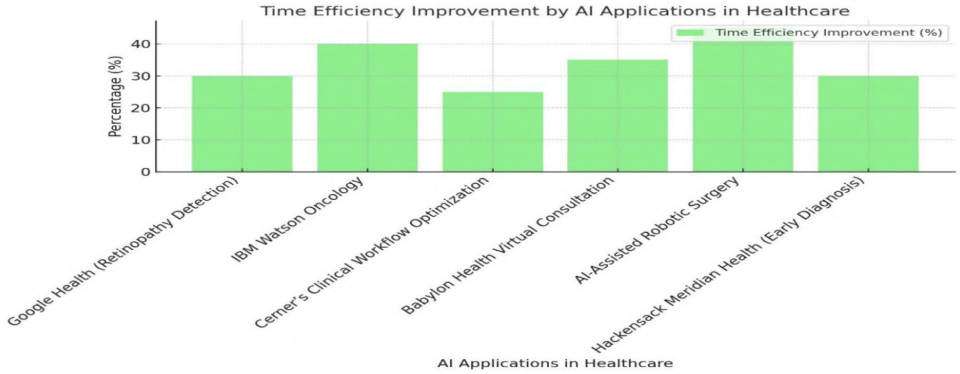
Case Studies: Successful Projects and Programs

Numerous AI-driven healthcare projects have significantly reduced costs while improving diagnostic accuracy, workflow efficiency, and overall healthcare delivery. For instance, AI-based early diagnosis systems

have facilitated early detection of chronic diseases, thereby reducing treatment costs and improving patient outcomes.

Figure 3

Time Efficiency Improvement via AI Applications in Healthcare (Zhang et al., 2022; Subramanian et al., 2020; Shaik et al., 2023; Davenport & Kalakota, 2019)



These estimates suggest that AI has a significant financial impact on healthcare management when effectively integrated with existing systems (Figure 3).

One notable example is Hackensack Meridian Health's AI-supported early detection program for kidney disease, which identified early-stage conditions in 20% of patients, leading to improved clinical interventions and cost savings (Subramanian et al., 2020). Similarly, AI-driven lung cancer screening technologies, such as those developed by Zhang et al. (2022), have enhanced the detectability of lung nodules through AI-assisted CT image analysis, significantly reducing the time required for radiological assessment and enabling faster decision-making.

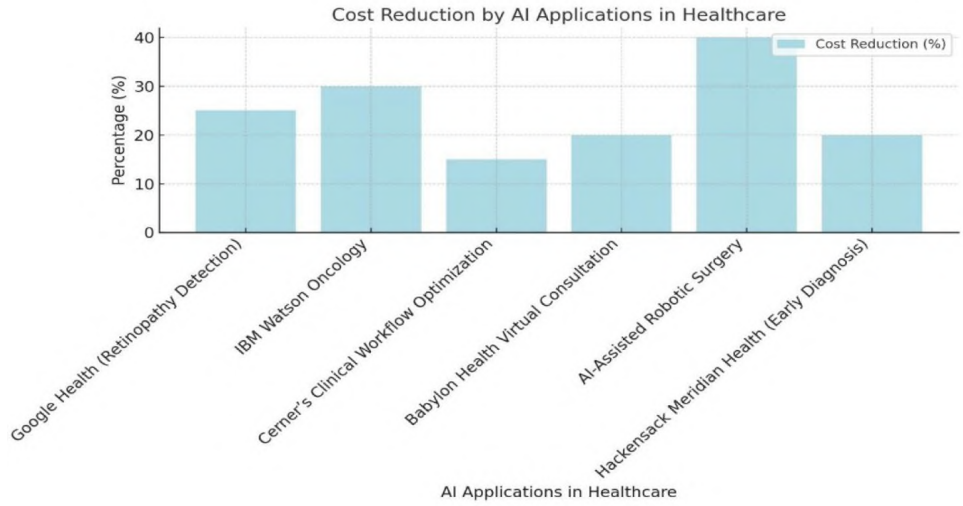
AI's impact on workflow efficiency and digital pathology has been exemplified in the Mayo Clinic's digital pathology project, where AI algorithms optimized cancer cell detection, reducing manual analysis time by 30% and lowering laboratory operational costs (Shaik et al., 2023). Such AI-driven solutions have streamlined diagnostic processes, thereby allowing for more efficient resource utilization and operational cost reduction for healthcare institutions.

AI in supply chain management has also demonstrated substantial improvements in healthcare cost efficiency. Several hospitals in the U.S. have implemented AI-powered demand forecasting algorithms, which have optimized medical supply inventories, preventing overstocking and reducing supply costs by 15% (Zhang et al., 2022). Additionally, robotic process automation (RPA) has automated administrative functions, including billing, appointment scheduling, and electronic health record (EHR) management, leading to significant labor cost reductions and enhanced workflow automation (Davenport & Kalakota, 2019).

Moreover, telecritical care (tele-ICU) has revolutionized intensive care management by facilitating real-time patient monitoring, early intervention, and streamlined clinical decision-making. Studies have indicated that AI-powered tele-ICU solutions have improved patient outcomes and reduced the burden on intensive care units, enhancing operational efficiency and reducing mortality rates (Subramanian et al., 2020).

Figure 4

Cost Reduction by AI Applications in Healthcare (Khanna et al., 2022; Hayyolalam et al., 2021; Ezechukwu, 2024; Ingole et al., 2024; Ingole et al., 2024)



The integration of artificial intelligence (AI) in healthcare has demonstrated significant cost-saving potential through clinical decision support systems, robotic surgery, and supply chain optimization, achieving cost reductions of 10-20%. Other AI-driven applications, including predictive analytics, digital health records, and telemedicine, have contributed to cost efficiency by reducing expenses by 8-12% (Khanna et al., 2022; Hayyolalam et al., 2021). These implementations highlight AI's capability to enhance healthcare financial sustainability while improving patient care quality (Figure 4).

Successful AI-Based Cost Reduction Projects

Google Health's Deep Learning Model (USA): Google Health developed an AI-powered deep learning model for diagnosing retinopathy and other eye diseases**,** significantly improving diagnostic accuracy while reducing screening time by 90%. This advancement has facilitated early disease detection, leading to reduced long-term treatment costs and improved patient outcomes (Ingole et al., 2024). The success of this initiative underscores the role of big data analytics in optimizing clinical efficiency and cost-effectiveness (Ezechukwu, 2024).

IBM Watson Oncology Project (India): IBM Watson, in collaboration with Manipal Hospitals in India, launched an AI-driven oncology decision support system to personalize cancer treatment plans. By reducing treatment duration by 30%, the system significantly reduced hospitalization and medication costs, improving overall cost efficiency and patient care quality (Khanna et al., 2022). The implementation of AI in cancer treatment planning has demonstrated its capacity to enhance decision-making, reduce medical errors, and optimize financial resources.

Cerner's AI-Based Clinical Workflow Optimization (USA): Cerner's AI-supported clinical workflow optimization system has been deployed across major U.S. hospital networks, reducing patient wait times by 25%. By optimizing patient flow, this initiative minimized staff workload and operational expenses, leading to lower resource utilization costs while enhancing patient satisfaction (Hayyolalam et al., 2021). AI-driven workflow optimization exemplifies how hospitals can increase efficiency without compromising service quality.

Babylon Health's Virtual Consultation Application (UK): Babylon Health introduced an AI-powered virtual consultation platform, offering preliminary diagnoses and directing patients to appropriate specialists. This AI-driven approach reduces healthcare costs by reducing unnecessary hospital visits, easing hospital congestion, and enhancing healthcare accessibility (Ingole et al., 2024). The platform's success highlights AI's ability to improve healthcare resource allocation while maintaining cost efficiency.

AI-Driven Robotic Surgery (Japan): Japan has pioneered AI-assisted robotic surgical systems, particularly in prostate cancer surgeries, where AI-driven robotic procedures have reduced surgical intervention times by up to 40%. These robotic-assisted surgeries are less invasive procedures, reducing complication risks, hospitalization durations, and overall surgical costs (Ezechukwu, 2024). The cost-effectiveness of robotic surgery underscores AI's transformative role in modern surgical interventions.

Lessons Learned and Strategies for Wider Implementation

Successful deployment of AI-driven healthcare cost reduction strategies provides valuable insights for future applications:

Data Quality and Security: The efficiency of AI-powered healthcare applications relies on high-quality, secure data processing. AI adoption must align with data security regulations to mitigate risks related to cybersecurity and patient privacy (Khanna et al., 2022).

Training and Adaptation: Successful AI integration requires continuous education and adaptation by healthcare professionals. Establishing AI-compatible workflows and providing specialized training ensure optimal use of AI-driven tools (Ingole et al., 2024).

Ethical and Legal Compliance: AI implementation in healthcare should adhere to ethical guidelines and regulatory frameworks to maintain transparency and accountability. Regulatory policies must evolve to accommodate AI's growing role in healthcare management (Hayyolalam et al., 2021).

AI-driven cost reduction strategies have transformed healthcare management, improved financial sustainability, patient care quality, and operational efficiency. By integrating AI-powered diagnostic models, workflow automation, virtual consultations, and robotic-assisted surgeries, healthcare institutions can achieve substantial cost savings while maintaining high standards of care (Ezechukwu, 2024). The continued evolution of AI in healthcare necessitates strategic investment, ethical oversight, and workforce adaptation to ensure long-term benefits and sustainable cost reductions (Khanna et al., 2022; Ingole et al., 2024).

Discussion and Conclusion

This study highlights the significant role of artificial intelligence (AI) in healthcare management, particularly in cost reduction strategies and operational efficiency. AI-driven automation, predictive analytics, and decision-support systems have demonstrated substantial benefits in optimizing administrative workflows, improving clinical decision-making, and enhancing resource allocation. The literature reviewed indicates that AI-based applications contribute to operational cost reductions, particularly through automated processes, robotic surgery, and AI-enhanced supply chain management (Sahni et al., 2024; Rossi et al., 2022). Additionally, AI integration into early diagnosis and preventive care services has improved intervention outcomes and reduced long-term healthcare expenditures (Prabhod, 2024).

Balancing Opportunities and Challenges

AI presents significant potential for cost efficiency in healthcare management by minimizing labor costs, optimizing workflows, and enhancing resource allocation. Empirical evidence suggests that AI applications can lead to 15-40% cost reductions, particularly in clinical decision support, robotic surgery, and supply chain optimization (Prabhod, 2024; Sahni et al., 2024). AI's role in early diagnosis and predictive analytics has also been instrumental in preventing costly late-stage treatments, achieving up to 20% savings in operational costs (Matheny et al., 2019; Rossi et al., 2022).

However, successful deployment of AI requires overcoming critical barriers, such as high initial costs, cybersecurity risks, workforce resistance, and ethical constraints (Rajkomar et al., 2018; Topol, 2019). For example, the IBM Watson Oncology Project, despite its success in reducing treatment duration by 30%, faced challenges due to its high implementation costs and technological infrastructure requirements (Rossi et al., 2022). Furthermore, data security risks remain a significant limitation, as potential data breaches can incur legal penalties and undermine trust in AI systems (Obermeyer et al., 2019).

Thus, a well-balanced strategic approach is necessary to maximize AI's potential while mitigating risks. Regulatory enforcement, data protection policies, and ethical AI integration frameworks must be prioritized to enable sustainable AI adoption in healthcare (Matheny et al., 2019).

Long-Term Impacts

In the long term, AI is poised to redefine cost efficiency in healthcare by streamlining workflows, reducing unnecessary expenditures, and improving patient outcomes (World Economic Forum, 2024). The financial sustainability of healthcare systems will largely depend on AI's continued technological advancement and its seamless integration into clinical and administrative functions (Sahni et al., 2024).

For instance, the Hackensack Meridian Health early detection initiative for kidney disease successfully reduced treatment costs by 20%, underscoring AI's potential in early intervention and preventive care (Rossi et al., 2022). Similarly, AI-powered robotic surgery in Japan demonstrated a 30% reduction in operational costs by reducing surgical durations and postoperative complications (Prabhod, 2024). These examples highlight the tangible financial benefits of AI implementation in healthcare.

However, to fully realize AI's long-term cost-saving potential, further case studies and large-scale real-world data assessments are required. Developing regulatory policies, enhancing data quality, and continuously training AI systems are crucial to ensuring AI's sustained impact on cost efficiency (Rajkomar et al., 2018; Topol, 2019).

Nevertheless, fully assessing AI's long-term impacts requires more case studies and real-world data analysis. AI's cost-effectiveness can further improve with the development of regulatory frameworks and data policies (Rajkomar et al., 2018; Topol, 2019). Moreover, enhancing data quality and comprehensive training of AI systems will ensure sustainable cost-effectiveness (Obermeyer et al., 2019).

The potential for AI to reduce long-term costs can make healthcare systems more integrated and efficient. However, healthcare managers must initiate the necessary strategic changes for AI adoption and update regulatory frameworks to realize this potential (Matheny et al., 2019).

Artificial intelligence (AI) is a revolutionary technology in healthcare management that is intended to reduce costs and enhance efficiency. The acceleration of diagnostic processes, improvement of clinical decision-making, and reduction of operational costs highlight the impact of AI on healthcare systems.

However, data security, high initial costs, and ethical compliance remain significant barriers that limit AI's full potential.

To optimize AI's cost-effectiveness and scalability, healthcare institutions must adopt the following strategic actions:

Strengthening Data Infrastructure and Security: Robust data governance and cybersecurity frameworks must be established to safeguard patient privacy and enhance AI accuracy in predictive analytics (Khanna et al., 2022).

Comprehensive Workforce Training: Healthcare professionals should undergo continuous AI training programs to improve technological adaptation and streamline AI-human collaboration (Ingole et al., 2024).

Developing Ethical and Regulatory Guidelines: AI integration must comply with stringent ethical standards and evolving legal frameworks to maintain transparency, accountability, and patient trust (Hayyolalam et al., 2021).

AI can enhance healthcare management by improving efficiency and reducing operational costs. However, successful implementation requires balanced integration strategies that consider economic feasibility, ethical compliance, and technological readiness. To maximize AI's cost-saving potential, healthcare institutions must invest in data security, staff training, and adaptive regulatory policies to support sustainable integration.

Future research should focus on evaluating the long-term economic impact of AI, strengthening data governance frameworks, and exploring AI's applications across diverse healthcare environments. Policymakers should develop adaptive regulations that align with technological advancements while ensuring patient-centric healthcare delivery. As AI continues to evolve, strategic planning and responsible deployment will become crucial for maintaining cost efficiency and high-quality healthcare services.

Peer Review Externally peer-reviewed.
Conflict of Interest The author have no conflict of interest to declare.
Grant Support The author declared that this study has received no financial support.

Author Details Canan Bulut
[1] İstanbul Kültür University, Vocational School, Department of Medical Services and Technics, İstanbul, Türkiye
0000-0001-5092-5261 c.bulut@iku.edu.tr

References

Bennett, C. C., & Hauser, K. (2013). Artificial intelligence framework for simulating clinical decision-making: A Markov decision process approach. *Artificial Intelligence in Medicine*, *57*(1), 9-19. doi: https://doi.org/10.1016/j.artmed.2012.12.003

Biswas, A., and Talukdar, W. (2024). Intelligent clinical documentation: Harnessing generative ai for patient-centric clinical note generation. *arXiv preprint*. doi: https://doi.org/10.48550/arXiv.2405.18346

Buch, V. H., Ahmed, I., & Maruthappu, M. (2018). Artificial intelligence in medicine: current trends and future possibilities. *British Journal of General Practice*, 68(668), 143-144.

Davenport, T., & Kalakota, R. (2019). The potential of artificial intelligence in healthcare. *Future Healthcare Journal*, 6(2), 94-98. doi: https://doi.org/10.7861/futurehosp.6-2-94

Ezechukwu, C. V. (2024). *Artificial intelligence and healthcare: the opportunities, risks and barriers of ai adoption in healthcare in Sub-Saharan Africa*. (Master thesis). Available at: https://sapientia.ualg.pt/entities/publication/6786b359-5d37-464f-8565-6b108440a419/full.

Hayyolalam, V., Aloqaily, M., Özkasap, Ö., & Guizani, M. (2021). Edge intelligence for empowering IoT-based healthcare systems. *IEEE Wireless Communications*, *28*(3), 6-14. doi: 10.1109/MWC.001.2000345

Hirani, R., Noruzi, K., Khuram, H., Hussaini, A. S., Aifuwa, E. I., Ely, K. E., Lewis, J. M., Gabr, A. E., Smiley, A., Tiwari, R. K., & Etienne, M. (2024). Artificial Intelligence and Healthcare: A Journey through History, Present Innovations, and Future Possibilities. *Life (Basel, Switzerland)*, *14*(5), 557. doi: https://doi.org/10.3390/life14050557

Ingole, B. S., Ramineni, V., Krishnappa, M. S., & Jayaram, V. (2024). AI-driven innovation in medicaid: enhancing access, cost efficiency, and population health management. *arXiv preprint. doi:* https://doi.org/10.48550/arXiv.2410.21284

Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., Wang, Y., Dong, Q., Shen, H., and Wang, Y. (2017). Artificial intelligence in healthcare: Past, present and future. *Stroke and Vascular Neurology*, 2(4), 230–243.

Khanna, N. N., Maindarkar, M. A., Viswanathan, V., Fernandes, J. F. E., Paul, S., Bhagawati, M., Ahluwalia, P., Ruzsa, Z., Sharma, A., Kolluri, R., Singh, I. M., Laird, J. R., Fatemi, M., Alizad, A., Saba, L., Agarwal, V., Sharma, A., Teji, J. S., Al-Maini, M., … Suri, J. S. (2022). Economics of Artificial Intelligence in Healthcare: Diagnosis vs. Treatment. *Healthcare*, *10*(12), 2493. https://doi.org/10.3390/healthcare10122493

Matheny, M., Israni, S. T., Ahmed, M., & Whicher, D. (2019). Artificial intelligence in health care: The hope, the hype, the promise, the peril. *Washington, DC: National Academy of Medicine*, *10*.

McGillion, M. H., Parlow, J., Borges, F. K., Marcucci, M., Jacka, M., Adili, A., … & Devereaux, P. J. (2021). Post discharge after surgery virtual care with remote automated monitoring technology (PVC-RAM): protocol for a randomized controlled trial. *Canadian Medical Association Open Access Journal*, *9*(1), E142-E148. doi: https://doi.org/10.9778/cmajo.20200176

National Academy of Medicine. (2022). Artificial Intelligence in Health Care: The Hope, the Hype, the Promise, the Peril.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447-453. doi:10.1126/science.aax2342

Paul, J., & Barari, M. (2022). *Meta-analysis and traditional systematic literature reviews—What, why, when, where, and how? Psychology and Marketing*, 39(6), 1099–1115. https://doi.org/10.1002/mar.21657

Prabhod, K. J. (2024). The role of artificial intelligence in reducing healthcare costs and improving operational efficiency. *Quarterly Journal of Emerging Technologies and Innovations*, *9*(2), 47-59.

Rajkomar, A., Dean, J. & Kohane, I. (2019). Machine learning in medicine. *New England Journal of Medicine*, *380*(14), 1347-1358. doi: 10.1056/NEJMra1814259

Rossi, J. G., Rojas-Perilla, N., Krois, J., & Schwendicke, F. (2022). *Cost-effectiveness of artificial intelligence as a decision-support system applied to the detection and grading of melanoma, dental caries, and diabetic retinopathy. JAMA Network Open, 5*(3), e220269. https://doi.org/10.1001/jamanetworkopen.2022.0269

Sahni, N., Stein, G., Zemmel, R., and Cutler, D. (2024). *The potential impact of artificial intelligence on healthcare spending. National Bureau of Economic Research*, doi: 10.3386/w30857

Shaik, T., Tao, X., Higgins, N., Li, L., Gururajan, R., Zhou, X., & Acharya, U. R. (2023). Remote patient monitoring using artificial intelligence: Current state, applications, and challenges. *WIREs Data Mining and Knowledge Discovery*, 13(2), e1485. https://doi.org/10.1002/widm.1485

Snyder, H. (2019). Literature review as a research methodology: An overview and guidelines. *Journal of Business Research*, 104, 333–339. https://doi.org/10.1016/j.jbusres.2019.07.039

Subramanian, S., Pamplin, J. C., Hravnak, M., Hielsberg, C., Riker, R., Rincon, F., … & Herasevich, V. (2020). Tele-critical care: an update from the society of critical care medicine tele-ICU committee. *Critical Care Medicine*, *48*(4), 553-561. doi : https://doi.org/10.1097/CCM.0000000000004190

Tiwari, R. K., and Etienne, M. (2024). Artificial intelligence and healthcare: a journey through history, present innovations, and future possibilities. *Life*, 14(5), 557-562.

Topol, E. (2019). Deep medicine: how artificial intelligence can make healthcare human again. Hachette UK.

Wang, Z., Wei, L., & Xue, L. (2024). Overcoming medical overuse with ai assistance: An experimental investigation. *arXiv preprint*. doi: https://doi.org/10.48550/arXiv.2405.10539

World Economic Forum. (2024). How AI Can Transform Healthcare and Treatment. https://www.ache.org/blog/2022/how-ai-can-transform-healthcare-management

Zhang, Y., Jiang, B., Zhang, L., Greuter, M. J., de Bock, G. H., Zhang, H., & Xie, X. (2022). Lung nodule detectability of artificial intelligence-assisted CT image reading in lung cancer screening. *Current Medical Imaging Reviews*, *18*(3), 327-334. https://doi.org/10.2174/1573405617666210806125953

İSTANBUL UNIVERSITY PRESS

Acta Infologica,
ACIN 2025, 9 (1): 147–166

https://doi.org/10.26650/acin.1670197
Submitted: 04.04.2025
Revision Requested: 29.04.2025
Last Revision Received: 03.05.2025
Accepted: 04.05.2025
Published Online 20.06.2025

Acta Infologica

Research Article

Open Access

Generative AI in Higher Education: Students' Perspectives on Adoption, Ethical Concerns, and Academic Impact

Gökçe Karahan Adalı [1] & Asiye Bilgili [2]

[1] Haliç University, Faculty of Business Administration, Management Information Systems, İstanbul, Türkiye
[2] İstanbul Beykent University, Faculty of Engineering and Architecture, Department of Computer Engineering, İstanbul, Türkiye

Abstract This study explores the adoption patterns of Generative Artificial Intelligence (GenAI) tools among 226 university students from various departments in Turkish higher education institutions, revealing unexpected relationships between age, discipline, and ethical concerns.Through quantitative analysis of university students' survey responses, three distinct clusters of GenAI users' re indentified: high adopters with low ethical concerns, moderate adopters with high ethical awareness, and low adopters with moderate ethical considerations. Notably, the findings challenge the prevalent assumption about younger students' technology adoption, revealing a strong positive correlation between age and AI tool preferences (r=.858, p<.01). The study also revealed significant gender and disciplinary variations, with female and non-STEM students expressing stronger ethical concerns (p<.05 and p<.01, respectively). While students recognized GenAI's potential to enhance academic productivity, they expressed concerns about misinformation, plagiarism, and AI-enabled inequalities. These findings suggest the need for differentiated approaches to AI integration in higher education, considering age-based adoption patterns and discipline-specific variations. The results call for targeted institutional policies addressing ethical literacy, disciplinary needs, and equitable AI access. This research contributes to the growing discourse on GenAI in higher education by providing evidence-based insights for developing more nuanced and effective AI integration strategies.

Keywords Generative AI · Higher Education · Ethical Concerns · Technology Adoption

Citation: Karahan Adalı, G. & Bilgili, A. (2025). Generative AI in higher education: Students' perspectives on adoption, ethical concerns, and academic impact. *Acta Infologica*, *9*(1), 147-166. https://doi.org/10.26650/acin.1670197

This work is licensed under Creative Commons Attribution-NonCommercial 4.0 International License.
© 2025. Karahan Adalı, G. & Bilgili, A.

Corresponding author: Gökçe Karahan Adalı gokceadali@halic.edu.tr

Introduction

Artificial intelligence (AI) has developed significantly over the years, and one of the most notable examples of these developments is Generative Artificial Intelligence (GenAI) (Chakraborty et al., 2024). GenAI is a field of AI that uses large language models (LLMs) and other deep learning models to generate new content like text, images, audio, and video. Technology has revolutionized several fields, including education, medicine, business, and the arts, by generating data based on human input (Eysenbach, 2023). The ability of GenAI to create high-quality and human-like content has raised both enthusiasm and concerns regarding its implementation in educational contexts (Moorhouse et al., 2023). While the first AI applications were based on algorithms that mimicked human intelligence and performed tasks that typically required human cognitive abilities, GenAI goes beyond these capabilities. Thanks to these capabilities, GenAI has found application in various fields, such as art, entertainment, and education (Kirk & Givi, 2025).

Generative AI in Education

GenAI has been introduced as a novel educational technology that supports the learning of students. GenAI, which pushes the boundaries of artificial intelligence and offers new possibilities for human-computer interaction, has managed to attract the attention of researchers and technology enthusiasts (Chen et al., 2024). Applications of GenAI, especially in the field of education, are of particular importance because of its potential to transform and enrich learning processes (Kumar et al., 2025).

Such technologies can simplify academic work, personalize instructional processes, and optimize learning efficiency (Chaudhry & Kazim, 2024; Zhai, 2023). However, the ethical, academic integrity, and pedagogical consequences of such technologies are controversial (Fisher & Haake, 2024).

Use of Generative AI in Education

The application of GenAI in the classroom can transform pedagogy. Students engage with such tools to complete homework, conduct research, brainstorm, and summarize information (Montenegro-Rueda et al., 2023; Sun et al., 2024). The primary benefits of GenAI in education include the following:

- Personalized Learning: AI-generated content can be tailored to an individual's own learning rhythm (Zhai, 2023).
- More Engaging and Interactive Learning: AI-powered chatbots and content generators can facilitate interactive interaction with students (Fisher & Haake, 2024).
- Development of Language and Writing Skills: AI can assist students in improving their academic writing (Ali et al., 2024; Moorhouse et al., 2023).

However, the universal use of GenAI in learning is accompanied by challenges and ethical concerns, such as the following:

- Academic Integrity: Excessive reliance on these tools may lead to academic dishonesty (Xames & Shefa, 2023; Sun et al., 2024).
- Information Reliability: AI-generated information may at times not be dependable, and this may affect the academic performance of learners (Clark et al., 2024; Jo, 2024).
- Teacher Roles and Pedagogical Approaches: Educators should redesign their instructional models to accommodate AI technologies effectively (Fisher & Haake, 2024; Sandu et al., 2024).

With the rapid development of technology, students need to quickly access the right information in the field of education and effectively support their learning processes. With the recent spread of generative artificial intelligence tools such as ChatGPT, Claude, and Bard, the integration of these technologies into educational environments and the level of their adoption by students have become critical research topics (Saaida, 2023). In order for these technological transformations to be successfully realized in higher education and to be used effectively by students, it is necessary to first understand students' perceptions and attitudes toward these tools.

While generative artificial intelligence tools add a new dimension to the interaction between computers and students, they create an interdisciplinary field that focuses on the integration of this interaction into academic processes. These tools are affected by the continuous development of technology because their structures and areas of use are gradually expanding. In this field, which aims to shape technology according to student needs, the aim is not to have students adapt to technology but to adapt technology to student needs and expectations.

Literature Review

The evolution of Generative AI is deeply rooted in the foundational advancements of the 20th century. Alan Turing's theoretical contributions, particularly the concept of the Turing Machine and the Turing Test, laid the philosophical and computational groundwork for artificial intelligence (Turing, 1950). In the 1950s, the development of the Perceptron by Frank Rosenblatt introduced one of the first neural network models, enabling machines to learn from input data (Rosenblatt, 1958). During the 1960s and 1980s, symbolic AI, also known as "Good Old-Fashioned AI," dominated the field, relying on hand-coded rules and logical inference (Russell & Norvig, 2021). However, the limitations of symbolic systems led to a shift toward connectionist approaches, especially after the introduction of backpropagation algorithms in the 1980s (Rumelhart, Hinton, & Williams, 1986). These milestones not only expanded AI's computational capacity but also set the stage for today's data-driven, generative models that use deep learning and large-scale language processing.

Albadarin et al. (2024) systematically examined 14 empirical studies on the use of ChatGPT in education, and their research results showed that ChatGPT has a bidirectional effect on educational processes. While ChatGPT has positive contributions, such as providing instant feedback to students as a virtual intelligent assistant, improving writing and language skills, and providing personalized learning support, it also points out that if used excessively, it can weaken students' creativity, collaborative learning, and critical thinking skills. In the study examining the benefits and limitations of using ChatGPT in higher education, we found that ChatGPT can provide personalized educational experiences thanks to its natural language processing, text generation capacity, and performance evaluation opportunities. However, important problems such as content quality, possible bias in responses, risk of plagiarism, and content originality were also emphasized (Ali et al., 2024). In the study examining the negative effects of privacy concerns, fear of technology, and guilt on the intention to use chatbots, it was determined that this privacy concern is of critical importance due to the intensive sharing of personal data in interactions with AI chatbots. This study examined how fear of technology (technophobia), fear, and anxiety toward advanced technologies affect the intention to use such technologies, and it was determined that negative attitudes toward advanced technologies, such as AI chatbots, prevent their use. It was observed that the feeling of guilt arising from the thought of being dependent on AI in learning or work processes reduces the intention to use (Jo, 2024). This systematic literature review revealed the multifaceted effects of ChatGPT in the field of education. According to the

research findings, appropriate use of ChatGPT significantly improves students' academic performance and acts as a motivating tool for students in terms of practical application of information and communication technologies. In the context of communication departments in German universities, Henke (2025) observed a significant increase in the adoption of GenAI tools between 2023 and 2024, particularly for tasks such as text generation and translation, indicating a shift toward mainstream integration. However, it was determined that there is a need for comprehensive training to prevent misuse and that it has led to new discussions in traditional teaching methods, methodology and evaluation processes. It was emphasized that ethical use and appropriate supervision are of critical importance in this process (Montenegro-Rueda et al., 2023). In the study examining the possibility of transformation of large language models such as ChatGPT in the learning environment in higher education, it was found that despite some challenges, students viewed ChatGPT as an effective academic aid and that it can potentially create more inclusive learning environments. In addition, there was emphasis on how institutional backing and ethical issues play a key role in the effective adoption of AI tools in educational environments, and how matters such as data safety, plagiarism, and potential biases must be tackled proactively (Sandu et al., 2024).

Johnston et al. (2024) conducted a vast survey at Liverpool (n=2555) and discovered that, while 93% of students were familiar with generative AI tools, the intention to implement them was not identical.
While 70.4% of them opposed the use of AI to compose complete essays, the majority appreciated tools like Grammarly for assistance. Confidence in writing was most prominent, where students who were confident in writing were less likely to utilize AI tools. The study sought to address the need for specific institutional policies rather than bans so that the usage of AI technologies was made accessible to all student groups equitably. A study by Chan and Hu (2023) involving a survey of 399 Hong Kong university students found a predominantly positive attitude toward GenAI, with students recognizing its potential for personalized learning, research assistance, and writing help. Worries were raised about data privacy, ethical implications, and AI-generated misinformation. The study recommends that policymakers develop AI integration plans that weigh technological benefits against ethical use in higher education.

Similarly, research by Song and Wang (2024) on 487 Chinese design college students and identified that over 60% of them used AI tools primarily to collect data, brainstorm, and offer conceptual design assistance. Students identified AI as a tool for enhancing creativity and productivity although in-depth integration was limited. Furthermore, AI anxiety was rather prevalent among students who were not exposed to such technologies. The writers highlighted the importance of literacy in AI in creative endeavors, arguing that education in AI can prepare students to employ AI as a design tool rather than as a substitute. A global study by Ravšelj et al. (2025) revealed that while students recognize the utility of ChatGPT for tasks like summarization and idea generation, they also express concerns regarding its reliability and ethical implications in academic settings. Khlaif et al. (2024) explored university instructors' perspectives in the Middle East, finding that while educators acknowledge the potential of GenAI tools to enhance assessment practices, they also raise concerns about academic integrity and the necessity for clear institutional policies.

Recent studies have explored the determinants of higher education students' adoption of Generative AI technologies. For instance, Sergeeva et al. (2025) employed the Unified Theory of Acceptance and Use of Technology 2 (UTAUT2) framework to empirically investigate these factors, highlighting the roles of performance expectancy and social influence in students' adoption behaviors.

When the studies conducted in this field are examined; it is seen that the focus is on the use of generative artificial intelligence tools in higher education, student perceptions, attitudes, purposes of use, ethical concerns, and academic honesty (Michel-Villarreal et al., 2023; Ogunleye et al., 2024; Xames & Shefa, 2023).

With the use of GenAI tools in higher education, a new form of relationship has emerged between learning-teaching processes and artificial intelligence. This relationship is important in terms of understanding how students use these tools to fulfill their academic duties and their attitudes toward these tools. In this context, examining university students' perceptions and attitudes toward generative artificial intelligence tools can contribute to the development of strategies for the effective and ethical use of these technologies in higher education. At the same time, it is believed that understanding students' motivations and concerns for using these tools will guide the shaping of future education policies and practices. Accordingly, the research questions determined within the scope of the study are as follows:

Main Research Question:

What are the perceptions, attitudes, and purposes of university students' use of ChatGPT and similar generative AI tools?

RQ1: How do university students perceive generative AI tools in academic processes?

RQ2: What are the students' positive and negative attitudes toward the use of these tools?

RQ3: For what academic purposes do students use generative AI tools?

RQ4: What kind of experiences do users have when using these tools?

RQ5: What are students' views on the ethical use of generative AI tools?

RQ6: How did the respondents evaluate the use of these tools in the context of academic honesty?

These research questions provide a comprehensive framework to fill the gaps identified in the literature and to understand the educational dimension of generative AI tools from a student perspective.

Method

This research was designed using the descriptive survey model, which is a quantitative research method. A cross-sectional research design was used in the study, which aimed to determine the perception, attitude, and usage patterns of university students toward generative artificial intelligence tools.

Study Group

The study group of the research was determined using maximum diversity sampling, a purposeful sampling method. This method aimed to understand the change in attitudes toward artificial intelligence tools among different disciplines and demographic characteristics by ensuring the participation of students from different departments and demographic groups. The study group consisted of university students who were studying at different universities and faculties and actively used generative artificial intelligence tools. A total of 226 students participated in the study. Table 1 provides the demographic information about the participants.

Table 1
Demographic Information

		N	%
Gender	Female	126	56
	Male	100	44
Age	18-20	108	47.8
	21-23	87	38.5
	24–26	21	9.3
	27 and older	8	3.6

When the demographic characteristics of the participants are examined, it is seen that 56% (n=126) are female and 44% (n=99) are male. When evaluated in terms of age distribution, 47.8% (n=108) of the students were in the 18-20 age range, 38.5% (n=87) were in the 21-23 age range, 9.3% (n=21) are in the 24-26 age range, and 3.6% (n=8) are 27 years old and over. These demographic data show that most sample (86.3%) is concentrated in the traditional university age group of 18-23.

Students from 16 different associate and undergraduate departments participated in the study. The fact that 55.2% of the participants were from the informatics field reflects the more widespread use of artificial intelligence tools in these disciplines. However, this study aimed to compare the attitudes and perceptions of students in various disciplines toward artificial intelligence tools. For this reason, the research did not solely focus on information technology students but included participants from diverse academic backgrounds. While information technology students are known to use these tools more frequently, including students from other disciplines, was critical for evaluating the impact of artificial intelligence tools on general education processes from a broader and interdisciplinary perspective.

When examining the distribution of students by department, the highest participation rate was observed in the Management Information Systems department (33.6%, n=76), followed by Computer Engineering (15%, n=34), Business Administration (11.9%, n=27), Public Relations and Promotion (10.2%, n=23), and International Trade and Business (6.2%, n=14). Departments such as Electrical and Electronics Engineering (4.4%, n=10), Political Science and International Relations (4%, n=9), and Software Engineering (3.5%, n=8) showed medium-level participation. Meanwhile, departments like Industrial Engineering (2.2%, n=5), Computer Technology and English Translation and Interpreting (1.8%, n=4), and Computer Programming (1.3%, n=3) had lower participation rates. Lastly, Radio, Television and Cinema, Mechanical Engineering, and American Culture and Literature had the least representation, each with 0.4% (n=1). This distribution highlights the strong participation from information technology and business-related fields, while also ensuring representation from various academic disciplines to achieve a comprehensive perspective.

Data Collection Tools

The questions asked the participants were prepared within the scope of the study as a result of the literature review. This consists of questions covering basic dimensions such as students' purposes and frequency of use of generative artificial intelligence tools, perceived benefits and difficulties, impact on academic processes, ethical concerns and security concerns. The survey form comprised sections that included demographic information and various dimensions regarding the use of generative artificial intelligence, in line with the purposes of the research. In terms of the ethical compliance of the research, an application was made to the Haliç University Social and Human Sciences Ethics Committee, and the ethics committee

approval was obtained with the decision numbered (10) dated (25.12.2024). The voluntary participation of the participants was taken as a basis in the research process, and the confidentiality of their personal data was guaranteed.

Data Analysis

In order to ensure methodological transparency and to clearly demonstrate how each research question is addressed, the following table presents the alignment between the research objectives, corresponding data sources, and statistical techniques employed in the study. Table 2 shows the mapping of each research question to the relevant data and analysis techniques.

Table 2

Mapping of Research Questions to Data Sources and Analysis Methods

Research Question	Relevant Survey Items	Analysis Method
RQ1	Q12, Q14, and Q16:	Descriptive, Correlation
RQ2	Q17–Q20	Factor, Cluster
RQ3	Q8, Q9, and Q10:	Descriptive
RQ4	Q11, Q13	Correlation
RQ5	Q21–Q24	Factor, ANOVA
RQ6	Q25–Q27	Factor

The persuasiveness of the findings obtained in the study is directly related to validity and reliability. All questions asked to the participants were included in the validity and reliability analyses conducted within the scope of the study. Kaiser-Meyer-Olkin (KMO) and Bartlett's sphericity test were used to assess the suitability of the data set for factor analysis. The test results are shown in Table 3.

Table 3

KMO and Bartlett's Test

KMO and Bartlett's test		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.713
Bartlett's test of sphericity	Approx. Chi-Square	1497.971
	df	435
	Sig.	.000

The KMO value was found to be 0.713. This value is in the range of 0.70-0.79, it shows that the sample adequacy is at a "good" level. When the Bartlett sphericity test results were examined, the Chi-square value was found to be 1497.971 (p<.001). The significance of the Bartlett test (p<.001) shows that the relationships between the variables are suitable for factor analysis. According to these results, the data set met the necessary assumptions for factor analysis, and the analysis was continued. In the study, a grouping factor analysis was performed to examine the relationship between several variables measuring a certain phenomenon. The Total Variance Explained Table indicates how many factors the variables are grouped under. The results are shown in Table 4.

Table 4

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings	
	Total	% Variance	Cumulative %	Total	% Variance
1	5.479	18.263	18.263	5.479	18.263
2	3.289	10.965	29.227	3.289	10.965
3	2.199	7.330	36.557	2.199	7.330
4	1.907	6.358	42.915	1.907	6.358
5	1.398	4.661	52.546	1.398	4.661
6	1.281	4.269	56.815	1.281	4.269
7	1.070	3.565	64.437	1.070	3.565
8	.962	3.205	67.643		
9	.925	3.083	70.725		
10	.917	3.057	73.783		
11	.806	2.687	76.470		
12	.779	2.596	79.067		
13	.687	2.290	81.357		
14	.617	2.058	83.415		
15	.599	1.996	85.410		
16	.552	1.841	87.251		
17	.508	1.692	88.943		
18	.461	1.538	90.482		
19	.422	1.408	91.890		
20	.394	1.313	93.203		
21	.342	1.139	94.342		
22	.323	1.077	95.419		
23	.287	.958	96.377		
24	.274	.914	97.291		
25	.258	.859	98.151		
26	.224	.746	98.896		
27	.180	.600	99.496		
28	.151	.504	100.000		
29					
30					

Because of the factor analysis, 7 factors with eigenvalues greater than 1. These factors explain 64.437% of the total variance. The first factor explains 18.263% of the total variance, the second factor explains 10.965%, and the third factor explains 7.330%. The contributions of other factors to the variance are 6.358%, 4.970%, 4.661%, 4.269%, 4.057%, and 3.565%, respectively.

Table 5 shows the Rotated Component Matrix indicating which question statement falls under each factor.

Table 5

Rotated Component Matrix

	Component								
	1	2	3	4	5	6	7	8	9
Which grade are you studying?				.846					
How do you perceive the use of Large Language Models (LLM) and AI tools and their impact on academic integrity and the quality of education?					.475			-.565	
On average, how many hours do you spend studying or working on academic assignments per week?			.405						
Which device do you most frequently use for academic purposes?				-.404					
Have you used Generative AI tools previously?								.521	
Indicate the reasons for not using Generative AI tools.									.895
How often do you use generative AI tools?			-.715						
How familiar are you with generative AI tools?			.763						
Which generative AI tools do you most frequently use?						-.697			
In which of the following ways do you use LLMs or other Generative AI tools? [Problem-solving and brainstorming (e.g., generating ideas, finding solutions to specific challenges)]	.582					.463			
In which of the following ways do you use LLMs or other Generative AI tools? [Learning about AI (e.g., experimenting with AI tools to understand how they work)]	.815								
In which of the following ways do you use LLMs or other Generative AI tools? [Academic purposes (e.g., writing essays, generating study materials, summarizing texts)]	.471					.426	.425		
In which of the following ways do you use LLMs or other Generative AI tools? [Creative content creation (e.g., writing stories, creating art, generating music or poetry)]	.767								
In which of the following ways do you use LLMs or other Generative AI tools? [Programming and coding assistance (e.g., debugging code, generating scripts, learning programming languages)]			.531			.429			
In which of the following ways do you use LLMs or other Generative AI tools? [Data analysis and visualization (e.g., generating graphs, summarizing datasets, statistical analysis)]	.660								
In which of the following ways do you use LLMs or other Generative AI tools? [Language learning (e.g., improving grammar, practicing conversational skills, translating texts)]	.765								
In which of the following ways do you use LLMs or other Generative AI tools? [Professional tasks (e.g., writing emails, preparing reports, brainstorming ideas)]	.770								
In which of the following ways do you use LLMs or other Generative AI tools? [Personal use (e.g., exploring AI capabilities, entertainment, chatting with AI for fun)]	.713								
In which of the following ways do you use LLMs or other Generative AI tools? [Social media content (e.g., creating posts, captions, hashtags, or managing online presence)]	.730								

	Component								
	1	2	3	4	5	6	7	8	9
How do students perceive the use of LLMs and AI tools and their impact on academic integrity and the quality of education? [As a helpful tool]					.723				
How do students perceive the use of LLMs and AI tools and their impact on academic integrity and the quality of education? [As a double-edge sword]		.578			.413				
How do students perceive the use of LLMs and AI tools and their impact on academic integrity and the quality of education? [As a way to level the playing field]					.693				
How do students perceive the use of LLMs and AI tools and their impact on academic integrity and the quality of education? [As a means for students to cheat]		.773							
How do students perceive the use of LLMs and AI tools and their impact on academic integrity and the quality of education? [As an unfair advantage]		.768							
How do students perceive the use of LLMs and AI tools and their impact on academic integrity and the quality of education? [With concern for the human element]		.739							
How do students perceive the use of LLMs and AI tools and their impact on academic integrity and the quality of education? [With risk of misleading information]		.768							
Indicate the extent to which you agree or disagree with the following statements.								.621	

Because of the factor analysis, 7 factors were identified. The first factor includes 9 items covering different usage purposes of AI, and the factor loadings range from .471 to .815. The second factor includes 5 items covering ethical concerns regarding AI use, and the factor loadings range from .578 to .773. The third factor includes 3 items regarding AI usage competence, and the factor loadings range from .405 to .763. The fourth factor includes AI tool preference and academic use (factor loadings: .426-.697), and the fifth factor includes 2 items reflecting positive perceptions toward AI (factor loadings: .413-.723). The sixth and seventh factors include AI usage experience and reasons for not using it. Table 6 shows the themes of the factors and Cronbach's Alpha values to show whether any variable creates distrust of the factor.

Table 6

Themes of the Factors

Factor	Theme	Number of Items	Cronbach's Alpha
F1	AI Usage Purposes	9	.874
F2	Ethical Concerns	5	.821
F3	AI Usage Competence	3	.421
F4	AI Tool preferences	3	.691
F5	Positive AI Perception	2	.450
F6	AI Usage Experience	1	Reliability analysis is not performed for single-item factors.
F7	Reasons for not using AI	1	

Reliability analysis of the structure obtained because of factor analysis was performed. While F1 (α = .874) and F2 (α = .821) showed high reliability, F4 (α = .714) had acceptable reliability. The reliability coefficients of F3 were below the expected level. Because F6 and F7 were single-item factors, reliability analysis was not performed. The Cronbach's Alpha value was still lower than the critical value (.70) in case of deletion of questions that decreased reliability for F6 and F7 factors were removed from the analysis.

One-Sample Kolmogorov-Simirnov test was performed to determine whether the tests to be performed after this stage were parametric or nonparametric, and the results are shown in Table 7.

Table 7

Tests of Normality

	Kolmogorov-Smirnov			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
F1	.085	194	.002	.981	194	.009
F2	.083	194	.002	.982	194	.016
F4	.196	194	.000	.919	194	.000

According to the analysis results, both Kolmogorov-Smirnov (p=.002) and Shapiro-Wilk (p=.009) test results were found to be significant for the F1 (AI Usage Purposes) factor. Similarly, Kolmogorov-Smirnov (p=.002) and Shapiro-Wilk (p=.016) tests for the F2 (Ethical Concerns) factor gave significant results. Normality distribution was also examined for F6 (AI Usage Experience) and F7 (Reasons Not to Use AI), and all test results had p values less than .05 and the data did not conform to a normal distribution. For this reason, we decided to use nonparametric tests in further analyses.

Results

Spearman's correlation analysis was performed to determine the relationship between the variables, as shown in Table 8.

Table 8

Spearman's correlation analysis

		AI Usage Purposes	Ethical Concerns	AI Tool preferences	Age	Gender	Which department are you studying in?	On average, how many hours do you spend studying or working on academic assignments per week?
AI Usage Purposes	Correlation Coefficient	1.000	-.005	-.011	-.024	.072	-.125	.041
	Sig. (2-tailed)	.	.949	.873	.735	.301	.074	.562
	N	209	195	207	207	208	207	206
Ethical Concerns	Correlation Coefficient	-.005	1.000	.186**	.164*	-.101	-.177*	.164*
	Sig. (2-tailed)	.949	.	.008	.019	.150	.012	.020
	N	195	204	203	203	204	202	202
AI Tool preferences	Correlation Coefficient	-.011	.186**	1.000	.858**	-.090	-.573**	.130

		AI Usage Purposes	Ethical Concerns	AI Tool preferences	Age	Gender	Which department are you studying in?	On average, how many hours do you spend studying or working on academic assignments per week?
Age	Sig. (2-tailed)	.873	.008	.	.000	.182	.000	.055
	N	207	203	224	224	223	221	220
	Correlation Coefficient	-.024	.164*	.858**	1.000	-.157*	-.455**	.106
Gender	Sig. (2-tailed)	.735	.019	.000	.	.019	.000	.116
	N	207	203	224	224	223	221	220
	Correlation Coefficient	.072	-.101	-.090	-.157*	1.000	.104	-.133*
Which department are you studying in ?	Sig. (2-tailed)	.301	.150	.182	.019	.	.121	.048
	N	208	204	223	223	225	222	222
	Correlation Coefficient	-.125	-.177*	-.573**	-.455**	.104	1.000	-.018
On average, how many hours do you spend studying or working on academic assignments per week?	Sig. (2-tailed)	.074	.012	.000	.000	.121	.	.796
	N	207	202	221	221	222	223	220
	Correlation Coefficient	.041	.164*	.130	.106	-.133*	-.018	1.000
	Sig. (2-tailed)	.562	.020	.055	.116	.048	.796	.
	N	206	202	220	220	222	220	222

The analysis findings demonstrate significant relationships between various variables. When examining the relationships between students' ethical concerns and other variables, it was found that as ethical concerns increased, AI tool preferences also increased (r= .164, p<.05). Similarly, ethical concerns were found to increase with age (r= .164, p<.05) and weekly study hours (r= .164, p<.05). Additionally, ethical concerns varied significantly across different academic departments (r= -.177, p<.05). Regarding AI tool preferences, the most notable finding is the strong positive correlation with age (r= .858, p<.01), indicating that AI tool usage preferences significantly increase with age. Conversely, a strong negative correlation was observed between academic department and AI tool preferences (r= -.573, p<.01) suggests substantial variations in AI usage preferences across different academic disciplines. Analysis of demographic variables revealed a weak negative correlation between age and gender (r= -.157, p<.05) and a moderate negative correlation between age and academic department (r= -.455, p<.01). Furthermore, a weak negative correlation was identified between gender and weekly study hours (r= -.133, p<.05), indicating gender-based differences in study duration patterns.

Notably, the AI usage purposes variable showed no significant correlation with any other variables (p>.05), suggesting that AI usage purposes develop independently of other factors. These findings indicate

that attitudes and behaviors toward AI usage are associated with various factors, including age, academic department, ethical concerns, and study habits. The strong influence of age and academic department variables on AI tool preferences suggests that these factors should be considered when developing strategies for AI technology integration into educational processes.

The factor and correlation analyses revealed the fundamental factors influencing students' attitudes and behaviors in AI usage while also uncovering the relationships between these factors. To gain a deeper understanding of these relationships and determine students' AI usage profiles, cluster analysis was performed. K-means cluster analysis using factor scores revealed three distinct student groups. The Elbow method and Silhouette score tests confirmed three clusters as the optimal number (Silhouette score = 0.421).

The relationships between age, gender, and ethical concerns identified in the correlation analysis showed similar patterns in the cluster analysis. For instance, the positive correlation between age and ethical concerns (r= .164, p<.05) aligns with the older age and high ethical concerns profile observed in Cluster 0. Similarly, the strong positive correlation between AI tool preferences and age (r= .858, p<.01) was reflected in the clustering results, particularly in Cluster 1, where younger students showed greater propensity for AI usage. These two analytical methods provided complementary findings, enabling a more comprehensive understanding of students' AI usage profiles. While factor and correlation analyses revealed relationships between variables, cluster analysis demonstrated how these relationships manifested across specific student groups.

Cluster Method

Factor scores were computed by adding Likert-scale values for each of the selected constructs: AI Usage Purposes, Ethical Concerns, AI Tool Preference, Demographic Characteristics, Gender, AI Usage Experience, and Reasons Not to Use AI. Standardization was performed using Z-score normalization to enable variable comparison. The optimal number of clusters was determined using both the Elbow Method and Silhouette Score. The Elbow Method indicated a significant decline in WCSS up to K=3, after which the rate of decrease flattened. The highest Silhouette Score (0.421) was also observed at K=3, confirming that this number provided the most distinct and well-separated clusters. The k-means algorithm with K=3 was used and resulted in the following characteristics for each cluster: Table 9 shows the statistical characteristics of the three identified clusters.

Table 9

Summary Statistics of Cluster Characteristics

Cluster	AI Usage Score	Ethical Concerns Score	AI Tool preferences	Predominant Demographic
0	−0.105	0.147	−0.519	Older, Female
1	0.031	−0.531	1.438	Younger, Balanced Genders
2	0.078	0.197	−0.421	Predominantly Male

Cluster 0: Moderate AI usage with high ethical concerns

Students in this group possessed the highest average AI utilization score (-0.105) but the highest ethical concerns (0.147), showing that such students, particularly among a bit older and female participants, are defensive toward AI use, most likely due to expected threats to academic honesty. Their AI tool preference was the lowest among clusters (-0.519) and is consistent with their cautious attitude.

Cluster 1: High AI Usage-Low Ethical Concerns

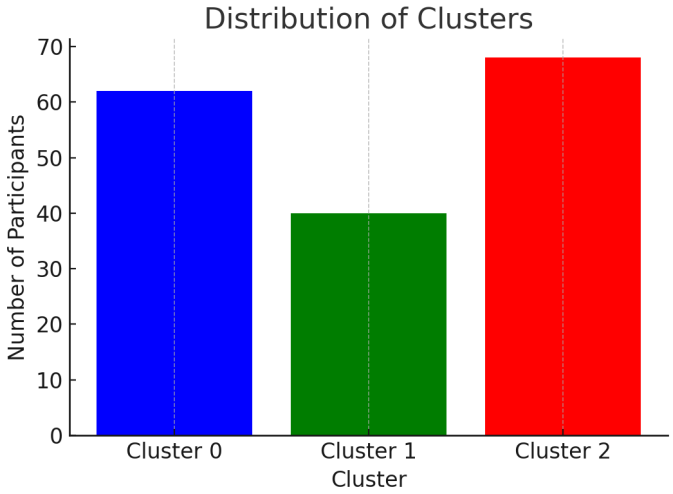
This cluster represents the student group with a moderate usage score of AI (0.031) coupled with greater ethical issues (-0.531) about the influence of AI on academic honesty. Indicating that these students are most enthusiastic about AI and view it as a useful tool, not an ethical matter. Their favor for AI tools was very high (1.438), which indicates frequent usage. Stronger possibility of female and more senior learners. This cluster has wary adoption of AI due to concerns about academic integrity.

Cluster 2: Low AI Usage-Moderate Ethical Concerns

This cluster had the lowest AI use score (0.078) and moderate ethical concerns (0.197), reflecting a less active but not necessarily anti-AI group. Their preference for AI tools (-0.421) is quite low, reflecting little reliance on AI-based tools, with a predominance of male participants. The students in this group represent the most engaged users of AI and have the lowest ethical concerns about using AI tools. They actively utilized AI for intellectual and professional purposes and were easier to use with AI technology. This cluster includes young students with an equivalent gender split. These results reflect diverse perspectives on AI adoption and ethical concerns among students.

Figure 1

Distribution of Clusters



The distribution of participants across the three clusters indicates varying trends in AI adoption (Figure 1). This suggests that although a large majority of students possess ethical concerns and conservative AI adoption (Cluster 0), an equally large majority are not engaged with AI (Cluster 2). The relatively high presence in Cluster 1 reflects the growing number of students actively employing AI tools with limited ethical issues. To understand these trends further, it is essential to examine them from a demographic perspective.

AI Usage Concerns and Preferences Across Demographic Levels

To further investigate whether there is a difference in design students' use of AI at the demographics level, statistical methods were applied to explore AI participation patterns and ethical concerns by groups of different demographics. The data were divided by gender and academic discipline, and one-way ANOVA and independent sample t-tests were conducted to compare means to establish differences. Findings indicated that female students had statistically higher ethical issues regarding AI application, particularly on the

topic of misinformation and cheating (p<.05, Table 10). Male students, on the contrary, would be more likely to answer that insufficient knowledge or technical expertise prevented the implementation of AI. These tendencies suggest that AI literacy courses must be designed to address ethical concerns and skill deficits.

Table 10

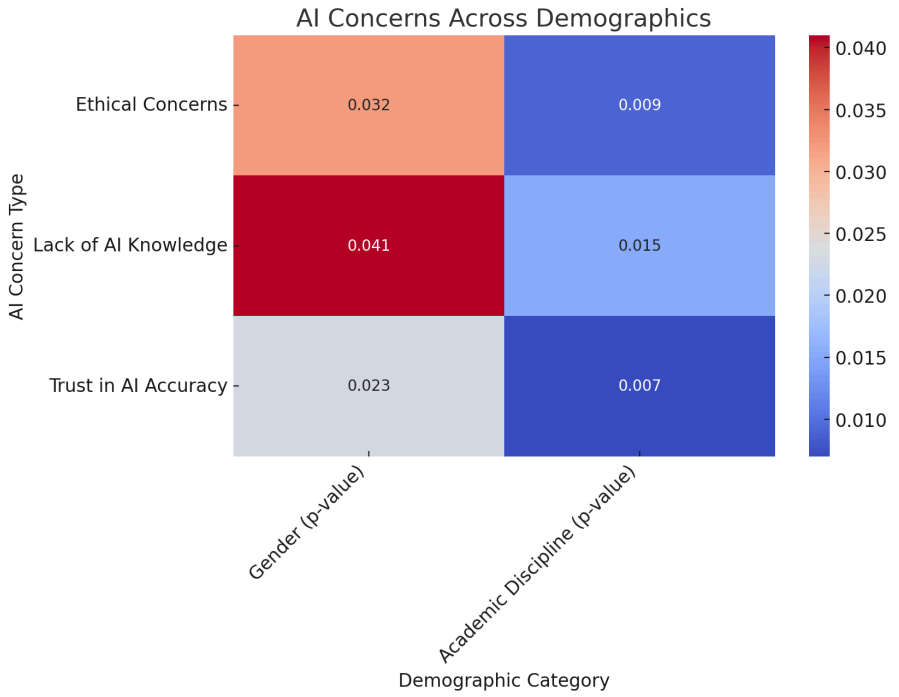
Statistical Significance of AI Concerns Across Demographics

Variable	Gender (p)	Academic Discipline (p)
Ethical Concerns	.032	.009
Lack of AI Knowledge	.041	.015
Trust in AI Accuracy	.023	.007

Academically, the findings revealed that non-technical students (e.g., social sciences and humanities) are more interested in the authenticity and ethical issues of AI-created content, and there are considerable differences found in their replies compared to STEM students (p<.01, Table 10). Technology and computer science students, however, showed extreme confidence in AI tools with few ethical issues. These variations highlight the importance of discipline-specific AI training to guarantee accountable and informed AI use across all areas of scholarship.

Figure 2

AI Concerns Across Demographics



The heatmap (Figure 2) illustrates the statistical significance of the variation in AI concerns by academic discipline and gender. The p-values for gender indicate significant variation in ethical concerns (p=.032), lack of AI knowledge (p=.041), and trust in AI accuracy (p=.023), suggesting that men and women perceive these concerns differently. Similarly, academic disciplines also show statistically significant differences in all AI issues, with the lowest p-value observed for AI trust in accuracy (p=.007), which describes participants from different fields having varying trust in AI systems. In total, these findings illustrate the influence of

demographic variables on AI-related issues and highlight the need for tailored approaches to AI education and policy adoption.

Figure 3

AI Usage Preferences according to Demographics and Discipline

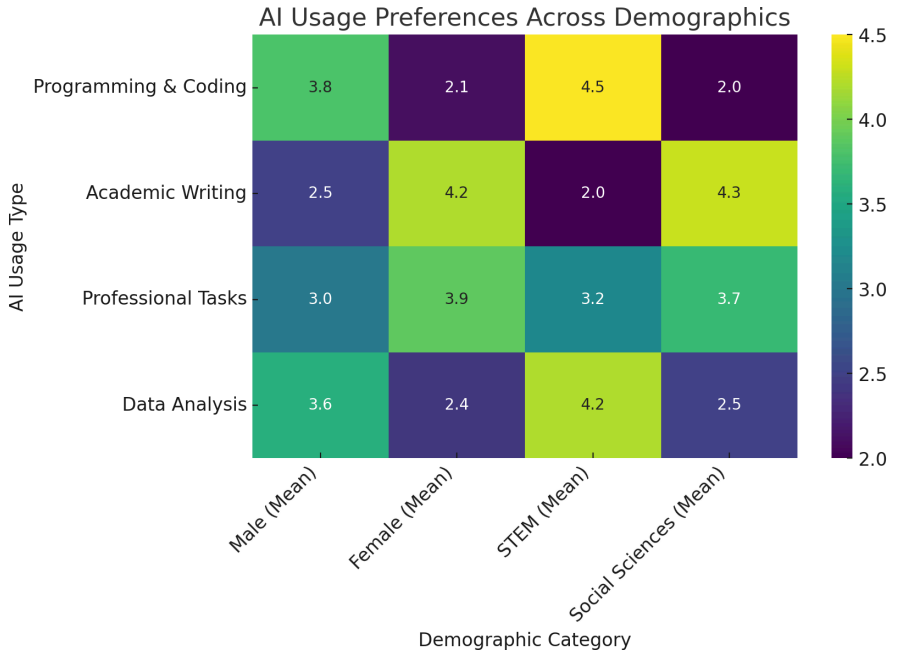


Figure 3 shows AI usage preference among different groups of demographics, including gender and stream of study (STEM and Social Sciences). These values are the mean scores for the four themes of AI usage: programming and coding, university-level writing, tasks at work, and data analysis. The results indicate clear demographic variations in AI preference. Men are more inclined toward AI in programming and coding (mean=3.8) and data analysis (mean=3.6) compared to women, who score much lower in these areas (mean=2.1 and 2.4). Women are more interested in AI for academic writing (mean=4.2) and professional work (mean=3.9) than men.

In academic discipline terms, STEM participants are most interested in using AI for programming and coding (mean=4.5) and data analysis (mean=4.2), whereas Social Sciences participants are most active in using academic writing (mean=4.3) but show lower interest in AI for technical work like programming (mean=2.0) and data analysis (mean=2.5). The results suggest that trends in AI use are shaped by gender as well as academic discipline, reflecting differences in requirements and comfort levels with AI according to discipline.

Discussion and Conclusion

This study provides comprehensive insights into university students' perceptions, attitudes, and usage patterns toward generative AI tools. In addition, it also reveals important findings that contribute to our understanding of AI integration in higher education.

Correlation analysis revealed high correlations between ethical concerns and the selection of AI tools, age, and other variables such as work habits. More precisely, the significant positive correlation between ethical awareness and preference for AI tools (r=.164, p<.05) implies that students with higher ethical aware-

ness are more careful when applying AI tools. This finding is supported by Johnston et al.'s (2024) research that confirmed that students' attitudes toward AI tools are shaped primarily by their ethical concerns. The strong positive correlation between age and preferred AI tools ($r=.858$, $p<.01$) is one of the interesting results, i.e., older learners are more likely to use AI tools. This differs from some previous studies showing that young learners are more sensitive to technology, but it is similar to Chan and Hu's (2023) findings that demonstrate that adult learners learn more deliberately from AI tools. The cluster analysis demonstrated three distinct student profiles, which are quite valuable for revealing various student groups in relation to interacting with AI technology. High ethical problems and medium levels of AI usage appeared together in Cluster 0 as a vigilant pattern, while Cluster 1 indicates an open personality represented by high use and low levels of ethical concerns. The outcome aligns with the work by Song and Wang (2024), where equal diverging student uses of AI were revealed.

Demographic differences in patterns of AI usage showed that female students had more ethical concerns, in particular, misinformation and academic dishonesty. This finding contributes to the growing body of evidence on gender differences in technology uptake and suggests critical implications for designing inclusive AI learning strategies.

The differences among disciplines, where STEM students showed more technical inclination and fewer ethical concerns than non-STEM students, suggest the need for discipline-based approaches to AI integration in education. This aligns with Montenegro-Rueda et al.'s (2023) research on the differential impacts of AI by academic discipline.

This study provides significant information about the complex adoption of generative AI in tertiary education. The findings revealed that students' use of AI tools was influenced by several factors, such as age, gender, study field, and ethical considerations. The classification of different user profiles using cluster analysis provides a foundation for different AI adoption strategies among students.

The results of this study are as follows:

- The relationship between ethical concerns and AI use is more complex than previously described. However, despite this complexity, high ethical awareness does not reduce the use of AI.
- Although existing literature generally suggests that younger students are more prone to using artificial intelligence, our findings demonstrate that older students exhibit higher levels of AI tool adaptation.
- Significant disciplinary and gender differences in AI adoption patterns suggest the need for customized approaches to AI integration in education.
- The emergence of different user profiles suggests that a single AI policy approach to education may be ineffective.

These findings have important implications for education policy and practice. Accordingly, institutions should:

- Develop discipline-specific guidelines for AI integration.
- Address ethical concerns while promoting responsible AI use.
- Consider demographic factors when designing AI policies.
- Provide targeted support to different student groups based on their usage patterns and concerns.

While this study focuses on students' present involvement with Generative AI, these opinions must be interpreted within the history of AI advances as a whole. The shift from symbolic systems of reasoning to statistical and neural ones over the last few decades has, in essence, changed what is possible with AI. Today's GenAI technologies, such as ChatGPT, are the end point of a lengthy development beginning in theoretical abstractions and evolving into productive, generative systems through cumulative innovation. Understanding this history not only serves to contextualize the educational relevance of GenAI today but also highlights the importance of historical awareness in shaping future AI policy and pedagogy. Future research should examine longitudinal changes in student attitudes toward AI, investigate the effectiveness of various AI integration strategies across disciplines, and examine how ethical frameworks can be developed to guide AI use in academic settings.

The limitations of this study include its cross-sectional nature and its focus on a specific geographic context. The findings also provide valuable insights for educators, administrators, and policymakers seeking to integrate AI technologies into higher education while maintaining academic integrity and promoting effective learning outcomes.

Ethical approval	Ethics committee approval for this research was obtained from the Haliç University Social and Human Sciences Research Ethics Committee (Date: 25.12.2024, No: 10).
Informed Consent	Written informed consent was obtained from all participants who participated in this study.
Peer Review	Externally peer-reviewed.
Author Contributions	Conception/Design of Study- G.K.A., A.B.; Data Acquisition- G.K.A., A.B.; Data Analysis/Interpretation- G.K.A., A.B.; Drafting Manuscript- G.K.A., A.B.; Critical Revision of Manuscript- G.K.A., A.B.; Final Approval and Accountability- G.K.A., A.B.
Conflict of Interest	The authors have no conflict of interest to declare.
Grant Support	The authors declared that this study has received no financial support.

Author Details

Gökçe Karahan Adalı

[1] Haliç University, Faculty of Business Administration, Management Information Systems, İstanbul, Türkiye

0000-0001-8567-4626 gokceadali@halic.edu.tr

Asiye Bilgili

[2] İstanbul Beykent University, Faculty of Engineering and Architecture, Department of Computer Engineering, İstanbul, Türkiye

0000-0002-9776-5119

References

Albadarin, Y., Saqr, M., Pope, N., & Tukiainen, M. (2024). A systematic literature review of the empirical research on ChatGPT in education. *Discover Education*, https://doi.org/10.1007/s44217-024-00138-2

Ali, D., Fatemi, Y., Boskabadi, E., Nikfar, M., Ugwuoke, J., & Ali, H. (2024). ChatGPT in Teaching and Learning: A Systematic Review. *Education Sciences*, *14*(6), 643. https://doi.org/10.3390/educsci14060643

Chakraborty, D., Kar, A. K., Patre, S., & Gupta, S. (2024). Enhance online grocery shopping trust through generative AI chatbots. *Journal of Business Research*, https://doi.org/10.1016/j.jbusres.2024.114737

Chan, C. K. Y., & Hu, W. (2023). Students' voices on generative AI: Perceptions, benefits, and challenges in higher education. *International Journal of Educational Technology in Higher Education*, *20*(1), doi:10.1186/s41239-023-00411-8

Chaudhry, S., & Kazim, H. (2024). AI-driven learning: Enhancing student engagement and critical thinking. *Journal of Educational Technology*, *35*(2), 98-112. https://doi.org/10.1016/j.jet.2024.02.005

Clark, T., Johnson, R., & Wang, M. (2024). Generative AI in learning environments: A framework for ethical use. *Computers and Education*, *180*, 104432. https://doi.org/10.1016/j.compedu.2024.104432

Eysenbach, G. (2023). The Role of ChatGPT, Generative Language Models, and Artificial Intelligence in Medical Education: A Conversation With ChatGPT and a Call for Papers. *JMIR Medical Education*, *9*(1), e46885. https://doi.org/10.2196/46885

Fisher, R., & Haake, J. (2024). The ethics of AI integration in higher education: Emerging trends and concerns. *AI & Society*, *39*(3), 567-580. https://doi.org/10.1007/s00146-023-01567-9

Henke, J. (2025). New normal: The increasing adoption of generative AI in university communication departments. *Journal of Science Communication*, *24*(2), A07. https://doi.org/10.22323/2.24020207

Jo, H. (2024). From concerns to benefits: a comprehensive study of ChatGPT usage in education. International *Journal of Educational Technology in Higher Education*, https://doi.org/10.1186/s41239-024-00471-4

Johnston, B., Strachan, R., & Mangat, M. (2024). Student perspectives on the use of generative artificial intelligence technologies in higher education. *International Journal for Educational Integrity*, *20*(1), Article 149, doi:10.1007/s40979-024-00149-4

Khlaif, Z. N., Ayyoub, A., Hamamra, B., Bensalem, E., Mitwally, M. A. A., Hattab, M. K., & Shadid, F. (2024). University teachers' views on the adoption and integration of generative AI tools for student assessment in higher education. *Education Sciences*, *14*(10), 1090. https://doi.org/10.3390/educsci14101090

Kirk, C. P., & Givi, J. (2025). The AI-authorship effect: Understanding authenticity, moral disgust, and consumer responses to AI-generated marketing communications. *Journal of Business Research*, https://doi.org/10.1016/j.jbusres.2024.114984

Kumar, A., Shankar, A., Hollebeek, L. D., Behl, A., & Lim, W. M. (2025). Generative artificial intelligence (GenAI) revolution: A deep dive into GenAI adoption. *Journal of Business Research*, https://doi.org/10.1016/j.jbusres.2024.115160

Michel-Villarreal, R., Vilalta-Perdomo, E., Salinas-Navarro, D. E., Thierry-Aguilera, R., & Gerardou, F. S. (2023). Challenges and Opportunities of Generative AI in Higher Education as Explained by ChatGPT. *Education Sciences*, *13*(9), 856. https://doi.org/10.3390/educsci13090856

Moorhouse, B. L., Kohnke, L., & Zou, D. (2023). ChatGPT in education: Opportunities, challenges, and future directions. *Educational Technology Research and Development*, *71*(1), 85-99. https://doi.org/10.1007/s11423-023-10109-8

Montenegro-Rueda, M., Fernández-Cerero, J., Fernández-Batanero, J. M., & López-Meneses, E. (2023). Impact of the Implementation of ChatGPT in Education: A Systematic Review. *Computers*, *182*, 104463. https://doi.org/10.1016/j.compedu.2023.104463

Ogunleye, B., Zakariyyah, K. I., Ajao, O., Olayinka, O., & Sharma, H. (2024). A Systematic Review of Generative AI for Teaching and Learning Practice. *Education Sciences*, *14*(6), 636. https://doi.org/10.3390/educsci14060636

Ravšelj, D., Keržič, D., Tomaževič, N., Umek, L., & Brezovar, N. (2025). Higher education students' perceptions of ChatGPT: A global study of early reactions. *PLOS ONE*, *20*(3), e0267890. https://doi.org/10.1371/journal.pone.0267890

Rosenblatt, F. (1958). The Perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*(6), 386–408. https://doi.org/10.1037/h0042519

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533–536. https://doi.org/10.1038/323533a0

Russell, S., & Norvig, P. (2021). Artificial intelligence: A Modern Approach* (4th ed.). Pearson.

Saaida, M. B. (2023). AI-Driven transformations in higher education: Opportunities and challenges. *International Journal of Educational Research and Studies*, 29-36, DOI 10.5281/zenodo.8164414

Sandu, R., Gide, E., & Elkhodr, M. (2024). The role and impact of ChatGPT in educational practices: insights from an Australian higher education case study. *Discover Education*, *2*(1), 14. https://doi.org/10.1007/s44217-024-00014-7

Sergeeva, O. V., M. R. Zheltukhina, T. Shoustikova, L. R. Tukhvatullina, D. A. Dobrokhotov, & S. V. Kondrashev. (2025). Understanding higher education students' adoption of generative AI technologies: An empirical investigation using UTAUT2. *Contemporary Educational Technology*, *17*(2), ep. 571. https://doi.org/10.30935/cedtech/16039

Shoufan, A. (2023). Students' perception of AI-generated content: A critical review. *International Journal of Educational Technology in Higher Education*, *20*(1), 12. https://doi.org/10.1186/s41239-023-00358-9

Song, Y., & Wang, M. (2024). A Survey and Research on the Use of Artificial Intelligence by Chinese Design- College Students. *Buildings*, *62*(3), 567–589, doi:10.3390/buildings14092957

Sun, H., & Li, P. (2024). AI, academic integrity, and pedagogical shifts: The new landscape of higher education. *Journal of Applied Learning and Teaching*, *5*(3), 45-60. https://doi.org/10.37074/jalt.2024.5.3.9

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, *59*(236), 433–460. https://doi.org/10.1093/mind/LIX.236.433

Xames, M. D., & Shefa, J. (2023). ChatGPT for research and publication: Opportunities and challenges. *Journal of Applied Learning and Teaching*, 390-395, 4(2), 123-134. https://doi.org/10.37074/jalt.2023.4.2.7

Zhai, X. (2023). Exploring the role of AI in personalized learning: A systematic review. *International Journal of Artificial Intelligence in Education*, *33*(1), 1-27. https://doi.org/10.1007/s40593-023-00291-4

İSTANBUL UNIVERSITY PRESS

Acta Infologica,
ACIN 2025, 9 (1): 167–182

https://doi.org/10.26650/acin.1566305
Submitted: 13.10.2024
Revision Requested: 02.03.2025
Last Revision Received: 11.05.2025
Accepted: 14.05.2025
Published Online 15.05.2025

Acta Infologica

Research Article

Open Access

An Investigation into the Impact of Influencers Producing Technological Content on Instagram on Customers' Purchasing Decisions

Kamil Çelik [1] & Ahmet Taş [2]

[1] Bartın University, Faculty of Economics and Administrative Sciences, Department of Management Information Systems, Bartın, Türkiye
[2] Gümüşhane University, Siran Dursun Keleş Health Services Vocational School, Department of Pharmacy Services, Gümüşhane, Türkiye

Abstract Nowadays, social media influencers play a significant role in the world of marketing. Social media influencers reach large audiences by producing content on social media platforms, and they exert considerable impact on these audiences. The literature suggests that social media users form parasocial relationships, and these relationships develop based on the influencer's characteristics, such as attractiveness, expertise, credibility, and similarity. It has also been proven in various studies that social media influencers have a strong effect on increasing the purchasing intentions of their followers. In this study, the effects of social media influencers on parasocial relationships and the purchasing intentions of their followers were examined. The research was conducted among Instagram users in Turkey, and the data were collected using a survey method. The survey included statements aimed at measuring elements such as parasocial relationships, attractiveness, expertise, credibility, and similarity. Because of the data analysis, it was found that characteristics such as influencers' attractiveness and expertise strengthen parasocial relationships, and these relationships positively influence followers' purchasing intentions. Additionally, the perception of fairness also had a significant impact on these relationships. These findings suggest that social media influencers can be an effective tool in marketing strategies.

Keywords Social Media Influencers · Purchase Intention · Social Media Communication

Citation: Çelik, K. & Taş, A. (2025). An investigation into the impact of influencers producing technological content on instagram on customers' purchasing decisions. *Acta Infologica*, *9*(1), 167-182. https://doi.org/10.26650/acin.1566305
This work is licensed under Creative Commons Attribution-NonCommercial 4.0 International License.
© 2025. Çelik, K. & Taş, A.
Corresponding author: Ahmet Taş ahmettas@gumushane.edu.tr

Introduction

Thanks to advances in information and communication technologies, "ordinary people" can become "social media influencers" by creating and sharing content on social media. A social media influencer is someone who has a loyal social following and exerts social influence over such followers (Yuan & Lou, 2019). In today's marketing world, influencer marketing has rapidly gained importance as a powerful strategy (Bastrygina & Lim, 2023). Brands are increasingly investing in this area because they recognize its vast potential (Koay et al., 2023).

Popularity has a significant impact on marketing, and this effect can be enhanced by influencer-product compatibility (Çelik & Taş, 2021; Kim & Kim, 2021). Many studies have highlighted the role of parasocial relationships, which are emotionally-based connections formed with social media influencers (Vrontis et al., 2021). Internet users with large numbers of followers on social media are often seen as well-known micro-influencers whose behaviors and attitudes do not influence large-scale attitudes. Whether online or offline, these individuals are trusted, and followers frequently seek their advice, underscoring the importance of social media influencers. People also view them as trendsetters in one or more fields and are particularly influential over individuals who share specific interests (De Veirman et al., 2017). In many cases, this influence triggers consumers' desire to emulate them by narrating their personal lives, lifestyles, and choices through user-generated content, both textually and visually, or through electronic word-of-mouth communication (Ki & Kim, 2019). In addition to product promotion, influencers also endorse certain lifestyles and behaviors. In this regard, similar to traditional media, influencers can shape their followers' attitudes and behaviors beyond purchasing decisions. However, studies examining the role of social media influencers in promoting positive behavioral changes outside purchasing decisions are rare (Casaló et al., 2020).

The importance of influencer marketing continues to grow, making it an indispensable component of integrated marketing communications (Childers et al., 2019). Brand managers increasingly rely on social media influencers, also known as "micro-celebrities," to promote their products. According to one study, approximately 49% of followers trust an influencer's recommendations, and approximately 40% purchase products endorsed by social media influencers (Delbaere et al., 2021). The widespread use of social-media platforms has enabled certain accounts to reach large audiences, turning these accounts into social-media influencers. Influencers have attracted the attention of marketing researchers due to their ability to influence their followers' purchasing intentions (Cabeza-Ramírez et al., 2022). The mechanism by which followers develop loyalty to social media influencers remains unclear. Loyalty to social media influencers differs from loyalty to websites, and these differences stem from the essence of various interpersonal interactions. According to social contagion theory, emotions are highly contagious among people. Interactions between followers and social media influencers can trigger psychological responses in followers (Yuan & Dennis, 2019).

Parasocial relationships are derived from parasocial interaction theory and consider the connection between individuals and influencers in terms of self-compatibility, language and interest similarity, interaction frequency, and the perception of friendship, in relation to physical attractiveness and credibility. Parasocial relationships are supported by shared content that fosters perceptions of intimacy, access to influencers' personal lives, and the presence of a dialog, thus validating the idea that parasocial relationships in the context of social networks can be seen as a form of advertising (Lueck, 2015). Social media provides influencers with the opportunity to build parasocial relationships by engaging in two-way communication with followers, sharing authentic and valuable content, responding to followers' questions, and offering

giveaways to encourage interaction and engagement. However, this relationship is not perfect, as it is largely controlled by the social media influencers themselves (Lou, 2022).

Parasocial interaction is a process that forms through viewers' psychological responses to media figures. Over time, this interaction leads to the development of parasocial relationships. The one-way relationships established between media figures and viewers make viewers feel as though they are in a close relationship with the characters. Viewers develop bonds with characters similar to the one they form with their friends and are influenced by their behavior. Social media influencers can also unknowingly initiate parasocial interaction processes through the bonds they establish with their followers (Akdeniz & Uyar, 2021). The concept of parasocial interaction was first used in the 1950s to describe one-way communication between media audiences and media figures. However, the development of mass communication tools over time has required a reevaluation of the interaction between media audiences and media figures. In particular, media figures now acknowledge and respond to viewers' emotions and reactions through social media. This interaction is related to how viewers perceive media figures and can result in internal or behavioral responses. With the innovations brought by technology, the dimension of parasocial interaction has changed, and individuals with high social anxiety try to resolve this interaction with imaginary media characters. A close relationship between parasocial interaction and psychological health has also been identified (Keskin, 2020).

The marketing opportunities provided by Instagram have enabled businesses operating in the technology sector to establish a presence on this platform (Çelik & Bora, 2021). Companies have exploited the opportunity to promote their products by collaborating with influencers to create technology-related content. This study investigates the impact of the following technology-focused influencers on users' intentions to purchase such products. Thus, this research aims to contribute to both theoretical and practical knowledge in the literature. Additionally, it seeks to inform businesses about the variables that they should focus on when developing their Instagram marketing strategies.

The study explored and evaluate the influence of Instagram influencers who produce technological content on customers' purchasing decisions. Specifically, this study aims to understand how the credibility, engagement, and perceived expertise of these influencers affect consumer behavior, brand perception, and product selection in the technological sector. By examining the relationship between influencer content and customer decision-making, this research aims to identify key factors that drive consumer trust and the role that influencer marketing plays in shaping purchase intentions within a digital platform. Additionally, the study provides insights into how companies in the technology industry can collaborate strategically with influencers to enhance their marketing effectiveness and drive sales.

Literature Review

Current studies on social media influencer marketing have focused on how viewers perceive the alignment between their own image and the perceived image of the influencer (personality-based self-image congruence) and how the influencer's image aligns with the image of the brand or destination that they endorse (brand-based congruence) (Xu & Pratt, 2018). When a social media influencer follows the personal brand sales model, they generate income by directly selling products to their followers. Influencers attract audiences by showcasing their personal tastes on social media (McQuarrie, Miller, & Phillips, 2013), and if followers appreciate the esthetic taste of an influencer, that influencer gains leadership in taste and

is considered a consumption role model (Ki & Kim, 2019). After acquiring the power to influence target consumers, influencers profit by selling personalized products that align with their image to their followers.

Evidence shows that the persuasive power of social media influencers (SMIs) is a significant factor and argument quality should be seen as just as important as source credibility in predicting the effectiveness of SMI marketing campaigns. In the context of social commerce platforms, users are more likely to engage with other users who are professionals in a particular field. As experts, social media influencers create communities around themselves in niche markets and become specialty experts within these targeted communities (Xiang et al., 2016). Research conducted on a social-media platform revealed that users prefer to use the application for entertainment and to gain knowledge about their interests, and they find the influencers they follow convincing. A study on mom influencers found that these influencers guide brand preferences and influence followers' purchasing behavior. Furthermore, sincerity was identified as the top trait followers seek in influencers (Akdeniz & Uyar, 2021).

If customers spend an unusually long time on a preferred site, their likelihood of purchasing increases, as does their exposure to advertisements (Koh, Kim, & Kim, 2014). Existing literature reveals that psychological and personal factors, such as social norms, self-identity, social marketing, environmental concern, intrinsic awareness, consumer value, and emotional intelligence, influence sustainable attitudes and consumption (Dhandra, 2019). Changing public consumption behavior is difficult because many other situational and psychological factors also come into play (Joshi & Rahman, 2019). Environmental factors on social media (e.g., parasocial relationships, celebrity endorsements, product advertisements, bundle deals, online reviews, and time pressure) trigger users' desire to purchase (Chen et al., 2016).

Social networks such as Twitter and Facebook facilitate increased parasocial interaction between media audiences and media figures today. When popular individuals/groups or celebrities respond to their followers on social media, this strengthens the parasocial interaction/relationship between the media audience and the media figure. One study pointed out that the positive direction of the relationship between fans and celebrities on Twitter, where reciprocal interaction is prominent, strengthens parasocial relationships (Pennington et al., 2016: 69). Another study found that when celebrities share various posts on Twitter, their fans perceive celebrities as social beings and feel as though these social beings are part of their own lives. Recognizing celebrities as social entities increases the positive direction of parasocial interaction. Fans tend to develop interpersonal relationships with celebrities as if they were friends in real life (Kim & Song, 2016. p. 574).

Studies have revealed that admiration for media figures may stem from psychological needs that are lacking in social life (Çelik, 2021; Batıgün & Sunal, 2017). The transformation of sociability, a fundamental topic in sociology, is revealed through the changes in the scope of interaction that individuals experience as they transition from social life to social networks. Since the 2000s, the increasing use of the internet and the resulting changes and transformations in social relationships have shown that the definition of sociability has begun to evolve with digitalization (Keskin, 2020).

Because consumers can communicate directly with people on social media and evaluate their lives as though they are friends, social media is an excellent platform for developing meaningful and intimate social relationships between media celebrities and audiences (Chung & Cho, 2017). In their study of luxury brands' video blogs on YouTube, it was found that the social attractiveness and physical attractiveness of video bloggers increased parasocial interaction (Lee & Watkins, 2016). They also found that the antecedents of parasocial interaction consist of physical attractiveness, social attractiveness, motivation for entertainment,

and motivation for relationship building (Liu et al., 2019). Parasocial interaction between video bloggers and their viewers helps viewers develop positive attitudes and behaviors (Lee & Watkins, 2016).

Conseptual Model

Studies in the literature have revealed that the time social media users spend on social media and their personality traits impact the degree of parasocial relationships. Attractiveness refers to the appearance and physical beauty of the source in terms of facial features and other physical characteristics (Ismagilova et al., 2020). Many studies have concluded that communication effectiveness is achieved through factors such as attractiveness, expertise, similarity, and credibility, and it is predicted that well-known individuals on social media will establish parasocial interaction with their followers (Munnukka et al., 2016; Yuan & Lou, 2020).

Viewers can engage in parasocial interaction with a character in a program or TV series that impacts them. The character's charm and attractiveness are important factors in this interaction (Keskin, 2020). The concept of social attractiveness, which refers to having positive feelings toward the character with whom one interacts socially, is considered a strong element in many studies in the field of social media, particularly in examining individuals' behaviors (Zheng et al., 2019). Expertise is related to the knowledge that well-known individuals on social media have about the products they promote and the domain of these products (Xiao et al., 2018). Expertise is an important factor in encouraging parasocial communication among followers of well-known social media figures (Aw & Chuah, 2021). The expertise of the source from which the message is received has always been considered an important element from the consumer's perspective. Expertise is also closely related to credibility, which is defined as the degree of credibility. Studies in the literature have shown that expertise has a positive impact on both attitude and purchase intention (Akdeniz & Uyar, 2021). Credibility is related to the honesty and accuracy of a source when conveying a message (Jin et al., 2019). Studies on the characteristics of media personalities suggest that their credibility is a crucial factor in determining the persuasiveness of their messages (Ecker & Antonio, 2021). The more trust the target audience has in the media figure, the stronger their parasocial interaction. Another key component of the relationship between the media figure and the audience, similar to interpersonal interaction, is the extent to which the audience can form judgments about the person with whom they are interacting (Aytulun & Sunal, 2020). Similarity is defined as the common qualities between well-known individuals on social media and their target audience (Schiappa et al., 2007). The concept of similarity is concerned with shared values, education, social status, beliefs, and attitudes between the source and receiver. The perceived similarity of users with a media character positively influences parasocial interaction (Akdeniz & Uyar, 2021). Similarity enhances the credibility of influencers with their target audience and increases the persuasiveness of messages (the Yuan & Lou, 2020). Based on this information, the following hypotheses are formulated:

- H1: Attractiveness has a significant and positive effect on parasocial relationships.
- H2: Expertise has a significant and positive effect on parasocial relationships.
- H3: Credibility has a significant and positive effect on parasocial relationships.
- H4: Similarity has a significant and positive effect on parasocial relationships.

Justice concerns what and how people perceive fairness. This perception can affect attitudes and behaviors (Dahanayake et al., 2018). Many studies have mentioned that justice is an important factor in increasing the effectiveness of online brands (Hernández Soto et al., 2021). Distributive justice refers to the fair distribution of rewards and resources (Uwa, 2022). Interpersonal justice refers to being treated with respect and dignity (Brockner et al., 2021). Interpersonal justice has been identified as a key precursor to parasocial

interactions with well-known individuals on social media (the Yuan & Lou, 2020). Procedural fairness is concerned with the fairness of the overall decision-making process (Choi et al., 2022). Justice strengthens personal bonds among members of online brand communities and creates a positive perception of the brand (Usoro & Majewski, 2011). Informational justice is concerned with how well information is communicated (Brockner et al., 2021). Justice fosters participation among members of online brand communities and leads to deeper brand understanding (Majewski et al., 2011). Considering this information, the following hypotheses are formulated:

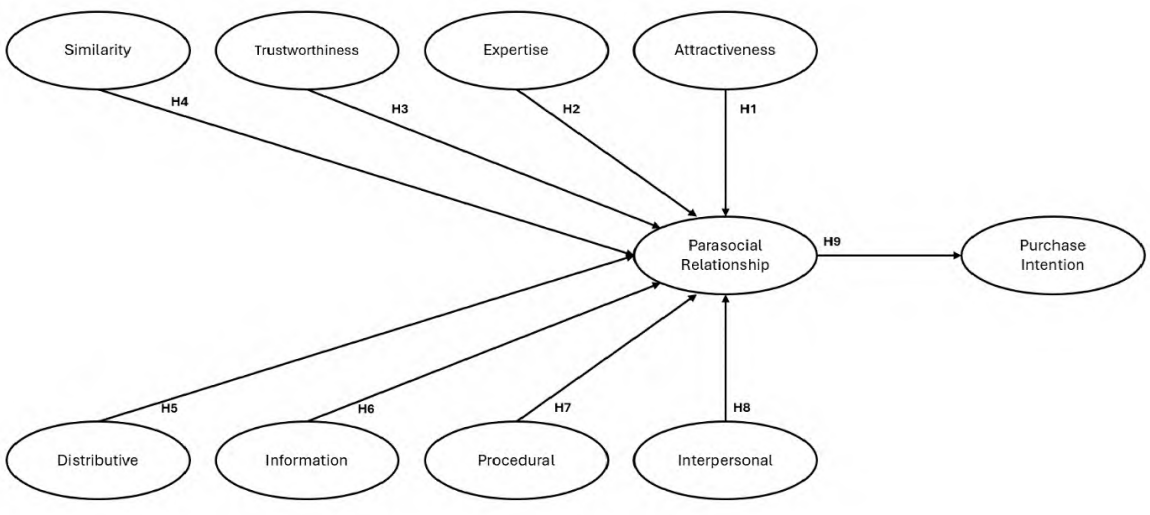
- H5: Distributive justice has significant and positive effects on parasocial relationships.
- H6: Interpersonal justice has significant and positive effects on parasocial relationships.
- H7: Procedural fairness has a significant and positive effect on parasocial relationships.
- H8: Informational justice has a significant and positive effect on parasocial relationships.

Marketing companies have been interested in whether the money spent on social media influencers yields results and whether influencers increase their followers' purchase intentions. Marketing-related outcomes are important indicators for measuring the success of marketing efforts (Leite & Baptista, 2021). Social media influencers are different from celebrities because their communication with followers is stronger (Lou & Kim, 2019). Numerous studies have been conducted on marketing activities conducted through social media (De Veirman et al., 2017; Hughes et al., 2019; Lou & Kim, 2019). There are also studies on the parasocial relationships of social media influencers and their impact on followers' purchase intentions (Farivar et al., 2021). Accordingly, the following hypothesis is formulated:

- H9: Parasocial relationships have a significant and positive effect on purchase intention.

Figure 1

The Conceptual Model for the Study



Research Methodology

Samples

The empirical part of this study was conducted through a survey in Turkey. An online survey was developed using Google Forms to collect the study data. The survey link was distributed through various channels

such as Instagram, WhatsApp, and email. While determining the sample, specific eligibility criteria were considered, and selection was made in line with demographic targets. Participants were deemed eligible if they used Instagram, and those who followed an influencer were included in the analysis. Individuals who did not follow social media influencers were excluded from the study. Individuals aged 18 years and older were randomly included in the study. The necessary data were gathered from a sample of 765 Instagram users who follow influencers who produce technology-related content on Instagram. To assess the participants' eligibility, a screening question was used to determine whether they were interested in technological products and whether they followed any Instagram influencers who share content related to such products. The first question of the survey was, "Do you follow a social media influencer on Instagram who posts about technological products?" Data from 32 participants who answered "No" to this question were excluded from the analysis. Consequently, data from 733 Instagram users were included in the analysis. Since participants who did not follow influencers producing technology-related content were excluded from the study, our sampling technique can be classified as purposive sampling (Cash et al., 2022). Ethical permission was obtained for this study at the meeting of Gümüşhane University Scientific Publication and Ethics Board dated 20/09/2024 and numbered 2024/7.

Because the study was conducted in Turkey, the scales used in the study were adapted to Turkish. To minimize the effects of cultural and linguistic differences, the survey was translated into Turkish, the official language of Turkey, using Brislin's (1976) back-translation method. Additionally, the Turkish version of the survey was reviewed by two experts specializing in Management Information Systems and Marketing. Furthermore, a pilot study was conducted with 32 undergraduate students at Bartın University, and most participants indicated that the language used in the survey was clear and understandable.

Measures

The survey consists of three sections. The first section provides information about the purpose of the survey and its conduct. Participants were informed that the data obtained from the study would be used for academic purposes and that participation was voluntary. Additionally, information was provided regarding the numerous technology-related accounts currently available on social media platforms.

The second section of the survey included questions that sought to obtain demographic information about the participants. In the third section, statements related to the scales of attractiveness, expertise, credibility, similarity, distributive justice, information justice, procedural justice, interpersonal justice, parasocial relationship, and purchase intention were included. All items were measured using a five-point Likert scale, ranging from 1 (Strongly Disagree) to 5 (Strongly Agree).

The scales for attractiveness, expertise, credibility, and similarity were adapted from Munnukka et al. (2016), while the scales for distributive justice, information justice, procedural justice, and interpersonal justice were obtained from Yuan and Lou (2020). The parasocial relationship scale was sourced from the study by Yuan and Lou (2020). The purchase intention scale was derived from Liu and Brock (2011).

Data Analysis

Because R is an open-source and free program, it was used for data analysis (Çelik et al., 2018). Initially, Confirmatory Factor Analysis (CFA) was conducted in R. Following this, Structural Equation Modeling (SEM) analysis was performed. The validity and reliability analysis of the data was conducted using R.

Results

The results section begins by providing a preliminary analysis of the demographic statistics. This section presents the two-step structural equation modeling (SEM) adopted to validate the conceptual model and test the related hypotheses. The results of the confirmatory factor analysis (CFA) in the first phase will be presented in Section 4.2, while Section 4.3 will present the results of the structural model, which is the second phase of the SEM.

Table 1

Demographic Characteristics of Individuals Participating in The Study

Characteristic	Measurements	Quantity	(%) Percentage
Gender	Male	325	45
	Female	412	55
Age	17-19	143	20
	19-21	189	26
	21-23	206	28
	23 and more	195	26
Monthly income (minimum wage)	Less than 1 minimum wage	519	71
	Between 1 and 2 minimum wages	123	17
	More than 2 minimum wages	91	12
How many years have you been using Instagram?	Less than 3 years	62	9
	3-6 years	292	40
	7-10 years	278	38
	11 years and more	101	13
How many hours a day do you spend on Instagram on average?	Less than 1 hour	107	15
	1-3 hours	359	49
	4-6 hours	197	27
	7 hours or more	70	9
How many social media influencers do you follow who share posts about technological products on Instagram?	1	214	29
	2-5	356	49
	6-9	125	17
	10 and more	38	5
TOTAL		733	100

As seen in the Table 1, more than half of the 733 participants in the survey were women (55%). The majority of the participants (28%) were in the 21-23 age group, 26% were in the 19-21 age group, another 26% were aged 23 and older, and 20% were in the 17-19 age group. When examining the income level distribution, nearly all participants (71%) reported having a monthly income of less than one minimum wage. Participants were asked how many years they have been using Instagram. 40% reported using Instagram for 3-6 years, 38% for 7-10 years, 13% for more than 11 years, and 9% for less than 3 years. 49% of participants use Instagram for 1-3 hours a day, while 27% spend 4-6 hours on Instagram. Participants were also asked how many social media influencers they followed who shared content related to technological products. 49% of participants reported following between 2 and 5 influencers, 29% followed 1 influencer, 17% followed between 6 and 9 influencers, and 5% followed 10 or more.

Confirmatory Factor Analysis (CFA)

In this section, confirmatory factor analysis (CFA) was conducted to examine the model fit indices. For this purpose, fit indices such as CMIN/DF (Normed Chi-Square), CFI (Comparative Fit Index), RMSEA (Root Mean Square Error of Approximation), RNI (Relative Non-Centrality), and IFI (Incremental Fit Index). As shown in the table, all goodness-of-fit indices indicated an acceptable level of fit. This suggests that the data obtained are consistent with the proposed research model.

Table 2

Results of Confirmatory Factor Analysis to Test Hypotheses' Validity

Goodness of fit values	Good Fit	Acceptable Compliance	Value	Conclusion
CMIN/DF	<3	<5	2895.760/657=4.4	Acceptable fit
CFI	>0.95	>0.90	0.901	Acceptable fit
RMSEA	<0.050	<0.080	0.068	Acceptable fit
RNI	>0.95	>0.90	0.901	Acceptable fit
IFI	>0.95	>0.90	0.901	Acceptable fit

According to Table 2, all goodness-of-fit indices were within acceptable fit levels. The results indicate that the model exhibits a satisfactory fit to the data. Therefore, the fit indices confirmed the adequacy of the proposed measurement model.

Validity and Reliability Analysis

Table 3

Correlation matrix and Validity and Reliability analysis

	α	AVE	CR	1	2	3	4	5	6	7	8	9	10
1-AT	.80	.53	.81	(.72)									
2-EX	.91	.72	.91	.47*	(.84)								
3-TR	.91	.74	.92	.48*	.76*	(.86)							
4-SM	.85	.66	.85	.52*	.51*	.63*	(.81)						
5-DI	.87	.70	.87	.46*	.70*	.75*	.65*	(.83)					
6-INT	.87	.69	.86	.47*	.57*	.67*	.66*	.69*	(.83)				
7-PRO	.80	.57	.80	.41*	.51*	.55*	.61*	.58*	.64*	(.75)			
8-INF	.86	.68	.86	.44*	.70*	.77*	.62*	.78*	.68*	.63*	(.82)		
9-PR	.89	.58	.89	.51*	.61*	.72*	.67*	.70*	.69*	.62*	.75*	(.76)	
10-PI	.89	.60	.90	.51*	.50*	.56*	.63*	.59*	.62*	.58*	.59*	.70*	(.77)

α: Cronbach's Alpha Confidence Coefficient; CR: Construct Reliability; AVE: Average Variance Extracted; Values in parentheses are the square roof of AVE values; AT: Attractiveness; EX: Expertise; TR: Credibility; SM: Similarity; DI: Distributive justice; INT: Interpersonal justice; PRO: Procedural fairness; INF: Informational justice; PR: Parasocial relationships; PI: Purchase intention, *: p<0.01

Table 3 presents the correlation matrix for each scale along with the Average Variance Extracted (AVE), Composite Reliability (CR), and Cronbach's Alpha Reliability Coefficient. The results indicate that Cronbach's Alpha coefficients for all scales ranges between 0.80 and 0.91. To ensure the reliability of the scales, Cronbach's Alpha should exceed 0.7 (Iacobucci & Duhachek, 2003; Nunnally, 1978). The findings suggest that the scales exhibit a high level of reliability.

CR should be higher than AVE, and for convergent validity, AVE should be greater than 0.5 (Hair et al., 2010; Fornell & Larcker, 1981). Table 3 shows that all AVE values exceeded 0.5, confirming the validity of the constructs. Additionally, the CR values for all scales surpass the AVE values, further supporting the validity of the scales. Discriminant validity was also examined. The square root of the AVE should be greater than the correlation coefficients between the respective variables and other variables to establish discriminant validity (Hair et al., 2010). Table 3 demonstrates that for each scale, the square root of the AVE is higher than the inter-component correlation.

A correlation analysis was conducted to examine how the variables are related to each other. The results indicate that all variables exhibit a significant (p<.01) positive relationship. Among the constructs, informational justice and distributive justice showed the highest correlation (r=0.78, p<.01), whereas attractiveness and procedural fairness exhibited the lowest correlation (r=0.41, p<.01).

Following the recommendation of Podsakoff et al. (2023), Harman's single-factor test was employed to assess common method variance (CMV). Accordingly, an unrotated exploratory factor analysis was performed for all variables. As indicated by the single-factor solution explaining 47% of the variance, no dominant component was present in the study, which is below the 50% threshold. Based on the results of Harman's single-factor test, common method variance was not a concern.

Structural Model

Table 3 presents the results of the structural equation model analysis, including the standardized β values, p-values, and the status of hypothesis support or rejection. Accordingly, H1, H4, H6, H8, and H9 were supported, whereas H2, H3, H5, and H7 were rejected.

The analysis revealed that the R^2 value for the parasocial relationship variable was 0.829. This indicates that the variables used in the study explain 83% of the parasocial relationships. Additionally, it was found that the purchase intention variable is explained by the related variable at a rate of 62%. The 83% explanation of the parasocial relationship variable indicates that a strong model has been established, whereas the 62% explanation of purchase intention suggests that consumer behavior is also influenced by other variables.

Table 4

Acceptance and Rejection Distributions of The Hypotheses

Hypotheses	Std. BETA	*p*	Accept/Reject
H1	0.112	.001	Accept
H2	−0.041	.427	Reject
H3	0.073	.299	Reject
H4	0.209	.000	Accept
H5	−0.095	.297	Reject
H6	0.145	.008	Accept
H7	0.056	.319	Reject
H8	0.562	.000	Accept
H9	0.786	.000	Accept

Because of the structural equation model analysis, H1, H4, H6, H8, and H9 were accepted, whereas H2, H3, H5, and H7 were rejected.

The analysis revealed that the R^2 value for the parasocial relationship variable was 0.829. This indicates that the variables used in the study explain 83% of the parasocial relationships. Additionally, it was found that the purchase intention variable is explained by the related variable at a rate of 62%.

Discussion and Conclusion

Based on the results of the structural equation model analysis, variables influencing parasocial relationships and purchase intentions were evaluated. The analysis supported H1, H4, H6, H8, and H9, whereas H2, H3, H5, and H7 were rejected. These findings indicate that some variables have a statistically significant effect on parasocial relationships and purchase intentions, whereas others do not have the expected impact.

The R^2 value obtained for the parasocial relationship variable was 0.829, indicating that the variables used explained 83% of the parasocial relationship. This finding suggests that the parasocial relationships were largely explained by the variables included in the model. Furthermore, the fact that the purchase intention variable is explained at a rate of 62% indicates that the factors influencing this intention are adequately represented within the model.

The findings from the structural equation model (SEM) analysis reveal significant results regarding parasocial relationships and purchase intentions. Acceptance of some hypotheses indicates that certain variables have a meaningful impact on these relationships, whereas rejection of other hypotheses suggests that those variables do not have the expected effect. This indicates that parasocial relationships and purchase intentions are complex constructs, and understanding them fully requires considering more factors and variables (Brown, 2019; Smith, 2020).

Studies on parasocial relationships have revealed that these relationships are defined as one-sided social bonds that individuals form with media figures (Horton & Wohl, 1956). The R^2 value of 83% in this study suggests that parasocial relationships can be largely explained. This high percentage confirms that influencers, especially those producing technology-focused content on social media platforms, can establish strong parasocial relationships with their followers, and these relationships can significantly impact consumer behavior (Gil de Zúñiga et al., 2018; Lee & Watkins, 2016).

Moreover, the fact that purchase intentions are explained by 62% of respondents indicates that the factors influencing these intentions are largely well-represented. In the marketing of technology products, the role of influencers emerges as a key factor that influences purchase decisions (Freberg et al., 2011). However, the rejection of some hypotheses suggests that not all variables have the expected effect, and further research is needed on this topic (Park & Lee, 2021). For instance, factors such as social interaction and content quality may have a different impact on parasocial relationships than expected, and these aspects should be examined more deeply in future studies. Rejection of H2 and H3; the lack of the expected impact of the "expertise" and "credibility" variables on parasocial relationships may be due to users perceiving technology influencers not as sources of information but rather as sources of entertainment and inspiration. While previous research (e.g., Yuan & Lou, 2020) has emphasized the effect of expertise, the characteristics of the sample group in this study (such as age and duration of social media use) may have led to different results. Rejection of H5 and H7; the lack of an effect of distributive justice and procedural justice on parasocial relationships suggests that users place less importance on perceptions of justice in their relationships with influencers. Individual consumer experiences and influencers' marketing strategies may particularly influence this outcome.

In conclusion, this study makes a significant contribution to the literature on understanding the effects of social media influencers on parasocial relationships and purchase intentions. However, the rejection of some hypotheses indicates that more variables and factors need to be considered to fully comprehend influencer effects (Smith, 2020). Future research focusing on different social media platforms and product categories may help to better understand these relationships in a broader context.

In conclusion, the accepted hypotheses in the structural equation model analysis provide significant contributions to understanding parasocial relationships and purchase intentions, whereas the rejected hypotheses suggest the need for further examination of the model with different variables. The findings of this study offer a solid foundation for future research into consumer behavior and parasocial relationships.

Theoretical contributions and practical implications

Given the purpose of this study, important relationships among the variables were identified as a result of the analyses. According to these results, the existence of individuals' parasocial interactions and purchase decisions was confirmed. From this perspective, the ease of access provided by social media influences people's behaviors. Overall, this study contributes to theoretical and practical knowledge in the literature and complements the results of other studies. Additionally, it can guide businesses, individuals, and potential entrepreneurs toward developing better marketing strategies.

This research theoretically extends the impact of social media influencers on individual purchase intentions. According to the results, the standardized regression weights for all statements related to the structures are above 0.50 and are consistent with the values suggested by Hair et al. (2010). This study contributes to a new approach that emphasizes the importance of information shared by social media influencers with their followers. The Average Variance Extracted (AVE) values for all structures were above 0.50 and were consistent with those suggested by Fornell and Larcker (1981). The findings of this study can help companies actively promoting products on social media by identifying information that aids in understanding consumer purchasing behavior and controlling the content of online reviews.

Businesses must be meticulous in selecting the right influencers to optimize their influencer marketing strategies. Factors such as the influencer's content type, alignment with the target audience, and engagement rate should be carefully evaluated. Rather than focusing solely on follower count, brands should analyze the influencer's relationship with their audience, credibility, and content style before making collaboration decisions. Micro-influencers, despite having smaller followings, can be valuable for brands due to their high engagement rates and strong credibility among their audience.

The content strategy should also be well-planned in influencer collaborations. It is crucial to present advertising content organically to capture followers' interest. Rather than seeing direct promotional messages, followers tend to find influencer-led, experience-based product introductions more trustworthy. Therefore, storytelling and user experience-driven content can generate higher engagement. Additionally, actively using video formats such as Instagram Reels, TikTok, and YouTube Shorts can enhance audience interaction with the content. Campaigns that encourage user-generated content can further expand a brand's reach and engagement levels.

Building marketing strategies that build trust and credibility is of utmost importance. Because transparency directly impacts consumer trust in a brand, influencers should explicitly disclose their sponsorship agreements in promotional content. When followers perceive that influencers share honest opinions about a product, their trust in the brand increases. At the same time, businesses should avoid collaborating with

influencers who have fake followers or low engagement rates. Engagement metrics and genuine audience reach should be among the top criteria when selecting influencers.

Limitations and future research

Participants in the study generally come from a specific demographic profile. This situation may limit the generalizability of the findings to the general population. For instance, a study conducted only with users from a specific age group or geographical area may struggle to represent the overall social media users. Online shopping is often subject to individuals' subjective approaches and attitudes. Assessing participants' personal experiences, expectations, and shopping habits can be challenging. Social media platforms are subject to rapidly changing features and algorithmic updates. Users are often selective when sharing information on social media platforms. Differences between actual shopping behaviors and what is shared on social media can affect the results of this study. There are various independent variables that influence online shopping experience. It may be challenging to exert complete control over these factors, which could affect the process of establishing accurate causal relationships. Understanding these limitations and being cautious regarding the mentioned constraints is crucial for interpreting the study findings.

The results indicate that parasocial interactions alter or influence individuals' purchasing behaviors. Future research could investigate different scenarios to broaden our understanding of the effects of parasocial interactions on people. Encouraging influencers to share their experiences may provide significant advantages for both large and small brands. Therefore, future research could explore how businesses using social media as a tool for promoting their products can use parasocial communication more effectively.

Ethics Committee Approval: Ethics committee approval was received for this study from the ethics committee of Gümüşhane University (Date: 20.09.2024, Number: 2024/7).
Informed Consent: Written informed consent was obtained from all participants who participated in this study.
Peer Review: Externally peer-reviewed.
Author Contributions: Conception/Design of Study- A.T.; Data Acquisition- K.Ç.; Data Analysis/Interpretation- A.T., K.Ç.; Drafting Manuscript- A.T., K.Ç.; Critical Revision of Manuscript- A.T., K.Ç.; Final Approval and Accountability- A.T., K.Ç.
Conflict of Interest: The authors have no conflict of interest to declare.
Grant Support: The authors declared that this study has received no financial support.

Author Details

Kamil Çelik
[1] Bartın University, Faculty of Economics and Administrative Sciences, Department of Management Information Systems, Bartın, Türkiye
0000-0002-4530-1048

Ahmet Taş
[2] Gümüşhane University, Siran Dursun Keleş Health Services Vocational School, Department of Pharmacy Services, Gümüşhane, Türkiye
0000-0002-1444-6068 ahmettas@gumushane.edu.tr

References

Akdeniz, P. C., & Uyar, K. (2021). Tüketicilerin satın alma niyetlerinde sosyal medya fenomenleri ile kurulan parasosyal etkileşimin rolü. *Erciyes Akademi, 35*(4), 1669-1688.

Aw, E. C.-X., & Chuah, S. H.-W. (2021). "Stop the unattainable ideal for an ordinary me!" This study fosters parasocial relationships with social media influencers: The role of self-discrepancy. *Journal of Business Research, 132*, 146–157.

Aytulun, G., & Sunal, A. B. (2020). Medya karakterleriyle kurulan parasosyal etkileşim. *Psikiyatride Güncel Yaklaşımlar, 12*(4), 494-506.

Bastrygina, T., & Lim, W. M. (2023). Foundations of consumer engagement with social media influencers. International Journal of Web Based Communities, 19(2-3), 222-242.

Batıgün, A. D., & Sunal, Y. (2017). Parasocial interaction with media characters. Current Approaches in Psychiatry, 12(4), 444–456. https://www.cappsy.org/archives/vol12/no4/cap_12_04_05_en.pdf

Brislin, R. (1976). Comparative research methodology: Cross-cultural studies. *International Journal of Psychology, 11*(3), 215–229.

Brockner, J., de Cremer, D., van Dijke, M., de Schutter, L., Holtz, B., & van Hiel, A. (2021). Factors affecting supervisors' enactment of interpersonal fairness: The interactive relationship between managers' informational fairness and supervisors' sense of power. *Journal of Organizational Behavior, 42*(6), 800–813.

Brown, S. (2019). Parasocial interaction and influencer marketing: Examining the influence of media personas on consumer behavior. *Journal of Marketing Research, 56*(4), 145-159.

Cabeza-Ramírez, L. J., Sánchez-Cañizares, S. M., Santos-Roldán, L. M., & Fuentes-García, F. J. (2022). Impact of the perceived risk in influencers' product recommendations on their followers' purchase attitudes and intention. *Technological Forecasting and Social Change*, *184*, 121997.

Casaló, L. V., Flavián, C., & Ibáñez-Sánchez, S. (2020). Influencers on Instagram: Antecedents and consequences of opinion leadership. *Journal of Business Research*, *117*, 510-519.

Cash, P., Isaksson, O., Maier, A., & Summers, J. (2022). Sampling in design research: Eight key considerations. *Design Studies, 78*, 101077. https://doi.org/10.1016/j.destud.2021.101077

Chen, Y., Lu, Y., Wang, B., & Pan, Z. (2016). Understanding self-disclosure on social networking sites: An integrated model based on social cognitive theory and the technology acceptance model. *Information & Management*, *53*(3), 381–393. https://doi.org/10.1016/j.im.2015.09.003

Childers, C. C., Lemon, L. L., & Hoy, M. G. (2019). # Sponsored# Ad: Agency perspective on influencer marketing campaigns. *Journal of Current Issues & Research in Advertising*, *40*(3), 258-274.

Choi, W.-S., Kang, S.-W., & Choi, S. B. (2022). Creativity in the South Korean workplace: Procedural justice, abusive supervision, and competence. *International Journal of Environmental Research and Public Health, 19*(1), 500–510.

Chung, S., & Cho, H. (2017). Fostering parasocial relationships with celebrities on social media: Implications for celebrity endorsement. *Psychology & Marketing*, *34*(4), 481–495. https://doi.org/10.1002/mar.21001

Çelik, K. (2021). The effects of information, entertainment, and relationship maintaining motivations on social media use. *Jomelips-Journal of Management Economics Literature Islamic and Political Sciences, 6*(2), 28-45. https://doi.org/10.24013/jomelips.1020599

Çelik, K., & Bora, C. (2021). Investigation of Instagram postings of hotel businesses using a content analysis method. *Journal of Tourism and Gastronomy Studies, 9*(4), 2813-2836. https://doi.org/10.21325/jotags.2021.921

Çelik, K., & Taş, A. (2021). Investigation of factors affecting consumer behaviors before purchase: A research on Instagram shoppers. *Journal of Business Research-Turk, 13*(4), 3821-3834.

Çelik, K., Özköse, H., & Güleryüz, S. (2018). Structural equation modeling with R. *Eurasian Journal of Researches in Social and Economics (EJRSE), 5*(10), 38-48.

Dahanayake, P., Rajendran, D., Selvarajah, C., & Ballantyne, G. (2018). Justice and fairness in the workplace: A trajectory for managing diversity. *Equality, Diversity and Inclusion: An International Journal, 37*(5), 470-490.

De Veirman, M., Cauberghe, V., & Hudders, L. (2017). Marketing through Instagram influencers: The impact of number of followers and product divergence on brand attitude. *International Journal of Advertising, 36*(5), 798–828.

Delbaere, M., Michael, B., & Phillips, B. J. (2021). Social media influencers: A route to brand engagement for their followers. *Psychology & Marketing*, *38*(1), 101-112.

Dhandra, T. K. (2019). Achieving triple dividend through mindfulness: More sustainable consumption, less unsustainable consumption and more life satisfaction. *Ecological Economics*, *161*, 83-90.

Ecker, U. K. H., & Antonio, L. M. (2021). Can you believe it? An investigation into the impact of retraction source credibility on the continued influence effect. *Memory & Cognition, 49*(4), 631–644.

Farivar, S., Wang, F. & the Yuan, Y. (2021). Opinion leadership vs. para-social relationships: Key factors in influencer marketing. *Journal of Retailing and Consumer Services, 59*, 102371.

Fornell C. & Larcker D.F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research*, *18*(1), 39–50. https://doi.org/10.2307/3151312.

Freberg, K., Graham, K., McGaughey, K. & Freberg, L. A. (2011). Who are the social media influencers? A study of public perceptions of personality. *Public Relations Review, 37*(1), 90-92.

Gil de Zúñiga, H., Jung, N., & Valenzuela, S. (2018). Social media use for news and individuals' social capital, civic engagement, and political participation. *Journal of Computer-Mediated Communication, 17*(3), 319-336.

Hair J.F., Black W.C., Babin B.J., & Anderson R.E. (2010). *Multivariate data analysis* (7th ed.). Upper Saddle River, NJ: Pearson Education.

Hernández Soto, R., Gutiérrez Ortega, M., & Rubia Avi, B. (2021). Key Factors in Knowledge-Sharing Behavior in Virtual Communities of Practice: A systematic review. *Education in the Knowledge Society, 22*(2), 215-227.

Horton, D., & Wohl, R. R. (1956). Mass communication and parasocial interaction: Observations on intimacy at a distance. *Psychiatry*, *19*(3), 215-229.

Hughes, C., Swaminathan, V., & Brooks, G. (2019). Driving brand engagement through online social influencers: An empirical investigation of sponsored blogging campaigns. *Journal of Marketing*, *83*(5), 78–96.

Iacobucci, D., & Duhachek, A. (2003). Advancing alpha: Measuring reliability with confidence. *Journal of Consumer Psychology*, *13*(4), 478–487. https://doi.org/10.1207/s15327663jcp1304_14

Ismagilova, E., Slade, E., Rana, N. P., & Dwivedi, Y. K. (2020). The effect of characteristics of source credibility on consumer behavior: A meta-analysis. *Journal of Retailing and Consumer Services*, *53*(3), 1–9.

Jin, S. V., Muqaddam, A., & Ryu, E. (2019). Instafamous and social media influencer marketing. *Marketing Intelligence and Planning*, *37*(5), 567-579.

Joshi, Y., & Rahman, Z. (2019). Consumers' sustainable purchase behaviour: Modeling the impact of psychological factors. *Ecological Economics*, *159*, 235-243.

Keskin, A. (2020). Parasosyal etkileşim açısından K-Pop fanlığı: Army hayran kültürü üzerinde bir inceleme. *İnönü Üniversitesi İletişim Fakültesi Elektronik Dergisi (İNİF E-Dergi)*, *5*(1), 25-38.

Keskin, S. (2020). The social anxiety scale for e-learning environments (SASE). *Interactive Learning Environments*, *28*(7), 1–15. https://www.tandfonline.com/doi/abs/10.1080/10494820.2020.1769681

Ki, C.-W. C., & Kim, Y.-K. (2019). The mechanism by which social media influencers persuade consumers: The role of consumers' desire to mimic. *Psychology & Marketing*, *36*(10), 905–922. https://doi.org/10.1002/mar.21244.

Kim, D. Y., & Kim, H. Y. (2021). Influencer Advertising on Social Media: The Multiple Inference Model on Influencer-Product Congruence and Sponsorship Disclosure. *Journal of Business Research*, *130*, 405-415. https://doi.org/10.1016/j.jbusres.2020.02.020

Kim, J., & Song, H. (2016). Celebrity's self-disclosure on Twitter and parasocial relationships: A mediating role of social presence. *Computers in Human Behavior*, *62*, 570–577. https://doi.org/10.1016/j.chb.2016.03.083

Koay, K. Y., Lim, W. M., Kaur, S., Soh, K., & Poon, W. C. (2023). How and when social media influencers' intimate self-disclosure fosters purchase intentions: the roles of congruency and parasocial relationships. *Marketing Intelligence & Planning*, *41*(6), 790-809.

Koh, J., Kim, Y. G., & Kim, Y. G. (2003). Sense of virtual community: A conceptual framework and empirical validation. *International Journal of Electronic Commerce*, *8*(2), 75-94.

Lee, J. E., & Watkins, B. (2016). YouTube vloggers' influence on consumer luxury brand perceptions and intentions. *Journal of Business Research*, *69*(12), 5753–5760. https://doi.org/10.1016/j.jbusres.2016.04.171

Lee, J. E., & Watkins, B. (2016). YouTube vloggers' influence on consumer luxury brand perceptions and intentions. *Journal of Business Research*, *69*(12), 5753-5760.

Leite, F. P., & Baptista, P. d. P. (2021). The effects of social media influencers' self-disclosure on behavioral intentions: The role of source credibility, parasocial relationships, and brand trust. *Journal of Marketing Theory and Practice*, *30*(3), 295–311.

Liu, M. T., & Brock, J. L. (2011). Selecting a female athlete endorser in China: The effect of attractiveness, match-up, and consumer gender difference. *European Journal of Marketing*, *45*(7), 1214–1235. https://doi.org/10.1108/03090561111137688.

Liu, M. T., Liu, Y., & Lin, C. P. (2019). Vlog and brand evaluations: The influence of parasocial interaction. *Asia Pacific Journal of Marketing and Logistics*, *31*(1), 1–17.

Lou, C. (2022). Social media influencers and followers: Theorization of a trans-parasocial relation and explication of its implications for influencer advertising. *Journal of Advertising*, *51*(1), 4-21.

Lou, C., & Kim, H. K. (2019). Fancying the new rich and famous? Explicating the roles of influencer content, credibility, and parental mediation in adolescents' parasocial relationships, materialism, and purchase intentions. *Frontiers in Psychology*, *10*(2), 256–276.

Lueck, J. A. (2015). Friend-zone with benefits: The parasocial advertising of Kim Kardashian. *Journal of Marketing Communications*, *21*(2), 91-109.

Majewski, G., Usoro, A. & Khan, I. (2011). Knowledge sharing in immersive virtual communities of practice. *Vine*, *41*(1), 41–62.

McQuarrie, E. F., Miller, J., & Phillips, B. J. (2013). The megaphone effect: Taste and audience in fashion blogging. *Journal of Consumer Research*, *40*(1), 136-158.

Munnukka, J., Uusitalo, O., & Toivonen, H. (2016). Trustworthiness of peer endorsers and advertising effectiveness. *Journal of Consumer Marketing*, *33*(3), 182–192.

NunnallyI J.C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill

Park, H., & Lee, J. (2021). Impact of social media influencers on consumers' purchase intentions: The role of parasocial interaction. *Computers in Human Behavior*, *115*, 106576.

Pennington, N., Hall, J. A., & Hutchinson, A. (2016). To tweet or not to tweet: Explaining fan–celebrity interaction on Twitter. *Iowa Journal of Communication*, *48*(1), 55–75. https://scholarworks.uni.edu/ijc/vol48/iss1/7.

Podsakoff, P.M., MacKenzie, S.B., Lee, J.Y., & Podsakoff N.P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, *88*(5), 879–903. https://doi.org/10.1037/0021-9010.88.5.879

Schiappa, E., Allen, M., & Gregg, P. B. (2007). Parasocial relationships and television: A meta-analysis of the effects. *Mass Media Effects Research: Advances Through Meta-Analysis*, 301-314.

Smith, A. (2020). Social media influencers and consumer behavior: The power of parasocial relationships. *International Journal of Digital Marketing*, *12*(2), 120-135.

Usoro, A., & Majewski, G. (2011). Intensive knowledge sharing: Finnish Laurea laboratory case study. *Vine*, *41*(1), 7–25.

Uwa, K. L. (2022). Organizational fairness and organizational citizenship behavior: A study of selected fast food industries in Uyo metropolis, Akwa Ibom State, Nigeria. *International Journal of Business and Management Review*, *10*(2), 33–43.

Vrontis, D., Makrides, A., Christofi, M., & Thrassou, A. (2021). Social media influencer marketing: A systematic review, integrative framework and future research agenda. *International Journal of Consumer Studies*, *45*(4), 617-644.

Xiang, L., Zheng, X., Lee, M. K., & Zhao, D. (2016). Exploring consumers' impulse buying behavior on social commerce platform: The role of parasocial interaction. *International Journal of Information Management*, *36*(3), 333-347.

Xiao, M., Wang, R., & Chan-Olmsted, S. (2018). Factors affecting YouTube influencer marketing credibility: A heuristic-systematic model. *Journal of Media Business Studies*, *15*(3), 188–213.

Xu, X., & Pratt, S. (2018). Social media influencers as endorsers to promote travel destinations: an application of self-congruence theory to the Chinese Generation Y. *Journal of Travel & Tourism Marketing*, *35*(7), 958-972.

Yuan, L., & Dennis, A. R. (2019). Acting like humans? Anthropomorphism and consumer's willingness to pay in electronic commerce. *Journal of Management Information Systems*, *36*(2), 450-477.

Yuan, S., & Lou, C. (2020). How social media influencers foster relationships with followers: Roles of source credibility and fairness in parasocial relationships and product interest. *Journal of Interactive Advertising*, *20*(2), 133–147. https://doi.org/10.1080/15252019.2020.1769514.

Zheng, X., Men, J., Xiang, L., & Yang, F. (2019). The role of technology attraction and parasocial interaction in social shopping websites. *International Journal of Information Management*, *51*, 1-13.

İSTANBUL UNIVERSITY PRESS

Acta Infologica

Acta Infologica,
ACIN 2025, 9 (1): 183–207

https://doi.org/10.26650/acin.1670469
Submitted: 05.04.2025
Revision Requested: 11.04.2025
Last Revision Received: 14.05.2025
Accepted: 22.05.2025
Published Online 04.06.2025

Research Article

Open Access

Designing a Large Language Model-Based AI System for Dynamic Difficulty Adjustment in Digital Games

Onur Aşkın [1]

[1] Doğuş University, Faculty of Art and Design, Department of Digital Game Design, İstanbul, Türkiye

Abstract This study investigates how large language models (LLMs) can serve as dynamic agents in game-based interactions by comparing two prototypes of a color-guessing game. One model (Cohere Command) operates on a zero-shot prompt-based mechanism, while the other (FLAN-T5) is fine-tuned on a semantically structured dataset. A total of 20 participants were divided into two experimental groups to evaluate the models' ability to generate semantically coherent yes/no questions, maintain flow, and perform accurate predictions. Quantitative data, including session durations, number of interactions, and AI outputs, were analyzed, along with a post-game user experience survey grounded in Flow Theory. Results show that while both systems achieved task completion, the fine-tuned FLAN-T5 model significantly outperformed the other models in terms of semantic clarity, user engagement, and perceived fluency. The findings highlight the potential of LLM-based DDA systems in creating meaningful, adaptive player experiences and underscore the importance of semantic alignment and interaction transparency in game-based AI design.

Keywords Artificial Intelligence · Machine Learning · Digital Game Design · Dynamic Difficulty Adjustment

Citation: Aşkın, O. (2025). Designing a large language model-based AI system for dynamic difficulty adjustment in digital games. *Acta Infologica*, *9*(1), 183-207. https://doi.org/10.26650/acin.1670469

This work is licensed under Creative Commons Attribution-NonCommercial 4.0 International License.

© 2025. Aşkın, O.

Corresponding author: Onur Aşkın oaskin@dogus.edu.tr

Introduction

When a game's difficulty level does not match the player's skill level, the experience becomes tedious or overly challenging. To achieve this balance, Dynamic Difficulty Adjustment (DDA) systems aim to make the game flow more motivating and sustainable. Traditional DDA approaches are usually based on numerical data such as the player's success rate and score. However, these methods do not adequately consider a player's decision-making style, contextual strategies, and in-game interaction patterns.

Integrating big language models (LLMs) into game systems provides a new dimension to the concept of LDA. Such models enable the development of AI units that are not only data-driven but also able to read context, analyze previous interactions, and make strategic decisions (Brown et al., 2020; Radford et al., 2019). Thanks to the flexible nature of the language, AI can offer player-specific difficulty levels and enable more human-like decisions. Game scenarios with short yes/no answers provide a suitable ground to test the contextual analysis power of language models. This kind of experimental setup allows us to observe how AI performs at different difficulty levels while rethinking the role of these systems in game design.

However, in the existing literature, the use of LLMs in the context of dynamic difficulty tuning has been addressed in a limited number of studies; in particular, the effects of different LLM architectures on in-game strategic decision making, semantic output quality, and flow experience have not been systematically compared. Aiming to fill this gap, this paper empirically compares two different LLM architectures (zero-shot and fine-tuned) in a game-based interaction setting and presents a multi-layered evaluation of semantic directiveness, flexibility, and user experience metrics.

Purpose of the Study

This study investigates how artificial intelligence using large language models (LLMs) contributes to adjusting the difficulty level of games according to the player. For this purpose, a question-answer-based guessing game between two artificial intelligences was designed, and their performances were compared at different difficulty levels: easy, medium, and hard.

Research Question and Hypotheses

Research Question: Can large language models optimize dynamic difficulty adjustment in digital games with humanlike strategic decision-making?

Hypothesis 1 (H1): AI-assisted dynamic difficulty adjustment (DDA) systems can be optimized using large language models (LLM).

Sub-Hypotheses:

- H1a: LLM-based prediction strategies enable AI systems to dynamically adjust in-game difficulty levels in real time.
- H1b: LLM-supported AI systems can make strategic decisions at different difficulty levels (easy, medium, hard).
- H1c: LLM-based AI predictions show lower response times and higher decision flexibility than traditional machine learning algorithms.
- H1d: The output entropy of LLMs increases the accuracy and adaptability of AI-based prediction strategies.

- H1e: LLM-based DDA systems provide more consistency in game flow and player interaction than traditional methods.

Hypothesis 0 (H0): LLM-based AI systems do not provide significant improvements in terms of dynamic difficulty adjustment.

Variables

Independent Variables:

- Natural Language Processing (NLP) Model: Language processing capacity that drives AI's prediction strategies.
- Difficulty Level (Easy, Medium, Hard): The complexity level of the AI prediction strategies.
- Prediction Time: The prediction time taken by the AI algorithm.
- Success Rate: The rate at which the AI correctly assumes the hidden color.

Dependent Variable:

- Prediction Strategy Success: The proportion of correct predictions based on the language model used by AI in dynamic difficulty adjustment strategies.

Scope and Limitations

This research used an AI-assisted color prediction game to evaluate the performance of large language models (LLMs) in dynamic difficulty adjustment (DDA) systems. After playing the game, the participants completed a user experience survey with questions generated by two different artificial intelligence models (Cohere AI and FLAN-T5) at different difficulty levels. Thus, a multi-layered evaluation was performed in line with the system data and user opinions.

However, the study has certain limitations.

Within the scope of the research, only two LLM architectures were examined; models from other platforms, such as OpenAI, IBM, and Azure, were not included in the study. The Cohere model was used as a routing-based model from abrasion, while the FLAN-T5 model was fine-tuned and implemented on the Hugging Face platform with a custom dataset. However, different model versions, advanced training configurations, or alternative task protocols offered by Hugging Face were excluded from the scope of this study. In the implemented models, the AIs were not allowed to learn the difficulty levels independently; the difficulty settings were determined according to predefined rules.

Data analysis was limited to key performance indicators such as accuracy rate, number of questions, and playing time; however, no advanced methods such as long-term learning processes, time series analysis, multiple user interactions, or in-game social dynamics, were used. Interactions were text-based, with no visual or auditory supportive components.

Ethically, no personal data were collected in the study; all sessions were conducted based on participant consent and volunteerism. The research was conducted with the approval of Doğuş University Scientific Research and Publication Ethics Committee (Date: 28.05.2025, No: 82320). When similar systems are planned to be tested with real player data in the future, basic ethical principles, such as data privacy, user consent, algorithmic transparency, and system auditability, must be rigorously observed.

Methodological Justification and Design Rationale

Although the number of participants in this study (n = 20) seemed limited, the experimental design was deliberately divided into two independent groups (n_1 = 10, n_2 = 10). This partitioning eliminates factors that threaten the internal validity of comparing AI systems, such as the learning effect, ranking effect, and surrogate learning. The experimental design ensured that all participants interacted with only one model, preventing cross-contamination and allowing for a more precise effect size measurement.

Furthermore, another prominent aspect of this study is the multi-layered data collection and analysis strategy. Different layers of data were collected and analyzed, including game session data, user experience surveys, Shannon Entropy analyses, and independent sample t-tests for statistical significance. This increased the study's contextual depth and validity despite the quantitative sample's limitations.

Another study strength is that the game prototypes are reproducible through open-source platforms. This ensures the experiment's replicability and provides a scalable frame of reference for future work. Therefore, the study goes beyond a pretest and presents a transparent model in terms of experimental design and offers added value for practice.

This methodological justification section provides defensible arguments that emphasize the empirical nature of the study, data depth, and analyzability despite superficial criticisms about the number of participants. On the other hand, only two major language models (Cohere Command and FLAN-T5) were compared in this study; other models, such as OpenAI GPT, Claude, and Gemini, were excluded. This choice was made deliberately to ensure methodological control of the study and to perform an in-depth comparison. The selected models represent different production approaches: Cohere Command offers a prompt-based architecture, while the FLAN-T5 model has a fine-tuned architecture with a specialized dataset. Thanks to this structure, we can directly compare the advantages and limitations of different language model-based dynamic difficulty tuning methods. In similar studies, the number of models is usually limited, and factors such as methodological consistency, parametric control, and technical resource requirements are considered in comparative analyses. In future research, comparisons involving more language models can be performed to increase generalizability and analyze different architectures' production behavior more comprehensively.

Theoretical Framework

This study's theoretical basis is the Flow Theory developed by Csikszentmihalyi (1990). Flow is a special state of consciousness in which the individual fully focuses their attention on the activity loses the perception of time, and experiences high motivation during the activity (Keller & Bless, 2008). In order to sustain this experience in game design, an ideal balance between the player's skill level and the game's difficulty is recommended.

Sweetser and Wyeth (2005) highlighted three key elements that support the flow state: clear goals, immediate feedback, and skill-appropriate challenges (Cowley et al., 2008). These elements strengthen the player's sense of control and emotional connection to the game, thus satisfying the experience. Chen (2007) took these principles to an algorithmic level and developed the concept of "Dynamic Difficulty Adjustment" (DDA). DDA systems aim to maintain the continuity of the flow state by changing the game difficulty in real time according to the player's performance.

Difficulty setting and balancing cognitive load are critical for sustaining the Flow experience. This is where Sweller's (1988) Cognitive Load Theory comes into play. Systems that allow the player to make meaningful decisions within the game without interrupting their attention with excessive stimuli increase the quality of learning and experience.

Ryan and Deci's (2000) proposed the self-determination theory, which provides an important structure that supports the flow in a motivational framework. According to this theory, intrinsic motivation increases when individuals' basic psychological needs, such as competence, autonomy, and relationship building, are satisfied. Dynamic challenge systems sustain this motivation to the extent that they provide the player with a sense of competence.

Unlike classical rule-based LLM frameworks, artificial intelligence systems based on big language models (LLM) not only generate responses to player inputs and have the capacity to develop semantic contextual guidance, meaningful question generation, and prediction strategies (To & Brusilovsky, 2021).

In this study, we compare two different big language models, prompt-based (Cohere AI) and fine-tuned (FLAN-T5), and evaluate their impact on semantic production quality, guidance capacity, and flow experience. The findings revealed significant differences in technical accuracy and experiential parameters, such as player perception, time perception, and cognitive continuity. In this respect, this study proposes a multi-layered rethinking of how dynamic difficulty adjustment systems can be made semantically guided and personalized at the individual interaction level through LLM-based production behaviors in the context of flow theory.

Literature

This section provides a literature review on Dynamic Difficulty Adjustment (DDA) systems, structured around several key themes: their application in games, their impact on player experience, integration with artificial intelligence (AI) and machine learning (ML), personalization strategies, and their use in education and rehabilitation. In addition, this review highlights a significant gap in studies that integrate Large Language Models (LLMs) with DDA systems. This review aims to establish a foundational understanding of DDA's technical and experiential dimensions by outlining these domains.

DDA refers to techniques that dynamically adjust game difficulty in response to the player's performance, skill level, or emotional state. This adaptive approach is often employed to support player motivation and facilitate flow experiences. Early studies, such as Gilleade et al. (2004), examined how adaptive difficulty could sustain players in a flow state, while Kowlessar (2020) demonstrated that optimal difficulty settings enhance player engagement. However, some findings suggest that relying solely on performance metrics is insufficient; psychological states must also be considered to avoid unintended player frustration. The application of AI and ML techniques has significantly advanced the design of DDA systems. For instance, Garcia-Ruiz et al. (2023) introduced a deep learning-based system that increased player satisfaction by adjusting difficulty based on skill progression. Roohi et al. (2021) combined Deep Reinforcement Learning (DRL) with Monte Carlo Tree Search (MCTS) to create an adaptive system that performs well at higher difficulty levels. Similarly, Yannakakis and Togelius (2018) and Lopes and Bidarra (2011) emphasized the central role of adaptive AI in dynamic gameplay experiences. Fisher and Kulshreshth (2024) contributed to this discussion by demonstrating that performance- and emotion-based DDA strategies produce varying outcomes across player profiles. In addition to performance optimization, DDA systems have been employed in motivational and cognitive frameworks. For example, Colwell et al. (2018) demonstrated that adaptive

difficulty calibrated using the ARCS motivational model improved trust and engagement. Smulter and Porsbjer (2019) examined how cognitive workload affects players under dynamically changing difficulty, suggesting that task complexity and attentional demand should be central parameters in adaptive systems. The growing interest in personalization has led to emerging studies using large language models (LLMs) in DDA. However, work on integrating LLMs with DDA remains limited, revealing a crucial research gap. In applied contexts, DDA has also shown potential in educational and therapeutic settings. Wu, Chen, and Chen (2017) demonstrated that an adaptive e-learning system based on dynamic scaffolding significantly enhanced student learning performance by aligning difficulty levels with individual progression. Cameirão et al. (2009) highlighted the use of the Rehabilitation Gaming System (RGS) as an example of how adaptive game-based environments can support cognitive and motor recovery in clinical neurorehabilitation. From a technical standpoint, Zheng (2024) reported that the integration of deep reinforcement learning into DDA systems enhanced the adaptability and precision of difficulty scaling, enabling more strategic and skill-responsive level design. These insights reinforce the role of DDA in structuring challenges in a way that aligns with individual player trajectories. In conclusion, the literature collectively demonstrates the effectiveness of DDA systems in terms of enhancing gameplay, motivation, and learning. However, integrating LLM-based AI systems into DDA remains an underexplored area. Although personalization through AI-driven adaptation has been considered in recent work, existing research still lacks comprehensive, comparative, and experimentally validated frameworks that assess how different LLM architectures influence dynamic difficulty adjustment across gameplay variables. In particular, little is known about the strategic decision-making capacity, semantic output quality, and flow facilitation of LLMs under different difficulty tiers. This study addresses this critical gap by experimentally comparing a zero-shot prompting-based model (Cohere Command) with a fine-tuned LLM (FLAN-T5) in a controlled, game-based interaction context. By examining semantic richness, adaptability, and user engagement metrics, the research aims to provide a multilayered evaluation of how LLM-driven DDA systems can be optimized for more intelligent and humanlike gameplay experiences.

Methodology

Experimental Structure and Setup

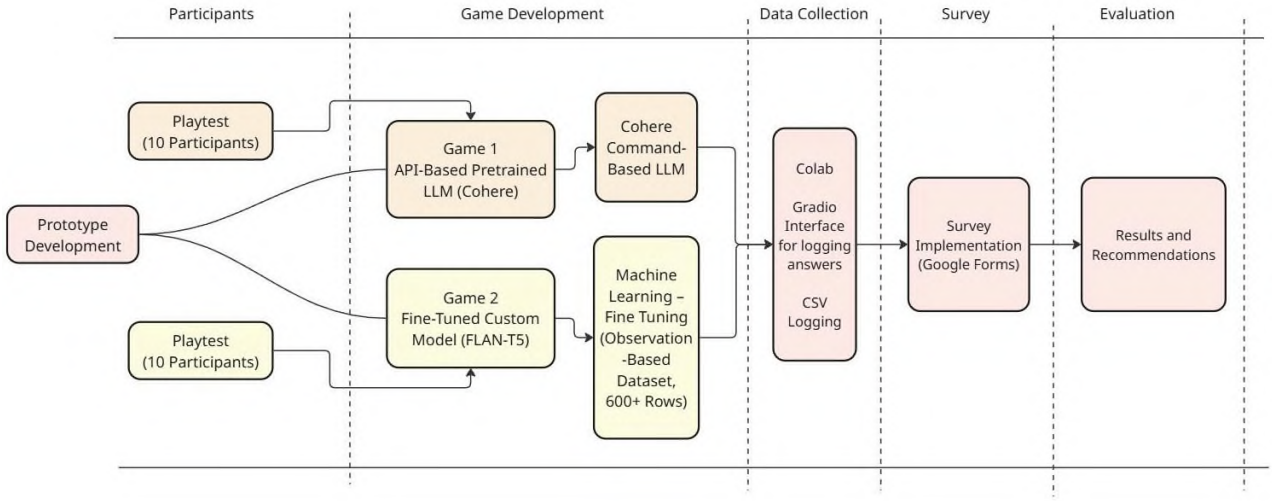
This study used an experimental method to examine the behavior, semantic, and experiential performance of large language models (LLMs) in game-based interactions. Within the scope of the research, two different game prototypes were developed, and the plar in their minds through yes/no questions. The first of these prototypes is based on the zero-shot prompting technique with the Cohere AI model, while the second one is fine-tuned and customized on the FLAN-T5 architecture on the Hugging Face platform. Both games were structured with three different difficulty levels (easy, medium, and hard), and the participants' responses to the semantic prompts were evaluated. Participants were divided into two groups. Each group experienced only the game prototype directed by an artificial intelligence model. In this way, internal validity issues that may be observed in comparative experiments, such as learning and ranking effects, were avoided (n_1 = 10, Cohere; n_2 = 10, FLAN-T5).

Accessibility, architectural diversity, and the feasibility of fine-tuning were considered in the model selection. The FLAN-T5 model stands out due to its open-source nature, easy trainability on the Hugging Face platform, and semantic generation success in the literature. The Cohere Command model, on the other hand, is known for its zero-shot performance based on routing and provides a suitable counterexample for comparative testing in terms of natural language generation without the need for fine-tuning. The fact

that both models represent different production approaches provides a meaningful framework for the experimental comparison of their production behavior and the analysis of their impact on dynamic difficulty management (Figure 1).

Figure 1

Experimental Workflow of AI-Supported Game Prototype Development and Evaluation



Technical Implementation and Game Prototypes

The Cohere AI-based game version is based on the Command model, accessed via a ready-made API, and driven only by system prompts. The model was expected to generate short questions focusing on color properties that could only be answered with yes or no answers. At the three difficulty levels, production variance and semantic complexity gradually increased. The FLAN-T5 model was fine-tuned on a custom dataset comprising more than 600 color-based question-answer pairs. This process was run on Google Colab Pro in the Hugging Face environment using the Transformers and Seq2SeqTrainer classes (Vaswani et al., 2017). The model was optimized for strategic question generation by considering factors such as semantic context, psychological connotation, and color categories (Goodfellow, Bengio, & Courville, 2016).

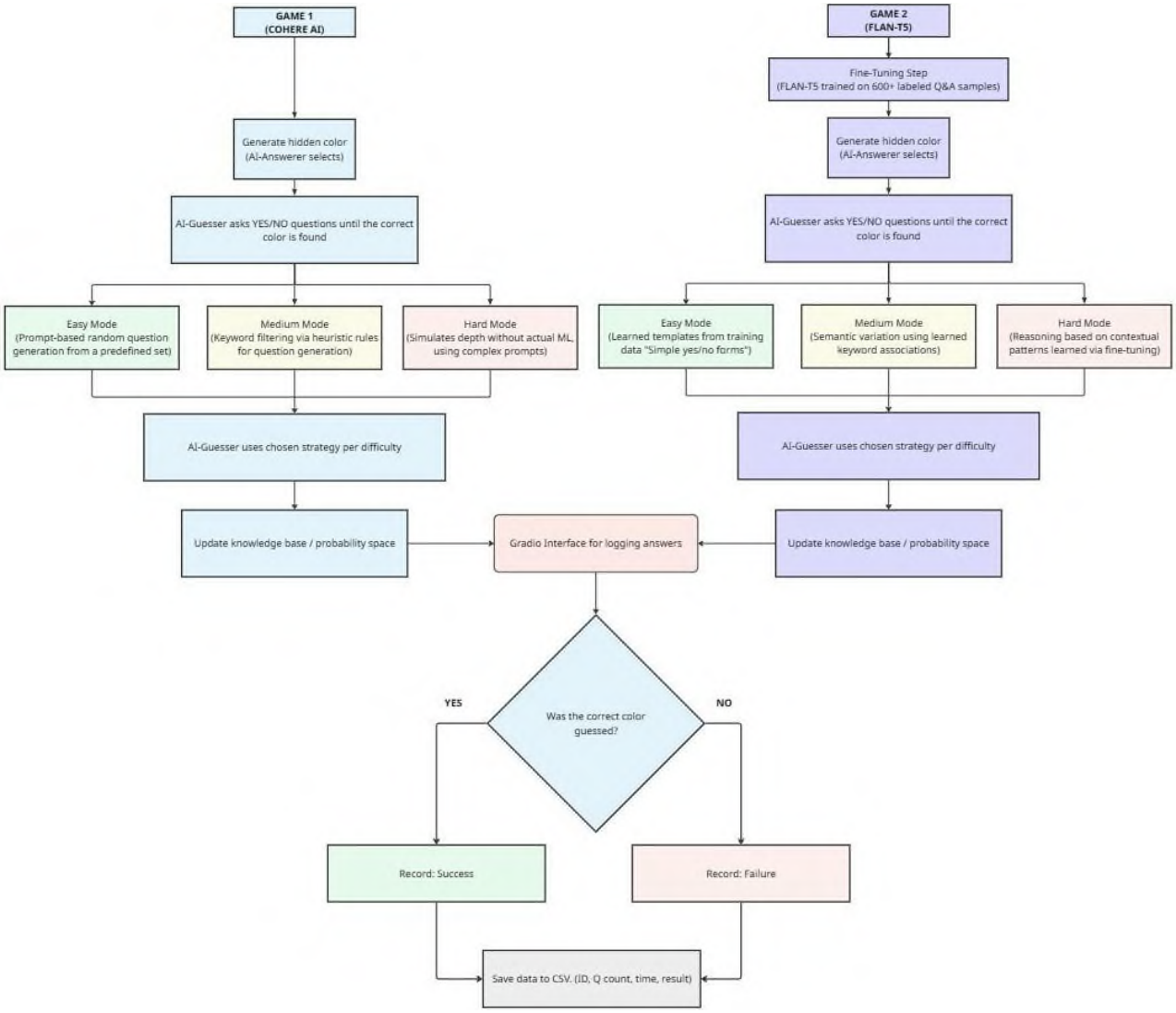
Both games were presented to the user through a Gradio-based interface, and all sessions were recorded throughout the production process. In the color prediction process, the system had the user predict a fixed target color; thus, the artificial intelligence was directed to make meaningful inferences based only on user responses.

The training dataset used in the fine-tuning process of the FLAN-T5 model was created using a large language model (ChatGPT) that supported text generation according to the semantic rules created by the researcher. The dataset contains over 600 yes/no questions in total, each labeled with a target color, difficulty level (easy, medium, hard), and semantic category (e.g., nature, temperature, symbolism). For example, the question "Is this color often linked with temperature in nature?" is in the category "temperature" under the color "burgundy" and the difficulty level "Easy". In the production process, multi-layered conceptual areas, such as the hot-cold perception of colors, the frequency of their occurrence in nature, cultural connotations, and emotional effects, were considered. The researcher manually reviewed the generated data samples, and the reliability of the training set was ensured by removing content containing semantic

inconsistencies or physical meaning errors. The full dataset is presented in the appendix of this article (Figure 2).

Figure 2

Comparative Flowchart of Two LLM-Based Color Guessing Games



Data Collection, Survey, and Analysis Process

Interaction data were saved in CSV format during all game sessions. These data includes session IDs, question texts asked by the AI, user responses, remaining color probabilities, and prediction times.

At the end of the game, each participant was administered a Likert-type user experience questionnaire designed in accordance with Flow Theory. The questionnaire aimed to measure experiential elements such as AI question clarity, content repetition rate, prediction success, time perception, and overall immersion.

The collected data were analyzed using Python. The following libraries were used in this analysis: pandas, seaborn, scipy, and collections (Harris et al., 2020; Virtanen et al., 2020). The key metrics computed include the semantic repetition rate, number of out-of-context productions, session duration, and entropy. The performance comparison between models was performed using independent samples t-tests.

Entropy Analysis: Diversity in User Opinions

The user survey data were evaluated not only in terms of average scores but also in terms of the diversity of responses. For this purpose, Shannon Entropy calculations were performed. The entropy is particularly meaningful for assessing the subjective dimensions of the player experience. The more diverse user responses to a metric, the more flexible and multilayered the interpretation of that experience metric.

The entropy value is calculated using the following formula:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \cdot log_2 p(x_i)$$

Here$p(x_i)$denotes the probability of each response category (Likert options 1-5) being observed. The higher the entropy value, the more diverse the respondent responses are; the lower the entropy value, the more homogeneous are their responses.

These calculations were performed using collections. Counter and math.log2 functions in Python (Harris et al., 2020). Thus, user experience was evaluated not only in terms of score averages but also in terms of differences, level of unpredictability, and variety of cognitive awareness.

Evaluation of Outlier Duration Values

When the interaction data of the Game 1 (Cohere AI) game prototype were analyzed, it was found that there were some unusual (outlier) values, especially in session durations. For example, some sessions exceeded 135,000 seconds (approximately 37 hours), which is not a realistic interaction duration in the context of the experimental task. Upon further investigation, it became clear that such data were due to the game being left open (idle state) without the users actively participating.

This is a common problem, particularly in web-based experimental studies. The participant can walk away from the computer, tab to the background, or do something else without closing the game screen. The game interface does not close during such passive situations; thus, the system continues to record the session duration without interruption. In this context, in this study conducted on the Gradio interface, long-opening the browser tab was recorded as game time; however, these values do not reflect the actual user interaction.

To avoid this problem and ensure the validity of the data, interaction time alone was not used as an evaluation criterion; instead, we focused on more reliable indicators, such as the number of questions, production entropy, and user experience surveys. Furthermore, only active and consistent interaction patterns were included in the statistical analysis, ignoring such outliers. This approach ensured that the results accurately represented the interaction between the user and the AI.

Results

In the process of evaluating the game prototypes developed with two different large language models, both in-game interaction data and survey results based on user tests were analyzed in multiple layers. The playability performance of both games was measured using CSV-formatted game data obtained during plates sessions conducted with real users; AI-generated questions, player responses, and processing times were recorded in detail. These in-game data were systematically analyzed in terms of the AI model's guidance, production consistency, and prediction strategy. Furthermore, cognitive, semantic, and flow feedback on the game experience was collected through structured questionnaires that the participants

answered after each game session. By evaluating the quantitative and qualitative data sources together, the findings presented below were reached, and the performance characteristics of both models were presented comparatively in the context of game interaction.

Game 1 (Cohere AI)

The interaction data of 10 individuals participating in the color guessing game were analyzed at three different difficulty levels (easy, medium, hard) based on the total game time and the number of questions generated. This analysis provides important findings on the productivity, semantic guidance capacity, and response generation tempo of AI-supported interactions.

Table 1

Detailed Interaction Time and Question Count Across Difficulty Levels in Game 1 (Cohere AI)

Participant ID	EASY		MEDIUM		HARD		TOTAL	
	SEC	Q COUNT	SEC	Q COUNT	SEC	Q COUNT	SEC	Q COUNT
1	1636.79	30	13802.49	158	420.72	22	15860.00	210
2	297.81	39	135772.98	112	1237.32	63	139982.11	214
3	628.17	21	32390.66	189	12321.56	79	45340.39	289
4	1177.12	20	19103.48	161	1243.99	34	21524.59	215
5	2590.70	39	10868.40	124	4336.61	61	17795.71	224
6	157.43	12	106001.85	220	2564.08	43	108723.36	275
7	4623.27	39	159620.68	310	2826.37	50	167070.32	399
8	1329.26	24	39339.04	212	4672.47	46	45340.77	282
9	1867.80	30	214292.23	199	21295.94	91	237455.97	320
10	815.26	21	147767.02	316	3675.33	28	152257.61	365

When the overall distribution is analyzed, it can be seen that the sessions at the medium difficulty level lasted significantly longer (M = 1793 minutes, SD = 990). In contrast, the average duration at the easy level was 38.7 minutes. The difficult level was positioned between the two levels with an average of 106.9 minutes. These differences suggest that interactions with large language models are strongly influenced by technical parameters and contextual stability (Table 1).

Table 2

Summary of Average Duration and Question Count by Difficulty Level (in Minutes)

Difficulty Level	Average Duration (min)	Standard Deviation (min)	Average Question Count	Standard Deviation (count)
Easy	38.69	24.18	29.6	9.2
Medium	1793.28	990.34	202.0	67.0
Hard	106.95	97.17	51.7	22.1

There was a similar disproportion in terms of the number of questions. An average of 29.6 questions were produced in the easy mode; this number reached 202 in the medium mode. This dramatic difference suggests that the system has entered a kind of "contextual stuck" state. At the intermediate level, the model continued to generate questions without making any meaningful predictions, which caused the guiding interactions to be replaced by aimless production. The basis of this situation is that the model cannot end the game at any point and cannot make a prediction, such as "Is this color red?" to the user. The fact that

the standard deviation at this level reached 990 minutes supports the observation that in some sessions, the model exhibited serious inconsistencies in both semantic and temporal terms. As a matter of fact, in some examples, the model's generating questions contrary to the rules of the game such as "Is this color integrated with nature?" or "Does this color change according to the seasons?" caused the players to become indecisive and prolonged the interaction time (Table 2).

Furthermore, hallucinatory (unreal, contextually irrelevant, or physically contradictory) content generation was observed in all participants throughout the interaction process. Although grammatically coherent, the content did not provide meaningful guidance in terms of the game context; on the contrary, it interrupted the flow of the game experience by increasing the users' cognitive load. Similarly, physically contradictory questions such as "Is this color both warm and blue?" revealed that the model could not perform conceptual fact-checking and only generated content based on statistical patterns. The prevalence of such production indicates an overproduction and a serious lack of semantic control in the quality of production.

In hard mode, the number of questions was relatively limited (M = 51.7 questions, SD = 22.1), and the generation time was shorter and more consistent (M = 106.9 minutes, SD = 97.2 minutes) (Table 2). This may indicate that the AI is trying to execute production more carefully when faced with more complex prompts. However, this positive trend does not eliminate the system's tendency to fall again. For example, asking three questions with the same meaning in succession, such as "Is this color warm?" "Is it one of the warm tones?" and "Is it close to red?" clearly demonstrate the model's inability to use short-term context memory effectively.

In order to stabilize the production process of the model, a short-term micro-context memory was created, and the previous three questions were included in the prompt. However, this strategy was not statistically effective because of the context window limitation. It was systematically recorded that similar questions were repeated many times within the same session, meaningless jumps were made, and questions irrelevant to the task were produced. This, coupled with the non-deterministic production behavior of the model, led to a pattern in which uncontrolled variance was transformed into semantic distortion. In particular, the generation of different and often non-directional responses to the same prompt, despite the temperature parameter being fixed at 0.5, demonstrates the inherent inconsistency of the production. The hallucinatory productions not only prevented the natural ending of the game but also caused the players were unable to understand what the AI was aiming for.

These findings demonstrate that the production behaviors of LLM-based AI systems operate solely based on linguistic patterns and have significant limitations in contextual task adaptation. When the model operates without a specific contextual goal, it tends to engage in uncontrolled production, which directly impacts user experience. In such cases, even if the amount of production is high, the content remains weak in terms of guidance and semantic accuracy, and the semantic coherence of the system is sacrificed in favor of productivity. Production "abundance" was often inversely proportional to production "quality".

Large language models appear vulnerable to exhibit directionless and uncontrolled behavior in free production structures and game-like interaction environments. This behavior threatens not only the user experience and the logical flow of the game. Therefore, in the in-game use of LLM systems, the outputs and production process must be controlled and redesigned in a multi-layered manner. Otherwise, the players' experience will be semantically unsatisfactory, technically inefficient, and repetitive.

Game 1 Survey Evaluation

This study collected user experience data through a 5-point Likert-type questionnaire administered to participants who interacted with an AI-powered color guessing game (n = 10) (Table 3).

Table 3

User Experience Survey Results for Game 1 (Cohere AI)

Question	Number of Participants *(n)*	Mean *(M)*	Standard Deviation *(SD)*	Minimum *(Min)*	1st Quartile *(Q1)*	Median *(Md)*	3rd Quartile *(Q3)*	Maximum *(Max)*
I immediately understood the questions asked by the AI.	10	2.1	1.449	1.0	1.00	1.5	2.75	5.0
Similar questions were repeated repeatedly.	10	4.4	1.264	1.0	4.25	5.0	5.00	5.0
The AI tried guessing the color.	10	1.5	0.527	1.0	1.00	1.5	2.00	2.0
I lost track of time while playing the game.	10	1.4	0.699	1.0	1.00	1.0	1.75	3.0
The game was very engaging.	10	1.4	0.699	1.0	1.00	1.0	1.75	3.0

Participants' responses were interpreted in terms of the system's comprehensibility, content diversity, predictive capacity, and experiential immersion, and the statistical data obtained were supported by qualitative content and subjected to analysis.

"The mean M = 2.10, standard deviation SD = 1.45, median Md = 1.5, minimum = 1, maximum = 5, first quartile ($Q1$) = 1.00 and third quartile ($Q3$) = 2.75. This result indicates that a large proportion of the participants had difficulty understanding the AI questions. However, the high standard deviation reveals that there is significant variation in this perception among the participants. In this context, the fact that some participants had difficulty understanding the questions directly can be attributed not only to the semantic inadequacy of the AI but also to the naturalness of the form of questioning offered by the system, i.e., its capacity to engage in human-like interaction. Rather than a technical failure, such a "problem of comprehension" may be due to the artificial system producing a sufficiently humanoid form of language that renders its artificiality invisible. This reflects the success of large language models (LLMs) in interactive systems (Table 3).

The findings regarding the statement "Similar questions came over and over again." are M = 4.40, SD = 1.26, Md = 5.0, Min = 1, Max = 5, $Q1$ = 4.25, *and* $Q3$ = 5.00. The fact that 75% of the participants responded positively to this item creates the impression that the questions posed by the system are not sufficiently diversified in terms of content and fall into repetition at the perceptual level. The fact that 75% of the participants chose the two most positive options (4 and 5) indicates that the questions posed by the artificial intelligence have paternal similarities, and users can easily recognize this situation. This finding indicates that the system must improve semantic depth and variation generation. However, it should be noted that repetitive questions should not always be considered negative, as this may also be due to the fact that the system intentionally posed similar questions in a controlled manner while generating predictive hypotheses (Table 3).

The mean of the participants' responses to the statement "Artificial intelligence tried to guess the color" was found to be relatively low (*M* = 1.50, *SD* = 0.53, *Md* = 1.5, *Min* = 1, *Max* = 2, *Q1* = 1.00, *Q3* = 2.00). This value indicates that most participants either did not notice the AI's prediction behavior or the system failed to present this function clearly and distinctly enough. Thus, there is a serious lack of transparency in terms of how the prediction process is reflected to the user. The player's ability to intuit how the system works, from which information it makes predictions, and when it switches to generating a response is critical to the credibility of the artificial agent with which they interact. Therefore, the prediction process should not only be algorithmic but also be supported by clear and visible experiential feedback to the user.

Responses to the statements "I did not realize how time passed while playing the game." and "The game was very immersive." were evaluated with equal scores as *M* = 1.40, *SD* = 0.70, *Md* = 1.0, *Min* = 1, *Max* = 3, *Q1* = 1.00, *and Q3* = 1.75, respectively. These low averages indicate that users could not enter a flow state while experiencing the game and were not sufficiently involved in the game in terms of attention and time perception (Kaye, 2016). When evaluated within the framework of flow theory, the lack of cognitive intensity revealed that the system was limited in immersing the user and creating a sense of continuity. In particular, the low immersion and time perception parameters indicate that the game's interactional dynamics or esthetic experience is not at a level that satisfies the player.

It is also noteworthy that some participants subjectively perceived the experience as low, even though they had played the game for a long time in terms of session data. This is also important in terms of the disconnect between subjective user experience and objective playing time. Although the participants spent time, this time did not turn into a meaningful experience, suggesting that the in-game tasks or the responses received from the artificial intelligence did not have emotional or cognitive impact.

In conclusion, these statistical findings reveal that users perceive the game system as semantically repetitive, cognitively weak, and ambiguous. The main areas for improvement include increasing the semantic diversity of the AI-generated content, clearly visualizing and intuitively communicating the prediction process to the user, and designing experiential feedback mechanisms to make the interaction sustainable and satisfying. Such improvements will increase not only the system's technical capacity but also the value it offers in terms of user experience.

Game 2 (FLAN-T5)

The user interaction times and productivity levels were analyzed separately for each difficulty level (easy, medium, hard) of the AI-assisted color guessing game in Game 2. In this game version, a pretrained large language model (LLM) was used for fine-tuning to provide users with a more controlled, meaningful, and relevant question-answer flow. In this context, the dataset includes the total time (in seconds) and the number of questions for 10 participants at easy, medium, and hard levels. The total number of question interactions with the 10 participants ranged from 170 to 240 (Table 4).

According to the statistical analysis results, the average time spent at the easy level was only 1.0 minutes (SD = 0.5), and 10.1 questions (SD = 1.7) were asked at this level. The low duration and limited number of questions at this level indicate that the model attempts to reach the goal in a shorter time with less information. At the same time, because the structure of the questions at this level was more direct and goal-oriented, the repetition rate was low, and users generally found the questions understandable (Table 5).

Table 4

Detailed Interaction Time and Question Count Across Difficulty Levels in Game 2 (FLAN-T5)

Participant ID	EASY		MEDIUM		HARD		TOTAL	
	SEC	Q COUNT	SEC	Q COUNT	SEC	Q COUNT	SEC	Q COUNT
1	139.67	13	3634.08	116	489.24	41	4262.99	170
2	56.64	9	11251.97	152	477.20	47	11785.81	208
3	51.17	10	4052.21	22	338.27	35	4441.65	178
4	54.11	9	3954.56	168	290.98	41	4299.65	218
5	48.41	9	3026.96	136	277.91	31	3353.28	176
6	72.15	13	3387.60	150	202.96	28	3662.71	191
7	52.87	10	2836.00	133	314.31	32	3203.18	175
8	44.98	8	1407.68	89	271.18	31	1723.84	128
9	49.20	10	3786.60	187	224.25	30	4060.05	227
10	52.14	10	4111.97	176	454.31	54	4618.42	240

A significant jump was observed in the frequency of medium interactions. The average duration was 69.1 minutes (SD = 43.7), and the average number of questions posed was 132.9 (SD = 48.5). This high variance suggests that the model achieved a certain stable text flow through fine-tuning; however, its productivity remained flexible according to the players' responses. Participants reported that they encountered meaningful and contextually coherent questions at the medium difficulty level and that the AI's prediction process became more visible at this level, where the sense of repetition was low. This indicates that the model successfully generated differentiated questions while maintaining the semantic context (Table 4).

The hard level was completed in an average of 5.6 minutes (SD = 1.7) with an average of 37.0 questions (SD = 8.5). At this level, according to the data reported by the participants, the model maintained focused prediction behavior, but repetition was observed. These results indicate that the model may experience semantic contraction in scenarios requiring more specificity at the difficult level; however, it is nevertheless able to maintain its productivity (Table 5).

When the overall distribution is analyzed, the interaction times and the number of questions at the intermediate level are observed to be significantly higher than at the other levels. This shows that the fine-tuned LLM model provides a relevant, meaningful, and fluid game experience, especially at the medium difficulty level. At this level, where players experienced a change in their time perception, a flow experience was provided throughout the game. When evaluated within Mihaly Csikszentmihalyi's Flow Theory, the skill-difficulty balance was established at an optimum level. Participants became more involved in the game through clear goals and constant feedback and reported losing their perception of time.

In this context, the LLM-assisted version of Game 2 demonstrated both AI-controlled content generation and high player engagement. In particular, intermediate-level data demonstrate that a system structured to optimize productivity and interaction time functions successfully.

Table 5

Summary of Average Duration and Question Count by Difficulty Level (in Minutes)

Difficulty Level	Average Duration (min)	Std. Deviation (min)	Average Question Count	Std. Deviation (count)
Easy	1.0	0.5	10.1	1.7
Medium	69.1	43.7	132.9	48.5
Hard	5.6	1.7	37.0	8.5

Model Performance and Training Findings

The training outputs obtained because of the model's fine-tuning process demonstrate that the system has successfully acquired goal-oriented production competence. The training loss and validation loss values monitored during each training epoch demonstrate that the model improved steadily in terms of semantic generalization and task fidelity.

As shown in Table 1, the training loss value, which was 3.50 at the beginning, decreased significantly within only a few epochs, reaching 0.0317 at the end of the 10th epoch. Similarly, the validation loss, which was 2.17 in the first phase, decreased to 0.0197 at the end of the training, indicating that the model moved away from a rote structure and reached the ability to produce generalizable knowledge (Table 6).

In this process, the fact that the model does not show any signs of overfitting and that the training and validation losses are close to each other indicates that the system has gained production power for training data and wider probability sets. In addition, system information about the training process (sampling rate, FLOPs value, etc.) is summarized in Table 2, and these values demonstrate the technical efficiency of the process.

Table 6

Epoch-Based Training and Validation Loss of the Fine-Tuned FLAN-T5 Model

Epoch	Training Loss	Validation Loss
1	3.5011	2.1697
2	0.6172	0.2374
3	0.1344	0.0559
4	0.0639	0.0236
5	0.0417	0.0210
6	0.0329	0.0213
7	0.0338	0.0204
8	0.0275	0.0196
9	0.0316	0.0198
10	0.0317	0.0197

These results suggest that the contextual orientation goals described in the previous section were directly reflected in the educational outcomes of the model. The model has achieved the capacity to provide grammatically correct structures and semantically diverse and contextually coherent productions. This defines the main technical achievement behind the more meaningful, less repetitive, and more strategic question productions observed in the game flow (Table 7).

Table 7

Technical Summary of the FLAN-T5 Model Fine-Tuning Process

Metric	Value
Total Training Time	10,485.35 seconds
Training Sample Rate	0.858 samples/second
Training Step Rate	0.215 steps/second
Total FLOPs (Floating Point Ops)	1.54×10^{16}
Total Training Steps	2250

Game 2 Survey Evaluation

Following the game interactions, user experience data were collected through a structured questionnaire with a 5-point Likert-type scale (n = 10).

Table 8

User Experience Survey Results for Game 2 (FLAN T-5)

Question	Number of Participants (n)	Mean (M)	Standard Deviation (SD)	Minimum (Min)	1st Quartile (Q1)	Median (Md)	3rd Quartile (Q3)	Maximum (Max)
I immediately understood the questions asked by the AI.	10	2.7	1.1	1	2	2.5	3.8	4
Similar questions were repeated repeatedly.	10	3.8	0.8	2	4	4	4	5
The AI tried guessing the color.	10	3.4	0.8	2	3	3	4	5
I lost track of time while playing the game.	10	3.5	0.7	2	3	4	4	4
The game was very engaging.	10	3.5	0.7	2	3	4	4	4

The obtained data reveal the impact of artificial intelligence (AI) on user experience using quantitative indicators and make the system's strengths and areas open to improvement visible in a concrete way. "I understood the questions asked by artificial intelligence (AI) immediately." Mean $M = 2.7$, standard deviation $SD = 1.1$, median $Md = 2.5$, minimum = 1, maximum = 4, and $n = 10$. Although this value indicates that, in general, the participants had partial difficulty in understanding the questions generated by the AI, this can be interpreted as the result of an experience in which the artificiality was not felt and a language setup that was quite close to human, rather than the failure of the system. In other words, the fact that some participants could not fully understand the questions could also be attributed to the AI's questioning style being quite realistic. Therefore, they felt that they were interacting with a real human being, not the artificial system. This result indirectly indicates that the semantic production capacity of the system based on the big language model (LLM) is strong (Table 8).

"Similar questions came repeatedly." The mean values for this statement were $M = 3.8$, $SD = 0.8$, $Md = 4$, minimum = 2, and maximum = 5. The participants' high level of agreement with this item suggests that the AI system cannot provide sufficient diversity in question production and follows an inquiry strategy based

on certain patterns. While this finding reveals that the system can be further improved regarding contextual richness, it also demonstrates that the model establishes a consistent and systematic question structure. In this context, repetitive questions can be considered not a weakness of the system but a natural outcome of the hypothesis-based prediction process (Table 8).

"Artificial Intelligence (AI) tried guessing the color." The calculated mean for this statement is *M* = 3.4, *SD* = 0.8, *Md* = 3, minimum = 2, and maximum = 5. These data reveal that participants gave a mostly positive evaluation; however, some individuals did not find the system's prediction process sufficiently visible or understandable. In this case, it is important for AI to ask questions and to be able to reflect the prediction process and logic more clearly. In the context of cognitive guidance and process tracking experienced by participants during the game, the transparency of the prediction systems was a factor that directly affected the user experience (Table 8).

"I lost track of time while playing the game," and "The game was very immersive." The means of the responses to the items were measured as *M* = 3.5, *SD* = 0.7, and *Md* = 4 for both items, respectively. The minimum value = 2, and the maximum value = 4. These findings suggest that users developed high levels of attention and interest while experiencing the game and experienced a partial shift in their time perception.

These results are significant when evaluated in the context of flow theory. The AI interaction offered by the game supports cognitive continuity by generating a constant curiosity and desire to produce answers by the user. This situation restricts the player's perception of time. The positive evaluation of the immersion level of the game shows that the system offers a satisfying structure not only technically but also on an esthetic and experiential level.

The survey results demonstrate that the LLM-based AI system has gained a meaningful and substantial place in game experience. The participants' answers to the questions demonstrate that the system is successful in terms of *naturalness* in language production, continuity in interaction, and overall integrity of the game. However, the system could be improved to offer more variety in question generation and to express the prediction process more clearly, which would further enhance the user experience. This study provides a pioneering example of how LLMs can play an active role in text generation, in-game interaction, guidance, and meaning-making processes.

Entropy and User Experience Analyses

In this study, entropy is used to evaluate the production behavior of large language models (LLMs) regarding diversity and decision-making consistency; in particular, Shannon Entropy is considered a basic measure of information theory. Developed by Claude Shannon, it allows for calculating the average amount of information carried by a message or system output and the level of uncertainty in this output (Shannon, 1948). The higher the entropy value, the more diverse and unpredictable the distribution; the lower the entropy value, the more homogeneous and repetitive the user responses to the system outputs. In this context, Shannon Entropy was used in the dynamic challenge management (DDA) process to evaluate the semantic diversity and contextual richness of the content provided by artificial intelligence systems and their impact on the player experience.

Table 9

Comparison of Shannon Entropy Values for User Experience Criteria in Two AI-Supported Game Prototypes

Metric	Game 1: Cohere	Game 2: FLAN-T5
Question Clarity	1.846	1.846
Perceived Repetition	1.295	1.357
Prediction Clarity	0.971	1.685
Flow Experience	1.157	1.361
Engagement	1.157	1.361

The analysis findings based on Shannon entropy revealed that the two different AI-powered game prototypes (Game 1 - Cohere, Game 2 - FLAN-T5) exhibited significant differences in user experience diversity. For the "Question Clarity" criterion, equal entropy values (1.846) were obtained for both games. This indicates that although the FLAN-T5 model had a higher average score, the distributional diversity was similar in both models. On the "Perceived Repetition" criterion, the FLAN-T5 model demonstrated higher entropy (1.357 vs. 1.295), indicating that some users perceived the model as repetitive and diverse by others. The most striking difference was observed in the "Prediction Clarity" metric, where the FLAN-T5 model produced a more flexible and multi-layered experience than the Cohere model (0.971), with an entropy value as high as 1.685. Similarly, FLAN-T5 showed higher entropy values in the "Flow Experience" and "Engagement" criteria (1.361 vs. 1.157). This suggests that the system offers experiences of different intensities to a broader user profile, i.e., it creates a personalized level of engagement. These findings show that Shannon Entropy is not only a statistical measure of dispersion but also a powerful analysis tool that quantitatively measures diversity in player perception and differences in experiential impact (Table 9).

Comparative Evaluation of User Experience Metrics between Game 1 (Cohere AI) and Game 2 (FLAN-T5)

The user experience comparison between Game 1 (Cohere AI) and Game 2 (FLAN-T5) reveals how the two systems differ in terms of player perception, interaction quality, and emotional connection. The proposed FLAN-T5 model exhibits significant superiority in terms of all five user experience metrics. In particular, the +2.1 point difference recorded in experiential measures such as time perception (flow experience) and immersion shows that this model offers a more balanced and satisfying structure, technically and experientially (Table 10).

The questions presented by FLAN-T5 were clearer (+0.6) and less repetitive (-0.6), thus reducing the players' cognitive load and increasing the sense of novelty in the interaction. One notable difference was that users perceived the AI's prediction behavior more clearly (+1.9). This reflects the model's success in not only asking questions but also making the prediction strategy intuitive to the player (Table 10).

In particular, the high scores obtained in the flow experience and immersion scores indicate that the FLAN-T5 model can effectively provide dynamic difficulty management (DDA) and allow the player to have experience in accordance with the skill-difficulty balance. These findings confirm the potential of controlled, diversified language output to improve production quality, interactional continuity, and player satisfaction.

This analysis supports the table and demonstrates that the relationship between output entropy and user experience is multi-layered. There is a direct link between technical variety and meaningful, guided, and satisfying gaming experiences.

Table 10

Comparative Evaluation of User Experience Criteria Between Game 1 (Cohere AI) and Game 2 (FLAN-T5)

User Experience Metric	Game 1-Cohere AI (n=10)	Game 2 - FLAN-T5 (n=10)	Difference (Δ)	Interpretation
1. Question Clarity	2.1	2.7	+0.6	FLAN-T5 is more understandable
2. Perceived Repetition	4.4	3.8	−0.6	FLAN-T5 exhibited less redundancy
3. Predictive Behavior Clarity	1.5	3.4	+1.9	FLAN-T5 is significantly more transparent than
4. Flow Experience (Time Perception)	1.4	3.5	+2.1	FLAN-T5 facilitates more substantial flow experience
5. Engagement	1.4	3.5	+2.1	FLAN-T5 is more engaging

Analysis of t-Test Results for Independent Samples

The results of the independent samples t-test demonstrate significant differences between the two AI-powered game prototypes (Game 1-Cohere AI and Game 2 - FLAN-T5) in terms of user experience metrics. In comparing five main user experience metrics, statistically significant differences were observed for four metrics, while only one metric showed results on the borderline of significance.

In the first criterion, "The comprehensibility of the questions posed by the AI", the mean scores of the user group belonging to the FLAN-T5 prototype were higher (M = 2.7), and this difference was significant at the $t(18) = 2.17$, $p = .043$ level. This result suggests that the fine-tuned model produced clearer and more comprehensible questions, thus supporting user cognition more effectively (Table 11).

The t-value for the criterion "perception of repetition of similar questions" is 5.19 and $p < .001$. This difference indicates that FLAN-T5 can produce less repetitive, diversified semantic outputs. The contextual production strategy of the model prevents users from experiencing cognitive fatigue due to repetitive patterns.

Table 11

Independent Samples t-Test Results Comparing User Experience Metrics Between Game 1 and Game 2

User Experience Metric	t-value	p-value
I immediately understood the AI questions	2.17	.043
Similar questions were repeated	5.19	< .001
The AI attempts guessing the color	−4.57	.0002
I lost track of time while playing the game.	−5.33	< .001
The game was very engaging	−5.33	< .001

The highest effect size was observed for "Visibility of the AI's prediction behavior" (t (18) = −4.57, p = .0002). FLAN-T5 users observed the system's predictive behavior more clearly, suggesting that production transparency positively affects the user experience. This finding emphasizes that not only performance but also the understandability of the process is a critical parameter in DDA (Dynamic Difficulty Adjustment) systems (Table 11).

Similarly, strong differences were observed for experiential measures such as "change in time perception" and "immersion" ($t(18) = −5.33$, $p < .001$ for both). This suggests that the FLAN-T5 model created an intense experience that enabled participants to maintain attentional focus and change their perception of time

during the game. When considered in the context of Flow Theory, these findings confirm that FLAN-T5 supports the experience of "flow" by providing a skill-challenge balance (Kaye, 2016).

In summary, these statistical findings reveal that the retrained FLAN-T5 model outperforms not only in terms of production quality but also in terms of cognitive and emotional parameters that shape user experience. It is concluded that accuracy and multi-layered criteria, such as transparency, diversity, meaningfulness, and experiential satisfaction, should be optimized simultaneously in the design of LDA systems.

Discussion

The comparative analysis of two different artificial intelligence-based game prototypes developed in this research evaluates the contributions of systems based on large language models (LLM) in the context of dynamic challenge management (DDA) on the axis of each hypothesis. The obtained data are discussed below in line with the results of user experience surveys, entropy measurements, and independent sample t-tests.

H1a: LLM-based prediction strategies enable AI systems to dynamically adjust in-game difficulty levels in real-time.

This hypothesis is supported by the semantic guidance, time-balanced interaction, and high flow scores of the FLAN-T5 model, especially at the medium difficulty level. The fact that participants reported experiencing a shift in time perception (M = 3.5) and that the system could make meaningful transitions in question generation indicates that dynamically structured difficulty adjustment was successful. In this context, hypothesis H1a was confirmed.

H1b: LLM-supported AI systems can make strategic decisions at different difficulty levels.

The data for Game 2 showed that the model could change strategy at easy, medium, and hard difficulty levels thanks to the consistency in production and clarity in prediction behavior. In particular, the high productivity and low repetition rate observed at the medium difficulty level indicate that the system can strategize based on the participants' responses. Specific and goal-oriented production at difficult levels also indicates strategic flexibility. Therefore, hypothesis H1 b was also supported.

H1c: LLM-based AI predictions have lower response times and higher decision-making flexibility than traditional machine learning algorithms.

When the prediction process of both systems was analyzedT5 tended to reach the target with fewer questions and was perceived more visibly as a "guessing agent" by the participants. In particular, while the average score of the responses to the statement "AI tried to guess the color" was M = 1.5 in the Cohere model, it was M = 3.4 in the FLAN-T5 model. This demonstrates that FLAN-T5 both produces semantically clearer questions and makes more flexible decisions in the prediction process. These findings supp

H1d: The output entropy of LLMs increases the accuracy and adaptability of AI-based prediction strategies.

The analysis results based on Shannon entropy demonstrate that the FLAN-T5 model has higher entropy values in almost every criterion, and this diversity positively affects the game experience. Especially in the "Prediction Clarity" criterion, the entropy value was calculated as 1.685 for FLAN-T5, whereas it was only 0.971

for the Cohere model. This difference reveals that entropy is a determinant not only of technical diversity but also experiential diversity. Accordingly, hypothesis H1d is confirmed.

H1e: LLM-based LDA systems provide higher consistency in game flow and player engagement than traditional methods.

This hypothesis was supported by the high mean scores (M = 3.5 for both) and statistical significance ($t(18) = -5.33$, $p < .001$) of the user group of the FLAN-T5 model on the measures of "flow" (perception of time) and "engagement" (immersion). Flow experience was measured using indicators such as keeping users' attention in the game and not realizing that time had passed. At the same time, engagement was assessed through interaction continuity and attention intensity. These findings suggest that the FLAN-T5 model offers experiential consistency through its semantic production success and its more stable and flowing interaction structure with the user. Thus, hypothesis H1e is empirically confirmed.

H0: LLM-based AI systems do not provide significant improvement in dynamic challenge management.

Hypothesis H0 is rejected considering all findings. The survey data, t-test results, and entropy analysis revealed that LLM-supported systems provide a more guided, varied, and user-centered experience.

The empirical analysis revealed that artificial intelligence systems based on large language models (LLMs) can be effectively evaluated for dynamic difficulty adjustment (DDA) strategies in digital games. The comparison of two different models, Cohere Command and FLAN-T5, provided an opportunity to comprehensively examine not only technical production differences but also key parameters that determine the player experience, such as semantic cohesion, interaction quality, and flow continuity.

The primary user experience differences between the two prototypes are summarized in the table below. The comparative data in the table present directly observable results in line with the findings detailed in the discussion section (Table 1).

Table 12

Comparison of user experiences

Game 1 (Cohere AI) (n=10)	Game 2 (FLAN-T5) (n=10)
Questions were less clear; participants had difficulty understanding them.	The questions were clearer and easier to understand.
High sense of repetition; questions were very similar.	Less repetition; questions were more varied.
The prediction behavior was unclear; most participants did not notice it.	The prediction process was more visible; most participants noticed it.
The game was not fluent; participants did not lose track of time.	A flow experience occurred; participants lost track of time.
The game was less engaging; interaction felt limited.	The game was more engaging; it provided an attention-grabbing experience.

The Cohere Command model is based on zero-shot prompting; thus, its production behavior is limited. Entropy analyses show that this model offers low diversity, especially in terms of the visibility of prediction behavior (Prediction Clarity Entropy: 0.971) and high question repetition (Perceived Repetition Entropy: 1.295). In addition, the Flow experience performed poorly (Flow Entropy: 1.157), indicating that the model's interaction with the players was limited and far from satisfying. In some sessions, it was observed that

content that was disconnected from meaning, lacked contextual integrity, and was hallucinatory was generated; this negatively affected the players' decision-making processes and increased their cognitive load. The survey findings also support these results: Participants reported low clarity (M = 2.1), high repetition (M = 4.4), and poor flow (M = 1.4). It appears that a high production volume does not guarantee content quality but rather negatively affects the quality of interaction. In addition, users reported that they were often unable to notice the AI's predictive behavior, indicating a lack of transparency in the interaction process (see Table 1, row 3).

On the other hand, the FLAN-T5 model, customized by fine-tuning, produced more balanced, contextualized, and guided questions. In particular, the entropy value obtained at the medium difficulty level (0.489) indicates that the model produced meaningful, diverse, and semantically coherent questions. Participants perceived the prediction process as more open (M = 3.4), experienced a shift in time perception (M = 3.5), and found the game immersive (M = 3.5), suggesting that this model offered a stronger structure in terms of experience. These data suggest that users rated FLAN-T5 as a more understandable, fluid, and interactive system (see Table 1, rows 1 and 4).

Particularly noteworthy is that meaningful but not directly predictive questions in the FLAN-T5 game prototype (e.g., "Does this color occur frequently in nature?") maintained player focus and prevented hallucinatory content. By smoothing the transition between difficulty levels, such questions balanced the player's cognitive load and ensured a more controlled DDA process. This structure can be described as a "difficulty buffer" in the literature.

Moreover, the more explicit and intuitive realization of the prediction behavior in the FLAN-T5 model positively affected the game functioning and the strategic relationship that the player established with the system. This suggests that DBDA systems should be designed not only at the algorithmic level but also in a way compatible with player perception. On the other hand, in the Cohere Command model, the uncertainty of the prediction process, the inability to feel the difficulty transitions, and the low flow experience reveal that the system's lack of transparency weakens the game experience.

As a result, LLM strategies make sense not only in terms of the amount of content production but also in terms of semantic quality, contextual coherence, and player intuitiveness of the production process. In this context, the following conditions must be satisfied for LLM-based systems to provide a successful LDA experience:

1. Generate meaningful content with a balanced level of entropy.
2. Strategic management of contextual transitions
3. Balancing production processes with meaningful intermediate content
4. The prediction process is presented to the player clearly and intuitively.

Although the FLAN-T5 model met these conditions to a large extent and provided a more experientially successful structure, the Cohere Command model failed to provide interaction integrity despite its production volume. The research provides a powerful model of how LLM-based AI systems can be optimized in DDA processes with technical accuracy, semantic routing, flow continuity, and controlled entropy management centered on player experience (Table 12).

These findings are consistent with the theoretical framework of the study. Participants' feedback, such as shifts in time perception, concentration of attention, and feelings of productivity, fulfills the ideal experience conditions described in Csikszentmihalyi's (1990) Flow Theory (Cowley et al., 2008). In particular, the fact that

the FLAN-T5 model posed meaningful and goal-oriented questions to the player supported the continuity of the flow state by balancing difficulty levels with skill levels. However, when evaluated within the framework of Sweller's (1988) Cognitive Load Theory, repetitive, semantically incomplete, or hallucinatory productions reduced the quality of experience and caused distraction in the Cohere model. Moreover, when evaluated within Ryan and Deci's (2000) Self-Determination Theory, the FLAN-T5 model reinforced intrinsic motivation by offering players more autonomy, competence, and interaction. Overall, large language models can be repositioned not only as producers of technical output but also as dynamic actors providing semantic guidance, experience management, and flow continuity. However, besides these strengths, some practical limitations in integrating LLMs into game systems should also be considered. In particular, most LLMs work with credit-based systems through API access, and each user interaction incurs a certain cost. This makes economic sustainability difficult in highly interactive game scenarios and can be limiting, especially for independent developers and research-oriented projects. Furthermore, the dependency of LLM systems on technical components such as server response time, connection stability, and data security can lead to problems such as delays or response inconsistencies in real-time gameplay. For these reasons, the large-scale integration of LLM-based systems into games requires optimization of not only model performance but also economic accessibility and technical compatibility.

Conclusions and Recommendations

By comparing the game-based production performance of the zero-shot prompting Cohere Command model and the semantically fine-tuned FLAN-T5 model, this research demonstrates that large language models (LLMs) should be evaluated not only in terms of output accuracy but also in terms of contextual directiveness, semantic variety, and experiential impact. The findings demonstrate that the FLAN-T5 model, in particular, makes the interaction more sustainable, flowing, and satisfying by generating more meaningful and goal-oriented questions, reducing repetition, and making the prediction process more explicit and intuitive to the player. Accordingly, it is suggested that artificial intelligence, especially in narrative-driven or role-playing games (RPGs), should not be limited to text generation but should be integrated with high-level cognitive processes such as meaning-making, contextual guidance, prediction strategy development, and interaction continuity. Transparently perceivable prediction behavior of the AI by the player plays a critical role in system reliability and quality of experience. When evaluated in the context of Flow theory, production formats that provide flow, especially at medium difficulty levels, support high user engagement by balancing the experience. In this framework, game systems should be developed with technical success and structures that can immerse the user on cognitive and emotional levels. Future studies suggest that scenarios should be developed enriched with sensory inputs such as voice commands, gesture recognition, and emotional state analysis, where LLM-based artificial intelligence systems can be integrated with multi-modal data inputs. Integrating such systems into in-game narrative structures and dialog systems will enable more human, intuitive, and seamless interaction between AI and users. Furthermore, the design of visual interfaces that can intuitively convey the decision logic of artificial intelligence to the player, mechanisms that test player perception with gamified experimental scenarios, and adaptive system architectures that manage the production process with semantic filters will deepen the research in this field at both technical and experiential levels. In this context, this study fills an important theoretical and practical gap by opening up the behavioral and experiential performance of different LLM architectures in game-based DDA scenarios, which have not yet been systematically compared in the literature. The study strongly suggests that the

role of AI in digital game design is not only as a generative tool, but also as an active and redefined design component in the construction of meaning, guidance, and experience.



Ethics Committee Approval: Ethics committee approval was received for this study from the ethics committee of Doğuş University (Date: 28.05.2025, No: 82320).
Informed Consent: Written informed consent was obtained from all participants who participated in this study.
Peer Review: Externally peer-reviewed.
Conflict of Interest: The author has no conflict of interest to declare.
Grant Support: The author declared that this study has received no financial support.

Author Details
Onur Aşkın
[1] Doğuş University, Faculty of Art and Design, Department of Digital Game Design, İstanbul, Türkiye
0000-0002-1928-474X oaskin@dogus.edu.tr

References

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165.* https://doi.org/10.48550/arXiv.2005.14165

Cameirão, M. S., Bermúdez i Badia, S., Duarte, E., & Verschure, P. F. M. J. (2009). *The Rehabilitation Gaming System: A review. Studies in Health Technology and Informatics, 145*, 65–83. https://doi.org/10.3233/978-1-60750-018-6-65

Chen, J. (2007). Flow in games (and everything else). *Communications of the ACM*, 50(4), 31-34. https://doi.org/10.1145/1232743.1232769

Colwell, A. M., & Glavin, F. G. (2018). Colwell's castle defence: a custom game using dynamic difficulty adjustment to increase player enjoyment. *arXiv preprint arXiv:1806.04471.*

Cowley, B., Charles, D., Black, M., & Hickey, R. (2008). Toward an understanding of flow in video games. *Computers in Entertainment (CIE)*, 6(2), 1-27. https://doi.org/10.1145/1371216.1371223

Csikszentmihalyi, M. (1990). *Flow: The psychology of optimal experience*. Harper and Row.

Csikszentmihalyi, M. (1997). *Finding flow: The psychology of engagement with everyday life*. Basic Books.

Csikszentmihalyi, M. (2014). *Applications of flow in human life*. Springer.

Fisher, N., & Kulshreshth, A. (2024). Exploring dynamic difficulty adjustment methods for video games. *Virtual Worlds*, 3(2), 230-255. https://doi.org/10.3390/virtualworlds3020012

Garcia-Ruiz, M., Montesinos-López, O. A., & Anido-Rifón, L. E. (2023). The use of deep learning to improve player engagement in a video game through dynamic difficulty adjustment based on skills classification. *Applied Sciences*, 13(14), 8249. https://doi.org/10.3390/app13148249

Gilleade, K. M., Dix, A., & Allanson, J. (2005, January). Affective videogames and modes of affective gaming: assist me, challenge me, emote me. In *Proceedings of DiGRA 2005 Conference: Changing Views: Worlds in Play*.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Harris, C. R., Millman, K. J., Van Der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., ... & Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357-362. https://doi.org/10.1038/s41586-020-2649-2

Kaye, L. K. (2016). Exploring flow experiences in cooperative digital gaming contexts. *Computers in Human Behavior*, *55*, 286-291. https://doi.org/10.1016/j.chb.2015.09.023

Keller, J., & Bless, H. (2008). Flow and regulatory compatibility: An experimental approach to the flow model of intrinsic motivation. *Personality and Social Psychology Bulletin*, 34(2), 196-209. https://doi.org/10.1177/0146167207310026

Kowlessar, T. (2020). *How Difficulty Affects Player Engagement in Digital Games* (Doctoral dissertation, Flinders University, College of Science and Engineering.).

Lopes, R., & Bidarra, R. (2011). Adaptivity challenges in games and simulations: A survey. *IEEE Transactions on Computational Intelligence and AI in Games*, 3(2), 85-99. https://doi.org/10.1109/TCIAIG.2011.2152841

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8), 9.

Roohi, S., Guckelsberger, C., Relas, A., Heiskanen, H., & Takatalo, J. (2021). Predicting game difficulty and engagement using AI players. *arXiv preprint arXiv:2107.12061.* https://doi.org/10.48550/arXiv.2107.12061

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. American Psychologist, 55(1), 68-78. https://doi.org/10.1037/0003-066X.55.1.68

Shannon, C. E. (1948). A mathematical theory of communication. Bell System Technical Journal, 27(3), 379-423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Sweetser, P., & Wyeth, P. (2005). GameFlow: A model for evaluating player enjoyment in games. *Computers in Entertainment (CIE)*, 3(3), 1-24. https://doi.org/10.1145/1077246.1077253

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2), 257-285. https://doi.org/10.1016/0364-0213(88)90023-7

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (pp. 5998-6008). https://doi.org/10.48550/arXiv.1706.03762

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., ... & Van Mulbregt, P. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, *17*(3), 261-272. https://doi.org/10.1038/s41592-020-0772-5

Wu, C.-H., Chen, Y.-S., & Chen, T.-C. (2017). *An adaptive e-learning system for enhancing learning performance: Based on dynamic scaffolding theory*. *Eurasia Journal of Mathematics, Science and Technology Education, 14*(3). https://doi.org/10.12973/ejmste/81061

Yannakakis, G. N., & Togelius, J. (2018). *Artificial intelligence and games*. Springer. https://doi.org/10.1007/978-3-319-63519-4

Zheng, T. (2024). Dynamic difficulty adjustment using deep reinforcement learning: A review. *Applied and Computational Engineering*, *71*, 157-162. https://doi.org/10.54254/2755-2721/71/20241633

İSTANBUL UNIVERSITY PRESS

Acta Infologica,
ACIN 2025, 9 (1): 208–222

https://doi.org/10.26650/acin.1673277
Submitted: 10.04.2025
Revision Requested: 28.04.2025
Last Revision Received: 06.05.2025
Accepted: 03.06.2025
Published Online 04.06.2025

Acta Infologica

Research Article

Open Access

Investigation of User Comments on Videos Generated by Deepfake Technology

Sertaç Kaya [1]

[1] İstanbul Aydin University, Faculty of Communication, Department of Journalism, İstanbul, Türkiye

Abstract The ease of content production with developing technologies has caused deepfake-based videos to gain popularity and prevalence on social media platforms. Deepfake technology can be used for both manipulation and entertainment purposes because it can imitate people's voices and faces realistically. This study aims to evaluate user perceptions of content produced using deepfake technology. For this purpose, we captured the comments of the first three videos with the most views on the relevant theme on YouTube using the Python programming language. From each video, top-level comments that received 100 or more likes, along with 50 randomly selected comments with fewer than 100 likes, were analyzed through content analysis to ensure the inclusion of both prominent and less visible user perspectives. As a result of the analysis conducted on the data set collected from the YouTube platform, it was observed that users experienced feelings such as admiration and surprise toward the content produced with deepfake technology, as well as fear and anxiety regarding potential risks.

Keywords Deepfake · Artificial Intelligence · Synthetic Content · Content Manipulation · Social Media

Citation: Kaya, S. (2025). Investigation of user comments on videos generated by deepfake technology. *Acta Infologica*, *9*(1), 208-222. https://doi.org/10.26650/acin.1673277

This work is licensed under Creative Commons Attribution-NonCommercial 4.0 International License.

© 2025. Kaya, S.

Corresponding author: Sertaç Kaya sertackaya1@aydin.edu.tr

Introduction

New digital technologies and social media platforms allow users to create, share, and consume content whenever and wherever they want (Kaplan & Haenlein, 2010; Kietzmann et al., 2011). In this way, people have the opportunity to express themselves and share their ideas, access alternative content in online environments, and interact with other users. Although this situation has increased the channels for accessing information, the abundance of information causes an increase in false and fake content (Chen et al., 2015; Marin, 2021; Mintz, 2012), and therefore, it causes difficulties for people to verify and trust information (Zoone, Luoma-aho & Lievonen, 2024). Recent developments in the field of artificial intelligence have also affected the production and consumption of information in the digital environment.

In recent years, deepfake technology has come to the forefront of content production and consumption by users. This technology uses artificial intelligence to imitate an existing image or sound, creating content that appears real but is not real. Thus, it is possible to change the faces of the people in the video, shape their speech, or create completely new content (Agarwal et al., 2020; Farid, 2022). Such videos are circulated on the internet and social media platforms, gaining popularity and reaching a wide audience. This situation, along with its advantages, also carries some risks. While creative and entertaining content can be produced using deepfake technology (Mihailova, 2021; Murphy, 2023), fake and manipulative videos (Tahir et al., 2021) can also be produced. The ease of producing both types of content makes it increasingly difficult to distinguish between real and fake media content.

This study aims to reveal users' attitudes toward video content produced by deepfake technologies and their general approach to such technology. This study aims to determine whether people perceive this type of content production as a threat or opportunity. The examined video comments are used to identify clues about people's thoughts about such videos, their views on their possible effects, and whether they can distinguish between real and fake content.

Literature Review

Deepfake is a machine learning-based software tool (Albahar & Almalki, 2019). The word is derived from the combination of the words “deep learning” and “fake” (Mirsky & Lee, 2021). This technology refers to techniques that enable the production of realistic-looking audio and video recordings of people doing things they do not do or say (Johnson & Diakopoulos, 2021; Ruiter, 2021). It began to gain popularity in 2017 when an anonymous user placed celebrities on adult content and shared it on the Reddit platform (Hern, 2018; Kietzmann et al., 2020). Therefore, the use of these techniques produces digitally manipulated hyper-realistic videos (Taulli, 2019). The production of such videos is becoming easier and cheaper day by day, and with developing technologies, they are becoming more convincing and realistic (Metz, 2019; Rubenking, 2019). Today, the ease of performing operations such as face replacement or synthesis is due to reasons such as accessibility to large-scale public data and the development of deep learning techniques (Tolosana et al., 2020). This phenomenon causes people with a large digital footprint to become more vulnerable to the production of their own copies (Bohacek & Farid, 2024). Those who produce deepfakes consist of curious amateurs who use technology for entertainment and personal interest, actors who use them for political purposes for disinformation and propaganda, malicious individuals who commit crimes such as financial fraud and phishing, and legitimate organizations that use them for entertainment, advertising, or artistic projects within the legal framework (Westerlund, 2019).

In their study, Tolosana et al. (2020) explained the manipulation made via deepfake through the classifications of entire face synthesis, identity swap, attribute manipulation, and expression swap. The entire face synthesis creates the entirety of non-existent facial images. Identity swap refers to the replacement of a person's face in the video with the face of another person. Attribute manipulation emphasizes facial editing or retouching operations, such as hair or skin color, gender, and age. Expression swap, on the other hand, means changing a person's facial expression. In another study, deepfake uses in the context of human images were classified into four categories. These are reenactment, replacement, editing, and synthesis (Mirsky & Lee, 2021):

Reenactment: The person’s expression, mouth, gaze, or body are directed. In the film and video game industry, well-intentioned uses are encountered in educational media where actors’ performances are adjusted in post-production processes and historical figures are re-enacted.

Replacement: It is a transfer process. This involves replacing one content with another. It can be used in the fashion industry to visualize a person in different clothes or to produce entertaining outputs by changing the identity of an actor to that of a famous individual. However, it is also frequently encountered in malicious uses such as slander and blackmail.

Editing: It refers to the points where some characteristics of people are added, changed, or removed. Examples such as changing the target's clothes, age, weight, and beauty. In applications such as FaceApp, people can be given an example of changing their appearance for entertainment. The same process can be used to make a sick leader appear healthy, or fake identities can be created to mislead others.

Synthesis: Synthesis involves the use of human face and body synthesis techniques without any goal. It can be used to create characters for movies and games and to create fake personalities online.

As with many technological innovations, deepfakes bring opportunities and risks. While they can be used advantageously in sectors such as film, gaming, entertainment, and fashion (Kietzmann et al., 2020), the same technology can also be used for malicious purposes, such as blackmailing individuals, discrediting individuals or institutions, and manipulating fake news (Chesney & Citron, 2019; Fletcher, 2018; Verdoliva, 2020). For example, actors or singers can be placed in erotic content (Wang, 2019), used for revenge pornography (van der Nagel, 2020), or used to smear public figures in various ways (Maddocks, 2020). The use of deepfakes in this direction can also incite political or religious tensions, deceive the public, and create turmoil in financial markets by spreading false information (Heidari, 2024). Such fake content can be successful for basic reasons, such as credibility and accessibility (Kietzmann et al., 2020). Because it is easier to access such videos on social media platforms, the possibility of confusing viewers by sabotaging and imitating the facts because of problems with misinformation and conspiracy theories increases (Hasan & Salah, 2019; Yazdinejad et al., 2020). For example, a study has shown that deepfake videos are more vivid, convincing, and reliable than fake news articles (Hwang, Ryu & Jeong, 2021). The content produced using this technique creates uncertainty in the mind of the user; therefore, trust in the news on social media may decrease (Vaccari & Chadwick, 2020).

It is becoming easier to create disinformation by imitating political figures or by presenting pornographic content with images of celebrities (Heilweil, 2020). It has been revealed that 96% of deepfakes found online are pornographic, and almost all of them belong to women (Abram, 2020). Another study demonstrated that users who encounter deepfakes depicting politicians significantly worsen their attitudes toward that politician (Dobber et al., 2021). The finding that deepfake videos increased from 7,964 to 14,678 between December 2018 and July 2019 (Duffy, 2019), the number of these videos exceeding 85,000 in December 2020

and doubling every six months since December 2018 (Petkauskas, 2023) is also important in terms of showing the speed at which such content spreads. In addition, whether the person or persons in the deepfake will object to the way it is represented, whether the deepfake deceives the audience, and the purpose of creating the deepfake cause the moral discussion of the content produced with artificial intelligence (Ruiter, 2021). Therefore, the main threat is that deepfake-based content weakens our trust in the reality of what we see and hear (Floridi, 2018) and can easily lead individuals to false beliefs (Fallis, 2021). Because people, including educated ones, have difficulty distinguishing between real and fake (Matern et al., 2019). Although people trust their own abilities, the credibility created by having seen it results in difficulty in detecting deepfakes (Köbis, Doležalová & Soraperra, 2021).

In order to understand a technology like deepfake, it requires an examination from the perspectives of media and society, media production, media representation, media audiences, gender, law, and politics. Because it shapes social perception, users are involved in content production. Artificial intelligence is conveyed in the media with entertainment and horror themes, and the ability to interact with the audience comes to the fore. In addition, the first deepfake content has a gender-based analysis due to the creation of pornographic videos of fake female celebrities and revenge porn directed at women, the emergence of initiatives to reduce possible risks, and political contexts due to the potential to increase uncertainty and reduce trust through fake political videos (Karnouskos, 2020).

In line with all these definitions, approaches, and discussions, this study aims to better understand people's feelings and thoughts by examining deepfake videos and user perceptions toward this technology. The analysis will examine how different cognitive and emotional reactions to deepfakes are reflected in user comments.

Methodology

With the rise of artificial intelligence (AI) and deep learning techniques, the circulation of fake content in digital environments is rapidly increasing. Therefore, the main purpose of the study is to reveal users' feelings and thoughts about videos produced with deepfakes. In this context, this study seeks an answer to the research question, "What are the general perceptions of users toward deepfake videos?".

This study's sample selection process was carried out based on certain criteria. In the search made on the YouTube platform, among the videos that were reached with the keyword "deepfake video" and had over 10 million views, those that were open to user comments were preferred. Thus, only content that reached a large audience and received high interaction was included in the analysis. The aim was to make more meaningful inferences about the general reactions of the public to such content. However, two of the five videos that met the specified criteria were excluded from the study; one was closed to comments, and the other was excluded because its musical structure made user comments focus on artistic content, not deepfake technology. This limited the number of samples to three videos.

In the study, the comments of three videos on the YouTube platform were captured using the Python programming language. Since it was intended to examine directly the comments made on the videos determined as a sample, "top-level comments" were determined as the basic unit to be analyzed. In the evaluation process of the comments, those that received 100 or more likes were selected to better understand the impact of the content on large audiences. The primary reason for analyzing only "top-level comments" with over 100 likes in this study is that such comments are more likely to reflect views that have gained visibility, support, and resonance among a broader audience. In fact, in the three videos analyzed, the total number

of likes on comments with 100+ likes was 212,431, 133,555, and 87,733, respectively. These figures indicate that the analyzed comments are not merely individual expressions but also reflect opinions that have found broader social resonance. From this perspective, the selected comments serve as a strong sample to represent dominant perceptions and emotional responses among users, thereby contributing to a more meaningful and accurate interpretation of the public perception of deepfake technology. Nevertheless, in order to increase the diversity of viewpoints and ensure the inclusion of potentially underrepresented voices, an additional set of 50 randomly selected comments with fewer than 100 likes were incorporated into the analysis of each video. This approach aims to go beyond the most prominent comments and achieve a more balanced and comprehensive understanding of user perspectives, thereby increasing the representativeness of the findings. The user comments were analyzed using content analysis. Themes were created inductively on the basis of findings directly obtained from the data. The repetitive expressions in the comments and expressions that were similar to meaning were grouped under the same theme.

In order to ensure validity and reliability within the scope of content analysis, two academics independently conducted the coding process. The coding process was continued until 100% agreement was achieved among the academics. The aim of this approach is to ensure the accuracy of the coding process and the consistency of the obtained data. In this context, five themes were identified for the first two videos, and another five different themes were identified for the third video.

Results

The comments were taken from the YouTube platform on 17.03.2025. The video data corresponding to the relevant date are presented in Table 1:

Table 1
Content Information of Videos

Post Title	Number of Views	Number of Comments	Number of Likes > 100
Bill Hader impersonates Arnold Schwarzenegger [DeepFake]	22,565,151	23,170	214
Bill Hader channels Tom Cruise [DeepFake]	14,590,306	16,484	153
You Won't Believe What Obama Says In This Video! 😊	10,047,828	8,851	102

As part of the study, 214 comments from the first video, 153 from the second, and 102 from the third were analyzed based on their high number of likes. In addition, 50 randomly selected comments with fewer than 100 likes were included in each video. In total, 619 user comments were examined through content analysis. The first two videos were uploaded to the YouTube platform by the account named "Ctrl Shift Face," while the third video was uploaded by the account named "BuzzFeedVideo." This situation causes differences in the content and context of the videos; thus, the thematic distribution of user comments also differs. For this reason, the comments of the first two videos were categorized under the same themes, while the comments of the third video were categorized under different themes.

Table 2

Theme and Comment Information of the Video "Bill Hader impersonates Arnold Schwarzenegger [DeepFake]"

Theme and Comment Information of the Video "Bill Hader impersonates Arnold Schwarzenegger [DeepFake]"	Number of Comments > 100+ Likes	Number of Comments Under 100 Likes
The Distortion of Reality Perception and Creepy Effect	85	28
Admiration for and Amazement at Deepfake Technology	51	8
Entertainment and Humor	54	6
Bill Hader	5	2
Future Risks of Deepfake Technology	19	4

Table 2 presents the thematic categorization of user comments on the video, providing insight into the distribution of reactions based on their content and engagement levels. The comments on the theme of distortion of perception of reality and creepy effects focused on how users felt when they realized how realistic the deepfake technology in the video was. Commenters expressed their feelings of being disconnected from reality with sentences such as "I really questioned my sanity.", "I thought I was having a stroke.", "I was about to drive myself to the emergency room.", "My mind was blown for about 5 minutes.", "Sh*t am I drunk?". Comments such as "That might officially be the creepiest thing I've ever seen...", "Jesus christ, this technology is terrifying." and "Its creepy how seamlessly the face swap is working." stand out as examples of users' fear.

The comments on admiration for and amazement for deepfake technology emphasize that this technology has made great progress and that the level of reality is high. Users state that the face swap process looks extremely natural and express their admiration for the created content. For example, a comment such as "From the thumbnail I thought that this was Arnold's son." indicates that the user found the face swap realistic and thought that it was an image of Arnold Schwarzenegger's son. In addition, comments such as "How did u do this. This looks pretty amazing." and "Bloody hell that's an impressive DeepFake, the transitions are really subtle." express the admiration felt by those who watched the video.

The theme of entertainment and humor comments focuses on users writing about the possibilities offered by deepfake technology in an exaggerated, funny, and sometimes critical language. Comments such as "If he had the Arnold Schwarzenegger Body. He would be him 90%", "He is the Arnold who didn't choose bodybuilding as a career." and "This is what a non bodybuilding Arnold Schwarzenegger would look like." are seen as ironic reflections of an iconic appearance being more ordinary or unusual. Here, users are attempting to make funny inferences from the new perspectives revealed by deepfakes. In addition, the expressions "And they spent millions for mustache in JL..." and "They should have removed henry cavills moustache like this." refer to the removal of Henry Cavill's mustache with CGI (Computer-generated imagery) in the 2017 film Justice League. In the film, the actor's mustache was digitally removed, but the result was not very successful. This comment also humorously expresses that deepfake technology can be more successful in this type of visual effects.

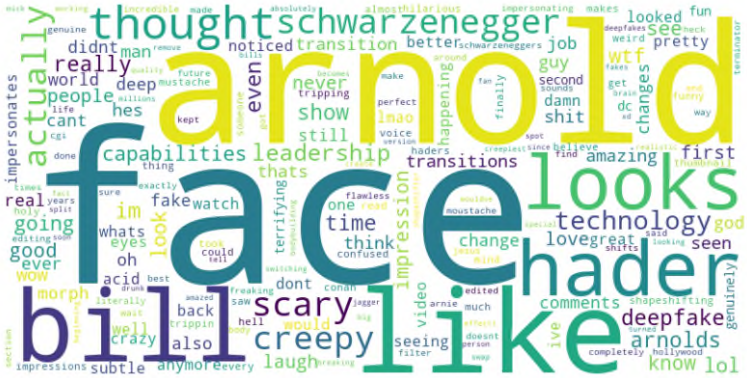
In the theme about Bill Hader, users highlight Bill Hader's ability to imitate a subject that is truly commendable, despite the impressiveness of deepfake technology. Comments such as "Huge fan of Bill Hader since Superbad." An even bigger fan now that he transformed into Arnold ;)", "I love how Bill Hader just enjoys himself. He's a real genuine guy." and "I love it when Bill giggles at himself. Always makes me laugh as much as when he's doing impressions." are prominent examples in this context. It can be seen that Hader's talent and approach, independent of deepfake technology, are praised in these comments.

Comments on the future risks of deepfake technology reflect serious concerns about its ethical, legal, and societal risks. Expressions such as “This is scary technology. Soon we won’t know if what we’re watching is real or fake.”, “Amazing and scary at the same time. It looks almost perfect. In 2-3 years max this will be beyond recognizable.”, “This is so absolutely creepy. Imagine how this will impact politics, especially now when you can’t trust anything the “media” says.” emphasize that it will cause people to question the accuracy of what they watch or hear. These expressions also indicate that we are entering a period in which the perception of reality will be shaken, trust in information will decrease, and people will be forced to question the content they watch. In addition, “The implications of possible misuse of this technology are terrifying!” “Deepfakes are going to be used to frame people one day. I think it's likely already happening.”, “Political deepfake video is what will cause world war 3.” comments like these reflect concerns about the potential misuse of this technology. These statements indicate that deepfake technologies can also be used for propaganda, false accusations, and disinformation. These increases concerns that the boundaries between real and fake will become increasingly blurred, which could have dangerous consequences for individuals and communities.

The comments that were randomly selected from the comments that received fewer likes in order to increase the variety of the comments also support the previous findings. It can be seen that users gave similar reactions in the comments. Especially in the comments under the title “Distortion of Reality Perception and Creepy Effect”, there are statements such as “Woah wtf I thought I was losing my shit...”, “What the heck did I just watch?”, “damn, it's both creepy and impressive.”, “fuckin scary...”, “This is... confusing... Stop doing that.”, “this is scary. but fun. but scary.” and “Trippy!” These statements demonstrate that users have difficulty believing what they see and the confusion this creates. The impressive aspect of the technology is emphasized with comments such as “this is the best deepfake i have seen”, and “I'm so impressed by the flawless camera trick.” in the category of “Admiration for and Amazement at Deepfake Technology”. Humorous comments and praise for Bill Hader's acting show that viewers find the content entertaining and admirable at the same time. However, some comments also raise serious ethical concerns by drawing attention to the risks that technology may pose in the future, especially manipulation and information security. Examples of this situation are comments such as “I think this technology will confuse the heck out of people rather soon..”, “Creepy shit .you could be framed for a crime you didn't do .fuck that ...”, “This was so funny but also so scary and just the thought of this tech being misused is very unsettling.”, “This is why you can't ever believe anything the news or TV tells u . It's really going to be bad for us.”. Two comments that were not directly related to the content were excluded from the analysis.

Figure 1

Word Cloud of Comments on “Bill Hader impersonates Arnold Schwarzenegger [DeepFake]” Video



As illustrated in Figure 1, the word cloud highlights the most frequently used words in user comments, reflecting viewers' emotional responses to the video. In addition, we determined that certain words reflecting the emotional reactions created by the content and deepfake technologies in the viewers were frequently used in the examined user comments. In particular, "creepy" and its derivatives (e.g. creepiest, creepiness, etc.) were used in a total of 23 comments, "scary" 18 times, "crazy" 5 times, and "terrifying" 4 times. These data demonstrate that viewers adopt emotional language when evaluating videos. In addition, these words, which are preferred to be used at a high frequency, also show that the video content and deepfake technology are perceived as disturbing, giving a sense of uncanny or unusual.

In addition, the use of words with positive value judgments, such as "Good" (9 times) and "Amazing" (7 times), demonstrates that users find content and technology esthetically and technically successful and gain their appreciation and admiration. Another striking point is the frequent use of expressions such as "Really" (11 times), "Wow" (8 times), "Wtf" (11 times) and "Confused" and its derivatives (7 times). This situation reveals that users feel astonishment, amazement, and sometimes even confusion toward the content and the technology.

Table 3

Theme and Comment Information of the Video "Bill Hader channels Tom Cruise [DeepFake]"

"Bill Hader channels Tom Cruise [DeepFake]" Video Themes	Number of Comments > 100+ Likes	Number of Comments Under 100 Likes
The Distortion of Reality Perception and Creepy Effect	39	16
Admiration for and Amazement at Deepfake Technology	51	11
Entertainment and Humor	42	8
Bill Hader	6	1
Future Risks of Deepfake Technology	15	10

Table 3 presents the thematic classification of user comments on the video 'Bill Hader channels Tom Cruise [DeepFake],' reflecting viewers' emotional and cognitive reactions, especially concerning the unsettling nature of deepfake technology. Commenting on the theme of distortion of perception of reality and creepy effect, expressions such as "Let's be real.....this can't be 'normal'", "That's crazy.", "Holy shit...this truly creepy.", "I instantly questioned my sanity." stand out. These comments demonstrate that the user's perception of reality has been shaken, and they are worried. One user's statement "If you hadn't put deep fake in the description I would be questioning reality because it is such a subtle transition." shows that they only grasped the truth through the comments of other users. This situation can be evaluated as an indicator that individual perception has reached a level that can be easily manipulated using deepfake technology. Another user's statement "Im a software engineer... im always amazed with new technology coming out...but this is scary." proves that even people with technical knowledge are worried about developing technologies.

Comments on the theme of admiration for and amazement toward deepfake technology include expressions such as "The way you seamlessly transition between Cruise and Hader's face is mind-bending, genuinely amazing stuff." and "This is without a doubt the most realistic deepfake I have seen so far." These comments can be evaluated as praise indicating how much deepfake technology has developed and how close it has come to reality. In addition, one user commented that "1:39 is such an impressive transition because it's so damn quick and still seamless. This is the best deepfake I've ever seen." indicates that the lines between real and fake are disappearing, and another user's statement "This is hands down, the

creepiest but the best and funnest and coolest shit ive seen in some time." shows both admiration for the technique and concern about its potential.

In entertaining and humorous comments, many users reference different series and movies and express their desire for this technology to be applied to such productions. Comments requesting that different actors' faces be added to productions such as Superman, Iron Man, Inglourious Basterds, and The Office highlight the entertaining aspects of deepfake technology and show that it is seen as a creative tool that can offer new and alternative versions to viewers. Another user commented, "Bill Hader has Mystique's powers, but only the face lol." Here, the user humorously emphasizes the effectiveness of deepfake technology by comparing Bill Hader's ability to skillfully change facial expressions to the powers of Mystique, the shapeshifting character in the X-Men series.

Comments on the theme about Bill Hader show that the actor is appreciated not only for the impressiveness of deepfake technology, but also for his extraordinary ability to imitate, modesty, and natural acting skills. Viewers express admiration for Hader, highlighting his talent regardless of the deepfake effects.

The theme of the future risks of deepfake technology highlights serious concerns about the potential for misuse of deepfakes and the risks they may pose in the future. "This technology should be destroyed, could ruin innocent people's lives.", "This is both amazing and creepy. And the technology is going to create a shit ton of distrust in media (or any) videos in the near future.", "We will never be able to believe anything we see on video.", "The potential uses for this technology are frightening.", "Just imagine how hard we've been already manipulated by this.", "This technology will be used to frame someone and most likely has. Scary future is in store for us.", "They're going to try to weaponize this for the 2020 election." These comments highlight the ethical and security concerns created by deepfake technology. Users believe that this technology can undermine media credibility, can be used to slander or manipulate people with fake content, and can have negative effects on political processes.

There are statements similar to the previous statements in the comments under 100 likes for this video. The comments under the title Distortion of Reality Perception and Creepy Effect illustrate the confusion, discomfort, and anxiety that technology creates in individuals. Comments such as "Yes ... Creepy", "That was creepy as hell!", "My brain is confused and hurt.", "This melted my brain", "This is so strange to watch" reflect these feelings. Comments such as "Awesome", "Incredible" and "How? How?" can be interpreted as indicators of feelings of admiration and surprise. On the other hand, the comments under the category of Future Risks of Deepfake Technology carry serious concerns about the ethical and social consequences of the technology: Comments such as "This is the beginning of the end of 'video proof'" or "This kind of technology is terrible and able to make a lot of disasters, should be forbidden" show that deepfake is perceived not only as a means of entertainment but also as a potential threat. The five comments examined in this context consist of expressions that were not directly related to the content.

Figure 2

Word Cloud of Comments on "Bill Hader channels Tom Cruise [DeepFake]" Video

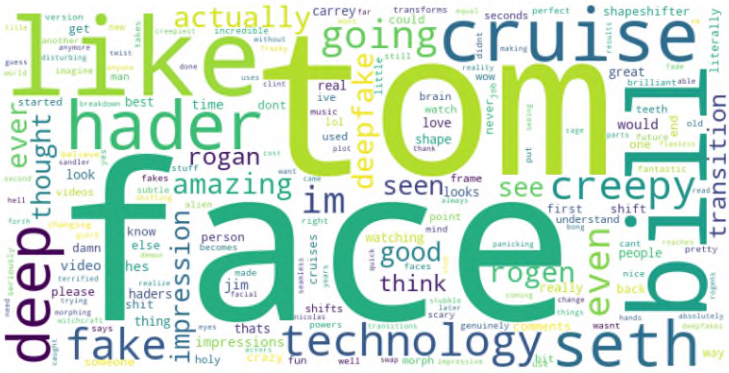


Figure 2 visualizes the most frequently used words in user comments on the video, illustrating the coexistence of both positive and negative emotional responses to deepfake technology. It is also noteworthy that expressions such as "Creepy" and its derivatives (15 times), "Crazy" (5 times), and "Scary" (6 times) were used in the comments. The preference for these adjectives that express discomfort and concern indicates that users find the deepfake technology in the video creepy or damaging to their perception of reality. On the other hand, it was determined that words with positive judgments such as "Good" (9), "Amazing" (10), "Love" (8), and "Great" (6) were also used with remarkable frequency for the same content and technology. This shows that users evaluate video and deepfake technology as technically successful and esthetically problem-free.

Table 4

Theme and Comment Information of the Video "You Won't Believe What Obama Says in This Video! 😊"

"You Won't Believe What Obama Says In This Video! 😊" Video Themes	Number of Comments > 100+ Likes	Number of Comments Under 100 Likes
Ironic and Critical Comments on BuzzFeed's Credibility	32	10
Obama-Focused Commentary	8	7
Internet Culture Humor	32	11
Media Literacy and Fake Content Warnings	15	15
Unexpected and Surprising Content	15	6

Table 4 presents the thematic distribution of user comments, revealing a range of responses including ironic critiques of source credibility, humorous engagement with internet culture, and concerns related to media literacy and misinformation. The theme of ironic and critical comments regarding BuzzFeed's reliability focuses on the source sharing of the content rather than the content of the video. In their comments, users express ironic, sarcastic, and critical reactions to BuzzFeed being presented as a "reliable news source." Comments such as "Is Buzzfeed reliable? That's a funny joke.", "*The last news source ill trust is buzz feed*", "It's ironic how it's coming from buzzfeed.", "Yea on the list of non trusted news sources buzzfeed is under cnn." stand out under this theme. In the comments, the source's history and content strategy are questioned, while its reliability as a source is criticized using a humorous and sarcastic tone.

In the Obama-focused comments theme, users express admiration for Obama's personal characteristics or general image. They are seen to express these feelings with comments such as "Please come back as a president!", "He still looks cool.", "We want you back Obama! We want you back Obama and I know i'm

honestly not alone ;) ♡♡." In addition, critical and humorous statements are also included among these comments and directly refer to Obama.

The comments on internet culture humor carry humorous, viral, and popular culture effects that reflect the typical characteristics of internet culture. In comments such as “Plot twist: That wasn't Jordan Peele either. It was actually Bernie Sanders.” and “And behind Jordan Peele was Kanye West, who was actually just Donald Trump, who was actually Beyonce doing an impression, who was just Putin, which was fake and it was just John McCain who was actually just played...” the distortion of the perception of reality is addressed in a humorous style. In addition, many users repeated a swear word from the video in their comments, creating a viral effect. In addition, the comment “Remember guys an apple is never anything but an apple - cnn” mocks the media’s oversimplified analyses. Another striking point is that users create absurd humor by adding historical figures to today's digital problems using fake quotes. Sentences such as "'Most Quotes On The internet Are Fake' Winston Churchill", "'My death was a deepfake' -John F. Kennedy" and "'Don't believe everything you see on the internet.'- Sun Tzu The Art of War" reveal this situation.

The comments on the theme of media literacy and fake content warnings focused on concerns about the reliability of information created by deepfake technology. Comments such as “People need to be more aware of what they see on the internet especially on Facebook, be more cautious about what they read and watch.”, “This is great. Thanks buzzfeed. This is important to show how this happens and what fake can look like.”, “As hilarious as this video is, it's actually very educational. It teaches us that we can’t believe anything we see or hear on the internet, or TV, (especially the news) even if the words appear to be spoken straight from the person's mouth.” also indicate concerns. These comments are important in terms of demonstrating how deepfakes and artificial intelligence can erode social trust and that digital reality is becoming increasingly uncertain.

Comments on the theme of unexpected and surprising content express users' admiration and surprise despite technology and video. Comments such as “Well that was unexpected lol.”, “Damn. that confused me.”, “I did NOT expect this.”, “Whahahahaaaat?! I'm Shook.”, “I am dazed and confused.” reveal users' emotional states.

User reactions with comments with less than 100 likes reflect an atmosphere of distrust focused on the platform’s media reputation and political positioning. Users appear to directly target BuzzFeed and adopt a sarcastic and critical tone toward its content. For example, the phrases “ya trusted news sources like BuzzFeed” and “Trusted news source? Definitely not BuzzFeed nor CNN.” ironically turns the discourse of trustworthiness upside down. The comments about Obama, on the other hand, show that he is a figure that has been targeted by both intense criticism and conspiracy theories. Expressions such as “One screwed up dude. Obama really fucked things up.”, and “insane”, “fake Muslim” reflect these thoughts. In addition, there are comments in which the discourses in the video are evaluated using an entertaining approach. Here, popular movie and series references and comments from the video stand out. In comments such as “OMGG”, “I love this”, “How did you do this?”, users’ appreciation and surprise are observed. There were also thoughts about how advanced AI tools like deepfake technology could manipulate content. While some users criticized the way in which they recognized deepfake technology and analyzed such content in these comments, others pointed out that this technology could be more dangerous in the future. Among the comments, one comment was out of context.

Figure 3

Word Cloud of Comments on "You Won't Believe What Obama Says in This Video! 😊" Video



Figure 3 illustrates the prominent word frequencies in user comments, highlighting a predominantly critical discourse centered on the credibility of the source sharing the video rather than the video content itself. The word frequencies in the comments of this video show that users are dominated by a critical discourse that focuses on BuzzFeed, which shared the video rather than the content and deepfake technology. Words such as "buzzfeed" (35), "news" (37), "trusted" (26), "sources" (17), and "source" (9) that occur at high frequencies in the comments demonstrate that users frequently emphasize the issue of trust in the institution that shared the video content. The word preferences of "fake" (15), "lol" (8), and "ironic" (6) support that this questioning is made in a sarcastic and skeptical tone.

Discussion and Conclusion

The creation and modification of visual and audio media content has become extremely easy with the use of artificial intelligence. The deepfake process represents the newest way to create misleading media content. For this reason, such technologies have paved the way for creating fake voices, videos, and photographs. Since social media platforms offer a significant photo archive, some programmers can produce the content they want from these images without the knowledge and consent of the other party. The creation of video content that includes imitations of many celebrities and politicians is important in terms of identifying possible risks. This situation can cause harm to the people in the videos on the one hand, and on the other hand, it can cause serious social consequences by damaging the public's trust in the content that is believed to be true.

The study shows that users often react to deepfake videos with a mix of surprise and fear, especially due to the high level of realism. Some find the technology so convincing that it makes them question their perception of reality. The ability of deepfakes to closely imitate real people and events leads to significant viewers' unease. Alongside these concerns, many users express admiration for the technology's sophistication. The realism achieved by deepfakes is seen as both impressive and creatively powerful, reflecting the rapid progress made in artificial intelligence and media generation tools. However, users also highlight the serious risks of misuse. They voice strong concerns about the social, ethical, and legal challenges deepfakes may cause, particularly the spread of false information. Many fears that such content could erode public trust and blur the line between reality and fabrication, making it increasingly difficult to verify the truth.

These findings highlight a broader issue: the blurring of boundaries between reality and simulation in the digital age. The increasing difficulty of distinguishing fake from real images threatens public trust, especially on social platforms where content is rapidly consumed and shared. This raises urgent questions about how

societies can adapt to a media environment in which perception can be manipulated so convincingly. To address these risks, several steps should be taken. Media literacy programs must be strengthened to help users critically evaluate digital content. Policymakers and technology developers must work together to establish transparent standards, improve detection tools and regulate the misuse of synthetic media. As Johnson and Diakopoulos (2021) suggested, limiting access to deepfake training data and designing technical verification mechanisms are key to mitigation. Similarly, as noted by Maras and Alexandrou (2019), counter-technologies capable of identifying fake content must evolve in parallel.

The European Union's General Data Protection Regulation (GDPR) and Turkey's Law No. 6698 on the Protection of Personal Data (KVKK) provide important frameworks for safeguarding individuals' visual and audio data; however, they may fall short in addressing AI-based threats such as deepfakes. Therefore, there is a need for additional legal provisions targeting biometric data manipulation, and supporting content verification systems are required. Moreover, the development of labeling mechanisms by social media platforms to indicate content authenticity can play a significant role in mitigating the negative impacts of deepfake technology.

Developing artificial intelligence technologies will cause more fake content to exist in the digital environment in the future due to the ease of access and use. Although fake content and disinformation are not new problems, deepfake and similar technologies significantly increase the difficulty of finding the truth and distinguishing lies.

Peer Review Externally peer-reviewed.
Conflict of Interest The author has no conflict of interest to declare.
Grant Support The author declared that this study has received no financial support.

Author Details Sertaç Kaya
[1] İstanbul Aydin University, Faculty of Communication, Department of Journalism, İstanbul, Türkiye
0000-0003-3483-572X sertackaya1@aydin.edu.tr

References

Abram, C. (2020). The most urgent threat of deepfakes isn't politics. It's porn. Retrieved from https://www.vox.com/2020/6/8/21284005/urgent-threat-deepfakes-politics-porn-kristen-bell

Agarwal, S., Farid, H., Fried, O., & Agrawala, M. (2020). Detecting deep-fake videos from phoneme-viseme mismatches. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops* (pp. 660-661).

Albahar, M., & Almalki, J. (2019). Deepfakes: Threats and countermeasures systematic review. *Journal of Theoretical and Applied Information Technology*, *97*(22), 3242-3250.

Bohacek, M., & Farid, H. (2024). The making of an AI news anchor—and its implications. *Proceedings of the National Academy of Sciences*, *121*(1), e2315678121. https://doi.org/10.1073/pnas.2315678121

Chen, X., Sin, S. C. J., Theng, Y. L., & Lee, C. S. (2015). Why students share misinformation on social media: Motivation, gender, and study-level differences. *The Journal of Academic Librarianship*, *41*(5), 583-592. https://doi.org/10.1016/j.acalib.2015.07.003

Chesney, R., & Citron, D. (2019). Deepfakes and the new disinformation war: The coming age of post-truth geopolitics. Retrieved from https://www.foreignaffairs.com/articles/world/2018-12-11/deepfakesand-new-disinformation-war

De Ruiter, A. (2021). The distinct wrong of deepfakes. *Philosophy & Technology*, *34*(4), 1311-1332. https://doi.org/10.1007/s13347-021-00459-2

Dobber, T., Metoui, N., Trilling, D., Helberger, N., & De Vreese, C. (2021). Do (microtargeted) deepfakes have real effects on political attitudes? *The International Journal of Press/Politics*, *26*(1), 69-91. https://doi.org/10.1177/1940161220944364

Duffy, R. (2019). New research finds 96% of deepfakes are pornographic. Retrieved from https://www.emergingtechbrew.com/stories/2019/10/09/new-research-finds-96-deepfakes-pornographic

Fallis, D. (2021). The epistemic threat of deepfakes. *Philosophy & Technology*, *34*(4), 623-643. https://doi.org/10.1007/s13347-020-00419-2

Farid, H. (2022). Creating, using, misusing, and detecting deep fakes. *Journal of Online Trust and Safety*, *1*(4). https://doi.org/10.54501/jots.v1i4.56

Fletcher, J. (2018). Deepfakes, artificial intelligence, and some kind of dystopia: The new faces of online post-fact performance. *Theatre Journal*, *70*(4), 455-471. https://doi.org/10.1353/tj.2018.0097

Floridi, L. (2018). Artificial intelligence, deepfakes and a future of ectypes. *Philosophy and Technology*, *31*, 317-321. https://doi.org/10.1007/s13347-018-0325-3

Hasan, H. R., & Salah, K. (2019). Combating deepfake videos using blockchain and smart contracts. *Ieee Access*, *7*, 41596-41606. https://doi.org/10.1109/ACCESS.2019.2905689

Heidari, A., Jafari Navimipour, N., Dag, H., & Unal, M. (2024). Deepfake detection using deep learning methods: A systematic and comprehensive review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *14*(2), e1520. https://doi.org/10.1002/widm.1520

Heilweil, R. (2020). How deepfakes could actually do some good. Retrieved from https://www.vox.com/recode/2020/6/29/21303588/deepfakes-anonymous-artificial-intelligence-welcome-to-chechnya

Hern, A. (2018). My May-Thatcher deepfake won't fool you but its tech may change the World. Retrieved from https://www.theguardian.com/technology/2018/mar/12/may-thatcher-deepfake-face-swap-tech-change-world

Hwang, Y., Ryu, J. Y., & Jeong, S. H. (2021). Effects of disinformation using deepfake: The protective effect of media literacy education. *Cyberpsychology, Behavior, and Social Networking*, *24*(3), 188-193. https://doi.org/10.1089/cyber.2020.0174

Johnson, D. G., & Diakopoulos, N. (2021). What to do about deepfakes. *Communications of the ACM*, *64*(3), 33-35. https://doi.org/10.1145/3447255

Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of social media. *Business Horizons*, *53*(1), 59-68. https://doi.org/10.1016/j.bushor.2009.09.003

Karnouskos, S. (2020). Artificial intelligence in digital media: The era of deepfakes. *IEEE Transactions on Technology and Society*, *1*(3), 138-147. https://doi.org/10.1109/TTS.2020.3001312

Kietzmann, J. H., Hermkens, K., McCarthy, I. P., & Silvestre, B. S. (2011). Social media? Get serious! Understanding the functional building blocks of social media. *Business Horizons*, *54*(3), 241-251. https://doi.org/10.1016/j.bushor.2011.01.005

Kietzmann, J., Lee, L. W., McCarthy, I. P., & Kietzmann, T. C. (2020). Deepfakes: Trick or treat? *Business Horizons*, *63*(2), 135-146. https://doi.org/10.1016/j.bushor.2019.11.006

Köbis, N. C., Doležalová, B., & Soraperra, I. (2021). Fooled twice: People cannot detect deepfakes but think they can. *Iscience*, *24*(11), 1-17. https://doi.org/10.1016/%20j.isci.2021.103364

Maddocks, S. (2020). 'A deepfake porn plot intended to silence me': Exploring continuities between pornographic and 'political' deep fakes. *Porn Studies*, *7*(4), 415-423. https://doi.org/10.1080/23268743.2020.1757499

Maras, M. H., & Alexandrou, A. (2019). Determining authenticity of video evidence in the age of artificial intelligence and in the wake of Deepfake videos. *The International Journal of Evidence & Proof*, *23*(3), 255-262. https://doi.org/10.1177/1365712718807226

Marin, L. (2021). Sharing (mis) information on social networking sites. An exploration of the norms for distributing content authored by others. *Ethics and Information Technology*, *23*(3), 363-372. https://doi.org/10.1007/s10676-021-09578-y

Matern, F., Riess, C., & Stamminger, M. (2019). Exploiting visual artifacts to expose deepfakes and face manipulations. In *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, (pp. 83-92). https://doi.org/10.1109/WACVW.2019.00020

Metz, R. (2019). The fight to stay ahead of deepfake videos before the 2020 US election. Retrieved from https://edition.cnn.com/2019/06/12/tech/deepfake-2020-detection/index.html

Mihailova, M. (2021). To dally with Dalí: Deepfake (Inter) faces in the art museum. *Convergence*, *27*(4), 882-898. https://doi.org/10.1177/13548565211029401

Mintz, A. P. (Ed.). (2012). *Web of deceit: Misinformation and manipulation in the age of social media*. Medford, N.J: CyberAge Books.

Mirsky, Y., & Lee, W. (2021). The creation and detection of deepfakes: A survey. *ACM Computing Surveys (CSUR)*, *54*(1), 1-41. https://doi.org/10.1145/3425780

Murphy, G., Ching, D., Twomey, J., & Linehan, C. (2023). Face/Off: Changing the face of movies with deepfakes. *Plos One*, *18*(7), e0287503. https://doi.org/10.1371/journal.pone.0287503

Nirkin, Y., Wolf, L., Keller, Y., & Hassner, T. (2021). Deepfake detection based on discrepancies between faces and their context. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10), 6111-6121. https://doi.org/10.1109/TPAMI.2021.3093446

Petkauskas, V. (2023). Report: number of expert-crafted video deepfakes double every six months. Retrieved from https://cybernews.com/privacy/report-number-of-expert-crafted-video-deepfakes-double-every-six-months/

Rubenking, N. J. (2019). How lab mice are helping detect deepfakes. Retrieved from https://www.pcmag.com/news/how-lab-mice-are-helping-detect-deepfakes

Tahir, R., Batool, B., Jamshed, H., Jameel, M., Anwar, M., Ahmed, F., ... & Zaffar, M. F. (2021). Seeing is believing: Exploring perceptual differences in deepfake videos. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1-16). https://doi.org/10.1145/3411764.3445699

Taulli, T. (2019). Deepfake: What you need to know. Retrieved from https://www.forbes.com/sites/tomtaulli/2019/06/15/deepfake-what-you-need-to-know/?sh=d9c8c23704da

Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., & Ortega-Garcia, J. (2020). Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion*, *64*, 131-148. https://doi.org/10.1016/j.inffus.2020.06.014

Vaccari, C., & Chadwick, A. (2020). Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social Media+ Society*, *6*(1), 1-13. https://doi.org/10.1177/2056305120903408

Van der Nagel, E. (2020). Verifying images: Deepfakes, control, and consent. *Porn Studies*, *7*(4), 424-429. https://doi.org/10.1080/23268743.2020.1741434

Van Zoonen, W., Luoma-aho, V., & Lievonen, M. (2024). Trust but verify? Examining the role of trust in institutions in the spread of unverified information on social media. *Computers in Human Behavior*, *150*, 107992. https://doi.org/10.1016/j.chb.2023.107992

Verdoliva, L. (2020). Media forensics and deepfakes: an overview. *IEEE Journal of Selected Topics in Signal Processing*, 14(5), 910-932. https://doi.org/10.1109/JSTSP.2020.3002101

Wang, C. (2019). Deepfakes, revenge porn and the impact on women. Retrieved from https://www.forbes.com/sites/chenxiwang/2019/11/01/deepfakes-revenge-porn-and-the-impact-on-women/

Westerlund, M. (2019). The emergence of deepfake technology: A review. *Technology Innovation Management Review*, *9*(11), 39-52. http://doi.org/10.22215/timreview/1282

Yazdinejad, A., Parizi, R. M., Srivastava, G., & Dehghantanha, A. (2020). Making sense of blockchain for ai deepfakes technology. In *2020 IEEE Globecom Workshops (GC Wkshps)*, 1-6. http://doi.org/10.1109/GCWkshps50303.2020.9367545

İSTANBUL UNIVERSITY PRESS

Acta Infologica

Acta Infologica,
ACIN 2025, 9 (1): 223–252

https://doi.org/10.26650/acin.1676351
Submitted: 16.04.2025
Revision Requested: 20.05.2025
Last Revision Received: 31.05.2025
Accepted: 03.06.2025
Published Online 05.06.2025

Research Article

Open Access

Voices from the Lungs: An Innovative Approach to Asthma Diagnosis using Machine Learning

Murat Gezer [1], Mehmet Atilla Uysal [2], Neval Alagöz [2], Can Ortak [3] & Elif Yelda Niksarlıoğlu [2]

[1] İstanbul University, Faculty of Science, Department of Computer Science, İstanbul, Türkiye
[2] University of Health Sciences Türkiye, Yedikule Hospital for Chest Diseases and Thoracic Surgery, Department of Chest Diseases, İstanbul, Türkiye
[3] Medcase Corp, İstanbul, Türkiye

Abstract Asthma is one of the most common chronic respiratory diseases worldwide, and early and accurate diagnosis is critical for effective clinical management. In this study, we evaluated the diagnostic potential of machine learning models based on voice analysis as a non-invasive approach for asthma diagnosis. Using audio samples containing seven different phonetic units, the performances of 13 different machine-learning algorithms were comprehensively analyzed. The StandardScaler and SMOTE techniques were applied in the data preprocessing stage, and a 5-fold cross-validation methodology was adopted to evaluate the models. Accuracy, F1-score, sensitivity, precision, specificity, and area under the curve (AUC) metrics were used for performance evaluation. The results demonstrate that ensemble learning approaches, particularly the stacking ensemble model, exhibit superior discriminative capacity for all phonetic units. Individual models, such as neural networks and support vector machines, also produced remarkable results, whereas simpler models were limited in terms of capturing complex patterns in audio data. This study demonstrated the promising diagnostic potential of voice analysis-based ensemble learning approaches for asthma diagnosis; however, it emphasizes the need for an optimal balance between sensitivity and specificity in clinical applications.

Keywords Asthma diagnosis · Sound analysis · Machine learning · Ensemble models · Acoustic feature extraction

Citation: Gezer, M., Uysal, M. A., Alagöz, N., Ortak, C. & Niksarlıoğlu, E. Y. (2025). Voices from the Lungs: An innovative approach to asthma diagnosis using machine learning. *Acta Infologica*, *9*(1), 223-252. https://doi.org/10.26650/acin.1676351

This work is licensed under Creative Commons Attribution-NonCommercial 4.0 International License.

© 2025. Gezer, M., Uysal, M. A., Alagöz, N., Ortak, C. & Niksarlıoğlu, E. Y.

Corresponding author: Murat Gezer murat.gezer@istanbul.edu.tr

Introduction

Asthma is a prevalent chronic respiratory condition that affects millions of people globally, leading to severe morbidity and mortality if not effectively managed. As recognized as a priority area for intervention by the World Health Organization's Global Action Plan for the Prevention and Control of Noncommunicable Diseases and the United Nations 2030 Agenda for Sustainable Development, asthma is a significant global health concern (World Health Organization, 2024). This heterogeneous respiratory disease is characterized by chronic inflammation and airway narrowing (bronchoconstriction) (GINA 2023). This pathological process manifests as clinical symptoms, such as recurrent wheezing, dyspnea, chest tightness, and coughing episodes (Barnes, 2008). Structural changes, including increased mucus production, smooth muscle hypertrophy, and epithelial damage, also contribute to asthma pathophysiology (Holgate et al., 2015). These symptoms are typically exacerbated at night or in the early morning and can be triggered by exposure to various factors, such as allergens, infections, exercise, and environmental stimuli (Lambrecht & Hammad, 2015).

The burden of asthma on the global health economy is considerable. The WHO estimates that 262 million individuals were living with asthma in 2019 (World Health Organization, 2024), and 455,000 deaths were due to the disease in the same year (Ministry of Health, General Directorate of Public Health, 2024). Most deaths occur in low- and lower-middle-income countries, emphasizing the critical need for improved diagnosis and management strategies in these regions (World Health Organization, 2024).

The data obtained from Türkiye also show a similar picture. In a 2017 study, 6.9% of adults aged 15 years and older were diagnosed with asthma, and this rate was higher in women (8.7%) than in men (5.0%) (Republic of Türkiye Ministry of Health, General Directorate of Public Health, 2024). Furthermore, data from the Turkish Statistical Institute (TUIK) for 2022 show that respiratory system diseases are the third leading cause of death, with asthma accounting for 0.3% of these deaths (Ministry of Health, General Directorate of Public Health, 2024).

In addition to statistics, asthma profoundly impacts an individual's quality of life, affecting their ability to participate in daily activities and significantly altering their general wellbeing. Speech is an area considerably affected by asthma. Speech production is a complex physiological process involving coordinated interactions among multiple systems. The air expelled from the lungs vibrates the vocal cords in the larynx to produce sounds. This sound is shaped and modulated by articulators in the oral cavity, including the tongue, teeth, and palate, and the resonance spaces of the sinuses and nasal cavity (Sezer & Akıl, 2020). Airway narrowing and inflammation caused by asthma disrupt this complex process (Fesci & Görgülü, 2005). Consequently, individuals with asthma experience difficulty breathing, resulting in changes in voice quality, decreased speech intelligibility, and voice fatigue.

Pulmonary function tests, such as spirometry and peak flowmetry, have long been used as standard methods in clinical practice for the diagnosis and follow-up of asthma. These methods are applied in clinical settings under the supervision of specialized healthcare professionals and are insufficient for continuously monitoring changes in patients' daily lives. These limitations in the follow-up of the disease highlight the need for more comprehensive and dynamic approaches to asthma management.

Technological developments in recent years have paved the way for innovative asthma prediction and monitoring approaches in the healthcare field (Saygılı, 2019). These new methods enable real-time or out-of-clinic monitoring of patients' clinical observations. Research has shown that the respiratory acoustic

characteristics of patients with asthma differ from those of healthy individuals (Schreur et al., 1994). These differences are manifested especially in the form of high-frequency wheezing and low-frequency crackles (Reichert et al., 2008; Schreur et al., 1994). In a study conducted by Malmberg et al. in 1994, spectral analyses showed an increase in power in certain frequency ranges of breath sounds in patients with asthma (Malmberg et al., 1994).

The unique contribution of this study is that it provides an innovative approach for asthma diagnosis by analyzing speech signals. While several studies have explored asthma diagnosis and risk analysis using speech signals (e.g., Alam et al., 2022; Chen et al., 2025; Yadav et al., 2018; Yadav et al., 2020), a comprehensive investigation comparable in scale and methodology to the present research, specifically, one utilizing data collected from Turkish-speaking individuals diagnosed in a hospital setting according to GINA criteria, and evaluating the performance of 13 different machine learning algorithms across seven distinct phonetic units, had not been reported in the literature at the time this study was conducted. This research aims to go beyond current diagnostic methods that are only applied in standardized clinical settings and develop a methodology that offers the potential for non-invasive and continuous monitoring of patients in their daily lives. The aim of our study was to determine whether a patient had asthma using acoustic analysis of voice recordings and to identify machine learning models that can achieve the highest accuracy.

The developed acoustic analysis method detects possible asthma symptoms by analyzing various parameters, such as frequency, amplitude, and spectral features (graphical representation of the power distribution of sound waves) in sound signals. The proposed approach is less invasive than conventional methods and offers continuous monitoring. For example, analyzing respiratory sounds recorded during the night may be useful for detecting asthma symptoms that occur during sleep.

These innovative sound analysis methods can potentially lead to a paradigm shift in asthma management. Continuous and non-invasive monitoring may enable early detection and prevention of asthma attacks. It can also improve patients' quality of life and reduce healthcare costs by helping create personalized treatment plans.

However, further research and validation studies are required to fully integrate these technologies into clinical practice. Issues such as data security, algorithm reliability, and cost-effectiveness should be carefully addressed before their widespread use. In addition, comprehensive training programs should be developed to enable healthcare professionals and patients to effectively use these new technologies.

Machine learning techniques have been widely used in the field of bioinformatics to exploit disease hallmarks in the gene, proteomic, and metabolic expression of samples to build accurate prediction models. In conclusion, the asthma prediction model developed in this study can help individuals who cannot immediately visit a diagnostic doctor due to a lack of infrastructure and other health resources. Furthermore, this model can advance the prediction of other diseases affecting the pulmonary system.

Literature Review

The use of machine learning algorithms for asthma diagnosis has recently attracted considerable attention recently. In particular, the effectiveness of artificial intelligence has been observed in the analysis of lung imaging and the interpretation of pulmonary function tests (Kaplan et al., 2021). A comprehensive literature review revealed that various researchers have examined different machine-learning models to improve the clinical diagnosis of asthma (Bolat, 2021). These studies have generally been conducted on large-scale patient datasets and have yielded promising results for the automatic diagnosis of asthma.

Yadav et al. (2018) investigated the classification of asthmatic patients and healthy subjects using sustained phonations (/A:/, /i:/, /u:/, /eI/, /oU/, /s/, /z/) and nonspeech sounds (cough, wheeze) from 47 asthmatic and 48 healthy participants. Using MFCC statistics and Support Vector Machines (SVM), wheeze achieved the highest classification accuracy of 90.25%, whereas among sustained vowels, /i:/ achieved 80.79% accuracy (Yadav et al., 2018).

In a subsequent study, Yadav et al. (2020) employed INTERSPEECH 2013 Computational Paralinguistics Challenge baseline (ISCB) acoustic features from a similar dataset for more detailed analysis. Using ISCB features, /oU/ phonemes achieved 75.4% accuracy (18.28% improvement over baseline), while exhale achieved the highest classification accuracy of 77.8%. The loudness and MFCC feature groups have been identified as the most important contributors to asthma classification (Yadav et al. 2020).

Alam et al. (2022) developed machine-learning approaches to predict lung function from voice recordings of patients with asthma. In total, 323 voice recordings were collected from 26 patients with asthma who underwent bronchoprovocation tests. A threshold-based mechanism was designed to separate speech and breathing segments, and 23 features were extracted. The random forest regression model showed the best performance, with the lowest root mean square error (RMSE = 10.86) for predicting FEV1%. The Random Forest algorithm achieved 85% accuracy in binary classification to predict abnormal lung function (Alam et al., 2022).

Sterling et al. (2014) developed an Automated System for Asthma Monitoring (ADAM) that operates on a mobile platform. The system uses an iOS application to detect cough sounds using an external microphone. Using the hidden Markov model (HMM)-based Viterbi decoder and MFCC features, the system achieved 63% sensitivity with three false positives per hour (Sterling et al., 2014). This study demonstrated the potential of mobile technology for the objective monitoring of asthma symptoms, particularly among adolescents.

Deep learning methods, particularly convolutional neural networks (CNNs) and long short-term memory (LSTM) architectures, are prominent approaches for the classification of lung sounds. Aykanat et al. (2020) obtained an 86% accuracy rate by combining Mel-Frequency Log Energy (MFCC) features with the (GB) algorithm (Aykanat et al., 2020). Similarly, Zhang et al. (2024) developed a hybrid CNN-LSTM model using spectrogram images as inputs and achieved 99.01% accuracy and 99.13% sensitivity (Zhang et al., 2024). These findings confirm that time-frequency representations (e.g., spectrograms) provide an effective framework for the automatic analysis of pathological respiratory sounds (wheezes and crackles).

Beyond the success of spectrogram-based analyses, Petmezas et al. (2022) achieved an accuracy of 76.39% using a hybrid CNN-LSTM architecture (Petmezas et al., 2022). Research emphasizes that visual representations of spectrograms in the frequency-time domain are critical sources of features for both traditional and deep learning models.

To provide practical solutions for asthma screening in resource-constrained areas, Gunawardana et al. (2024) evaluated 13 different machine-learning algorithms using data collected in the Sri Lanka Health and Aging Study (SLHAS). This study found that the hybrid model of Logistic Regression and LightGBM performed best with a sensitivity of 79.85% and AUC value of 0.9062. These results indicate that a low-cost model can be developed for community-based health screening and clinical settings can be developed (Gunawardana et al., 2024).

Topaz et al. (2022) conducted a study using audio-recorded patient-primary care provider encounters to evaluate the level of shared decision-making (SDM) and predict inhaled corticosteroid adherence in primary

care settings. Speech-to-text algorithms were used to automatically transcribe 80 audio-recorded encounters (ROUGE F-score = 0.9), and machine learning algorithms (Naive Bayes, Support Vector Machines, Decision Tree) were applied. The highest F-score achieved was 0.88 for SDM evaluation (Naive Bayes) and medication adherence prediction (Support Vector Machines) (Topaz et al., 2022). This study represents pioneering work demonstrating that speech data can be used to predict patient adherence to asthma treatment.

Rivas-Navarrete et al. (2025) developed an edge-computing-based system for detecting chronic respiratory diseases (COPD and asthma) by analyzing cough and breath sounds. The system operates on Raspberry Pi devices and smartphones by using MFCC and Chromagram features. Trained on a dataset collected from 86 participants (53 with respiratory conditions and 33 healthy), the system achieved 90% sensitivity, 93.55% specificity, and 91.75% balanced accuracy in detecting chronic respiratory diseases (Rivas-Navarrete et al., 2025). This study represents a significant step forward in the development of low-cost and portable diagnostic tools suitable for resource-limited areas.

Chen et al. (2025) developed an AI-based system that uses voice analysis to predict the risk of asthma. Using 1500 speech samples from the Saarbrucken Voice Database (high-pitch, normal-pitch, and low-pitch recitations of phonemes [i, a, u]), the Long-Term Average Spectrum (LTAS) and Single-Frequency Filtering Cepstral Coefficients (SFCCs) were extracted as features. Seven machine-learning algorithms (Decision Tree, Random Forest, Gradient Boosting, SVM, ANN, CNN, LSTM) were employed in this study. The decision Tree, CNN, and LSTM models achieved average accuracy values greater than 80%, and the decision tree model demonstrated the best accuracy for high pitch phonemes (accuracy: 98.66%) (Chen et al., 2025).

Mayr et al. (2025) analyzed the speech characteristics of patients with COPD during and after exacerbation. Using machine learning (SVM) on speech data from 50 patients with COPD, they achieved 84% accuracy in classifying patient status, which was significantly higher than the 65% accuracy obtained using CAT and BORG scores alone. After exacerbation, patients exhibit reduced voice breathiness, more stable loudness and phonation, fewer pauses, and more regular reading rhythms (Mayr et al., 2025).

Material and Methods

This study applied a comprehensive methodology for asthma diagnosis based on the analysis of audio signals. The methodological approach comprises three main stages: data collection, feature extraction, and machine learning model development.

Data Collection

This study was conducted in a single center with a cross-sectional design between January 2024 and July 2024 at the University of Health Sciences Istanbul Yedikule Chest Diseases and Thoracic Surgery Training and Research Hospital. Thus, 344 participants, comprising 284 patients diagnosed with asthma by a pulmonologist according to the GINA 2023 asthma guidelines and 60 healthy individuals without any history of chronic disease (Figure 1), were included in the study and were volunteers aged between 18 and 64 years (Figure 2). The study was conducted in accordance with the principles of the Declaration of Helsinki and was approved by the Clinical Research Ethics Committee of Istanbul Yedikule Chest Diseases and Thoracic Surgery Training and Research Hospital (Approval date: December 14, 2023, Approval number 2023-433).

Figure 1

Distribution of study subjects.

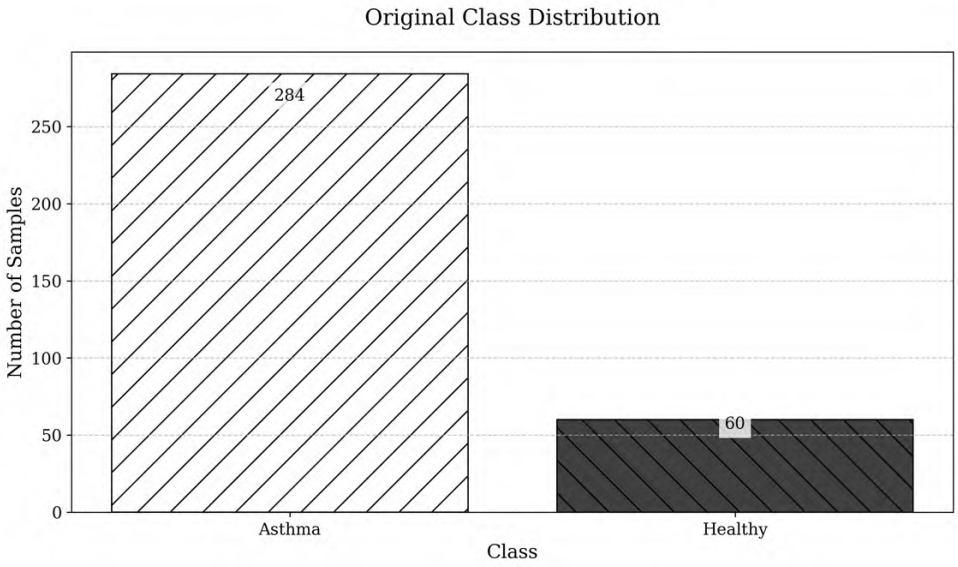
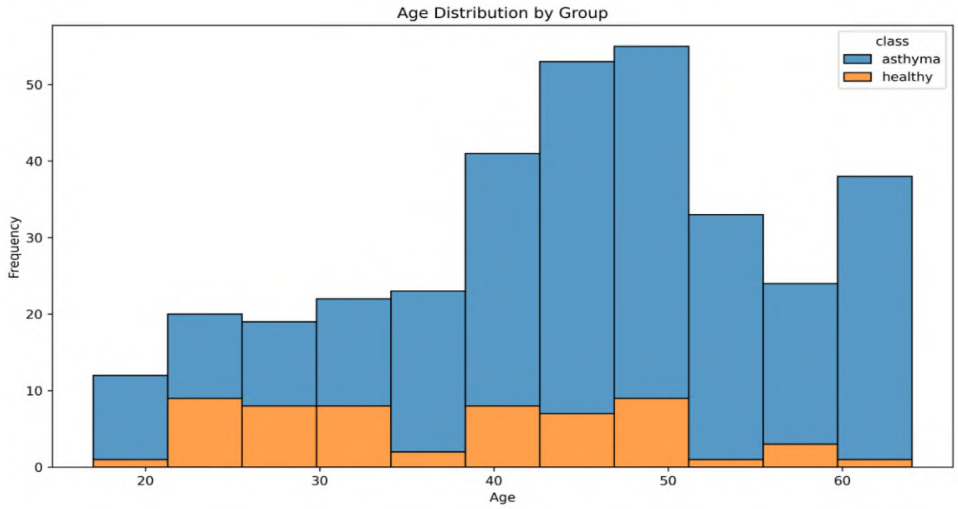


Figure 2

Distribution of Subject Age Groups



A physician detailed the purpose and processes of the study to all participants. Participants were matched by age and sex, and records of unmatched participants were excluded from the study. All individuals who agreed to participate signed a written informed consent form. During the enrollment process, the participants' age, sex, and additional diagnostic information, if any, were systematically recorded.

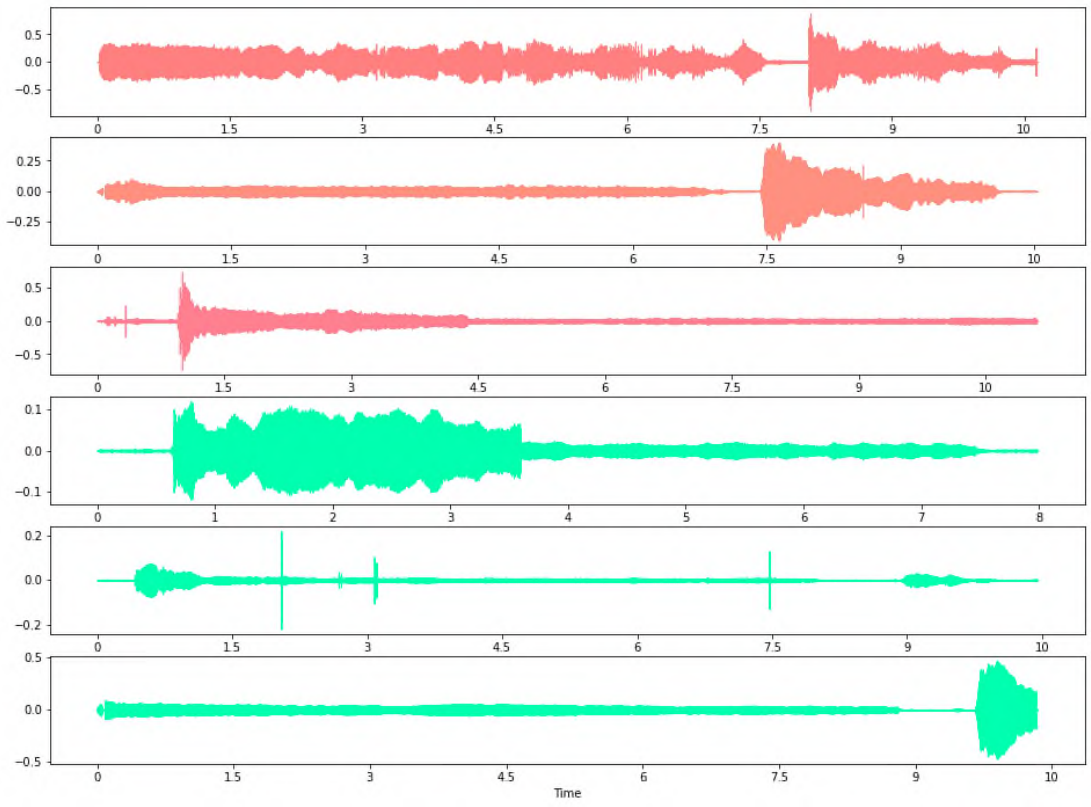
Data collection was carried out by asking participants to read predetermined sound phrases for 10 s on an iPhone 14 Pro mobile phone, positioned 10 cm away from their mouths, in the presence of an expert physician. The phrases vocalized by the participants consisted of the phonetic sound ‘aaa’ and the words ‘ana, ’ ‘araba, ’ ‘ordu, ’ ‘gelecek, ’ ‘titiz’ and’ ünlem.’ Each participant provided seven distinct 10-s voice recordings, resulting in seven audio samples per user, and the phonetic features of the selected words were specifically chosen to reveal acoustic differences in the voices of asthma patients. The data collection process was standardized by applying the same recording procedure to all participants. The collected voice

data were labeled by the attending physician as 'Asthma' or 'Healthy' according to the GINA criteria. Because a separate dataset was created for each of the seven phonetic units, each individual dataset comprised 344 samples (audio recordings), with 284 samples from patients with asthma and 60 from healthy controls. Machine learning models were developed on these datasets, and their performance was evaluated using the 5-fold cross-validation methodology detailed in section "2.5 Experimental Design." According to this methodology, the dataset for each phonetic unit was divided into 80% (approximately 275 samples) training data and 20% (approximately 69 samples) test data at each iteration.

To ensure the confidentiality of the participants, all data were anonymized and stored on a secure software platform developed by MedCase Yazılım Teknolojileri A.Ş. The data were grouped using unique identification numbers (ID) assigned to each participant and stored on a secure server. Only authorized administrators have access to the server. Figure 3 shows the time-domain data slice of the healthy and asthmatic data.

Figure 3

Sample data of subjects within the TIME domain: asthma patient (red) and healthy subject (green)



Feature Inference

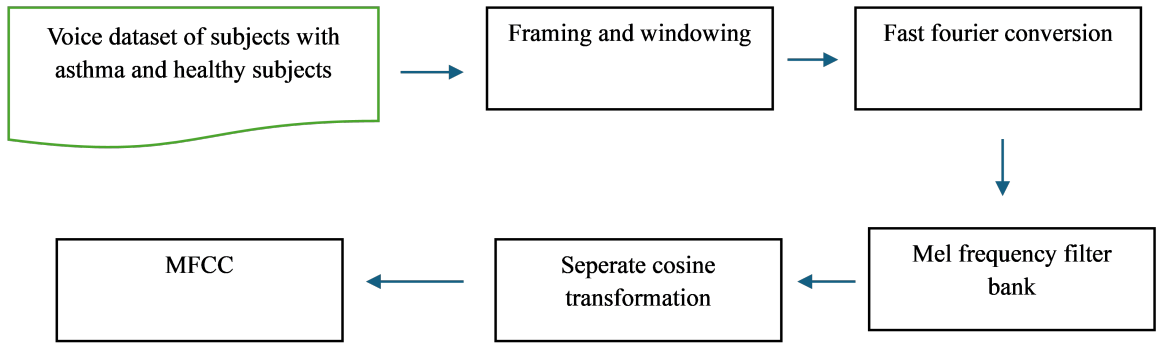
Participants provided audio data for seven distinct phonetic units ('aaa,' 'ana,' 'araba,' 'ordu,' 'gelecek,' 'titiz,' and 'ünlem'), which were specifically selected due to their inclusion of diverse Turkish vowels, potentially reflecting vocal tract alterations. These recordings were captured using an iPhone 14 Pro mobile device at a 44.1-kHz sampling rate and underwent a comprehensive pre-processing and feature extraction pipeline. The Python programming language and the Librosa library were utilized for this process. To ensure computational consistency and adhere to standard feature extraction practices, all audio files were resampled to

a 22.05-kHz sampling rate using the Librosa library before feature extraction. The initial pre-processing step before specific feature calculation involved steps such as silence removal from the signals and application of noise-reduction techniques to enhance signal quality. Figure 3 provides a visual example of a pre-processed audio waveform.

Subsequently, Mel-Frequency Cepstral Coefficient (MFCC) extraction was performed to convert the audio data into numerical features (Logan, 2000). MFCC is an effective technique widely used in audio processing and speech recognition; it models the frequency perception of the human ear and transforms audio signals into multidimensional feature vectors (Davis & Mermelstein, 1980; Rabiner & Schafer, 2007). The MFCC features were extracted by the following systematic steps (Figure 4):

1. Preprocessing: Silence segments were removed from the signals, and noise-reduction techniques were applied.
2. Windowing: The audio signal was divided into short time intervals suitable for analysis. For the MFCC calculation, the default parameters of the Librosa library were used: a frame size of 2048 samples (n_fft=2048), a hop length of 512 samples (hop_length=512), and a Hamming window
3. Fourier transform: The time-domain signals are transformed into frequency domains.
4. Mel Filter Application: The frequency spectrum was converted to the Mel scale to model the human hearing system.
5. Logarithmic Transform and Discrete Cosine Transform (DCT): Spectral information was processed by logarithmic transformation and DCT to obtain the final MFCC coefficients.

Figure 4
Algorithm for creating MFCC



Mel-Frequency Cepstral Coefficients (MFCCs) and Dynamic Features

For each audio signal, 12 basic MFCC coefficients were extracted and optimized to model the spectral characteristics of the human voice. These coefficients represent the amplitude information in the frequency spectrum of the audio signal (Young et al., 2006). To capture temporal dynamics within the audio signal, which is crucial as vocal characteristics may change over time due to disease, dynamic features, such as delta (first-order temporal derivative) and delta-delta (second-order temporal derivative) coefficients, were calculated from these 12 basic MFCCs and included in the analysis. This approach provided 36 MFCC-based features (12 basic coefficients, 12 delta coefficients, and 12 delta-delta coefficients).

Additional Acoustic Features and Statistics

To complement the MFCC features and enable more sensitive detection of potential disease markers, the following additional acoustic features were extracted:

- **Spectral Centroid:** Indicates the "brightness" center of the sound.
- **Spectral bandwidth: shows the energy spread** in the frequency spectrum.
- **Spectral Rolloff:** This parameter specifies the frequency below which a large portion (default 85%) of the spectral energy is concentrated, providing information about the high-frequency content.
- **Zero-Crossing Rate:** Indicates the frequency at which the signal crosses the zero axis, which is often related to the degree of noisiness or periodicity.

The mean and standard deviation values of the extracted additional acoustic features were calculated to enrich the feature vectors. For instance, instability or tonal shifts in the voice caused by respiratory illness might manifest as increases in the standard deviation of these features, whereas shifts in mean values could indicate overall changes in vocal characteristics. This resulted in 4*2 = 8 additional features from these four acoustic properties and their statistics. By combining the 36 MFCC-based features and the 8 features derived from additional acoustic properties, a total feature vector of 44 dimensions was created for each audio sample. This comprehensive and multidimensional feature extraction strategy aims to capture in more detail the spectral variations (diversity and changes) in the voices of patients with asthma, which might reflect the subtle and complex effects of diseases like asthma, thereby aiming to establish a more robust and discriminative basis for disease diagnosis. For each participant, recordings with seven different words (‘aaa,’ ‘ana,’ ‘araba,’ ‘ordu,’ ‘gelecek,’ ‘titiz,’ ‘ünlem’) were systematically pre-processed, and MFCC and other acoustic features were extracted and saved in CSV files with their labels (asthma or healthy). As a result, a separate dataset was created for each word. In addition, exploratory data analyses were performed, and class distributions were examined in detail to understand the statistical properties of the generated datasets and provide a basis for the classification models.

Performance Metrics

The precision, Recall, F1 Score, and accuracy metrics were used to evaluate the performance of the proposed system.

The inputs used in these equations reflect the relationship between the model predictions and actual situations.

- True Positive (TP - True Positive): Cases in which the model predicts asthma and is actually asthma.
- True Negative (TN - True Negative): Cases in which the model predicts as healthy and are actually healthy.
- False Positive (FP - False Positive): Cases in which the model predicts asthma but is actually healthy.
- False Negative (FN - False Negative): Cases that the model predicts as healthy but actually have asthma.

These classifications form the basis for calculating the metrics used to evaluate the performance of the model and allow us to measure its ability to detect asthma cases and correctly classify healthy individuals.

The metrics used to evaluate the performance of an asthma-detection system are crucial. Precision, Recall, Specificity, F1 Score, and Accuracy help us comprehensively evaluate the system’s success.

- **Precision**: Indicates the proportion of cases predicted by the model as asthma, which is actually asthma. Precision is important for minimizing false positives (FPs), and high-precision values are desirable.
- **Recall:** Refers to the proportion of actual asthma cases correctly detected by the model. A high sensitivity is important for reducing false negatives (FN).
- **Specificity:** The proportion of healthy individuals correctly identified by the model. Specificity is important for reducing false positives (FP). High specificity helps prevents healthy individuals from receiving unnecessary treatment.
- **F1 Score (F1 Score):** this is the harmonic mean of the sensitivity and specificity. This metric provides balanced assessment of sensitivity and specificity.
- **Accuracy:** Indicates the proportion of correct predictions among all predictions. A high-accuracy value represents the overall accuracy of the model. However, it can be misleading on unbalanced datasets; therefore, it should be evaluated together with other metrics.

In medical applications such as asthma diagnosis, there is often a trade-off between sensitivity and specificity. High sensitivity increased the ability to identify patients with asthma, whereas high specificity increased the ability to correctly identify healthy individuals. In a clinical context, the most important metric may vary depending on the potential risks of false-positive or false-negative diagnoses.

Models Used

In this study, various machine-learning models for asthma diagnosis via voice analysis were developed and evaluated. The classification algorithms used in this study and their specific parameters are described below.

1. **Naive Bayes (NB)**: The NB is a computationally efficient classifier based on probability theory that works under the assumption of feature independence.
2. **K-Nearest Neighbor (K-NN)**: This is a classification algorithm based on sample similarity. Predictions were made according to the classes of the five nearest neighbors of each sample using the k=5 parameter values.
3. **Decision Tree (Decision Tree)**: Based on hierarchical decision rules, this classifier was optimized with the parameter's maximum depth=5, minimum sampling split=5, and minimum leaf sample=2.
4. **Neural Network**: A neural network architecture consisting of four hidden layers with (256, 128, 64, 32) neuron numbers was used and trained with 'relu' activation function and adaptive learning rate.
5. **Extreme gradient boosting (XGBoost)**: An optimized version of the gradient boosting technique in terms of computational efficiency and performance.
6. **Categorical Boost (CatBoost)**: A gradient boosting algorithm designed for categorical variables was used with parameters of 100 iterations, five maximum tree depths, and a learning rate of 0.1.
7. **Adaptive Boosting (AdaBoost)**: An ensemble learning algorithm was applied with decision trees with a maximum depth of three, trained using 200 iterations, and a learning rate of 0.1.
8. **Gradient Boosting**: An ensemble learning algorithm configured with 100 trees, 0.1 learning rate, 3 maximum depth of three, and five minimum split value parameters.
9. **Random Forest**: An ensemble learning algorithm optimized with 100 trees, 10 maximum depths and 'balanced' class weights.

10. **Support Vector Machines (SVM)**: This classifier separates data points with hyperplanes, and is configured with regularization parameter C=1.0, kernel function 'rbf' and gamma parameter 'scale.'
11. **Logistic Regression**: A binary classification model configured with maximum iteration=2000, regularization intensity C=0.1, regularization l2, optimization algorithm 'saga' and class weights 'balanced' was applied.

In addition, various ensemble models were developed using the five best-performing models for each phonetic unit.

1. **Voting Ensemble**: An ensemble model was implemented in which the five best-performing models for each phoneme were combined with equal weights, and predictions were generated using a 'soft' voting strategy (averaging probability estimates).
2. **Stacking Ensemble**: This is an advanced meta-learner that uses the predictions of the top five models for each voice as input and applies Logistic Regression (maximum iteration=2000, C=0.5) as a meta-learner. It generates probability predictions with 5-fold cross-validation and the 'predict_proba' method.
3. **Weighted Voting Ensemble**: An ensemble model was developed in which the top five models for each voice were weighted according to their AUC performance metric; thus, the higher-performing models had more influence on the final classification decision.

The selection of these models was guided by the dataset size and the need for robustness in clinical applications. Deep-learning architectures (e.g., CNNs, LSTMs, Transformers) were excluded because of their requirement for large-scale datasets to achieve optimal performance (Goodfellow, Bengio, & Courville, 2016). With only 284 patients with asthma and 60 healthy controls, the sample size is significantly below the thousands or tens of thousands of examples typically required for effective deep learning training (Zhang et al., 2021). Furthermore, high-dimensional voice data combined with limited samples increases the risk of overfitting, where models memorize training data rather than generalize to new cases (Shorten, Khoshgoftaar, & Furht, 2019). Traditional machine learning and ensemble methodologies were chosen to prioritize the generalizability and reliability of medical diagnostics. These approaches have demonstrated robust performance on small- to medium-scale datasets (Dietterich, 2000; Ganaie et al., 2022). Ensemble models (Voting Ensemble, Stacking Ensemble, Weighted Voting Ensemble) were specifically developed using the top five base models for each phonetic unit, leveraging soft voting and meta-learning strategies.

Future studies should systematically assess deep learning architectures for asthma diagnosis using expanded and heterogeneous datasets.

Experimental Design

In this study, a 5-fold cross-validation methodology was adopted to evaluate the performance of the developed machine learning models. The proposed approach is based on the principle of dividing the dataset into five equal parts, training the model with four parts, and testing the remaining parts. In each iteration, a different piece of data was used as a test set; thus, all data points were evaluated in both the training and testing phases (Kohavi, 1995). By averaging the results of five iterations, the generalization performance of the models was measured in a more reliable and objective manner. The standard deviation of the results across these folds was also calculated to assess the stability of model performance.

In the data preprocessing stage, StandardScaler was applied to normalize the scale differences between the features. This standardization process optimized the performance of the algorithms by removing the

imbalance between the attributes of different sizes. In addition, the SMOTE (synthetic minority oversampling technique) algorithm was used to address the sample distribution imbalance between asthma and healthy classes (Chawla et al., 2002). By generating synthetic samples of the minority class, SMOTE ensures that classification models have equal learning opportunities for both categories. Importantly, in terms of methodological rigor, the SMOTE technique was applied only to the training data, while the test data were retained in their original form, preserving their true distribution. This strategic approach enables a more realistic assessment of model performance under real-world conditions in clinical settings.

The data preprocessing steps (standardization and SMOTE) and model training procedures were integrated into the cross-validation cycle. This integration was realized through the ImbPipeline structure of the Scikit-learn library, ensuring that the data preprocessing and model training phases were independently and methodologically consistent.

The model training and evaluation processes were performed on a MacBook Pro (2023) computer with an Apple M3 Pro processor (36 GB RAM) running on a macOS Sonoma operating system in a high-performance computing environment. The following specific libraries were used to implement the algorithms developed in the Python 3.11.9 programming language using Spyder Integrated.

Development Environment:

1. Scikit-learn 1.5: Used to implement machine learning models and cross-validation procedures.
2. Librosa 0.10, provides specialized audio data processing and acoustic feature extraction functions.
3. Matplotlib: Used for visualizing research findings and model performance.

This comprehensive technological and methodological framework maximized the scientific reproducibility of the study's experimental design and its potential for integration into clinical practice.

Results

This study comprehensively evaluated the diagnostic performance of voice analysis-based machine-learning models for asthma diagnosis. The classification efficiency of the developed algorithms was systematically tested on audio samples containing various phonetic units ('aaa, ' 'ana,' 'araba,' 'gelecek, ' 'ordu, ' 'titiz,' 'ünlem'), and the quantitative results obtained are presented in detail in Tables 1–7. The results are reported as mean ± standard deviation from the 5-fold cross-validation to reflect the consistency of model performance across different data folds. The analyses demonstrate that ensemble models perform better than individual classifiers, with the Stacking Ensemble and Voting Ensemble approaches in particular showing superior performance in terms of diagnostic accuracy.

Analyzing the sound "aaa"

When the performance metrics of the machine learning models applied on the 'aaa' sound presented in Table 1 and Figure 5 are analyzed, it can be seen that the Stacking Ensemble model exhibits superior classification capacity. This meta-learning model achieved the highest overall performance, with an accuracy of 80.2% ± 2.6% and an AUC of 0.627 ± 0.107. Similarly, the Neural Network model showed equivalent discriminative ability with an AUC value of 0.628 ± 0.141, followed by the Gradient Boosting algorithm with an AUC value of 0.607 ± 0.065.

When we evaluated model performance from the perspective of classification algorithms, it is observed that ensemble learning methods produced more consistent results than individual classifiers. The Voting

Ensemble and Weighted Voting Ensemble models exhibited almost identical performance profiles (AUC values of 0.649 ± 0.112 and 0.649 ± 0.113, respectively), suggesting that the model weighting strategy did not provide a clear advantage for the 'aaa' sound.

The simpler classification approaches, Naive Bayes (57.0% ± 4.6% accuracy, 0.564 ± 0.083 AUC) and K-NN (57.2% ± 7.8% accuracy, 0.620 ± 0.088 AUC) algorithms, demonstrated limited discrimination capacity compared with the complex models. This suggests that voice analysis for asthma diagnosis requires nonlinear models and capable of capturing complex relationships.

In terms of performance metrics, although the Random Forest model achieved high accuracy (75.8% ±5.5%) and an F1 score (0.856 ± 0.039), it showed potential limitations in clinical applications due to its low specificity (0.183 ± 0.062). This suggests that the model has difficulty correctly categorizing healthy individuals and, therefore, may produce a high rate of false positive results under real conditions.

In terms of the balance of sensitivity and specificity, the Stacking Ensemble model (sensitivity: 0.874 ± 0.035, specificity: 0.443 ± 0.096) was highly successful in detecting patients with asthma, but it performed moderately in discriminating healthy individuals. This profile has potential value for screening purposes by detecting high-risk individuals with minimal false negatives.

The results of the 'Aha' sound analysis demonstrated the feasibility of voice-based asthma diagnosis and revealed the limiting factors of a single phonetic unit. The AUC values of the models were generally in the range of 0.60-0.65, suggesting that the 'aaa' sound alone is not sufficient for a definitive diagnosis; however, in combination with other phonetic units, its diagnostic value may increase.

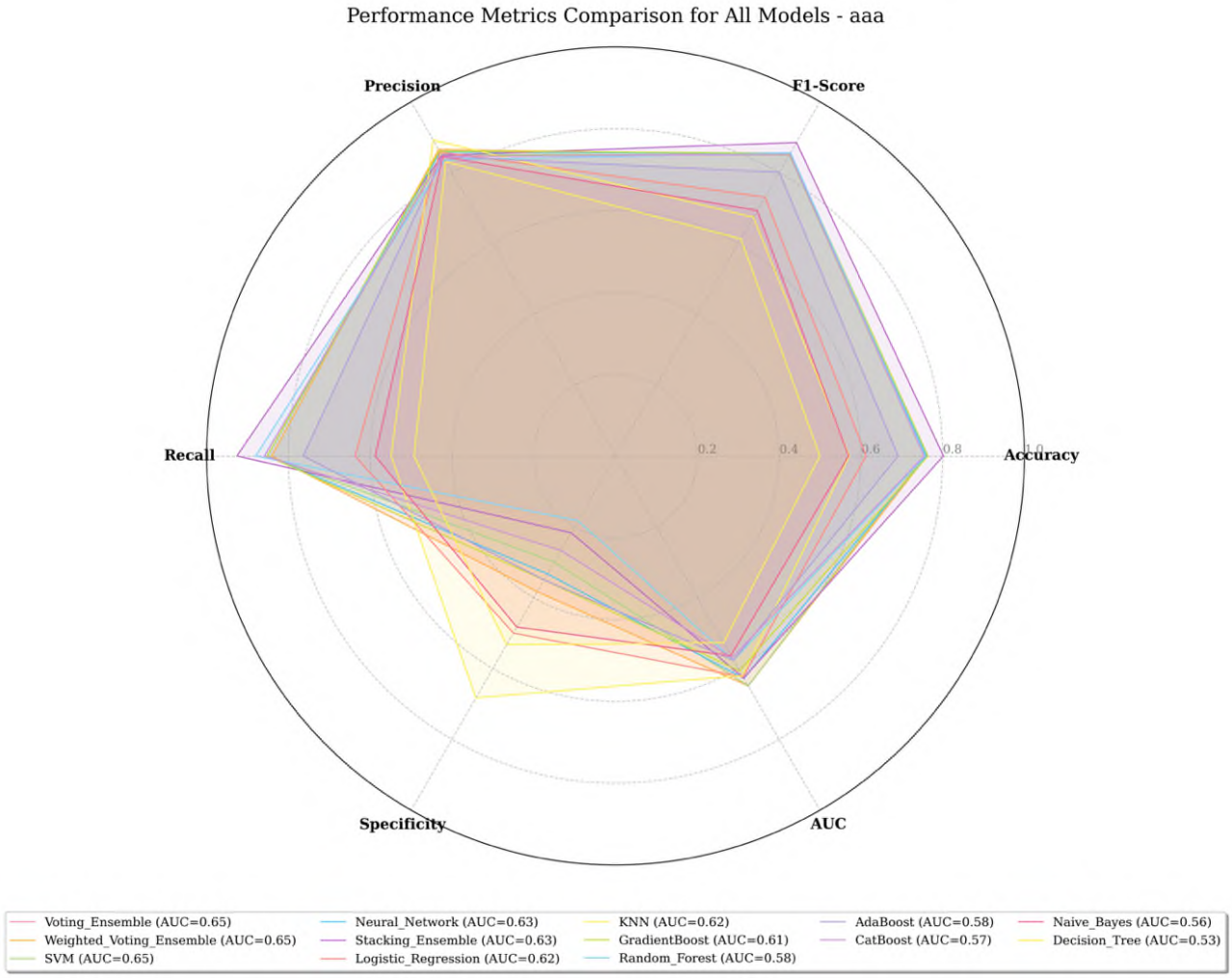
Table 1

Machine learning model performance evaluation metrics for the sound "aaa"

Model	Accuracy	F1-Score	Precision	Recall	Specificity	AUC
Naive_Bayes	0.570±0.046	0.693±0.033	0.845±0.045	0.588±0.035	0.483±0.170	0.564±0.083
KNN	0.572±0.078	0.674±0.081	0.892±0.040	0.549±0.100	**0.683±0.143**	0.620±0.088
Decision_Tree	0.500±0.080	0.613±0.091	0.831±0.029	0.493±0.103	0.533±0.100	0.528±0.067
Neural_Network	0.761±0.063	0.854±0.042	0.858±0.035	0.852±0.061	0.333±0.175	0.628±0.141
CatBoost	0.755±0.065	0.852±0.045	0.846±0.022	0.859±0.070	0.267±0.062	0.570±0.052
AdaBoost	0.691±0.087	0.801±0.066	0.845±0.033	0.764±0.094	0.350±0.097	0.577±0.086
GradientBoost	0.764±0.055	0.855±0.038	0.861±0.021	0.852±0.059	0.350±0.082	0.607±0.065
Random_Forest	0.758±0.055	0.856±0.039	0.836±0.011	0.880±0.073	0.183±0.062	0.577±0.076
SVM	0.755±0.066	0.850±0.047	0.851±0.019	0.852±0.082	0.300±0.085	**0.649±0.120**
Logistic_Regression	0.613±0.063	0.731±0.045	0.858±0.050	0.637±0.041	0.500±0.175	0.624±0.100
Voting_Ensemble	0.761 ± 0.081	0.850 ± 0.060	**0.865 ± 0.032**	0.841 ± 0.098	0.383 ± 0.163	0.649 ± 0.112
Stacking_Ensemble	**0.802 ± 0.026**	**0.885 ± 0.017**	0.849 ± 0.016	**0.926 ± 0.041**	0.217 ± 0.113	0.627 ± 0.107
Weighted_Voting_Ensemble	0.761 ± 0.081	0.850 ± 0.060	**0.865 ± 0.032**	0.841 ± 0.098	0.383 ± 0.163	0.649 ± 0.113

Figure 5

Radar graphic for the sound "aaa"



Analysis of the word 'ana'

In the performance evaluation of the classification models on the word 'ana' presented in Table 2 and Figure 6, the Stacking Ensemble algorithm showed the highest discriminative performance with an accuracy of 78.8% ± 1.6% and an AUC value of 0.639 ± 0.056. The Neural Network model ranked second, with an accuracy of 77.9 ± 3.4% and an AUC value of 0.600 ± 0.072.

In the model-based evaluation, the Random Forest algorithm showed a high capacity to correctly identify asthmatic patients, with a sensitivity of 87.0% ± 6.1%. However, this model has a significant limitation in identifying healthy individuals, with a specificity of 0.233 ± 0.082. This unstable performance profile, similar to the Random Forest results for the 'aaa' sound, suggests that the model is prone to producing a high rate of false positives in clinical applications.

The Logistic Regression algorithm showed moderate success in discriminating between asthmatic and healthy individuals, with an AUC of 0.628 ± 0.100 and specificity of 0.550 ± 0.113. This model provides a more balanced performance between sensitivity and specificity, making it a reliable option for clinical applications.

The K-NN algorithm demonstrated unbalanced performance with an accuracy of 55.8% ± 5.7%, precision of 0.887 ± 0.034, and sensitivity of 0.535 ± 0.084. These results indicate that a high proportion of cases classified as positive by the K-NN model were indeed asthmatic (high precision); however, it was only able to detect half of all asthmatic patients (moderate precision). This profile is consistent with the limited discriminative capacity of the simple algorithms observed in the 'aaa' sound.

Analyzing the ensemble learning approaches, the Voting Ensemble and Weighted Voting Ensemble algorithms showed almost identical performance values for the word 'ana' (both AUC 0.650 ± 0.080). This result suggests that for the phonetic unit 'ana,' as for the sound 'aaa,' different model-weighting strategies do not provide a clear advantage.

In terms of the F1-score metric, the Stacking Ensemble (0.879 ± 0.009) and XGBoost (0.858 ± 0.021) models yielded the highest values. These results demonstrate that these models can achieve an optimal balance between sensitivity and precision, particularly in datasets with class imbalance.

The results of the analysis of the word 'ana,' similar to the findings for the sound 'aaa,' show the limitations of the classification performance of a single phonetic unit but reveal that ensemble-based approaches in particular produce promising results. AUC values in the range of 0.600-0.650 indicate the potential value of this phonetic unit as an auxiliary biomarker for asthma diagnosis.

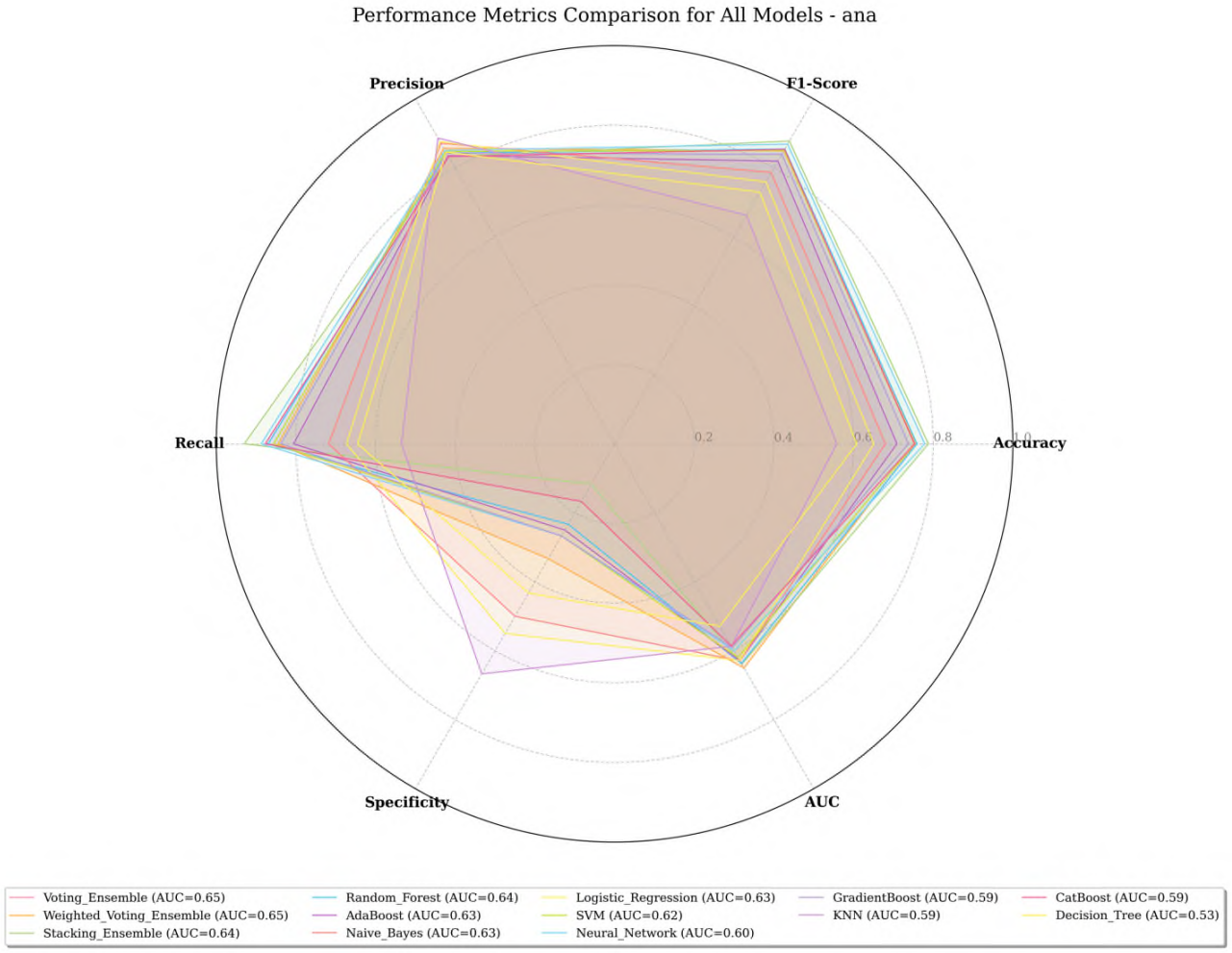
Table 2

Machine learning model performance evaluation metrics for the sound "ana"

Model	Accuracy	F1-Score	Precision	Recall	Specificity	AUC
Naive_Bayes	0.680 ± 0.030	0.787 ± 0.022	0.872 ± 0.019	0.718 ± 0.028	0.500 ± 0.075	0.629 ± 0.077
KNN	0.558 ± 0.057	0.663 ± 0.063	0.887 ± 0.034	0.535 ± 0.084	**0.667 ± 0.139**	0.587 ± 0.068
Decision_Tree	0.608 ± 0.029	0.730 ± 0.025	0.846 ± 0.037	0.645 ± 0.045	0.433 ± 0.162	0.529 ± 0.081
Neural_Network	0.779 ± 0.034	0.869 ± 0.023	0.851 ± 0.009	0.887 ± 0.040	0.267 ± 0.033	0.600 ± 0.072
CatBoost	0.753 ± 0.016	0.854 ± 0.011	0.833 ± 0.008	0.877 ± 0.022	0.167 ± 0.053	0.586 ± 0.052
AdaBoost	0.709 ± 0.029	0.820 ± 0.020	0.836 ± 0.015	0.806 ± 0.035	0.250 ± 0.091	0.629 ± 0.051
GradientBoost	0.738 ± 0.056	0.840 ± 0.038	0.843 ± 0.019	0.838 ± 0.059	0.267 ± 0.062	0.591 ± 0.052
Random_Forest	0.759 ± 0.050	0.855 ± 0.033	0.843 ± 0.017	0.870 ± 0.061	0.233 ± 0.082	0.637 ± 0.068
SVM	0.753 ± 0.025	0.851 ± 0.019	0.848 ± 0.021	0.856 ± 0.048	0.267 ± 0.143	0.620 ± 0.093
Logistic_Regression	0.651 ± 0.051	0.760 ± 0.040	**0.876 ± 0.030**	0.673 ± 0.051	0.550 ± 0.113	0.628 ± 0.100
Voting_Ensemble	0.756 ± 0.047	0.850 ± 0.033	0.857 ± 0.010	0.845 ± 0.066	0.333 ± 0.075	**0.650 ± 0.080**
Stacking_Ensemble	**0.788 ± 0.016**	**0.878 ± 0.010**	0.833 ± 0.011	**0.930 ± 0.027**	0.117 ± 0.085	0.639 ± 0.056
Weighted_Voting_Ensemble	0.756 ± 0.047	0.850 ± 0.033	0.857 ± 0.010	0.845 ± 0.066	0.333 ± 0.075	**0.650 ± 0.080**

Figure 6

Radar graphic for sound "ana"



Analysis of the word 'araba'

In the classification analyses of the word 'araba' presented in Table 3 and Figure 7, the Stacking Ensemble model exhibited the highest discriminative performance among all models, with an accuracy of 82.3 ± 1.5% and an AUC of 0.684 ± 0.076. This result confirms the superiority of ensemble learning approaches in generalizing complex acoustic patterns, as in previous phonetic units ('aaa' and 'ana'). In the model-based evaluation, the Random Forest algorithm showed superior sensitivity in detecting asthma cases (sensitivity: 95.1% ± 2.8%. However, the low specificity value of 0.167 ± 0.075% maintains the similarity limitation observed for the words 'aaa' and' ana.' This suggests that the model can produce a high rate of false-positive results in clinical applications. When the ensemble learning approaches were examined, the Voting Ensemble and Weighted Voting Ensemble models showed high discriminative performance for the word 'araba' with AUC values of 0.704 ± 0.081 and 0.703 ± 0.081, respectively. These results show that, unlike previous findings for phonetic units, the ensemble models achieved higher AUC values for the word 'araba.' This suggests that the acoustic features of the word 'araba' may be more discriminative for asthma detection. The Logistic Regression algorithm showed a balanced performance, with an AUC of 0.685 ± 0.078 and a specificity value of 0.600 ± 0.133. This model exhibits a more balanced profile between sensitivity and specificity than previous

word analyses, thereby providing a reliable option for clinical applications. The CatBoost algorithm demonstrated strong predictive capacity with an accuracy of 79.9 ± 3.4% and an AUC value of 0.673 ± 0.092. This result demonstrates that the proposed algorithm, which is optimized for processing categorical variables, can also effectively model complex patterns in audio data. In contrast, the K-NN model demonstrated the lowest performance among all models, with an accuracy of 48.8% ± 6.5% and an AUC value of 0.521 ± 0.048. This result confirms the limitations of neighborhood-based approaches in high-dimensional and complex datasets, such as voice analysis, as observed in previous "wording" analyses.

The results of the analysis of the word 'araba' showed that, compared with previous phonetic units, the ensemble models achieved higher AUC values (0.684-0.704), indicating that this word may be a stronger acoustic biomarker for asthma diagnosis. This suggests that phonetic units containing multiple vowel and consonant combinations may more clearly reflect the effects of asthma on the voice.

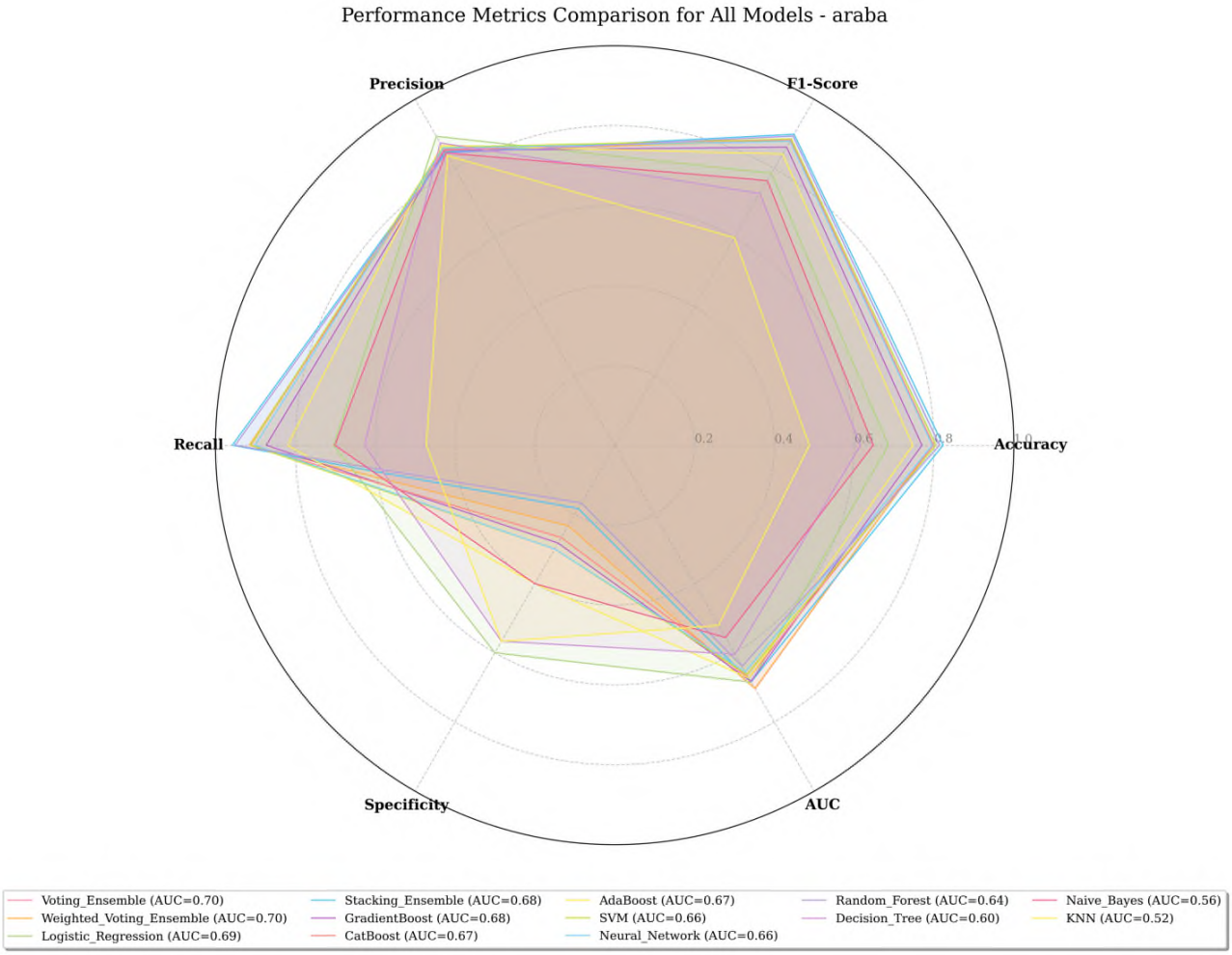
Table 3

Machine learning model performance evaluation metrics for the sound "araba"

Model	Accuracy	F1-Score	Precision	Recall	Specificity	AUC
Naive_Bayes	0.648 ± 0.065	0.765 ± 0.053	0.845 ± 0.022	0.701 ± 0.075	0.400 ± 0.062	0.556 ± 0.083
KNN	0.488 ± 0.065	0.600 ± 0.070	0.837 ± 0.044	0.472 ± 0.076	**0.567 ± 0.143**	0.521 ± 0.048
Decision_Tree	0.616 ± 0.045	0.728 ± 0.040	0.874 ± 0.035	0.627 ± 0.058	**0.567 ± 0.143**	0.604 ± 0.071
Neural_Network	0.796 ± 0.040	0.879 ± 0.025	0.859 ± 0.025	0.901 ± 0.041	0.300 ± 0.135	0.655 ± 0.041
CatBoost	0.799 ± 0.034	0.882 ± 0.020	0.855 ± 0.023	0.912 ± 0.020	0.267 ± 0.122	0.673 ± 0.092
AdaBoost	0.747 ± 0.045	0.842 ± 0.029	0.867 ± 0.028	0.820 ± 0.049	0.400 ± 0.143	0.673 ± 0.075
GradientBoost	0.770 ± 0.041	0.862 ± 0.026	0.852 ± 0.019	0.873 ± 0.037	0.283 ± 0.085	0.683 ± 0.056
Random_Forest	0.814 ± 0.029	0.894 ± 0.017	0.844 ± 0.014	0.951 ± 0.028	0.167 ± 0.075	0.639 ± 0.081
SVM	0.805 ± 0.020	0.885 ± 0.013	0.861 ± 0.014	0.912 ± 0.025	0.300 ± 0.085	0.663 ± 0.052
Logistic_Regression	0.686 ± 0.055	0.786 ± 0.044	**0.893 ± 0.034**	0.704 ± 0.063	0.600 ± 0.133	0.685 ± 0.078
Voting_Ensemble	0.796 ± 0.027	0.881 ± 0.016	0.850 ± 0.015	0.915 ± 0.023	0.233 ± 0.082	**0.704 ± 0.081**
Stacking_Ensemble	**0.823 ± 0.015**	**0.899 ± 0.009**	0.847 ± 0.010	**0.958 ± 0.018**	0.183 ± 0.062	0.684 ± 0.076
Weighted_Voting_Ensemble	0.796 ± 0.027	0.881 ± 0.016	0.850 ± 0.015	0.915 ± 0.023	0.233 ± 0.082	0.703± 0.081

Figure 7

Radar graphic for "araba"



Analysis of the word 'gelecek'

In the classification analyses performed on the word 'gelecek' presented in Table 4 and Figure 8, the Stacking Ensemble model exhibited the highest discriminative performance with an accuracy of 82.0% ± 3.1% and an AUC value of 0.702 ± 0.028. This result confirms the superior ensemble learning approaches observed in the previous phonetic units for the word 'gelecek.' The voting ensemble and weighted voting ensemble models demonstrated similarly high performance, with AUC values of 0.717 ± 0.062 and 0.716 ± 0.062, respectively.

In the model-based evaluation, the Logistic Regression algorithm exhibited remarkable discriminative power, with an AUC value of 0.709 ± 0.049. This result, similar to the Logistic Regression performance for the word 'araba,' suggests that this model can effectively discriminate vocal characteristics between the asthma and control groups. However, a specificity value of 0.583 ± 0.091 indicates that false-favorable rates may increase.

The SVM algorithm demonstrated strong performance with an accuracy of 79.1% ± 2.0% and an AUC value of 0.643 ± 0.069. This model is characterized by its capacity to effectively separate complex sound features

in high-dimensional spaces. Consistent with its performance in previous phonetic units, the discriminative power of the SVM for the word 'gelecek' was moderate to high.

Remarkably, the Decision Tree algorithm showed limited performance with an accuracy of 55.3% ± 7.4% and an AUC of 0.551 ± 0.053. This confirms the inability of a single decision tree to model complex audio features, which has been observed in previous studies. Similarly, the K-NN model also showed limited discriminative capacity, with an accuracy of 59.6 ± 6.5% and an AUC value of 0.606 ± 0.071. This result supports the limited effectiveness of neighborhood-based approaches in complex datasets, such as voice analysis, as observed for the words 'aaa', 'ana' and' araba.'

The analysis results of the word 'gelecek' show that compared with the previous phonetic units, the ensemble models reach higher AUC values (0.702-0.717), indicating that this word may be a stronger acoustic biomarker. In particular, the 0.717 AUC value of the Voting Ensemble model is the highest discriminative value among all the phonetic units analyzed thus far. This suggests that the variety of vowel and consonant sounds and the complexity of articulatory movements in the word 'gelecek' may capture the effects of asthma on the voice more clearly.

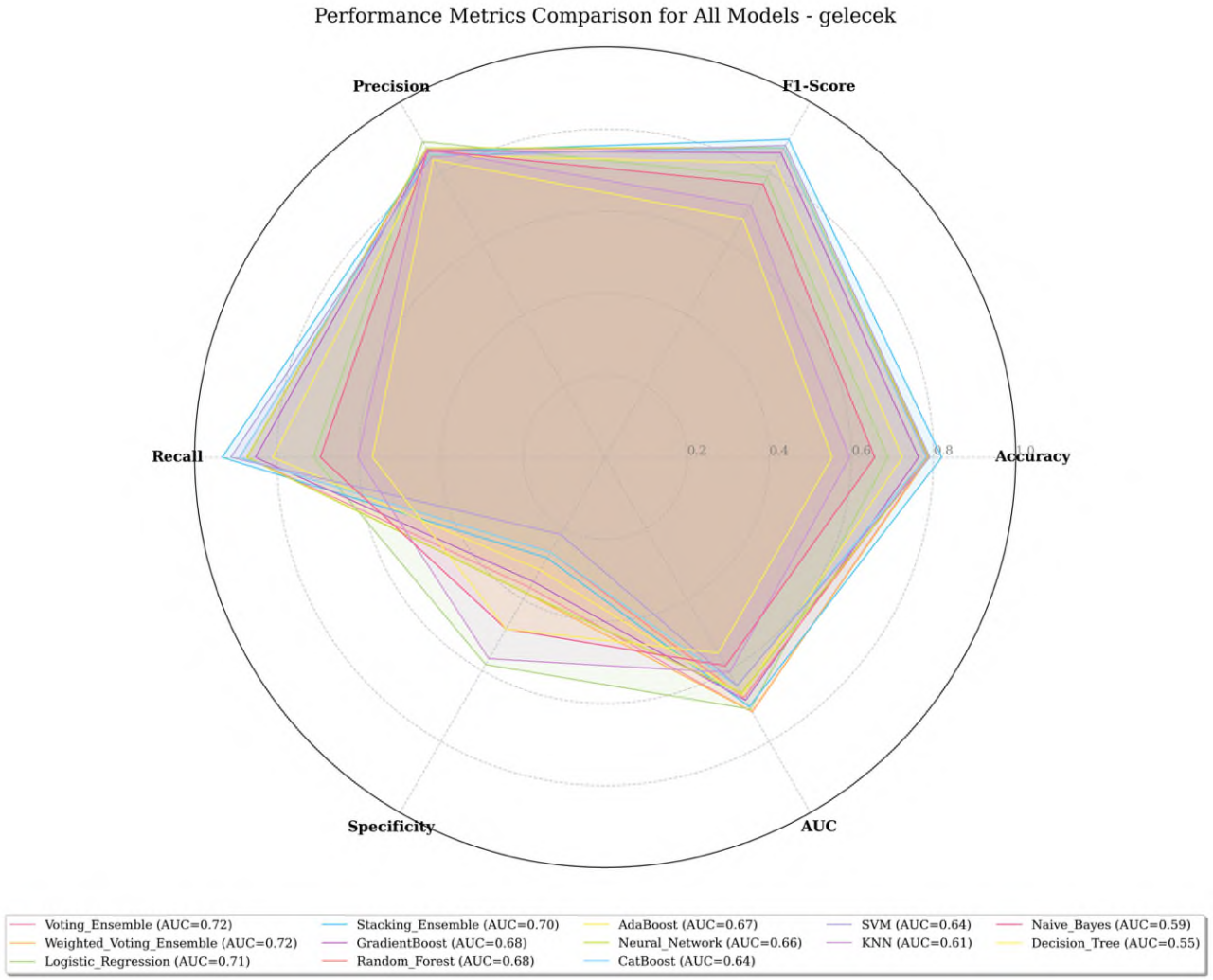
The high AUC values observed for the word 'gelecek' suggest that words with multiple syllables and different phonetic features are more effective biomarkers for asthma diagnosis. This finding emphasizes the diagnostic value of phonetic diversity in clinical practice.

Table 4

Machine learning model performance evaluation metrics for the sound "gelecek"

Model	Accuracy	F1-Score	Precision	Recall	Specificity	AUC
Naive_Bayes	0.657 ± 0.028	0.769 ± 0.020	0.864 ± 0.024	0.694 ± 0.022	0.483 ± 0.097	0.587 ± 0.062
KNN	0.596 ± 0.065	0.709 ± 0.060	0.867 ± 0.033	0.603 ± 0.077	**0.567 ± 0.097**	0.606 ± 0.071
Decision_Tree	0.553 ± 0.074	0.671 ± 0.079	0.838 ± 0.022	0.567 ± 0.104	0.483 ± 0.111	0.551 ± 0.053
Neural_Network	0.788 ± 0.031	0.872 ± 0.019	0.870 ± 0.018	0.873 ± 0.023	0.383 ± 0.085	0.664 ± 0.050
CatBoost	0.782 ± 0.032	0.871 ± 0.018	0.852 ± 0.023	0.891 ± 0.014	0.267 ± 0.122	0.644 ± 0.085
AdaBoost	0.724 ± 0.012	0.829 ± 0.007	0.850 ± 0.028	0.810 ± 0.029	0.317 ± 0.162	0.669 ± 0.031
GradientBoost	0.764 ± 0.017	0.857 ± 0.010	0.863 ± 0.030	0.852 ± 0.032	0.350 ± 0.186	0.684 ± 0.021
Random_Forest	0.782 ± 0.017	0.871 ± 0.008	0.853 ± 0.022	0.891 ± 0.017	0.267 ± 0.133	0.677 ± 0.047
SVM	0.791 ± 0.020	0.878 ± 0.011	0.847 ± 0.014	0.912 ± 0.015	0.217 ± 0.085	0.643 ± 0.069
Logistic_Regression	0.689 ± 0.063	0.789 ± 0.051	**0.889 ± 0.026**	0.711 ± 0.069	0.583 ± 0.091	0.709 ± 0.049
Voting_Ensemble	0.785 ± 0.050	0.870 ± 0.033	0.867 ± 0.024	0.873 ± 0.049	0.367 ± 0.113	**0.717 ± 0.062**
Stacking_Ensemble	**0.820 ± 0.031**	**0.895 ± 0.018**	0.861 ± 0.023	**0.933 ± 0.017**	0.283 ± 0.125	0.702 ± 0.028
Weighted_Voting_Ensemble	0.788 ± 0.047	0.871 ± 0.031	0.870 ± 0.021	0.873 ± 0.049	0.383 ± 0.100	0.716 ± 0.062

Figure 8
Radar graphic for "gelecek"



Analysis of the word 'ordu'

In the classification analyses performed on the word 'ordu' presented in Table 5 and Figure 9, the Stacking Ensemble model exhibited the highest overall performance with an accuracy of 79.6% ± 3.0% and an AUC of 0.678 ± 0.043. This result confirms the superiority of ensemble learning approaches for generalizing complex acoustic patterns, as observed in previous phonetic units. In the model-based evaluation, the SVM algorithm showed strong discriminative capacity with an accuracy of 79.0% ± 4.6% and an AUC value of 0.686 ± 0.077. This result confirms that SVM can discriminate effectively in high-dimensional acoustic feature space, as observed for the word 'gelecek.' However, the limitation in the specificity values is a factor to be considered in clinical applications. The Random Forest model showed high sensitivity in detecting asthma cases, with a sensitivity of 88.7 ± 5.4%. However, it showed a significant limitation in identifying healthy individuals, with a specificity of 0.200 ± 0.067%. This unbalanced performance profile is consistent with the low specificity problem observed for Random Forest in previous "wording" analyses. This highlights the risk of producing a high rate of false positive results in clinical applications. The Logistic Regression algorithm showed moderate success in discriminating between asthmatic and healthy individuals, with an AUC of 0.665 ± 0.035 and

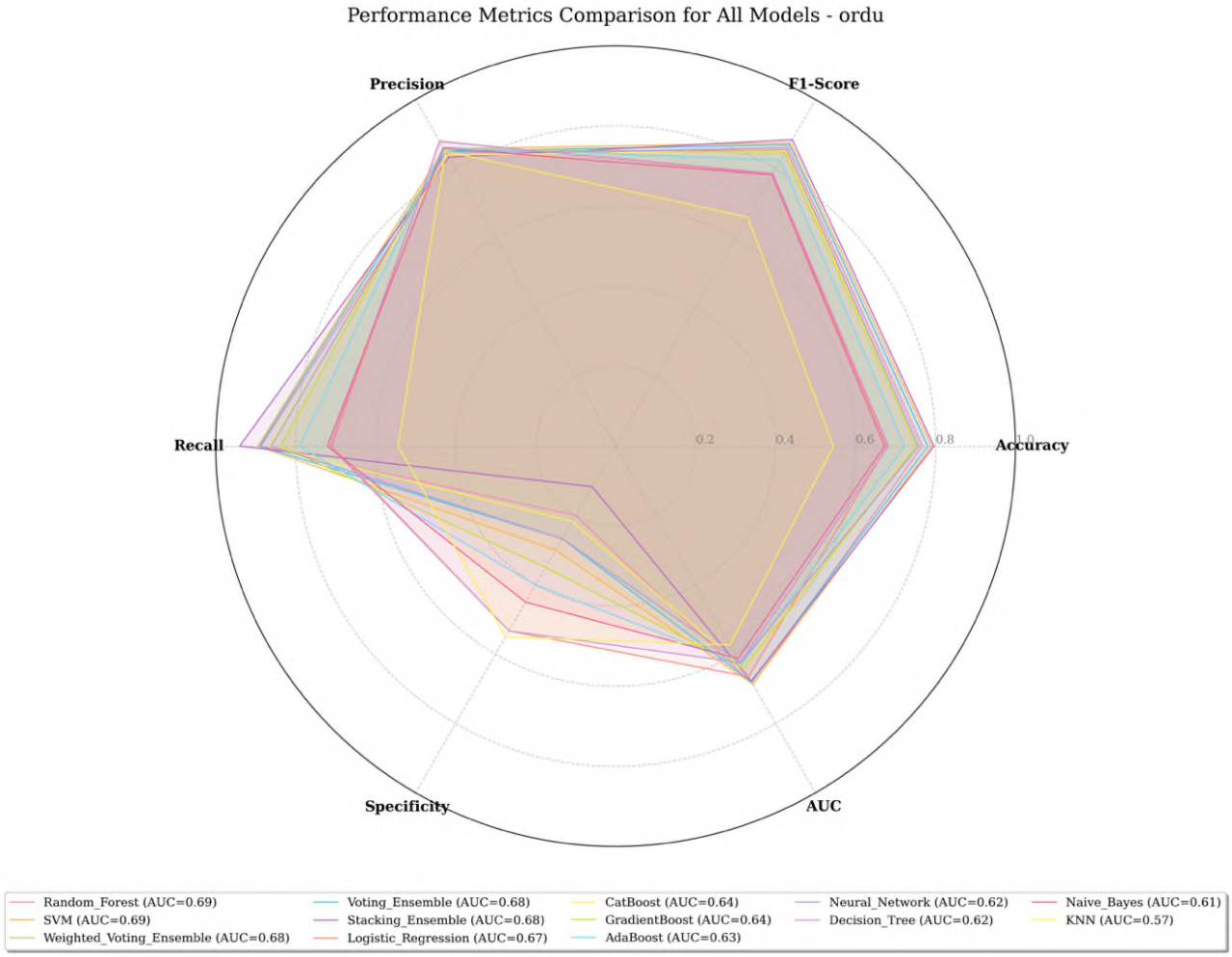
specificity value of 0.533 ± 0.135. This result is consistent with the Logistic Regression performance observed for the words 'gelecek' araba' araba,' confirming that the model has balanced discriminative power in voice-based biomarkers. When the ensemble learning approaches were analyzed, the Weighted Voting Ensemble model with an AUC value of 0.681 ± 0.041 minimally exceeded the AUC value of 0.680 ± 0.041 for the Voting Ensemble model. This demonstrates that, as in previous word analyses, the model weighting strategy for the word 'ordu' only maximizes classification performance. The K-NN algorithm exhibited limited discriminative capacity, with an accuracy of 54.5% ± 6.4% and an AUC value of 0.573 ± 0.106. This result confirms the limited effectiveness of neighborhood-based approaches in high-dimensional and complex datasets, such as voice analysis, which has been consistently observed in all previous word analyses. The results of the analysis of the word 'ordu' show that, compared to the last phonetic units, the community models exhibit slightly lower discriminative power than the word 'gelecek' (0.702-0.717), but higher discriminative power than the words 'aaa' (0.627-0.649) and 'ana' (0.639-0.650), with AUC values in the range 0.678-0.686. This suggests that the phonetic features of the word 'ordu' may reflect the effects of asthma on the voice at a medium-high level.

Table 5

Machine learning model performance evaluation metrics for the sound "ordu"

Model	Accuracy	F1-Score	Precision	Recall	Specificity	AUC
Naive_Bayes	0.673 ± 0.054	0.783 ± 0.044	0.860 ± 0.015	0.721 ± 0.064	0.450 ± 0.041	0.613 ± 0.076
KNN	0.545 ± 0.064	0.661 ± 0.059	0.851 ± 0.044	0.544 ± 0.073	**0.550 ± 0.145**	0.573 ± 0.106
Decision_Tree	0.682 ± 0.047	0.787 ± 0.035	0.879 ± 0.031	0.714 ± 0.044	0.533 ± 0.135	0.624 ± 0.070
Neural_Network	0.758 ± 0.026	0.855 ± 0.016	0.848 ± 0.017	0.862 ± 0.026	0.267 ± 0.097	0.625 ± 0.085
CatBoost	0.752 ± 0.027	0.852 ± 0.018	0.840 ± 0.018	0.866 ± 0.035	0.217 ± 0.113	0.642 ± 0.034
AdaBoost	0.723 ± 0.026	0.825 ± 0.021	0.862 ± 0.015	0.791 ± 0.041	0.400 ± 0.097	0.630 ± 0.042
GradientBoost	0.752 ± 0.044	0.847 ± 0.032	0.858 ± 0.012	0.837 ± 0.057	0.350 ± 0.062	0.641 ± 0.042
Random_Forest	0.767 ± 0.051	0.862 ± 0.033	0.839 ± 0.017	0.887 ± 0.054	0.200 ± 0.067	**0.687 ± 0.060**
SVM	0.790 ± 0.046	0.875 ± 0.027	0.858 ± 0.026	0.894 ± 0.033	0.300 ± 0.135	0.686 ± 0.077
Logistic_Regression	0.679 ± 0.027	0.784 ± 0.027	**0.880 ± 0.023**	0.710 ± 0.055	0.533 ± 0.135	0.665 ± 0.035
Voting_Ensemble	0.781 ± 0.025	0.870 ± 0.015	0.851 ± 0.015	0.890 ± 0.021	0.267 ± 0.082	0.680 ± 0.041
Stacking_Ensemble	**0.796 ± 0.030**	**0.884 ± 0.018**	0.834 ± 0.013	**0.940 ± 0.026**	0.117 ± 0.067	0.678 ± 0.043
Weighted_Voting_Ensemble	0.781 ± 0.025	0.870 ± 0.015	0.851 ± 0.015	0.890 ± 0.021	0.267 ± 0.082	0.681 ± 0.041

Figure 9
Radar graphic for "ordu"



Analysis of the word 'titiz'

In the classification analyses performed on the word 'titiz' presented in Table 6 and Figure 10, the Voting Ensemble and Weighted Voting Ensemble models exhibited the highest discriminative performance with AUC values of 0.736 ± 0.067 and 0.736 ± 0.068, respectively. These results confirm the superiority of ensemble learning approaches in generalizing complex acoustic patterns for the word 'fastidious,' as observed for all previous phonetic units. The Stacking Ensemble model showed superior performance, with an accuracy of 79.9% ± 4.0% and an AUC of 0.731 ± 0.064. In the model-based evaluation, the Neural Network algorithm exhibited a discriminative power close to that of the ensemble models with an accuracy of 79.9% ± 6.0% and an AUC value of 0.734 ± 0.056. This result shows that artificial neural networks can effectively model the acoustic features of the word 'titiz.' The SVM algorithm also demonstrated a strong classification capacity, with an accuracy of 76.4 ± 2.9% and an AUC of 0.718 ± 0.066. This performance profile confirms that the SVM can discriminate effectively in high-dimensional acoustic feature space, as observed for the words 'gelecek' and' ordu.' The Logistic Regression algorithm showed remarkable performance with an AUC of 0.710 ± 0.095 and specificity value of 0.567 ± 0.226. This result is very close to the AUC value of 0.709 obtained for the word'

palecek, demonstrating that the Logistic Regression model exhibits consistent discriminative power across different phonetic units.

As observed in all previous wording analyses, the Random Forest model showed high sensitivity in detecting asthma cases, with a sensitivity of 87.7 ± 3.7%. However, it had a marked limitation in identifying healthy individuals, with specificity of 0.200 ± 0.085%. This unbalanced performance profile was also consistently observed for the words 'aaa', 'ana, ' 'araba, ' 'gelecek' and 'ordu, ' indicating that the Random Forest model tends to produce a high rate of false positives in clinical applications.

Simpler models, such as K-NN and Decision Tree, exhibited limited discriminative capacity with AUC values of 0.583 ± 0.088 and 0.577 ± 0.051, respectively. This result confirms that simple algorithms are inadequate for complex and high-dimensional datasets, such as voice analysis, as consistently observed in all previous word analyses.

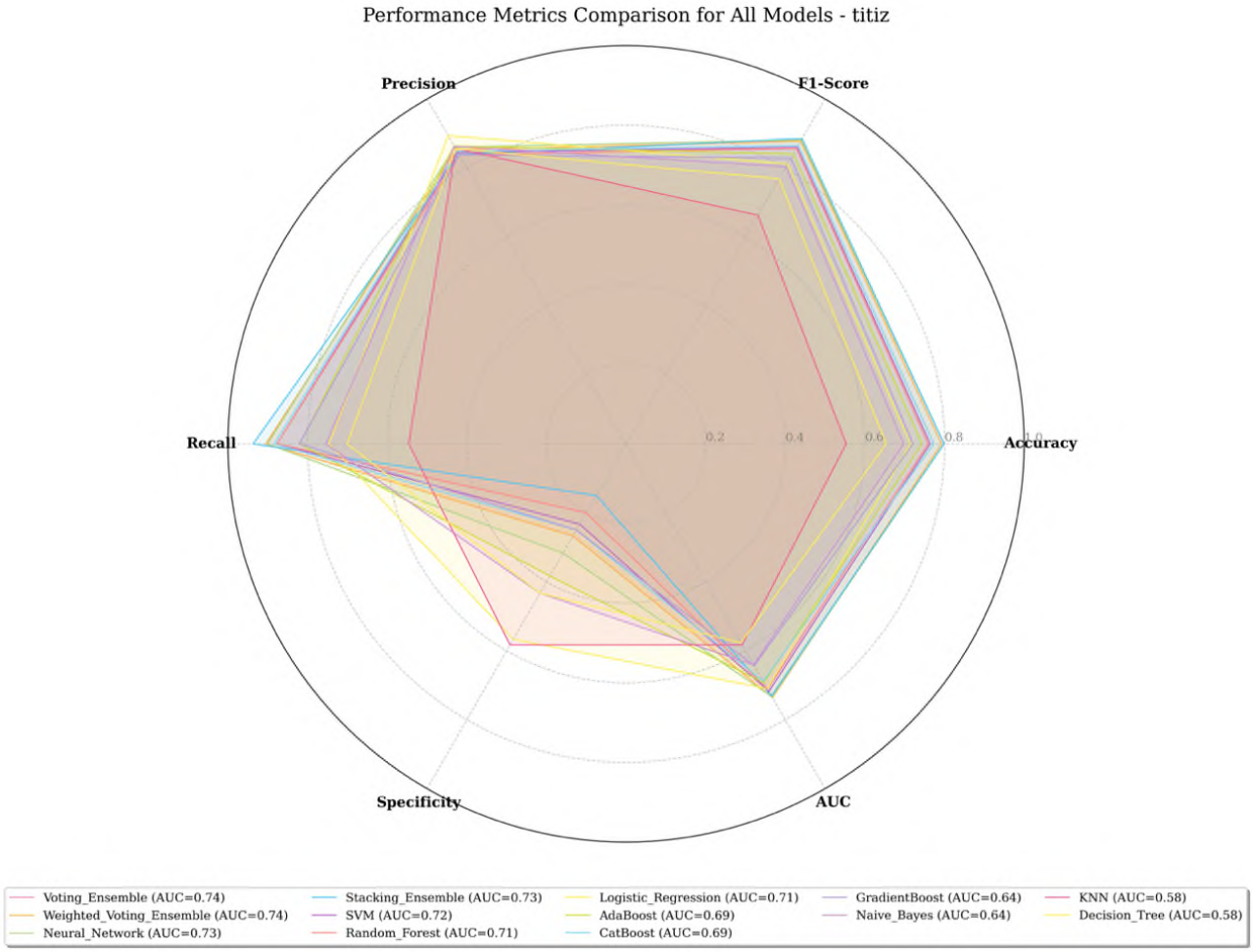
The analysis results of the word 'rigorous' show that, compared to the previous phonetic units, the ensemble models exhibit the highest discriminative power among all the words analyzed thus far, with AUC values in the range of 0.731-0.736. These values also exceed the AUC range of 0.702-0.717 obtained for the word 'gelecek,' indicating that 'titiz' may be the most powerful acoustic biomarker for asthma diagnosis. This suggests that the complexity of consonant sound combinations and articulatory movements in the word 'titiz' may reflect the effects of asthma on the voice more clearly.

Table 6

Machine learning model performance evaluation metrics for the sound "titiz"

Model	Accuracy	F1-Score	Precision	Recall	Specificity	AUC
Naive_Bayes	0.698±0.040	0.804±0.025	0.863±0.031	0.754±0.029	0.433±0.143	0.641±0.104
KNN	0.553 ± 0.088	0.663 ± 0.085	0.857 ± 0.035	0.546 ± 0.101	**0.583 ± 0.075**	0.583 ± 0.088
Decision_Tree	0.654 ± 0.033	0.769 ± 0.026	0.856 ± 0.031	0.701 ± 0.041	0.433 ± 0.143	0.577 ± 0.051
Neural_Network	**0.799 ± 0.060**	0.881 ± 0.037	0.861 ± 0.024	0.901 ± 0.052	0.317 ± 0.097	0.734 ± 0.056
CatBoost	0.773 ± 0.027	0.865 ± 0.019	0.848 ± 0.006	0.884 ± 0.041	0.250 ± 0.053	0.689 ± 0.066
AdaBoost	0.744 ± 0.019	0.841 ± 0.013	**0.864 ± 0.020**	0.820 ± 0.025	0.383 ± 0.113	0.693 ± 0.048
GradientBoost	0.721 ± 0.018	0.829 ± 0.014	0.838 ± 0.007	0.821 ± 0.029	0.250 ± 0.053	0.644 ± 0.019
Random_Forest	0.759 ± 0.033	0.857 ± 0.022	0.838 ± 0.015	0.877 ± 0.037	0.200 ± 0.085	0.711 ± 0.062
SVM	0.764 ± 0.029	0.860 ± 0.016	0.845 ± 0.024	0.877 ± 0.019	0.233 ± 0.133	0.718 ± 0.066
Logistic_Regression	0.715 ± 0.038	0.812 ± 0.025	0.894 ± 0.049	0.747 ± 0.041	0.567 ± 0.226	0.710 ± 0.095
Voting_Ensemble	0.793 ± 0.043	0.878 ± 0.028	0.853 ± 0.011	0.905 ± 0.048	0.267 ± 0.033	**0.736 ± 0.067**
Stacking_Ensemble	**0.799 ± 0.040**	**0.885 ± 0.025**	0.839 ± 0.012	**0.937 ± 0.047**	0.150 ± 0.062	0.731 ± 0.064
Weighted_Voting_Ensemble	0.793 ± 0.043	0.878 ± 0.028	0.853 ± 0.011	0.905 ± 0.048	0.267 ± 0.033	**0.736 ± 0.068**

Figure 10
Radar graphic for "titiz"



Analysis of the word 'ünlem'

In the classification analyses performed on the word 'ünlem' presented in Table 7 and Figure 11, the Stacking Ensemble model exhibited the highest overall performance with an accuracy of 81.4% ± 4.1% and an AUC of 0.716 ± 0.070. This result confirms the superiority of ensemble learning approaches in generalizing complex acoustic patterns for the word 'ünlem,' as observed for all previous phonetic units. The Voting Ensemble and Weighted Voting Ensemble models showed similarly high discriminative power, with AUC values of 0.726 ± 0.084 and 0.726 ± 0.083, respectively. In the model-based evaluation, the Logistic Regression algorithm performed remarkably well, with an AUC of 0.723 ± 0.085 and a specificity value of 0.633 ± 0.135. This result is similar to the values obtained for the words 'titiz' (0.710 AUC) and 'gelecek' (0.709 AUC), confirming that the Logistic Regression model can effectively discriminate voice characteristics between the asthma and control groups.

The SVM algorithm demonstrated strong classification capacity, with an accuracy of 79.4 ± 2.7% and an AUC value of 0.713 ± 0.041. This performance profile confirms that the SVM can discriminate effectively in high-dimensional acoustic feature space, as observed for the words 'titiz' (0.718 AUC) and 'gelecek' (0.643 AUC). The Random Forest model showed superior sensitivity in detecting asthma cases, with a sensitivity

of 88.4 ± 2.7%, as consistently observed in all previous wording analyses. However, it exhibited a marked limitation in identifying healthy individuals, with a specificity of 0.300 ± 0.145%. This uneven performance profile suggests that the Random Forest model tends to produce a high rate of false positives in clinical applications.

Remarkably, unlike the previous phonetic units, the Decision Tree algorithm showed a relatively higher specificity for the word 'ünlem' with an AUC of 0.659 ± 0.132 and a specificity value of 0.700 ± 0.215. This suggests that decision-tree-based approaches may be advantageous in terms of specificity for words with certain phonetic features.

The K-NN algorithm exhibited limited discriminative capacity, with an accuracy of 55.8 ± 4.8% and an AUC value of 0.652 ± 0.043. This result confirms the limited effectiveness of neighborhood-based approaches on high-dimensional and complex datasets, such as voice analysis, as consistently observed in all previous wording analyses.

The results of the analysis of the word 'ünlem' show that, compared to the previous phonetic units, the ensemble models exhibit a discriminative power slightly lower than the word 'titiz' (0.731-0.736), but close to the word 'gelecek' (0.702-0.717), with AUC values in the range 0.716-0.726. This suggests that the vowel and consonant combinations contained in the word 'ünlem' can significantly reflect the effects of asthma on the voice.

The overall performance profile of the models for the word 'ünlem,' particularly high sensitivity but limited specificity in most models, emphasizes the importance of balancing sensitivity and specificity in the diagnosis of asthma. This balance requires the development of optimized models that can detect a high proportion of true asthma cases while minimizing false-positives in clinical practice.

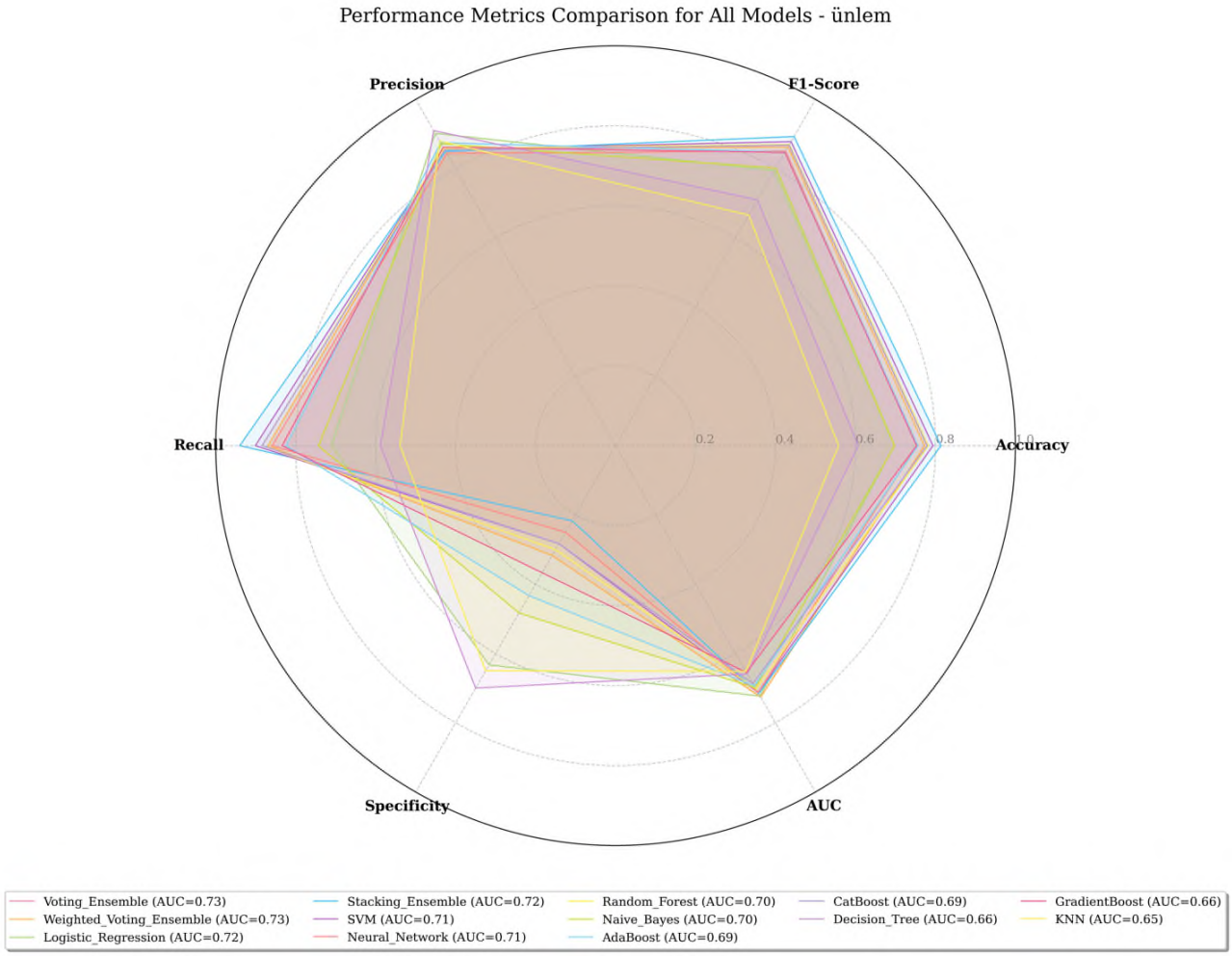
Table 7

Machine learning model performance evaluation metrics for the sound "ünlem"

Model	Accuracy	F1-Score	Precision	Recall	Specificity	AUC
Naive_Bayes	0.698±0.041	0.802 ± 0.027	0.872 ± 0.030	0.743 ± 0.027	0.483 ± 0.122	0.699 ± 0.085
KNN	0.558 ± 0.048	0.666 ± 0.048	0.880 ± 0.027	0.539 ± 0.064	0.650 ± 0.097	0.652 ± 0.043
Decision_Tree	0.607 ± 0.042	0.710 ± 0.040	**0.910 ± 0.062**	0.588 ± 0.063	**0.700 ± 0.215**	0.659 ± 0.132
Neural_Network	0.753 ± 0.030	0.851 ± 0.020	0.844 ± 0.016	0.859 ± 0.033	0.250 ± 0.091	**0.713 ± 0.063**
CatBoost	0.779 ± 0.042	0.868 ± 0.027	0.854 ± 0.023	0.884 ± 0.046	0.283 ± 0.125	0.687 ± 0.072
AdaBoost	0.756 ± 0.036	0.848 ± 0.021	0.875 ± 0.037	0.824 ± 0.019	0.433 ± 0.193	0.694 ± 0.085
GradientBoost	0.753 ± 0.055	0.847 ± 0.035	0.862 ± 0.036	0.834 ± 0.046	0.367 ± 0.172	0.655 ± 0.077
Random_Forest	0.782 ± 0.019	0.870 ± 0.012	0.858 ± 0.023	0.884 ± 0.027	0.300 ± 0.145	0.703 ± 0.087
SVM	**0.794 ± 0.027**	**0.878 ± 0.017**	0.856 ± 0.012	**0.901 ± 0.029**	0.283 ± 0.067	**0.713 ± 0.041**
Logistic_Regression	0.697 ± 0.041	0.795 ± 0.029	0.902 ± 0.036	0.711 ± 0.032	0.633 ± 0.135	0.723 ± 0.085
Voting_Ensemble	0.773 ± 0.043	0.863 ± 0.031	0.857 ± 0.010	0.869 ± 0.052	0.317 ± 0.033	0.726 ± 0.084
Stacking_Ensemble	0.814 ± 0.041	0.893 ± 0.025	0.850 ± 0.017	0.940 ± 0.037	0.217 ± 0.085	0.716 ± 0.070
Weighted_Voting_Ensemble	0.773 ± 0.043	0.863 ± 0.031	0.857 ± 0.010	0.869 ± 0.052	0.317 ± 0.033	0.726 ± 0.083

Figure 11

Radar graphic of "ünlem"



Discussion

In this study, the performance of 13 machine learning models for asthma diagnosis using different sound samples ('aaa, ' 'ana,' 'araba,' 'gelecek, ' 'ordu, ' 'titiz,' 'ünlem') was evaluated using Accuracy, F1-Score, Precision, Recall, Specificity and AUC (area under the curve) metrics. The findings revealed that ensemble models (Stacking Ensemble, Voting Ensemble, Weighted Voting Ensemble) generally demonstrate superior performance. For example, the Stacking Ensemble model achieved the best results with 82.3% accuracy and 0.684 AUC for the word 'araba' and 82.0% accuracy and 0.702 AUC for the word 'gelecek.' This result confirms that the ensemble models can better generalize complex patterns. In addition, complex models such as neural networks and random forests have yielded remarkable results. In particular, the neural network model demonstrated high sensitivity with 90.1% recall rates for the words 'araba' and 'titiz' respectively. However, the low specificity values of these models resulted in limitations in practical applications due to the increase in false-positive rates. Models such as SVM and Logistic Regression also performed well, with generally high precision and AUC values. For example, the SVM model ranked second for the word 'gelecek, ' with an AUC value of 0.643. In contrast, simpler models, such as the naive Bayes and K-NN models, demonstrated limited performance owing to low accuracy and specificity values. In particular, K-NN had

low values, such as 59.6% accuracy and an AUC of 0.606 for the word 'gelecek.' This indicates that they are inadequate for capturing complex patterns in audio data. In conclusion, the Stacking Ensemble model stood out as the most promising asthma diagnostic model, with high recall and AUC values.

However, the low specificity values of some models emphasize the need to control false-positive rates. Furthermore, the use of MFCC and additional acoustic features confirmed that MFCC is an effective feature extraction method for asthma diagnosis. On the other hand, the fact that deep learning techniques could not be evaluated in this study owing to the limited dataset provides an opportunity for future research.

Conclusions

In this study, the effectiveness of voice analysis-based machine learning models for asthma diagnosis was comprehensively evaluated. Within the scope of the study, a total of 13 different machine learning models were developed for different sound samples ('aaa, ' 'ana,' 'araba,' 'gelecek, ' 'ordu, ' 'titiz,' 'ünlem'), and their performances were compared. The findings showed that ensemble learning approaches (ensemble models) exhibit superior performance in asthma diagnosis.

The stacking ensemble model obtained the best overall performance. In particular, it achieved high AUC values (68.4% and 70.2%) for sound examples such as 'araba' and 'gelecek.' This result confirms that ensemble models can better generalize complex sound patterns. The Voting Ensemble and Weighted Voting Ensemble models showed similar performances; however, low specificity values were identified as a limitation to be considered in practical applications.

Complex models, such as Neural Networks and Gradient Boosting, have also demonstrated remarkable results. In particular, the Neural Network model provided high sensitivity with recall rates of 90.1% for the 'araba' and 'titiz' sound samples. However, the low specificity of these models should be considered a factor that may increase false-positive results in clinical applications.

The use of Mel-Frequency Cepstral Coefficients (MFCCs) and additional acoustic features (spectral center, spectral bandwidth, etc.) was confirmed to be an effective feature extraction method for asthma diagnosis.

The important limitations of our study include the limited and single-center dataset of 284 patients with asthma diagnosed using the GINA criteria and 60 healthy volunteers. This may have affected the generalizability of the models to larger and more diverse populations. Nevertheless, this study demonstrates the potential of voice analysis-based approaches as cost-effective screening tools for early detection, especially in areas with limited access to health services. The high-sensitivity values of the community models support this potential.

This study has important implications for clinical practice. In particular, the high sensitivity values of the community models suggest that they can be used as asthma screening tools. A voice analysis-based diagnostic approach may provide a cost-effective alternative for early diagnosis, particularly in areas with limited access to healthcare services. It may also enable remote and continuous monitoring of asthma symptoms in patients' daily lives through smartphone applications.

Several directions for future research are identified. First, validating the models in multicenter, larger, and more diverse populations (differing in age, ethnicity, and asthma phenotypes) is critical for enhancing generalizability, a need also highlighted by community-based studies, such as Gunawardana et al. (2024). Second, expanding the dataset could further improve model performance by enabling the application of deep learning techniques (e.g., CNN, RNN, Transformer architectures), as studies such as Chen et al. (2025)

demonstrated the potential of DNNs in related tasks, albeit with different datasets and objectives. Adapting existing models to other respiratory diseases (e.g., COPD, pneumonia, and bronchitis) using a transfer learning approach has emerged as a valuable research area.

Data Sharing: Contact the authors within three years of the date of publication to request the datasets used in the study. Data sharing will be performed while maintaining participant confidentiality.

Ethics Committee Approval: Ethics committee approval was received for this study from the ethics committee of İstanbul Yedikule Chest Diseases and Thoracic Surgery Training and Research Hospital (Date: 14.12.2023, No: 2023-433).

Informed Consent: Written informed consent was obtained from all participants who participated in this study.

Peer Review: Externally peer-reviewed.

Author Contributions: Conception/Design of Study- M.G., M.A.U.; Data Acquisition- N.A., C.O., E.Y.N.; Data Analysis/Interpretation- M.G.; Drafting Manuscript- M.G., M.A.U., N.A., C.O., E.Y.N.; Critical Revision of Manuscript- M.G.; Final Approval and Accountability- M.G., M.A.U., N.A., C.O., E.Y.N.

Conflict of Interest: The authors have no conflict of interest to declare.

Grant Support: This study was carried out with the financial support of MedCase Yazılım Teknolojileri A.Ş, which was used for data collection processes. The sponsoring organisation did not influence the design of the study, data analysis, interpretation of the results, or writing of the manuscript.

Acknowledgements: The author(s) would like to thank Oğuz Akbaş (MonitorCRO) for his invaluable contributions and support in the scientific writing of this article.

Author Details

Murat Gezer

[1] İstanbul University, Faculty of Science, Department of Computer Science, İstanbul, Türkiye

0000-0002-7286-3943 murat.gezer@istanbul.edu.tr

Mehmet Atilla Uysal

[2] University of Health Sciences Türkiye, Yedikule Hospital for Chest Diseases and Thoracic Surgery, Department of Chest Diseases, İstanbul, Türkiye

0000-0002-0430-498X

Neval Alagöz

[2] University of Health Sciences Türkiye, Yedikule Hospital for Chest Diseases and Thoracic Surgery, Department of Chest Diseases, İstanbul, Türkiye

0000-0002-4203-4562

Can Ortak

[3] Medcase Corp, İstanbul, Türkiye

0009-0001-3299-5562

Elif Yelda Niksarlıoğlu

[2] University of Health Sciences Türkiye, Yedikule Hospital for Chest Diseases and Thoracic Surgery, Department of Chest Diseases, İstanbul, Türkiye

0000-0002-6119-6540

References

Alam, M. Z., Simonetti, A., Brillantino, R., Tayler, N., Grainge, C., Siribaddana, P., Nouraei, S. A. R., Batchelor, J., Rahman, M. S., Mancuzo, E. V., Holloway, J. W., Holloway, J. A., & Rezwan, F. I. (2022). Predicting pulmonary function from the analysis of voice: A machine learning approach. *Frontiers in Digital Health*, 4, 750226. https://doi.org/10.3389/fdgth.2022.750226

Aykanat, M., Kılıç, Ö., Kurt, B., & Saryal, S. B. (2020). Lung disease classification using machine learning algorithms. *International Journal of Applied Mathematics Electronics and Computers, 8*(4), 125–132. https://doi.org/10.18100/ijamec.799363

Barnes, P. J. (2008). Immunology of asthma and chronic obstructive pulmonary disease. *Nature Reviews Immunology, 8*(3), 183–192. https://doi.org/10.1038/nri2254

Bolat, E. (2021). *Use of machine learning methods in the classification of respiratory system diseases* [Unpublished doctoral dissertation]. Istanbul University Institute of Health Sciences.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*, 321–357. https://doi.org/10.1613/jair.953

Chen, M.-H., Lee, G., & Hung, L.-P. (2025). AI-driven data analysis for asthma risk prediction. *Healthcare, 13*, 774. https://doi.org/10.3390/healthcare13070774

Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing, 28*(4), 357–366. https://doi.org/10.1109/TASSP.1980.1163420

Dietterich, T. G. (2000). Ensemble methods in machine learning. In *Multiple classifier systems* (pp. 1–15). Springer. https://doi.org/10.1007/3-540-45014-9_1

Fesci, H., & Görgülü, Ü. (2005). Asthma and life. *Journal of School of Nursing, 1*(1), 77–83. https://dergipark.org.tr/tr/download/article-file/88607

Ganaie, M. A., Hu, M., Tanveer, M., & Suganthan, P. N. (2022). Ensemble deep learning: A review. *Engineering Applications of Artificial Intelligence, 115*, 105151. https://doi.org/10.1016/j.engappai.2022.105151

Global Initiative for Asthma. (2023). *Global strategy for asthma management and prevention*. https://ginasthma.org/2023-gina-main-report/

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Gunawardana, J., Viswakula, S., Rannan-Eliya, R. P., & Wijemunige, N. (2024). Machine learning approaches for asthma disease prediction among adults in Sri Lanka. *Health Informatics Journal, 30*(3), 14604582241283968. https://doi.org/10.1177/14604582241283968

Holgate, S. T., Wenzel, S., Postma, D. S., Weiss, S. T., Renz, H., & Sly, P. D. (2015). Asthma. *Nature Reviews Disease Primers, 1*, 15025. https://doi.org/10.1038/nrdp.2015.25

Kaplan, A., Cao, H., FitzGerald, J. M., Iannotti, N., Yang, E., Kocks, J. W. H., Kostikas, K., Price, D., Reddel, H. K., Tsiligianni, I., Vogelmeier, C. F., Pfister, P., & Mastoridis, P. (2021). Artificial intelligence/machine learning in respiratory medicine and potential role in asthma and COPD diagnosis. *The Journal of Allergy and Clinical Immunology: In Practice, 9*(6), 2255–2261. https://doi.org/10.1016/j.jaip.2021.02.014

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence* (Vol. 2, pp. 1137–1143).

Lambrecht, B. N., & Hammad, H. (2015). The immunology of asthma. *Nature Immunology, 16*(1), 45–56. https://doi.org/10.1038/ni.3049

Logan, B. (2000). Mel frequency cepstral coefficients for music modeling. In *Proceedings of the 1st International Symposium on Music Information Retrieval* (pp. 1–13). https://doi.org/10.5281/zenodo.1416444

Malmberg, L. P., Sovijärvi, A. R. A., Paajanen, E., Piirilä, P., Haahtela, T., & Katila, T. (1994). Changes in frequency spectra of breath sounds during histamine challenge test in adult asthmatics and healthy control subjects. *Chest, 105*(1), 122–131. https://doi.org/10.1378/chest.105.1.122

Mayr, W., Triantafyllopoulos, A., Batliner, A., Schuller, B. W., & Berghaus, T. M. (2025). Assessing the clinical and functional status of COPD patients using speech analysis during and after exacerbation. *International Journal of Chronic Obstructive Pulmonary Disease, 20*, 137–147.

Ministry of Health of the Republic of Turkey, General Directorate of Public Health. (2024, August 1). *World Asthma Day May 7, 2024*. https://hsgm.saglik.gov.tr/tr/haberler/7-mayis-2024-duenya-astim-guenue.html

Petmezas, G., Cheimariotis, G.-A., Stefanopoulos, L., Rocha, B., Paiva, R. P., Katsaggelos, A. K., & Maglaveras, N. (2022). Automated lung sound classification using a hybrid CNN-LSTM network and focal loss function. *Sensors, 22*(3), 1232. https://doi.org/10.3390/s22031232

Rabiner, L. R., & Schafer, R. W. (2007). *Introduction to digital speech processing*. Now Publishers. https://doi.org/10.1561/2000000001

Reichert, S., Gass, R., Brandt, C., & Andrès, E. (2008). Analysis of respiratory sounds: State of the art. *Clinical Medicine: Circulatory, Respiratory and Pulmonary Medicine, 2*, 45–58. https://doi.org/10.4137/CCRPM.S530

Rivas-Navarrete, J. A., Pérez-Espinosa, H., Padilla-Ortiz, A. L., Rodríguez-González, A. Y., & García-Cambero, D. C. (2025). Edge computing system for automatic detection of chronic respiratory diseases using audio analysis. *Journal of Medical Systems, 49*, 33. https://doi.org/10.1007/s10916-025-02154-7

Saygılı, M. (2019). Mobile health. In M. S. Yıldız (Ed.), *Advanced technology applications in healthcare* (pp. 211–238). Nobel Academic Publishing.

Schreur, H. J. W., Vanderschoot, J., Zwinderman, A. H., Dijkman, J. H., & Sterk, P. J. (1994). Abnormal lung sounds in patients with asthma during episodes with normal lung function. *Chest, 106*(1), 91–99. https://doi.org/10.1378/chest.106.1.91

Sezer, E., & Akıl, S. (2020). Criteria used by speech-language pathologists in Turkey for diagnosing childhood apraxia of speech. *Journal of Language, Speech and Swallowing Research, 3*(2), 153–174.

Shorten, C., Khoshgoftaar, T. M., & Furht, B. (2021). Deep learning applications for COVID-19. *Journal of Big Data, 8*, 18. https://doi.org/10.1186/s40537-020-00392-9

Sterling, M., Rhee, H., & Bocko, M. (2014). Automated cough assessment on a mobile platform. *Journal of Medical Engineering, 2014*, 951621. https://doi.org/10.1155/2014/951621

Topaz, M., Zolnoori, M., Norful, A. A., Perrier, A., Kostic, Z., & George, M. (2022). Speech recognition can help evaluate shared decision making and predict medication adherence in primary care setting. *PLOS ONE, 17*(8), e0271884. https://doi.org/10.1371/journal.pone.0271884

World Health Organization. (2024, May 6). *Asthma*. https://www.who.int/news-room/fact-sheets/detail/asthma

Yadav, S., Keerthana, M., Gope, D., Maheswari, U. K., & Ghosh, P. K. (2020). Analysis of acoustic features for speech sound-based classification of asthmatic and healthy subjects. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 6789–6793). IEEE. https://doi.org/10.1109/ICASSP40776.2020.9054062

Yadav, S., Nk, K., Gope, D., Krishnaswamy, U. M., & Ghosh, P. K. (2018). Comparison of cough, wheeze and sustained phonations for automatic classification between healthy subjects and asthmatic patients. In *Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (pp. 1400–1403).

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., & Moore, G. (2006). *The HTK book*. Cambridge University Engineering Department.

Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2021). Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM, 64*(3), 107–115. https://doi.org/10.1145/3446776

Zhang, Y., Huang, Q., Sun, W., Chen, F., Lin, D., & Chen, F. (2024). Research on lung sound classification model based on dual-channel CNN-LSTM algorithm. *Biomedical Signal Processing and Control, 94*, 106257. https://doi.org/10.1016/j.bspc.2024.106257

İSTANBUL UNIVERSITY PRESS

Acta Infologica,
ACIN 2025, 9 (1): 253–274

https://doi.org/10.26650/acin.1616088
Submitted: 08.01.2025
Revision Requested: 06.03.2025
Last Revision Received: 02.05.2025
Accepted: 03.06.2025
Published Online 12.06.2025

Acta Infologica

Research Article

Open Access

The Impact of Social Networks on the Stock Market Using Sentiment Analysis and Machine Learning: Application to the Turkish Stock Market

Mustafa Kemal Mayuk [1] & Farid Huseynov [2]

[1] Gebze Technical University, Faculty of Management, Department of Management, Kocaeli, Türkiye
[2] Gebze Technical University, Faculty of Management, Department of Management Information Systems, Kocaeli, Türkiye

Abstract The proliferation of portable devices and social media has transformed opinion sharing, impacting individual behavior, particularly in financial markets. This research explores how online sentiments influence investors' decision-making, highlighting the complexities of sentiment measurement in behavioral finance. AI-driven techniques have been developed to quantify opinions from social media data, focusing on Twitter (rebranded as X). The transformer architecture, which is a cutting-edge deep learning method widely used in generative AI models, is employed for sentiment analysis. The relationship between digitized sentiment scores and share prices within Türkiye's Borsa İstanbul (BIST 30) index was analyzed using machine learning techniques. Social media activity, as indicated by tweet volume, was investigated in relation to stock prices. The dataset comprises nearly 1.9 million tweets related to BIST 30 stocks, collected from early 2021 to late 2022. Independent variables include tweet volume, sentiment (positive-negative), and tweet timing, whereas dependent variables comprise stock prices and index closures. The findings reveal that tweet volume effectively predicts stock prices. Positive sentiment demonstrates stronger predictive power for individual stocks, whereas overall tweet sentiment does not significantly affect index-wide prices. Conversely, tweet timing is ineffective for price prediction. This research exemplifies the growing application of AI and machine learning in the social sciences by quantifying human opinions. The proposed model offers both theoretical and practical contributions, serving as a model for future research while delivering new insights and recommendations. The insights gained underscore the potential to harness information systems to advance financial literacy, stimulate economic growth, and empower informed decision-making across diverse global contexts.

Keywords Artificial Intelligence · Machine Learning · Sentiment Analysis · Stock Market Forecasting · Social Networks

Citation: Mayuk, M. K. & Huseynov, F. (2025). The Impact of Social Networks on the Stock Market Using Sentiment Analysis and Machine Learning: Application to the Turkish Stock Market. *Acta Infologica*, *9*(1), 253-274. https://doi.org/10.26650/acin.1616088

This work is licensed under Creative Commons Attribution-NonCommercial 4.0 International License.
© 2025. Mayuk, M. K. & Huseynov, F.

Corresponding author: Mustafa Kemal Mayuk m.mayuk2022@gtu.edu.tr

Introduction

Forecasting is one of the fundamental activities that many professional traders in the stock market use in their decision-making processes. Over the years, these techniques have evolved considerably, and with the introduction of computers, both technical and fundamental analysis methods have come a long way. The rapid computability of statistical methods via computers has contributed to this. Although computers facilitate time-series price analysis and macro- and microeconomic parameter accounting, other challenges exist in stock market forecasting. Many people do not make purely rational decisions. In the stock market, as in the commodity market, investors' emotions, beliefs, and psychology can often be influential. The main difficulty in price forecasting arises from the difficulty in modeling human behavior (Schumaker et al., 2012).

Until the development of smartphones and social media, it was only possible to gain insight into human behavior through people's own statements. However, the rapid increase in the amount of data that people voluntarily create after recent developments makes new perspectives possible in the field of economics and social sciences (Rancho et al., 2015). Thanks to social networks such as Twitter (In 2023 the of platform was changed to X), the last few decades have seen the most idea-generating data generation period in human history. Many companies have been established, and many studies have been initiated in institutions to mine ideas on these data (Liu, 2012). Opinions shape our choices, and understanding them is crucial for both individuals and organizations. Traditionally, gathering opinions involved personal networks and focused research, but it has become a substantial business itself. Studying opinions can provide great opportunities in science, finance, marketing, business, and management (Liu, 2012).

This study investigates the relationship between Twitter/X data, a social network example, and the Turkish stock market. The effects of the quantity, sentiment, and time variables in tweet data on the closing prices of both stocks and the index are analyzed.

On a daily basis, tweets about stocks between the opening and closing hours of the stock exchange are included in the scope of this study. The number of daily tweets about a stock is a variable that can provide insight into the social media activity of that stock (Liu et al., 2023). Recent research supports the hypothesis that information in tweet data is related to financial indicators (Souza et al., 2015). The timing of tweets about stocks can also impact the closing value. It may take some time for social media to have an impact on stock prices, which indicates that the timing of tweets should also be considered (Yu et al., 2012).

In line with previous research, it is possible to examine the impact of the number of tweets, the sentiments in the content of the tweets, and the time between the time stamps of the tweets and the closing time of the stock market on the closing values of the index and stocks. As in related studies, it is common to choose price or closeness as the dependent variable when examining the effects or relationships of independent variables on a financial asset.

This study aims to measure the performance of social media activity and the influence of opinion on social media in predicting the closing value of stocks and the BIST 30 index. In this context, the following research questions are posed;

1. Does the number of tweets affect the closing value of the BIST 30 index, average closing prices of stocks, and closing prices of individual stocks?
2. Does the average sentiment score affect the closing value of the BIST 30 index, average closing prices of stocks, and closing prices of individual stocks?

3. Does the time remaining until the closing time of stock market have an effect on the closing value of the BIST 30 index, average closing prices of stocks, and closing prices of individual stocks?

Literature Review

Stock Analysis Methods

Stock analysis, whether on an individual stock basis or an index basis, aims to predict trends, find the stock of the company that will earn the most, and identify buying and selling zones. According to the efficient market hypothesis, the market price at any given moment contains all the information, so forecasting in financial markets is not possible to generate above-market returns. According to all three variants of the efficient market hypothesis, it is not possible to achieve high returns through forecasting (Shah et al., 2019).

i) There is no relationship between price changes, and price changes are random. Therefore, technical analysis cannot beat the market.
ii) Prices change rapidly according to information such as political developments, economic news, and balance sheet announcements. Therefore, numerous returns cannot be achieved by fundamental analysis.
iii) Price contains all information; thus, information is public, which is a barrier to high returns through forecasting.

Many researchers and investors disagree with these predictions of the efficient market hypothesis and think that forecasting can be done using financial market data (Shah et al., 2019). From a methodological point of view, two main methods are used in stock analysis; fundamental analysis and technical analysis (Bentes and Navas, 2013).

Both fundamental analysis and technical analysis have their advantages, disadvantages, and complementarities. These two methods are commonly used together. Many fundamental analysts use technical analysis to identify buy and sell points. Many technical analysts also use fundamental analysis to identify companies that can be profitable in portfolio construction (Bentes and Navas, 2013).

Artificial intelligence and machine learning can make new contributions to both of these types of analysis. Machine learning is widely used in technical analysis to detect patterns or anomalies in time-series price charts and in fundamental analysis to establish regression between company data and price. Artificial intelligence and machine learning techniques have proposed new methods to augment and combine existing analysis methods.

Behavioral Finance

The efficient market hypothesis (EMH) is one of the most important theories of the 21st century in the field of financial markets. The EMH assumes that all market information is quickly incorporated into the price, meaning that the price is all-inclusive.

The EMH is based on three main arguments (Naseer & Tariq, 2015):

i) Investors are rational and purely utility-motivated.
ii) If investors do not behave rationally, trades are assumed to be random and offset any effect on price.
iii) Rational arbitrageurs eliminate the influence of irrational investors on the market.

Despite the impact of EMH, many critical studies against EMH. Research shows that financial anomalies, which can be categorized as fundamental, calendar, or technical, allow for high returns, contrary to the EMH's predictions. In particular, studies in the field of behavioral economics and its extension, behavioral finance, reveal that the psychological and sociological conditions of investors are effective in making irrational decisions and that the market should not be seen as an area consisting of completely rational players (Naseer & Tariq, 2015).

Psychological empirical research shows that when decision-making is difficult and feedback is delayed or noisy, people tend to rely more on intuition and practice. Research shows that this is also true in financial trading. Investors appear to act on behavioral biases when stock valuations are difficult. Price movements in the market may cause different behaviors due to people's behavioral biases. Valuation uncertainty has different behavioral effects. It has been observed that it increases the disposition effect, which motivates some investors to buy during declines and sell during rises. The disposition effect generates behaviors such as frequently tracking prices to look for trading opportunities. In cases of valuation uncertainty, some investors may also experience a gambling effect. Investors who exhibit high-risk winning behaviors are found to consider more losses. On the other hand, the disposition effect is amplified for overconfident investors. Overconfident investors are less likely to accept their losses as mistakes (Kumar, 2009). In addition to the effects of overconfidence and disposition, biases about herding and blue-chip stocks also have an impact on risk perception (Almansour et al., 2023).

Behavioral finance studies have revealed that emotions and biases influence decision-making. Many studies have also shown that news flows impact the stock market. It is possible to assume that the emotional states of individuals affected by news feeds also impact stock market trading. Understanding the public sentiment that is used to require difficult and expensive studies. Moreover, the accuracy of the methods used was limited due to the low correlation of the selected indicators with public mood. The ability to measure public sentiment can be considered an important behavioral finance variable. Sentiment analysis studies, which began on blogs in recent years, have become common on social media data (Bollen et al., 2011).

Behavioral finance is a theory that proposes alternative views to the efficient market hypothesis. Although information spreads quickly, especially today, people do not always use that information rationally and may make irrational decisions. Examining the effect of sentiment analysis on financial asset prices can also be considered a behavioral finance study.

Use of AI and Machine Learning in Finance

With the development of computer technologies, computers have begun to be widely used in finance, as in all fields. Computer science, expert systems, and machine learning techniques have proven themselves many times over in finance, not only in stock analysis but also in computationally intensive areas such as portfolio management (Motiwalla & Wahab, 2000) and risk management (Ahn & Kim, 2009). Both expert systems using classical machine learning methods and artificial neural networks have led to many successful software solutions and research in the field of finance. Genetic algorithms and fuzzy logic are soft computing techniques used in financial applications other than machine learning (Dunis et al., 2016).

In addition to common areas such as stock price forecasting, credit risk scoring, and portfolio optimization, the use of artificial intelligence is expanding in relatively new areas such as cryptocurrencies and derivative financial assets. In highly volatile cryptocurrencies and derivative assets, deep learning architec-

tures such as Long Short-Term Memory (LSTM), provide important insights into analyzing long-term time-series data (Bahoo et al., 2024).

In recent years, soft computing methods have been widely used in both analysis and modeling activities for pricing and valuation purposes in the stock market. Prices in the stock market are highly volatile, and they are affected by many factors. Therefore, stock market forecasting is challenging. Forecasting daily stock prices is not a mere numbers problem; it is a real-world problem. Many non-numerical factors, such as political and international relations, and news, also affect forecasting. Many studies have taken these factors into account (Dunis et al., 2016).

In an early study using artificial neural networks in stock forecasting, the price of IBM stock was predicted, and the potential of artificial neural networks to increase the ability in forecasting was mentioned (White, 1988). Another study, which presented a forecasting model using an artificial neural network in a low-risk, high-return portfolio manager with a buy-and-hold strategy, demonstrated that artificial neural networks provide better results than other algorithms. However, like other studies, this study also pointed out that stock forecasting will remain a difficult issue due to the presence of non-numerical effects (Thawornwong & Enke, 2004).

Artificial intelligence and data science techniques contribute greatly to the study of the behavioral dimension of finance and economics disciplines where human emotions and opinions are important. Deep learning, which is a set of architectures and techniques commonly referred to as artificial intelligence today, is finding more and more applications in areas such as modeling the behavior of individual investors and groups, summarizing and interpreting global financial news flows, and time series analysis of financial assets. The use of deep learning to model the relationships in the data and their possible effects and obtain financial results is called deep financial modeling (Cao, 2022).

Non-quantitative factors such as politics, international relations, and news flows, pose challenges to stock forecasting. This highlights the importance of including public opinion as a variable in research. In recent years, the field of behavioral finance has been focusing more on studies that attempt to understand the impact of sentiment on decision-makers, markets, and institutions (Kearney & Liu, 2014). A study using news articles mentioned that sentiment analysis improves prediction accuracy, but the prediction time horizon should be reduced due to the dynamism of the stock market (Li et al., 2014). Instantaneous price changes in the stock market may be more strongly correlated with social media, which is naturally more dynamic than news articles. A study analyzing the relationship between sentiment analysis scores obtained from news and tweet data and their relationship with the prices of some selected company stocks revealed that tweet data have a more significant relationship with volatility than news data (Souza et al., 2015).

Twitter/X is an environment where instant news flow and social sentiment are instantly reflected; thus, it has begun to be used as a common data source in sentiment analysis-based financial asset forecasting research. A study by Rao and Srivastava (2012) shows that negative and positive dimensions of public mood have a strong causal relationship with the price movements of individual stocks/indices. Another study obtained sentiment analysis data on Bitcoin tweets using machine learning techniques and demonstrated that social media sentiment analysis can contribute to understanding Bitcoin price fluctuations and that machine learning models work with high accuracy in this area (Kına & Biçek, 2024). In financial sentiment analysis of news articles or social media data, deep learning algorithms and classical machine learning techniques are employed. Deep learning techniques are more successful in evaluating the context of a whole text than classical approaches based on the sentiment analysis score of individual words (Sohangir

et al., 2018). The psychological factors that trigger human behavior are often not obvious. By uncovering relationships in large amounts of data, deep learning can facilitate the study of relationships between investors' trading behavior and their emotions in the stock market (Ruan et al. 2020).

As in every field, artificial intelligence and machine learning technologies are widely used in the financial sector. In addition to the stock market, which is the scope of this study, new applications are increasingly being encountered in strategic areas, such as risk management and portfolio management.

Data and Methodology

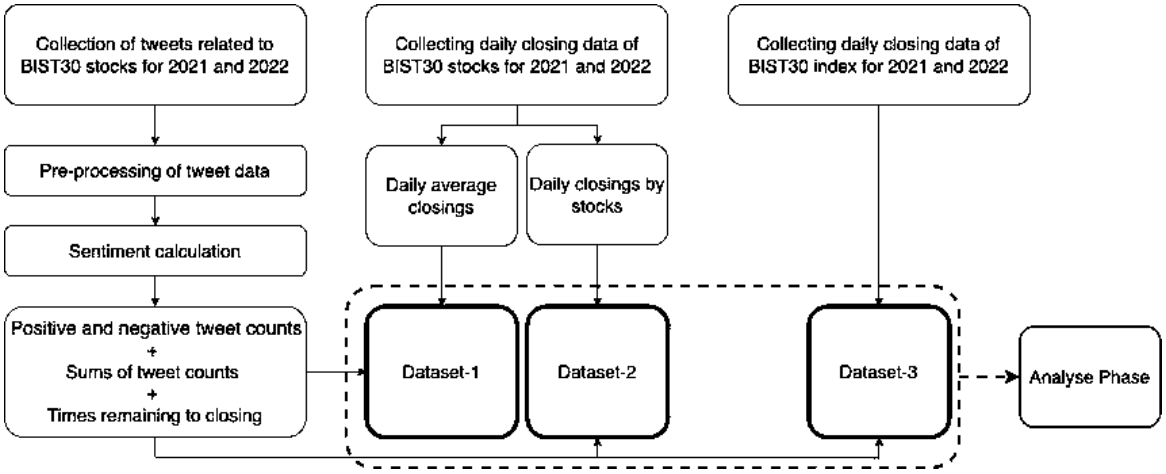
Creating Initial Data Sets

In this study, the Python programming language was used for data collection, cleaning, and preprocessing, sentiment score calculation, data exploration, and regression and forecasting analyses.

The collection of tweets and the daily closings of the BIST 30 components and the BIST 30 index can be considered separate processes. The obtained tweets were processed through data cleaning and sentiment analysis calculation stages. The datasets consisting of the index, stock, and sentiment analysis scores to be studied are created in this way. Figure 1 shows the three types of data initially obtained. Time information is critical to merging all data; thus, time data are included in all data. Tweet data contain three types of data: the user (who wrote the message), the text (i.e., the tweet), and the timestamp (i.e., the time when the tweet was sent).

Figure 1

Process of obtaining data from raw collected data before analysis.



The dataset contains daily closing data for the BIST 30 components and the BIST 30 index. The BIST 30 components data comprise the code of the stock, the date of the day, and the closing price.

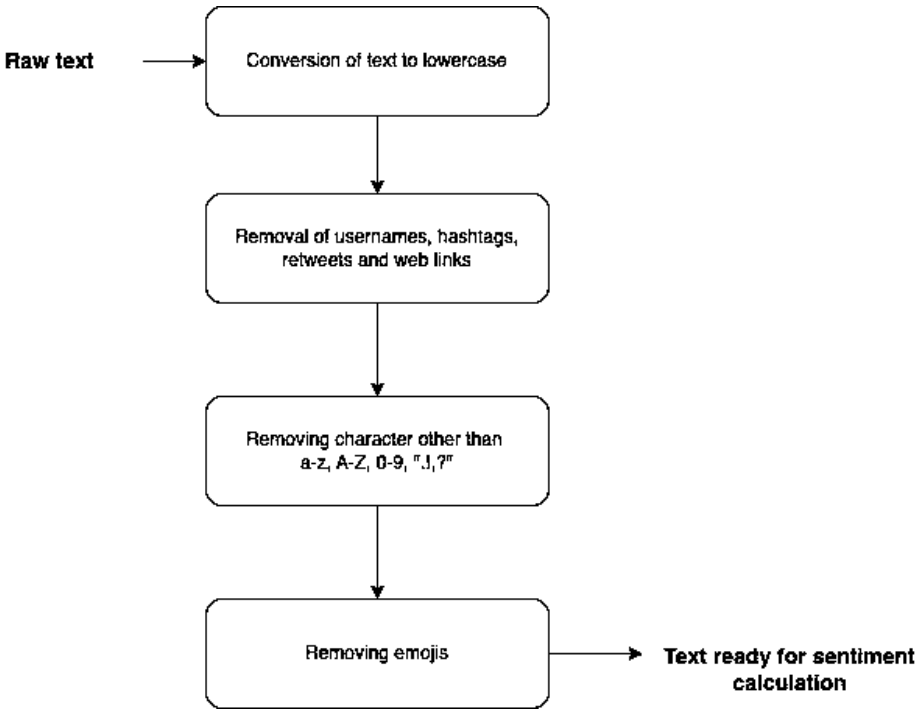
The tweet data from the beginning of 2021 to the end of 2022 used in this study were obtained when third-party libraries could access APIs. Day-to-day closing data for the BIST 30 index over the same date range were obtained using a public investing web platform (Investing, 2024). Daily closing data for the BIST 30 components for the same date range were obtained from the open website of the investment subsidiary of a private bank. Closing data can be obtained by filtering the date and stock code on this website (Is Yatirim, 2024).

Tweet data are the most unstructured data type that needs cleaning and pre-processing in tweet, index closings, and stock closings data. Index and stock closing data are numerical and noise-free. Commonly, in order to reduce the difficulty of sentiment analysis in microblog texts, some techniques, such as cleaning and, if necessary, modifications to the text before sentiment calculation, are applied. Preprocessing natural language texts to make them ready for sentiment analysis is a challenging task. Social media is not a formal correspondence platform, and people do not pay attention to spelling and grammar. In sentiment analysis studies, noise is removed as much as possible; however, it is impossible to remove it completely.

There is no general technique for reducing noise in social network user data. Although some operations are common, different preprocessing steps can be applied depending on the nature of the study or the choice of the researcher. In this study, all characters were minimized to homogenize the natural language text. Since the data are tweet data, usernames, hashtags, retweets, numbers, and web links, they do not contribute to the sentiment score (Effrosynidis et al., 2017). Exclamation marks, question marks and emojis, on the other hand, overly influence the polarity of the sentiment score. The elimination of these phrases and characters is summarized in the flow in Figure 2.

Figure 2

Data cleaning and pre-processing process before sentiment calculation.



Sentiment Calculation of Tweet Data

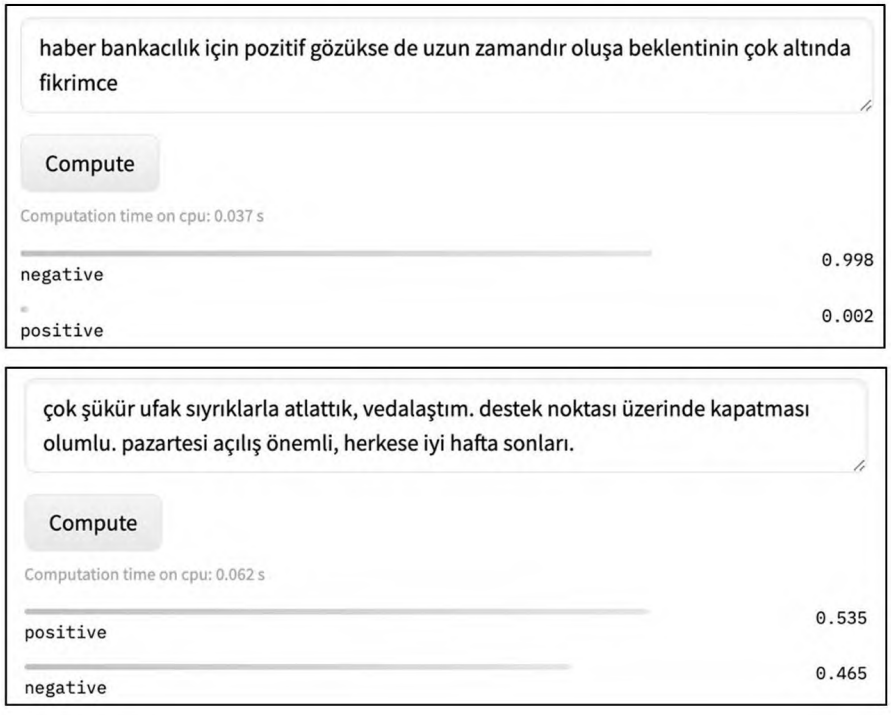
Artificial intelligence (AI), which has been a topic of computer science since the 1950s, mostly refers to rule-based algorithms and techniques until today. Today, the increase in data has enabled artificial neural networks, which date back to the 1950s, to work more effectively. AI mostly refers to deep learning architectures based on artificial neural networks, and transformers are the most advanced examples of this architecture for natural language processing. Machine learning (ML) is the general name of a discipline that includes artificial intelligence (AI) and other data-driven learning algorithms.

In this study, the BERT transformer model was used for sentiment analysis calculation. However, for this study, the model was not trained from scratch, but a trained model with the Turkish language was used to calculate the sentiment scores (Yıldırım, 2024). This model was trained on Turkish datasets from two different studies. One of these studies used film and shopping website reviews (Demirtas & Mykola, 2013), and the other one used tweet data (Hayran & Sert, 2017). The proposed model generates a sentiment score between −1 and 1 from the sentences given as input. Scores less than zero indicate negative feelings, and scores greater than zero indicate positive sentiment.

Many AI developers have published and made natural language processing models available on the Hugging Face platform established in 2016. This platform allows models to be retrained, tested through the interface, and used on platforms such as Amazon, Azure, or personal computers through APIs. The developer of the model used in this study also made the model available through Hugging Face, and the model was used in this study through Hugging Face libraries (Hugging Face, 2023). Hugging Face also allows models to be tested via a web interface. The manual analysis demonstrates that the proposed model can correctly classify the sentiment of tweets as positive or negative. Figure 3 shows examples of testing positive and negative sentiment tweets.

Figure 3

Manual testing of the sentiment analysis model



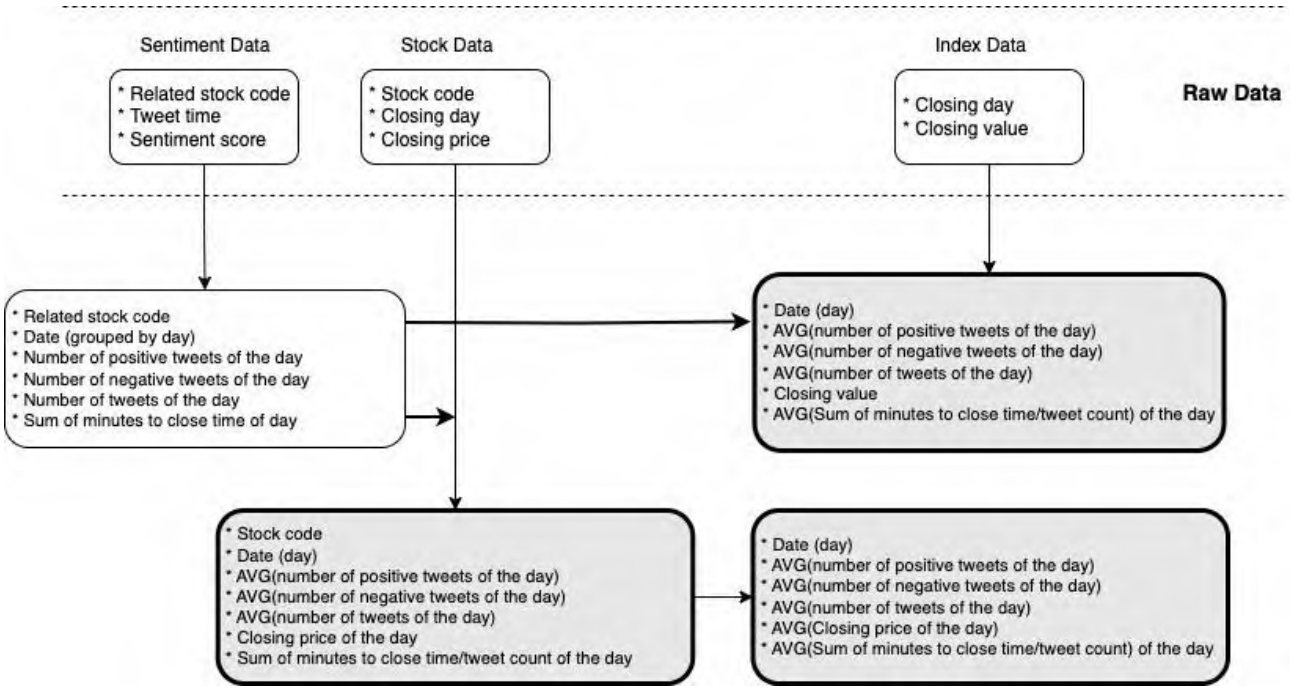
In this study, the sentiment calculation of pre-processed raw tweets was performed using Hugging Face libraries, which enable the use of this model in Python code. In this study, 1,906,966 tweets posted during the days and hours during which the stock exchange was open were collected. After data cleaning and preprocessing (Figure 2, the remaining 724,207 tweets were used for sentiment analysis.

Creation of Final Data Sets

The final data sets have features and targets or dependent and independent variables. The final datasets were obtained by merging the datasets among themselves by performing operations such as averaging and summing on the initial datasets created in Figure 1. Figure 4 shows the datasets and variables resulting from the merging.

Figure 4

Detailed process of creation of final datasets, features, and target fields.



In the merged datasets, the target (or dependent variable) is closure, and the features (or independent variable) are the number of tweets with positive sentiment, the number of tweets with negative sentiment, and the total number of tweets. Since the relationship between the data depends on the date and stock codes, the datasets were merged using these fields. For this reason, date and stock codes act as keys, not features, or targets.

The stocks dataset contains data on a stock-by-stock basis. The daily closing is already the closing value of the stock, so averaging is unnecessary. The consolidated stock data set is the daily average of the closing prices of all stocks. One of the research questions is related to the effect of the timing of tweets on price, and for this reason, the sum of the minutes remaining to the closing time of the tweets sent during the day is taken as an independent variable. However, this sum alone will be directly proportional to the number of tweets and will be dependent on the number of tweets. Therefore, we obtain a new variable by dividing this sum by the number of tweets.

In summary, features (independent variables) are the number of tweets, positive tweets, negative tweets, and time to close; target (dependent variable) is the closing of stocks or the index. All data sets were normalized before being input to the regression models, and 25% of the data were randomly assigned as test data. Normalization is the rescaling of features to a standard range, typically between 0 and 1. Normalization was performed programmatically. Throughout the training process, normalization ensures that each

feature contributes equally and prevents features that are too large in magnitude from overshadowing other features.

Table 1 lists the usages of all fields for each dataset. In this study, the effects of independent variables on the dependent variable are separately analyzed for each data set.

Table 1
Classification of variables as features or targets on a dataset basis

	Stocks	Consolidated Stocks	Index
Average number of positive tweets of the day	Feature	Feature	Feature
Average number of negative tweets per day	Feature	Feature	Feature
Average number of tweets per day	Feature	Feature	Feature
Closing price of the stock	Target	-	-
Average closing prices for the day	-	Target	-
(Sum of minutes to stock exchange close hour/Tweet counts) of the day	Feature	-	-
Average (Sum of minutes to stock exchange close hour/ Tweet Counts) of the day	-	Feature	Feature
Closing value of index	-	-	Target

Results

Time-Series Visualizations

Presenting time-series data visualization is important for understanding the general characteristics and main trends of the data. It will also make it easier to understand the trend relevance of the sections related to social media activity and sentiment analysis, which is the main topic of this study.

Figure 5 shows a graph of the BIST 30 index from the beginning of 2021 to the end of 2022, while Figure 6 shows a graph of the BIST 30 components for the same date range.

Figure 5
Chart of the BIST 30 index from the beginning of 2021 to the end of 2022

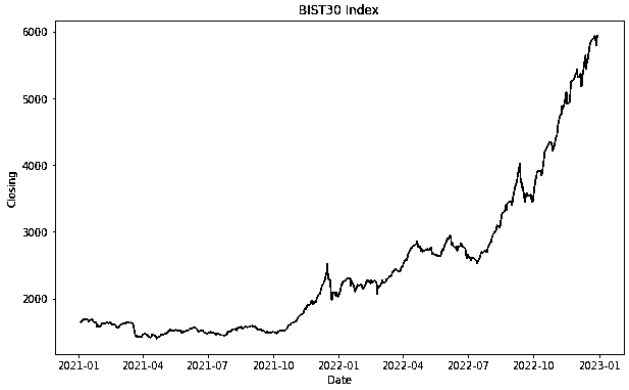
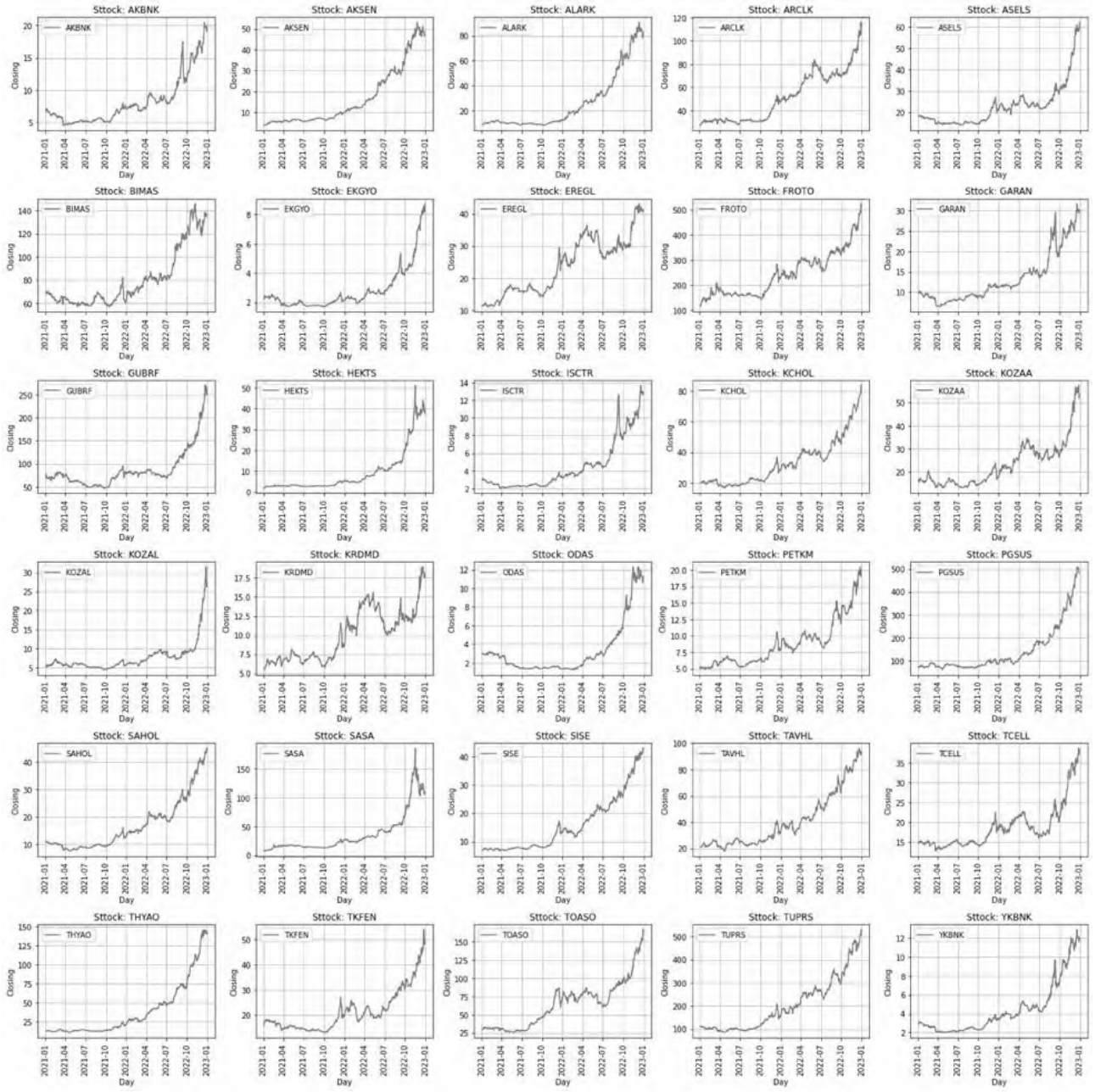


Figure 6

Chart of the BIST 30 components from the beginning of 2021 to the end of 2022



Time-series trend charts show common behavior for stocks and the index. The market, which remained stable toward the end of 2021, experienced an upward trend after 2022. This finding agrees with studies that examined stock market movements during the COVID-19 pandemic (Altınbaş, 2022).

Data Exploration

Figure 7 shows the number of tweets and stocks. The average line and the shares above it are shown. It is observed that a few stocks, in particular, have increased the average, while most of them have remained below the average.

Figure 7

Stock codes and number of tweets for each stock from the beginning of 2021 to the end of 2022

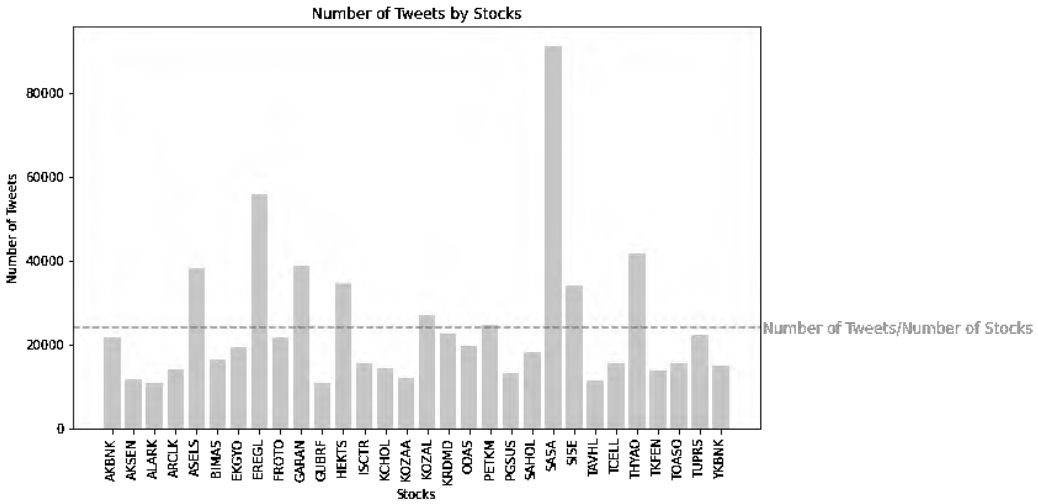
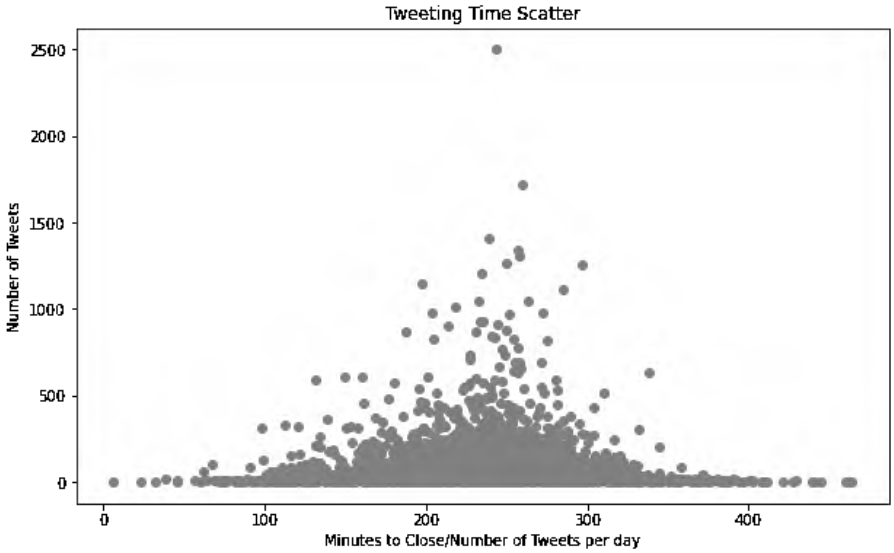


Figure 8 shows the same data as the distribution of the ratio of the number of tweets to the number of minutes before the closing time of the stock exchange. This graph clearly shows that tweet activity is low near the opening and closing hours, with most of the activity concentrated in the middle of the trading hours. This shows that investors' convictions are largely formed near the middle of the timeframe that can be traded.

Figure 8

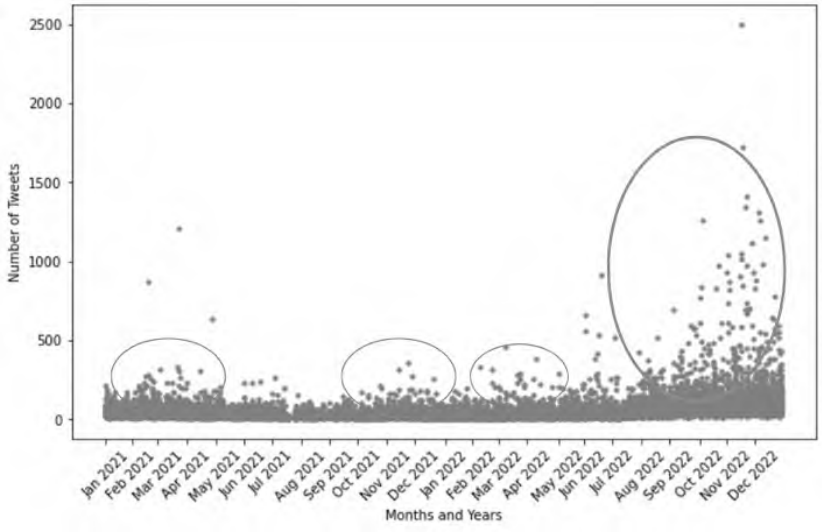
Distribution of number of tweets according to time remaining until stock exchange closing time.



The distribution of the number of tweets overlaps with the volatility of index closings at the beginning of 2021, and some at the end of 2021 and the first quarter of 2022. Likewise, it overlaps with the trend rise starting with the summer of 2022 in most of them. Figure 9 shows the tweeting activity distribution by time between the beginning of 2021 and the end of 2022. This scatterplot is quite parallel to the BIST 30 index graph shown in Figure 5. Tweeting activity is also high at times when there is an increase in the index.

Figure 9

Distribution of number of tweets about BIST 30 components by time from 2021 to 2022.



A low standard deviation in the sentiment score indicates a certain intensity of emotion in the tweets about that stock. Table 2 shows the stocks with some tweets above average, standard deviation of sentiment scores, and standard deviation rank in ascending order of standard deviations for all stocks.

Table 2

Most tweeted stocks and standard deviations of sentiment scores

Stock	Number of Tweets	Standard Deviation of Sentiment Score	Standard Deviation Rank in Ascending Order
SASA	91,16	0.13	1
EREGL	55,806	0.15	3
THYAO	41,815	0.14	2
GARAN	38,926	0.15	4
ASELS	38,336	0.17	6
HEKTS	34,707	0.19	8
SISE	34,212	0.16	5
KOZAL	27,224	0.19	9
PETKM	24,605	0.19	7

Regression Analysis

In this study, the dataset formed by combining the average sentiment score with the BIST 30 index data was analyzed using regression models. The number of tweets, the number of positive tweets, the number of negative tweets, and the ratio of the average of the minutes remaining until the closing time of the tweets to the number of tweets (this will be called as “time ratio”) are analyzed as independent variables, and the closing value is analyzed as a dependent variable. The relationship between the independent variables and dependent variable was analyzed one by one.

There are many machine algorithms for regression analysis. In this study, SVR (Support Vector Regression), XGBoost (Extreme Gradient Boosting), and random forest regression algorithms were used to create

models, and their metrics were compared. These evaluations are the rationale for testing the models built with these algorithms:

- SVR is useful for both linear and complex data and is tolerant of errors. Linear regression works well with more linear data and has a low tolerance for errors. Instead of linear regression, a model using the "linear" kernel of SVR is preferred.
- XGBoost creates a tree-based regression model that progressively reduces errors by gradient boosting. It offers high accuracy, speed, and efficient computational power via parallel computing using CPU.
- Random Forest Regression is an algorithm that works by averaging many decision trees, is less prone to overfitting, and is more tolerant of errors.

LSTM and other artificial neural network-based algorithms were not preferred in this regression study due to the fact that the data to be subjected to regression is not high volume, the correlation between the sequences is low, and they work slowly without GPU.

The R-squared value is an important metric for measuring the accuracy of regression models. However, it is not an indicator of model performance. Error metrics such as MAE (Mean Absolute Error), MSE (Mean Squared Error), RMSE (Root Mean Squared Error) should also be evaluated. The R-Squared value of a model that can explain the relationship between variables meaningfully should be between 0 and 1. A negative R-Squared value means that the reported predictive power of the model is lower than the average of the dataset. A model with a negative R-squared value cannot explain the relationship between variables and should be rejected. In this study, the relationship between the independent and dependent variables was analyzed one by one. Therefore, R-Squared and Adjusted R-Squared values are expected to be equal in all regression analyses.

Table 3 presents the R-Squared intervals and their interpretations (Ozili, 2023).

Table 3

R-Squared value ranges and interpretations

Value	Interpretation
< 0	The model does not explain the dependent variable at all and should be rejected
0.00-0.09	A model that is too weak should be rejected
0.10-0.50	Some or most explanatory variables are statistically significant.
0.51-0.99	Most of the explanatory variables are statistically significant

It should be considered that the risk of overfitting increases as the R-Squared value approaches 1, especially above 0.90. The tendency of overfitting can be observed by comparing the training and test data metrics separately and observing the difference.

MAE, MSE, and RMSE are indicators of the model's error rates in explaining the data; there is no ideal number for these metrics, and smaller values are always better. When comparing different models, a good approach is to consider models with lower error metrics. Models with significant R-Squared values but high error values also carry the risk of overgeneralization.

Table 4 presents the performance of the regression models SVR, XGB (XGBoost), and RF (Random Forest) constructed with the closing value for each feature using BIST 30 index test data. For variables with negative

R-squared values, error metrics are not included in the table because the regression cannot explain the relationship. Each triple column under the variable columns presents the metrics of the three algorithms.

Table 4

R^2 and error metrics of SVR, XGBoost, and random forest models for BIST 30 index test and training data

Data Type	Metric	Number of Tweets SVR XGB RF	Number of Positives SVR XGB RF	Number of Negatives SVR XGB RF	Time Ratio SVR XGB RF
Test	R^2	0.76 0.76 0.67	0.74 0.75 0.75	0.75 0.73 0.70	0.11, 0.19, 0.56
Test	MAE	0.09 0.09 0.10	0.09 0.09 0.09	0.09 0.09 0.10	• ‣ -
Test	MSE	0.01 0.01 0.01	0.01 0.01 0.01	0.01 0.01 0.01	• ‣ -
Test	RMSE	0.11 0.11 0.04	0.11 0.11 0.11	0.11 0.11 0.11	• ‣ -
Training	R^2	0.66 0.83 0.92	0.64 0.83 0.90	0.65 0.84 0.91	−0.05 0.45 0.79
Training	MAE	0.08 0.06 0.03	0.09 0.05 0.04	0.08 0.05 0.04	• 0.10 0.06
Training	MSE	0.01 0.005 0.002	0.01 0.005 0.003	0.01 0.004 0.002	• 0.01 0.006
Training	RMSE	0.10 0.07 0.13	0.10 0.07 0.05	0.10 0.07 0.05	• 0.13 0.08

By comparing the test and training data of these three models, it is possible to understand whether they overfit or underlit. A lower R-Squared metric for the test data than for the training data may indicate overfit. If the training score is excessively lower than the test score, the model may underfit. Table 4 shows that models trained with XGBoost and Random Forest are at risk of overfitting, whereas models trained with SVR are at risk of underfitting. On the other hand, for the time ratio variable, SVR produced a consistent R-Squared score between the training and test datasets, while XGBoost and Random Forest produced results with different signs.

Table 5 presents the training/test ratios of the R-Squared metrics of the models.

Table 5

R^2 training/test ratios of models for each variable

Model	Number of Tweets	Number of Positives	Number of Negatives	Time Ratio
SVR	0.86	0.86	0.86	0.45
XGBoost	1.09	1.10	1.15	−2.36
Random Forest	1.37	1.20	1.30	−1.41

As shown in Table 5, SVR does not show an overfitting risk for any variable. XGBoost shows a rate of approximately 10% for the number of tweets and the number of positive tweets, but it increases to 15% for the number of negative tweets. Rates above 10% indicate that the overfitting limit is exceeded. Random Forest exhibits overfitting for all variables. In the SVR, both R-Squared metrics for the time ratio variable are negative, indicating that the model trained with the SVR is consistent for both test and training data. In XGBoost and Random Forest, the result for this variable is negative, i.e., negative R-Squared for the test data and positive for the training data, indicating that these models are not consistent with the time ratio. There is a small risk of underfitting for SVR, but this may not be taken into account, especially when the dataset is small. The relationships in the data are simple, and all the variables of the model are included. For this reason, SVR is preferred in the following regression analyses of this study because similar datasets are used.

When a regression model is constructed with the same variables using the average closings of the BIST 30 components, the values in Table 6 are obtained.

Table 6

R^2 and error metrics of the SVR model for BIST 30 components average

Metric	Number of Tweets	Number of Positives	Number of Negatives	Time Ratio
R^2	0.7007	0.6798	0.6898	−0.0664
p-value	< 0.001	< 0.001	< 0.001	> 0.05
MAE	0.0876	0.0887	0.0891	0.1391
MSE	0.0105	0.0104	0.0115	0.0369
RMSE	0.1026	0.1024	0.1075	0.1921

Both tables show that the independent variables (number of positive and negative and all tweets) explain approximately 70% of the variance in the dependent variables (index close). The time ratio shows a very poor fit with the model. The MAE, MSE, and RMSE values are also higher than those of the other independent variables. This means that the error rate was higher than that of the other methods. The relationship between the number of tweets, the number of positive tweets, and the number of negative tweets with the closing is $p < 0.001$, which is 99.9% significant. The time ratio has a p-value >0.05 in both data sets, significance is less than 95%, and the regression coefficient is likely to be 0. The significance interval of the p-values was the same for the index data.

When the BIST 30 components are analyzed individually in regression analysis, the model can explain the relationship between the variables of most of the stocks and the closing. The data from which these results are obtained are the R-Squared metrics for individual stocks in Table 7. This table displays the R-Squared values of the effects of the variables for each stock separately.

Table 7

R^2 metric of the SVR model for individual stocks in the BIST 30 index

Stocks	Number of Tweets	Number of Positives	Number of Negatives	Time Ratio
AKBNK	0.3451	0.3699	0.2909	−0.0558
AKSEN	0.5797	0.5721	0.5222	−0.1221
ALARK	0.5025	0.5298	0.4149	−0.0922
ARCLK	0.0909	0.1145	0.0309	−0.1311
ASELS	0.3579	0.3482	0.3366	−0.0254
BIMAS	0.1669	0.2542	0.0866	−0.0228
EKGYO	0.1449	0.1690	0.1032	−0.0060
EREGL	0.2812	0.2883	0.2657	−0.0118
FROTO	0.3233	0.3085	0.3014	−0.0357
GARAN	−0.0101	0.0211	−0.0421	−0.0273
GUBRF	0.0486	0.0557	0.0337	0.0025
HEKTS	0.6445	0.6722	0.6051	−0.0200
ISCTR	0.6395	0.6192	0.5966	−0.0740
KCHOL	0.3808	0.3066	0.3907	−0.0478
KOZAA	0.3571	0.3438	0.3248	−0.0263
KOZAL	0.2534	0.2905	0.0221	−0.0130
KRDMD	0.2129	0.1698	0.2398	−0.0515
ODAS	0.4394	0.4691	0.0355	−0.0470

Stocks	Number of Tweets	Number of Positives	Number of Negatives	Time Ratio
PETKM	0.2940	0.2657	0.3111	−0.0449
PGSUS	0.2613	0.2289	0.2062	−0.0244
SAHOL	0.6213	0.5867	0.5745	−0.0547
SASA	0.6422	0.6548	0.6069	−0.0118
SISE	0.5421	0.5426	0.4939	−0.1199
TAVHL	0.0735	0.0540	0.0605	−0.0820
TCELL	0.2259	0.2365	0.1648	0.0049
THYAO	0.4071	0.3893	0.4081	−0.0677
TKFEN	0.2644	0.1999	0.2787	−0.0253
TOASO	0.1129	0.1370	0.0780	−0.0232
TUPRS	0.2920	0.2569	0.2873	−0.0715
YKBNK	0.5341	0.5064	0.4941	−0.0652
Average	0.3461	0.3322	0.2841	−0.464

It is observed that the model fails to explain significantly the ARCLK, GARAN, GUBRF, and TAVHL stocks because one of them is negative and the other three are less than 0.1 R-Squared values. For 11 stocks, the R-Squared value was below average for all stocks. The remaining 15 stocks were above average, and AKSEN, ALARK, HEKTS, ISCTR, SAHOL, SASA, SISE, YKBNK is above 50%. When stocks with insignificant R-Squared values are excluded, the R-Squared value of positive tweets is higher in 20 of 26 stocks. This represents approximately 76% of all significant stocks. The model performs better with negatives than positives for KCHOL, KRDMD, PETKM, TAHVL, TKFEN, and TUPRS. The model fails to explain the relationship between the time ratio and closings. For all stocks with a positive R-Squared value, there is $p < 0.001$, i.e., 99.9% significance between the number of all tweets, positive and negative tweets, and their closings. The time ratio has a significance of $p > 0.05$, i.e., less than 95% for all stocks.

Forecasting Models

Forecasting studies using stock market data are common. The forecasting activity aims to develop models to predict future values using only time-series data. ARIMA is a widely used approach for forecasting time-series data, and LSTM has also become widely used with developments in deep learning. Table 8 presents the test metrics of these two models on the BIST 30 index data.

Table 8

Test metrics of the ARIMA and LSTM models for forecasting study of BIST 30 index data

Metric	ARIMA	LSTM
R-Squared	−2.33	0.63
MAE	1337.39	436.94
MSE	2.55×10^6	249×10^3
RMSE	173.74	499.77

According to this table, it is clear that LSTM outperforms ARIMA in forecasting the BIST 30 index data. ARIMA assumes that the variance of the temporal data should be almost constant, and the graph in Figure 5 shows that the index data are not of this character, i.e., they are not stationary data (Namin & Namin, 2018).

The similarity in index and stock movements in Figures 5 and 6 show that forecasting models have similar metrics.

The R-Squared value of 0.63 for LSTM in Table 3 is very close to 0.71 for the number of tweets in Table 4. However, the error rates of the forecasting models were significantly higher than those of the regression models. This show that the relationship between the number of tweets and closing values is more significant than the relationship among the time-stamped closing values.

Discussion and Conclusion

This study investigates the impact of people's activities reflected on social media on the market by analyzing digital data using known machine learning techniques. All regression analyses revealed a significant relationship between positive and negative tweet counts and closing data. Tweet counts are an important indicator of social media activity. In this study, a significant relationship was observed between the number of tweets and the closing of the index and individual and consolidated stocks. In the analysis of index and consolidated stocks, there is not much difference between the effect of positive and negative tweets. In regression analysis, positive tweets fit the regression models more than negative tweets in individual stock analyses. This study demonstrates that stock prices and indexes can be predicted using social media activity by building a machine learning model. Activities categorized into two groups according to sentiment polarity can also be explanatory, although relatively less so.

These findings show that social media activity, which is a reflection of human psychology today, can provide an important study area for behavioral finance research. It is clear that the proposed method will make great contributions to the quantification of human behavior, which is an important challenge in social science research. In terms of the efficient market hypothesis, social media constitutes an important area of information diffusion. These and similar studies will enable modeling the possible effects of this field before it reaches people.

Practical Contributions

Individual investors, financial institutions, and portfolio managers use tools such as technical and fundamental analysis to evaluate stocks. Technical analysis focuses on price changes over time using statistical algorithms, whereas fundamental analysis relies on quarterly balance sheet data. Both methods often overlook organic data, such as news and psychology, which influence investor behavior. Technical analysis is typically used for short-term trading, whereas fundamental analysis is used for long-term strategies. Professional portfolio managers use tools that combine both analyses. This study reveals that these traditional techniques of modeling investor behavior and sentiment improve and strengthen stock analysis

The findings show the impact of public opinion on social media on stock prices and indices, indicating that the proposed approach can be integrated into existing tools as a software solution. Decision support systems that use short-term technical analysis can also be fed into decision support systems that use short-term technical analysis to measure investor sentiment in a continuous news feed and social networks. On the other hand, the metrics of these models can be added to tools that recommend long-term investment decisions based on parameter ratios in balance sheet data.

Sentiment analysis for financial forecasting requires significant computational resources and real-time data processing. High-frequency trading demands sophisticated AI systems; however, simpler models can

be used for long-term investments. Sentiment analysis can complement traditional fundamental analysis by providing insights into market and company perceptions. This combined approach can help investors make more informed decisions.

Theoretical Contributions

In this study, a transformer model based on deep learning was employed to obtain the sentiment scores. Transformers are architectures that are successful in modeling sequential inputs that have relationships with each other, like natural language. Deep learning matures later in natural language processing than in fields such as image processing. Prior to transformers, different approaches, such as long short-term memory (LSTM) and recurrent neural networks (RNN), were also proposed. A study using moving average, linear regression, autoregressive integrated moving average (ARIMA), and LSTM in stock forecasting with sentiment analysis compared the performances of the methods. LSTM, a deep learning approach, performed better than other techniques (Shubham et al., 2024). Another study examining pre-transformers deep learning models on the example of sentiment analysis and stock market forecasting also revealed that LSTM performed better than other models (Pattanayak et al., 2024). The ability of LSTM in forecasting is also observed in this study. While LSTM models have proven effective for forecasting, they have been widely superseded by transformers in the field of language processing.

The shortcomings of LSTM in modeling longer natural language texts can be overcome using the proposed transformer. The technical architecture of the model was revealed in 2017 (Vaswani et al., 2017); thus, it has been used as a method in fewer studies than other techniques. In this study, a transformer model called Bidirectional Encoder Representations from Transformers (BERT) was used. In a study using BERT for stock sentiment analysis, message data received from an investor platform were labeled as bearish or bullish, and the BERT model was tuned. Unlike this study, the model, which aims to classify messages rather than stock predictions, performed quite well (Lee et al., 2020).

This study investigates the impact of people's activities reflected on social media on the market by analyzing digital data using known machine learning techniques. The model in which the number of tweets representing social media activity is used as a feature can explain prices and closings clearly. A study examining whether social media or conventional media is more effective on companies found a stronger relationship between sentiment data and company returns and risks (Yu et al., 2013).

The number of tweets with positive and negative sentiments did not show significantly different performance in explaining the average closing prices of indices and stocks. However, when individual stocks were examined, it was found that the number of tweets with positive sentiment was able to explain price increases better than those with negative sentiment in most the tweets. A study examining the relationship between the frequency of words expressing happiness, fear, sadness, and excitement in tweet data and the closing prices of the Dow Jones, S&P 500, and NASDAQ indices found that, similarly, the index rose on days following days with a higher prevalence of positive sentiment expressions compared to days with a higher prevalence of negative sentiment expressions, specifically for the Dow Jones index (Zhang et al., 2011).

It performed slightly better at explaining the average number of negative tweets for the general index and stocks. In a study that studied the sentiment data of articles, the better performance of negative polarity was evaluated as an indicator that negative emotions may have a more dominant effect (Schumaker et al., 2012). However, in this study, the difference between the performances of the two polarities was not significant. Another study examining the impact of news feeds on stock prices using sentiment analysis concluded that

although sentiment analysis increases prediction accuracy, models that rely on polarity-based forecasting, scored as positive and negative, do not perform well (Li et al., 2014).

In this study, social media activity represented by the number of tweets was found to be a more effective factor than sentiment scores. A study that predicts NASDAQ stocks by obtaining sentiment scores using deep learning on tweet data has revealed that although positive sentiment is an important factor in explaining a price, it cannot be generalized to all stocks. It has been explained that having more news flow about some stocks results in more social media activity, and activity moves in parallel with positive sentiment (Guo & Xie, 2024).

Limitations and Suggestions

The process of calculating the sentiment analysis scores, which is the core task in the methodology of this study, is the most challenging phase due to the need for both high computational power and high-noise data processing. In order to obtain the sentiment scores, several tweets were captured, cleaned, and scores were calculated. The preprocessing of tweets is a challenging and difficult process, which causes the data to always have noise. It is not an easy task to extract numerical data from a structure as complex as natural language, where even punctuation marks have a meaningful effect. The main data on which this study is based are sentiment score data, and it is not possible to analyze hundreds of thousands of tweets individually. At this point, although tested, the artificial intelligence model that produces the sentiment score was relied upon.

Artificial intelligence is inherently a black box, and it is impossible to reverse engineer what it decides and why. Although regression analyses using less data have more explicit mathematical backgrounds, deep learning-based language models use billions of parameters and the networks undergo many weighting operations during learning. The manner in which the model generates the sentiment scoring value depends entirely on the quality and quantity of data used by the model. Re-tuning the model with very different data sets and comparing the performance between model versions requires specialized computer systems for artificial intelligence. Since it is not possible to determine how the model makes decisions, tuning the model is a way to obtain more successful models. Today, this topic remains a challenge in all research on deep learning. Enhanced results can be obtained by fine-tuning models for stock sentiment analysis using greater computational resources and advanced technical expertise.

The fact that stocks and the index move with very similar trends and volatility is low may have caused the study not to obtain more significant results between sentiment and closings. A significant relationship similar to the one between tweet activity and closings did not reveal the effect of positive or negative sentiment in a continuously rising market. This shows that in future AI-supported finance studies, in addition to solving the technical constraints mentioned above, taking macro- and even micro-economic factors into account may reveal more significant relationships. Studies using data lakes that combine investor sentiment with other economic and financial data may produce more accurate results.

In this study, the model failed to produce significant results in terms of predicting price using the duration between tweeting and closing times. Working with more granular price data in the form of a time series rather than a single price may provide more significant results.

Peer Review Externally peer-reviewed.
Author Contributions Conception/Design of Study- M.K.M., F.H.; Data Acquisition- M.K.M., F.H.; Data Analysis/Interpretation- M.K.M., F.H.; Drafting Manuscript- M.K.M., F.H.; Critical Revision of Manuscript- M.K.M., F.H.; Final Approval and Accountability- M.K.M., F.H.
Conflict of Interest The authors have no conflict of interest to declare.
Grant Support The authors declared that this study has received no financial support.

Author Details
Mustafa Kemal Mayuk
[1] Gebze Technical University, Faculty of Management, Department of Management, Kocaeli, Türkiye
0009-0002-5698-0813 m.mayuk2022@gtu.edu.tr

Farid Huseynov
[2] Gebze Technical University, Faculty of Management, Department of Management Information Systems, Kocaeli, Türkiye
0000-0002-9936-0596

References

Ahn, H., & Kim K. J. (2009). Bankruptcy prediction modeling with hybrid case-based reasoning and genetic algorithms approach. *Applied Soft Computing*, 9(2), 599–607. https://doi.org/10.1016/j.asoc.2008.08.002.

Almansour, B. Y., Elkrghli, S., & Almansour, A. Y. (2023). Behavioral finance factors and investment decisions: A mediating role of risk perception. *Cogent Economics & Finance*, 11(2). https://doi.org/10.1080/23322039.2023.2239032.

Altınbaş, H. Y. (2022). COVID-19's Impact on Global Stock Market Movements and an Examination on Turkish Market. *Yıldız Social Science Review*, 8(1). https://doi.org/10.51803/yssr.1146870.

Bahoo, S., Cucculelli, M., Goga, X., & Mondolo, J. (2024). Artificial intelligence in Finance: a comprehensive review through bibliometric and content analysis. *SN Bus Econ*, 4(23). https://doi.org/10.1007/s43546-023-00618-x.

Bentes, S. R., & Navas, R. (2013). The Fundamental Analysis: An Overview. *International Journal of Latest Trends in Finance and Economic Sciences*, 3(1), 389-393.

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, 2(2011), 1-8. https://doi.org/10.1016/j.jocs.2010.12.007.

Cao, L. (2022). AI in Finance: Challenges, Techniques, and Opportunities. *ACM Computing Surveys (CSUR)*, 55(3), 1-38. https://doi.org/10.48550/arXiv.2107.09051.

Dunis, C. L., Middleton, P. W., Theofilatos, K., & Karathanasopoulos, A. (2016). *Artificial Intelligence in Financial Markets: Cutting Edge Applications for Risk Management, Portfolio Optimization and Economics*. Palgrave Macmillan

Effrosynidis D., Symeonidis S., & Arampatzis A., (2017 September). *A Comparison of Preprocessing*

Techniques for Twitter Sentiment Analysis. 21st International Conference on Theory and Practice of Digital Libraries, Thessaloniki, Greece, 2019.

Guo, K., & Xie, H. (2024). Deep learning in finance assessing Twitter sentiment impact and prediction on stocks. *PeerJ Computer Science*, 10(2018). https://doi.org/10.7717/peerj-cs.2018.

Hayran, A., & Sert, M. (2017 May). *Sentiment Analysis on Microblog Data based on Word Embedding and Fusion Techniques*. IEEE 25th Signal Processing and Communications Applications Conference (SIU 2017), Antalya, Türkiye.

Hugging Face. (2023). *BERT Base Turkish Sentiment*. Retrieved from https://huggingface.co/savasy/bert-base-turkish-sentiment-cased. Accessed on March 10, 2023.

Investing. (2024). *BIST 30 Index*. Retrieved from https://tr.investing.com/indices/ise-30. Access Date: April 30, 2024.

Is Yatirim. (2024). *Historical prices*. Retrieved from https://www.isyatirim.com.tr/tr-tr/analiz/hisse/Sayfalar/Tarihsel-Fiyat-Bilgileri.aspx. Access Date: April 30, 2024.

Kearney, C., & Liu, S. (2014). Textual sentiment in finance: A survey of methods and models. *International Review of Financial Analysis*, *33*(2014), 171-185. https://doi.org/10.1016/j.irfa.2014.02.006.

Kına, E., & Biçek, E. (2024). Machine Learning Approach for Emotion Identification and Classification in Bitcoin Sentiment Analysis. *Yuzuncu Yil University Journal of the Institute of Natural & Applied Sciences, 29*(3), 913-926. https://doi.org/10.53433/yyufbed.1532649

Kumar, A. (2009). Hard-To-Value Stocks, Behavioral Biases, and Informed Trading. *Journal of Financial and Quantitative Analysis*, 44, 1375-1401. https://dx.doi.org/10.2139/ssrn.903820.

Lee, C., Gao, Z., & Tsai, C. (2020 September). *BERT-Based Stock Market Sentiment Analysis*. 2020 IEEE International Conference on Consumer Electronics, Taoyuan, Taiwan.

Li, X., Xie, H., Chen, L., Wanga, J., & Deng, X. (2014). News impact on stock price return via sentiment analysis. *Knowledge-Based Systems*, *69*(2014), 14-23. https://doi.org/10.1016/j.knosys.2014.04.022.

Liu, Z., Yang, L., & Takada, T. (2023 July). *Predicting Stock Prices Using Tweet Frequency and AI: Leveraging Social Media Insights to Forecast Tomorrow's Market Trends*. 12th International Conference on Digital Image Processing and Vision, London, United Kingdom.

Liu, B. (2012). *Sentiment Analysis and Opinion Mining*. Morgan and Claypool Publishers.

Motiwalla, L., & Wahab, M. (2000). Predictable variation and profitable trading of US equities: A trading simulation using neural networks. *Computers & Operations Research*, *27*, 1111–1129. https://doi.org/10.1016/S0305-0548(99)00148-3.

Namin, S. S., & Namin, A. S. (2018). Forecasting Economic And Financial Time Series: ARIMA vs. LSTM. Lubbock, TX, USA: Texas Tech University; 2008. doi:https://arxiv.org/ftp/arxiv/papers/1803/1803.06386.pdf

Naseer, M., & Tariq, Y. B. (2015). The Efficient Market Hypothesis: A Critical Review of the Literature. *The IUP Journal of Financial Risk Management*, 12(4), 48-63.

Ozili, P. K. (2023). The acceptable R-square in empirical modelling for social science research. *Munich Personal RePEc Archive*. https://mpra.ub.uni-muenchen.de/id/eprint/116496

Pattanayak, A. M., Swetapadma, A., & Sahoo, B. (2024). Exploring Different Dynamics of Recurrent Neural Network Methods for Stock Market Prediction - A Comparative Study. *Applied Artificial Intelligence*, *38*(1), https://doi.org/10.1080/08839514.2024.2371706.

Rao, T., & Srivastava, S. (2012 August). *Analyzing Stock Market Movements Using Twitter Sentiment Analysis*. International Conference on Advances in Social Networks Analysis and Mining, Istanbul, Türkiye.

Ruan, Q., Wang, Z., Zhou, Y., & Lv, D. (2020). A new investor sentiment indicator (ISI) based on artificial intelligence: A powerful return predictor in China. *Economic Modelling*, *88*(2020), 47-58. https://doi.org/10.1016/j.econmod.2019.09.009.

Schumaker R.P., Zhang Y., & Huang C.H. (2012). Evaluating sentiment in financial news articles. *Decision Support Systems, 53*(3), 458-464. https://doi.org/10.1016/j.dss.2012.03.001.

Shah, D., Isah, H., & Zulkernine, F. (2019). Stock Market Analysis: A Review and Taxonomy of Prediction Techniques. *International Journal of Financial Studie, 7*(2)*, 26–33*. https://doi.org/10.3390/ijfs7020026.

Shubham, A., Kumar, N., Rathee, G., Kerrache, C., Calafate, C., & Bilal, M. (2024). Improving stock market prediction accuracy using sentiment and technical analysis. *Electronic Commerce Research,* 1-24. https://doi.org/10.1007/s10660-024-09874-x.

Sohangir, S., Wang, D., Pomeranets, A., & Khoshgoftaar, T. M. (2018). Big Data: Deep learning models for financial sentiment analysis. *Journal of Big Data*, *5*(1), 1-25. https://doi.org/10.1186/s40537-017-0111-6.

Souza, T. T. P., Kolchyna, O., Treleaven, P.C., & Aste, T. (2015). Twitter Sentiment Analysis Applied to Finance: A Case Study in the Retail Industry. ArXiv:1507.00784.

Thawornwong, S., & Enke, D. (2004). The adaptive selection of financial and economic variables for use with artificial neural networks. *Neurocomputing*, *56*, 205–232. https://doi.org/10.1016/j.neucom.2003.05.001.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017 December). *Attention Is All You Need*. 31st International Conference on Neural Information Processing Systems, Long Beach, California, USA, 2019.

White, H. (1988 July). *Economic prediction using neural networks: The case of IBM daily stock returns*. Second IEEE annual conference on neural networks, San Diego, CA, USA.

Yıldırım, S. (2024). *Fine-tuning Transformer-based Encoder for Turkish Language Understanding Tasks* (ArXiv:2401.17396).

Yu, Y., Duan, W., & Cao, Q. (2012). The impact of social and conventional media on firm equity value: A sentiment analysis approach. *Decision Support Systems, 55*(4), 919-926. https://doi.org/10.1016/j.dss.2012.12.028.

Zhang, X., Fuehres, H., & Gloor P. A. (2011). Predicting Stock Market Indicators Through Twitter "I hope it is not as bad as I fear". *Procedia - Social and Behavioral Sciences*, *26*(2011), 55-62. https://doi.org/10.1016/j.sbspro.2011.10.562.

İSTANBUL UNIVERSITY PRESS

Acta Infologica

Acta Infologica,
ACIN 2025, 9 (1): 275–292

https://doi.org/10.26650/acin.1634472
Submitted: 07.02.2025
Revision Requested: 27.02.2025
Last Revision Received: 12.05.2025
Accepted: 03.06.2025
Published Online 16.06.2025

Research Article

Open Access

Enhancing Software Requirement Classification via Dataset Fusion and Machine Learning

Muhammad Owais Raza [1], Vajeeha Mir [2], Jawad Rasheed [1,3,4], Mirsat Yeşiltepe [5] & Shtwai Alsubai [6]

[1] İstanbul Sabahattin Zaim University, Department of Computer Engineering, İstanbul, Türkiye
[2] Mehran Univerisity of Engineering and Technology, Department of Software Engineering, Jamshoro, Pakistan
[3] Applied Science Private University, Applied Science Research Center, Amman, Jordan
[4] İstanbul Nisantaşı University, Department of Software Engineering, İstanbul, Türkiye
[5] İstanbul University, Department of Intelligent Transportation System, İstanbul, Türkiye
[6] Department of Computer Science, College of Computer Engineering and Sciences in Al-Kharj, Prince Sattam Bin Abdulaziz University, Al-Kharj, Saudi Arabia

Abstract Software engineering involves numerous steps; a successful software product follows these guidelines to the core. One such step is gathering requirements for a software product. This step is quite expensive in terms of time and money; a potential solution is to automate the requirement collection process. Automating the process of gathering software requirements requires separating requirements into types. An approach to predict the type of requirement is using text classification and machine learning; however, the problem with this approach is that it requires a large amount of data, which is not available for this use case. In this study, we perform dataset fusion to create a large dataset. We applied vertical fusion, which increased the number of instances in the dataset. Once a fusion-based dataset is created, machine learning algorithms are applied, and based on empirical results, the performance of the machine learning model after fusion drastically improved to 87.78% f1score with support vector machine (SVM). This improvement shows the efficacy of data fusion in improving the performance of a text classifier and demonstrates that it can overcome the limitations of small datasets by combining data from diverse sources. Our study demonstrated the robustness of our approach in software requirement classification by surpassing the highest recall scores from the previous four years, achieving 94.20% with fusion-based SVC and outperforming previous models even in non-fusion settings.

Keywords Software Engineering · Machine Learning · Software Requirement Engineering · Hybrid Models · Ensemble Modeling

Citation: Raza, M. O., Mir, V., Rasheed, J., Yeşiltepe, M. & Alsubai, S. (2025). Enhancing software requirement classification via dataset fusion and machine learning. *Acta Infologica*, *9*(1), 275-292. https://doi.org/10.26650/acin.1634472

This work is licensed under Creative Commons Attribution-NonCommercial 4.0 International License.

2025. Raza, M. O., Mir, V., Rasheed, J., Yeşiltepe, M. & Alsubai, S.

Corresponding author: Muhammad Owais Raza 6210024002@std.izu.edu.tr

Introduction

A successful computer program (i.e. Software) is the result of robust software requirements engineering processes; thus, software requirements (SR) are the foundation of software systems (Laplante & Kassab, 2022). The requirement gathering step is one of the most critical and difficult parts of SDLC (The Software Development Life Cycle) (Okesola et al., 2020). Risks associated with software requirements can be predicted using classification algorithms during the requirement collection stage (Handa, Sharma, & Gupta, 2022; Luitel, Hassani, & Sabetzadeh, 2024; Mullis et al., 2024).

There are two types of software requirements: functional requirements (FRs) and nonfunctional requirements (NFRSs) (Eckhardt, Vogelsang, & Fern'andez, 2016). The FRs are considerably easier to grasp and extract from the user narratives at (requirement gathering process) RGP. Non-functional requirements (NFRs) are also vital; they play a key role in improving product quality and determining the acceptability of a proposed system. Overly high NFRs can rapidly increase costs, whilst low non-functional demands can result in poor user experiences. It has been noted that NFR corrections are far more costly and challenging to perform.

Classifying the SR for each group is challenging, although automated requirement extraction and categorization can save time. It can be achieved through the use of machine learning (ML) (Kaur & Kaur, 2023; Rahman et al., 2024; Rahimi, Eassa, & Elrefaei, 2021), and various ML algorithms find the best results by learning from such data (Textual data). Supervised learning is a suitable choice for categorization problems, such as determining which category the data belong to (Khatian et al., 2021). To predict risks in new software requirements, a dataset that includes software requirements requirements is necessary.

One of the key problems in SR prediction is limited data availability, and the available datasets are relatively small for text classification (Catal & Diri, 2009). To address this issue, a potential solution is dataset fusion i.e. (combining two datasets to either increase number of features or number of instances). Therefore, in this study we are proposing a data fusion technique and apply machine learning to improve the performance of ML models. The main contribution of this study are:

According to the study, data fusion can successfully address the drawbacks of small datasets by merging information from several sources. In this study the Support Vector Classifier (SVC) attains an F1-score of 87.78\% which is quite exceptional for this task.

A thorough comparative analysis was performed to assess how well the machine learning models performed on separate datasets as opposed to the fused datasets and with the literature.

This study outperformed prior research in both fusion and non-fusion contexts, demonstrating consistent superiority in software requirement classification using machine learning models with recall scores of 94.20\% (fusion) and 77.77\% (non-fusion).

Literature Review

. Supervised and unsupervised machine learning techniques have been effectively used to categorize the non-functional requirements of software applications (Handa, Sharma, & Gupta, 2022). They created a model to categorize NFRs from the input text documents, and they evaluated several machine learning techniques using various criteria, including accuracy, recall, performance, and precision (Younas et al., 2020). Suggested a method for recognizing and categorizing NFRs into 14 different groups using an unsupervised machine-learning strategy based on developer- and user-oriented quality factors. Their method demonstrates how unsupervised machine learning can be used to properly categorize NFRs. (Hey et al., 2020) Suggested the

Norbert transfer learning model, which is a developed form of the BERT model. Applying natural language processing techniques, Norbert has demonstrated promising results in categorizing non-functional needs by using the best aspects of the BERT model.

A classification approach based on ontologies was employed to categorize nonfunctional needs into their respective subclasses. This study produced a fresh classifier using a Support Vector Machine (SVM) and a gold standard corpus to automatically classify requirement words from Software Requirement Specifications (SRS). Individuals populate the later ontology by connecting sentences with the correct classes; this allows SPARQL to be used to obtain sentences of a certain class, such as security (Rashwan, Ormandjieva, & Witte, 2013).

According to a method developed in (Haque, Rahman, & Siddik, 2019), user evaluations of a few app shops, such as Play Store, Apple Store, and others, were performed using 4 feature selection methods. These methods divided the requirements into NFR (Non Functional Requirements), functional requirements, and others. They performed this by combining the ML methods J48, Bagging, and Naive Bayes with the following four classification techniques: When comparing the F1 scores of various permutations of methods and algorithms, they found that the integration of AUR-BoW and bagging produced better results. In addition, they discovered that bagging is more appropriate for classifying NRFs from user feedback than the J48 and Naive Bayes results (Khatian et al., 2021). This research shows that supervised ML algorithms are suitable options for extracting and identifying NFRs in SR. To categorize NFR, empirical research contrasted 4 feature engineering techniques and multiple ML methods (Binkhonain & Zhao, 2019). In (Li & Nong, 2022), NFR from SRE papers was automatically categorized using a DNN model named NFRNet. The network is divided into two sections. To understand the context representations of the requirements descriptions, one method uses enhanced BERT based on ngram masking, and the other method uses a BILSTM classification network together with the context information. Paper (Khan et al., 2023) provides a unique methodology using transfer learning (TL) techniques to perform NFR detection and classification to reduce resource consumption and development time, which will eventually result in increased efficiency in software development. For this purpose, they assessed many cutting-edge pretrained models, such as XLNet, BERT, Distill Roberta, and Electra small (Devlin, Chang, & Toutanova, 2019; Khan et al., 2023).

The goal of this study (García, Fernández-y-Fernández, & Pérez, 2023) is to employ a convolutional neural network to enhance the performance of the categorization of non-functional requirements (NFR). In addition, it aims to demonstrate the significance of preprocessing, sampling strategy implementation, and the utilization of pre-trained matrices like Word2vec, Glove, and Fasttext (García, Fernández-y-Fernández, & Pérez, 2023). . An Agile methodology was explained and proposed for managing nonfunctional needs (Sherif, Helmy, & Galal-Edeen, 2023). The requirements gathering, assessment, documentation, and validation are among the main requirements engineering processes that the framework facilitates. The framework also manages recommendations that are not functional (Sherif, Helmy, & Galal-Edeen, 2023).

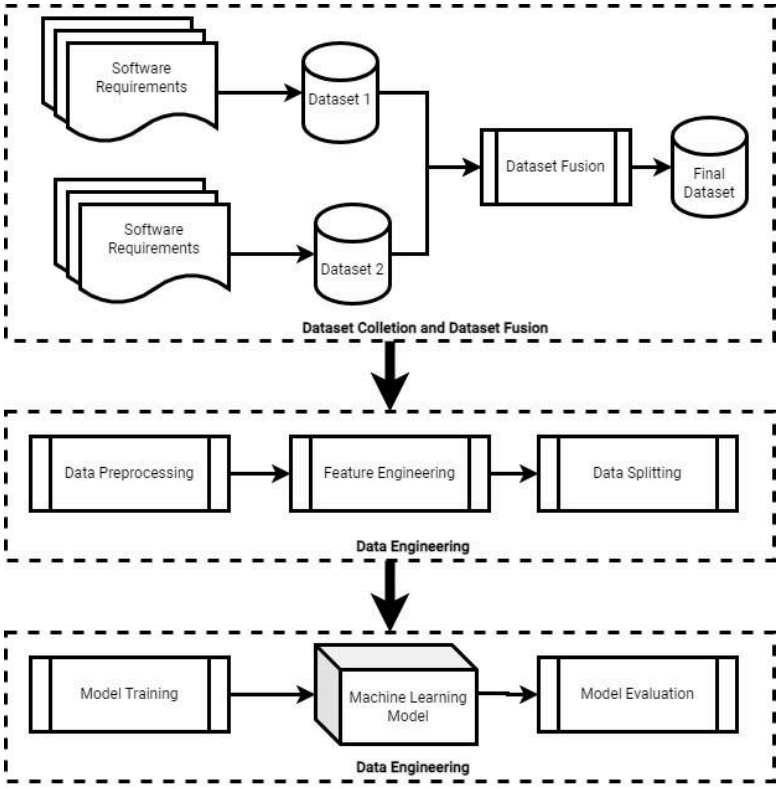
Based on the literature, very limited work is done for issue of limited data availability; therefore, in this study, data fusion was performed to create an augmented dataset. This dataset was trained on machine learning models, and the performance of machine learning models trained on individual and fused datasets was compared. Based on the comparison, the usefulness of data fusion for the use case of SR prediction was evaluated.

Methodology

A three-layered methodology was applied in this study. The first layer performs data collection and fusion, the second layer performs data engineering (i.e cleaning, feature selection and splitting), and the third layer performs machine learning modeling. Figure 1 shows the flow chart diagram of the methodology. In the following section, each step is discussed in detail.

Figure 1

Methodology Flow Diagram



Dataset Collection and Dataset Fusion

The first step in the methodology is the collection of dataset and performing data fusion. For this study, two publicly available datasets were used, and the fusion algorithm was applied to create a new large dataset for better performance. Details regarding datasets and fusion process are described in the following.

The datasets D_1 (Owais, 2025) and D_2 (Owais, 2025) both contain the software requirements in textual for and classification of those requirements into specific classes. D_1 contains 1005 rows, and D_2 has 978 rows. The independent variable in this study was the requirements from the datasets, and the dependent variable was the category or class of the requirement both dataset D_1, and D_2 contain different numbers of classes, and each dataset is evaluated based on the number of classes it contains. Table 1 lists the classes of requirement in each dataset and their number of instances.

Table 1
Class Distribution of Datasets D_1 and D_2

Class Distribution		
Requirements	# of Instance D_1	# of instances D_2
F	265	209
SE	147	56
US	118	63
O	96	58
PE	89	54
LF	78	34
A	56	21
MN	51	17
SC	43	21
PO	24	2
FT	23	10
FR	N/A	312
NFR	N/A	110
L	15	10

The class distribution for D_1 and D_2 is shown in Table 1, with D_1 exhibiting a greater diversity of cases. The majority of classes, including PO (D_1:24, D_2:2), have fewer instances in D2 even though F (D_1:265, D_2:209) is still the greatest in both classes. Indicating dataset differences, FR (D_2:312), NFR (D_2:110), and L (D_2:10) are unique to D2.

There are two primary ways to fuse the dataset, vertical and horizontal fusion. In horizontal dataset fusion, researchers merge two datasets side by side by aligning them on a common key or index. In horizontal dataset fusion, we merged two datasets side by side by aligning them on a common key or index.. The horizontal fusion assumption is This assumes that datasets D_1 and D_2 have the same number of rows and are aligned on a common key or index. As this type of fusion is not possible for this case, the dataset in this study do not coincide in terms of features but instances; thus, instances need to be merged not the features so that there is another type of fusion called vertical fusion, which is used in this study. In vertical data fusion, two datasets are combined by stacking one over another. The assumption in this type is that both datasets have similar columns. In this case, vertical data fusion is used. The mathematical representation of vertical data fusion is given as follows. D_1 and D_2 have m and n rows, respectively, and mathematically, it is defined as follows:

$$D_1 = \{(x_{11}, x_{11}, \ldots, x_{1k}), \ldots, (x_{m1}, x_{m2}, \ldots, x_{mk})\} \tag{1}$$

$$D_2 = \{(y_{11}, y_{11}, \ldots, y_{1k}), \ldots, (y_{n1}, y_{n2}, \ldots, y_{nk})\} \tag{2}$$

The vertical dataset fusion is performed by obtaining the union between the rows of both datasets, which is represented in this study as D_{vf} i.e. vertically fused dataset. Mathematically, it is expressed as follows.

$$D_{vf} = D_1 \cup D_2 \tag{3}$$

$$D_{vf} = \{\{(x_{11}, x_{11}, \ldots, x_{1k}), \ldots, (x_{m1}, x_{m2}, \ldots, x_{mk})\}, \{(y_{11}, y_{11}, \ldots, y_{1k}), \ldots, (y_{n1}, y_{n2}, \ldots, y_{nk})\}\} \tag{4}$$

Vertical data fusion is only possible when different datasets have consistent and compatible column structures. In this study, vertical data fusion combined two datasets to satisfy software requirements to form a large dataset. This increases the sample size and improves the performance of the machine learning models.

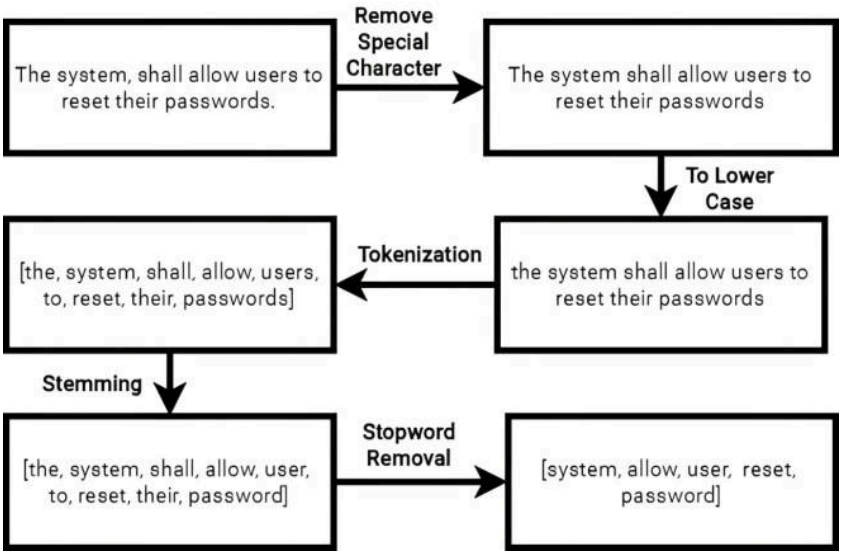
In this study, before feature extraction and classification by vertically concatenating the datasets D_1 and D_2. This method guarantees that both dataset's original feature space of both datasets is preserved while using all available data. This allows the model to learn intricate relationships between datasets. This research approach provides a deeper understanding of the best fusion strategies for enhancing model accuracy..

Data Engineering

The second layer of the methodology involves data preprocessing, feature engineering, and training test split. The following subsection elaborates each step in detail. In the data preprocessing stage, the first thing that occurs is that special characters are removed from the text; only letters are exits in the text from this step forward. Once only the English letter exists, the next step is to convert it into lower case letters for consistency. After making the text consistent throughout the dataset, the next step is tokenization, which involves breaking sentences into words. The next step is to perform stemming. Stemming converts all variation of words into root words, which makes the text consistent and reduces the dimensionality of the data. In the last preprocessing step all the stopwords from the text are removed. Stopwords are common words like 'a', 'the' which are in huge amounts but contribute very little during the classification process. Figure 2 shows the complete life cycle of preprocessing a text.

Figure 2

Preprocessing Step During Data Engineerinng Process



Once data is preprocessed and next step is to perform feature engineering. This is a crucial step in the text classification machine learning pipeline. In this study TF-IDF (Term Frequency-Inverse Document Frequency) vectorization was used. It converts textual data (i.e software requirements) into numerical features that can be fed to machine learning algorithms. Mathematically TF-IDF is shown in equation 5.

$$TF - IDF(t, d) = TF(i, d) \times IDF(t) \tag{5}$$

$$TF(t,d) = \frac{\text{No of times term t appears in document d}}{\text{Total No of terms in document d}} \tag{6}$$

$$TF(t,d) = \log \frac{\text{Total No documents}}{\text{No of documents containing term t}} \tag{7}$$

TF-IDF improves classification performance by incentivizing informative features and down-weighting common but insignificant ones. Because it captures the importance of features well, especially in text-based or high-dimensional datasets, this approach was selected.

For machine learning modeling and evaluation phase two sets of datasets are required, one for training and the other for testing or validation. In this study, a stratified split was used to break the dataset into training and testing sets. The split ratio used in this study is 80%-20%, 80% is the part of dataset used for model training, and the remaining 20% of the dataset is used for testing the model. The specific training and testing details are given in the next section.

Modeling and Evaluation

In this phase, models are developed using the training set, which comprises 80% of the dataset, and the remaining 20% of the dataset is used for testing the model. The models were tested on various performance metrics discussed as follows. We use a different of ML algorithms to build predictive engines during this phase. The training dataset was used to fit the models. To identify various patterns and relationships in the textual data (Software Requirements in this case). Python is the programming language used to implement the machine learning models in this study. Scikit-learn (sklearn) is a key library used for hyperparameter tuning and model building, training, and evaluation. Numpy manages array operations and numerical computations, and pandas handles data handling and preparation. The Matplotlib tool is used to build visualizations showing data distributions and model performance. The scikit-learn methods were used to compute the performance metrics, which include accuracy, precision, recall, and F1-score. The following are the algorithms used in this study:

Logistic Regression (LR): The LR model is a statistical machine learning model that uses logistic regression as its dependent variable. Using logistic regression, data and the relationship between one or more independent variables and dependent variables are described. Nominal, ordinal, or interval variables could be independent variables. Logistic regression is a term that comes from the concept of a logistic function. The logistic function is sometimes referred to as the sigmoid function. The range of values for this logistic function is 0–1. A logistic function is used by a binary classification linear model (LR) to predict the probability of a binary outcome (Dias Canedo and Cordeiro Mendes (2020).

Support Vector Machine (SVM): SVM is a popular ML algorithm. SVM algorithms find the best decision boundary to break the N-dimensional vector space into target labels or classes. This approach makes it easy for SVM to classify new data instances in the future. A hyperplane is a term used to describe this ideal decision boundary. SVM chose the extreme points and vectors in order to construct the hyperplane. According to (Wang, & Wang 2023), SVM works well for datasets with distinct margins of separation and in higher dimensions.

K-Nearest Neighbor (KNN): A KNN is a basic ML algorithm based on a supervised ML approach. The KNN algorithm takes the assumption that the new instances and previous data instances are similar for the existing cases, and it assigns the new instance to the label that is most similar to the existing labels. The KNN algorithm saves the available data and then employes similarity matrics to classify a new data instance.

This KNN algorithm simplifies the process of assigning recently discovered data to an appropriate category. When working with large datasets, a processor can demand substantial processing power despite being simple to use (Syriopoulos et al., 2023).

Decision Tree (DT): Regression problems can be solved using DT; however, classification problems are the main application of these supervised learning techniques. The features of a dataset are represented by internal nodes in this tree-structured classifier, the decision rules are represented by branches, and the outcome is represented by each leaf node. There are two nodes in a DT: the decision node and the leaf node. The result of these choices is a leaf node, which does not have further branches, unlike decision nodes that have multiple branches and are used to make specific decisions based on feature values. The proposed method manages categorical and numerical data and is easy to understand (Yuvaraj et al., 2021).

Random Forest (RF): DTs are used in RF, also called random decision forest, a supervised machine learning algorithm for classification, regression, and other applications. To generate a set of DTs, a random subset of the training set was selected. The votes from these decision trees are then tallied to determine the ultimate prediction. It reduces overfitting and enhances predictive performance by using majority voting (Yuvaraj et al., 2021).

Gaussian Naive Bayes (GB): A subset of Naive Bayes called GB accounts for continuous attributes and distributes data features across the dataset in a Gaussian manner. GB is a classification algorithm that uses continuous normally distributed features and is based on the Naive Bayes algorithm, according to the Sklearn library. It works effectively and efficiently on large datasets (Luo, 2021).

AdaBoosting (AB): AdaBoosting is an ML technique that employs an ensemble learning approach. Using the training dataset, it iteratively trains a weak classifier. AB's basic principle is that each succeeding classifier assigns more weight to incorrectly classified data points. The final AdaBoost model combines all of the training weak classifiers (Carreira-Perpiñán & Zharmagambetov, 2020).

Gradient Boosting (GRB): GRB is a very strong algorithm that transforms a number of weak classifiers or regressors into strong classifiers or regressors ones. GRB minimizes the loss function, for example the MSE (mean squared error or cross-entropy) of the previous models, and gradient descent is used to train each new model. The algorithm first calculates the gradient of the loss function in relation to the current ensemble's predictions for each epoch, after which it trains a new weak model to minimize the gradient. (Burhan & Ali, 2023).

Once the models were trained on the training set of the split test set, the next step was to evaluate the performance of the model in the test set. The instances in the test sets are new to the model i.e. data not seen by the model. Numerous performance metrics are available for classification problems; however, the selected metrics for this study are accuracy, precision, recall, and the F1 Score. Each of these metrics is described as follows.

Accuracy: Accuracy is the ratio of True Positive (TP) and True Negative TN to all instances (TP, TN, FP, FN). The result is shown in equation 8:

$$\text{Accuracy} = \frac{T_p + T_n}{T_p + T_n + F_p + F_n} \tag{8}$$

Precision: The ratio of True Positive (TP) to all positive instances (TP and FP) is referred to as precision. Equation 9 represents precision

$$\text{Preicsion} = \frac{T_p}{T_p + F_p} \tag{9}$$

Recall: Recall is the ratio of True positive to True positive and True negative; Equation 10 represents recall mathematically.

$$\text{Accuracy} = \frac{T_p}{T_p + F_n} \tag{10}$$

F1 Score: The F1 score is the hormonic mean of precision and recall and is considered a robust metric to evaluate a classification task. Equation 11 represents F1 Score.

$$\text{Accuracy} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{11}$$

Results

Table 2 shows the accuracy, precision, and recall for all 8 algorithms over the three datasets (D_f, D_1, and D_f) considered in this study. For D_1, the best performing algorithm was gradient boosting, which achieved accuracy, precision, and recall of 78.11%, 65.24%, and 76.87%, respectively, followed by RF and GB with accuracy of 78% and 76%. Gradient Boosting also provided the best results for D_2 with an accuracy of 70.92%. The second best model for D_2 is Naive Bayes with an accuracy of 70.92%. The top three algorithms for D_f were SVC, RF, and GRB with accuracy scores of 89.67%, 89.67%, and 89.17%. respectively. Overall, these accuracies were obtained for D_f because the overall number of instances of D_f is twice the number of rows in D_1 and D_2.

Table 2

Performance measure of different datasets with different algorithms

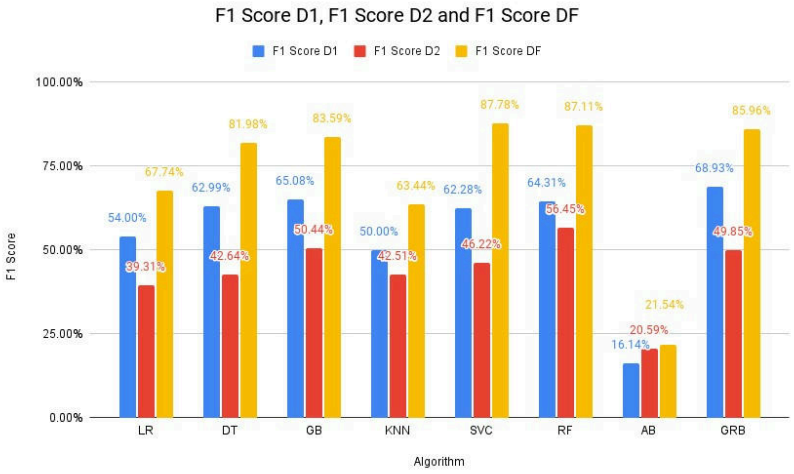
Accuracy–precision recall			
D1			
Algorithm	Accuracy	Precision	Recall
LR	70.65%	49.37%	64.70%
DT	71.14%	66.67%	63.11%
GB	76.62%	63.65%	69.52%
KNN	66.67%	50.15%	53.77%
SVC	73.63%	55.71%	77.77%
RF	78.11%	63.20%	74.88%
AB	33.83%	17.66%	23.57%
GRB	78.11%	65.24%	76.87%
D2			
LR	69.39%	37.53%	48.67%
DT	70.41%	40.11%	48.95%
GB	70.92%	49.30%	57.69%
KNN	70.41%	41.79%	54.27%
SVC	69.90%	41.07%	69.75%
RF	72.96%	50.84%	77.78%
AB	54.08%	22.45%	28.05%
GRB	70.92%	46.70%	57.18%

Accuracy–precision recall			
	Df		
LR	79.60%	62.29%	82.73%
DT	85.64%	83.05%	82.13%
GB	84.38%	82.65%	85.27%
KNN	73.30%	63.36%	68.22%
SVC	89.67%	83.86%	94.20%
RF	89.67%	83.80%	92.35%
AB	40.55%	22.83%	29.76%
GRB	89.17%	83.61%	89.54%

Figure 3 shows a bar chart that compares the F1 Scores of all algorithms across datasets D_1, D_2, and D_f. For D_1, GRB is able to achieve the best performance with an F1 score of 68.93%, although in case D_2, RF outperformed the algorithm with an F1 score of 56.45%. For D_f, the best performing algorithm was SVC with a top score of 87.78%. AdaBoost consistently received the lowest scores in every scenario (D_1, D_2 and D_f), indicating notable performance disparities among the algorithms. SVC, RF, and GRB consistently performed well across all datasets, especially on D_f. The superior performance of all algorithms on D_f was due to the increased number of instances due to data fusion.

Figure 3

F1 Score Comparison Between Different Datasets



The confusion matrix shows the number of correct predictions per class. This indicates how well or poorly a classifier performs in a given class. Figures 4, 5, and 6 show the confusion matrices for all the algorithms on dataset D_1, D_2, and D_f. The diagonal represents the correct prediction rest shows the incorrect predictions. The classes with smaller number of instances have poor performance. Classes with a large number of instances demonstrate better performance.

Based on Figure 4, the best performing algorithms for dataset D_1 are random forest and gradient boosting because they have the highest number of predictions along the diagonal and every small number of predictions outside the diagonal. The KNN shows poor performance across all classes for dataset D_1

Figure 4

Confusion Matrices For D_1

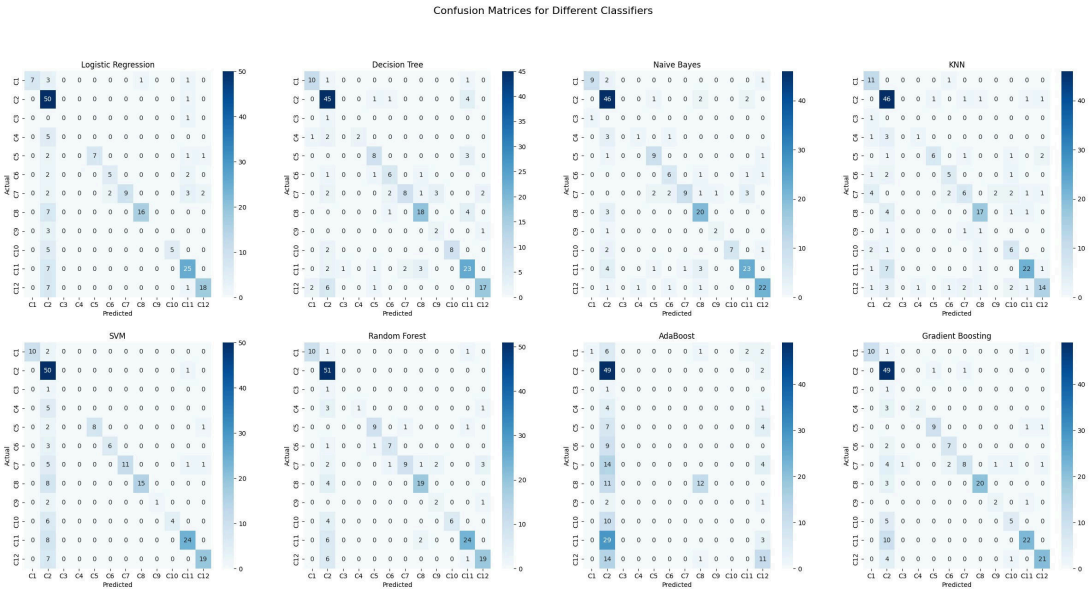


Figure 5 shows confusion matrices for all algorithms for D_2.The results indicate that the Random Forest and Gradient Boosting classifiers perform fairly well, especially for classes 3 and 4, as these have higher values along the diagonal that point to higher classification accuracy for these classes. The SVC and logistic regression classifiers exhibit comparable performance and higher accuracy, especially regarding class 3 and class 6 predictions. Compared to the other models, the Decision Tree, Gaussian Naive Bayes, and AdaBoost models exhibit higher rates of misclassification across several classes.

Figure 5

Confusion Matrices For D_2

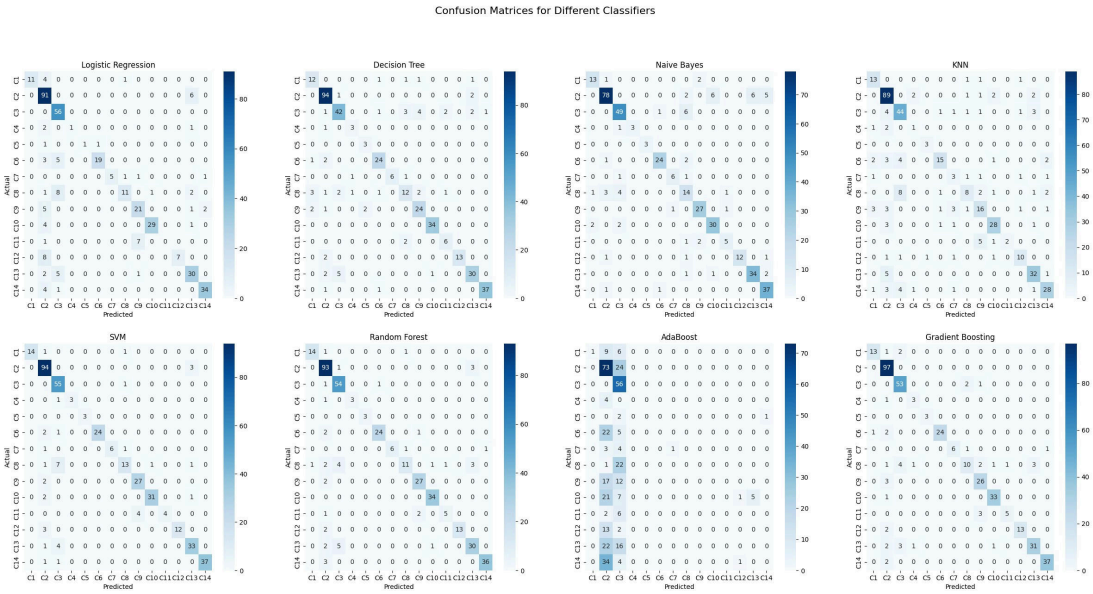


Figure 6 shows the confusion matrices for the fused dataset D_f. As shown in the results, gradient boosting focuses more on correct predictions overall and for key classes such as 12. In addition to gradient boosting,

random forest, logistic regression, KNN, SVC, also show higher level of performance. But model such as DT and GB show misclassification and hence poor classification accuracy. Compared to D_1, D_2, and D_f provided superior accuracy for all classes.

Figure 6

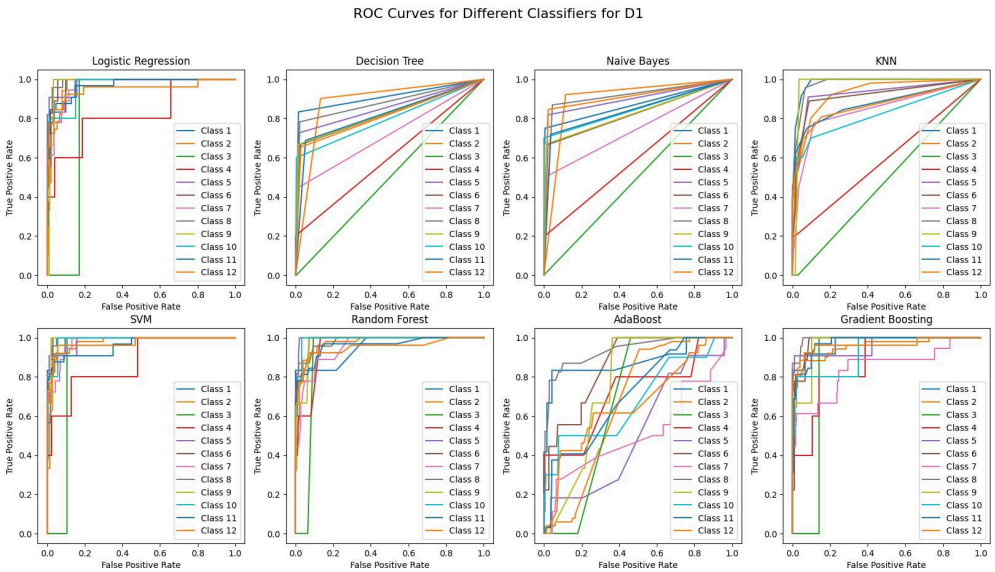
Confusion Matrices For D_f



Figures 7, 8, and 9 show the ROC curve for all algorithms used in this study for all datasets D_1, D_2, and D_f. If the line on the ROC AUC curve follows a straight diagonal line, then the model has no ability to discriminate between right and wrong predictions. Results in Figure 7 show that models such as GB, LR, and RF have good predictive ability as their curves follow a solid convex curve with a higher AUC score for all classes. Althought model such as KNN lack the ability to predict the curve. Through out the algorithms, class 3 is difficult to predict because its curve in some models is closer to a diagonal line.

Figure 7

ROC curve for D_1



Similar to Figure 7 and Figure 8 follows the same convex curve pattern for most of the algorithms, including the LR, SVC, GB, and RF algorithms. The models with weaker ability to predict and differentiae among classes include KNN and GNB with lesser AUC scores and lower convex curves

Figure 8

ROC curve for D_2

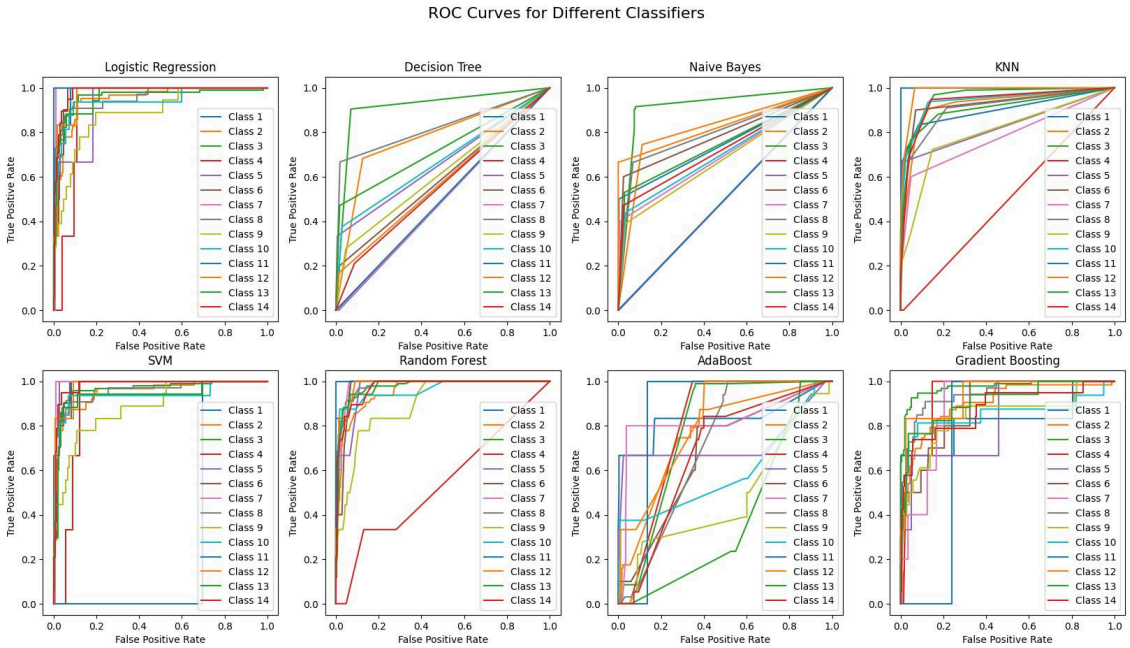
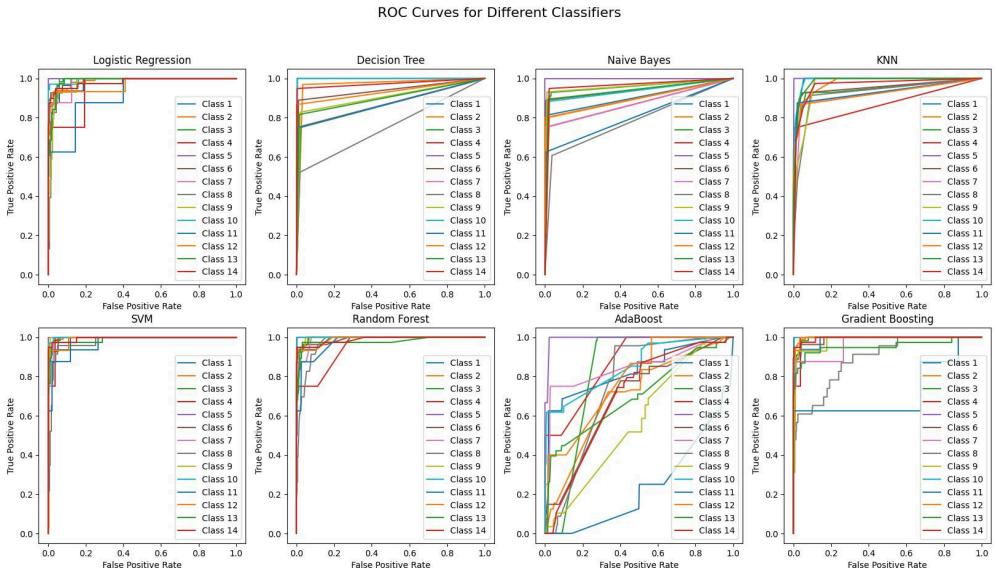


Figure 9 shows the models developed using D_f, which exhibit superior performance compared to models developed on D_1 and D_2. The curve in Figure 11 show higher AUC score and a higher convex curve. These results are evident from the confusion matrix and other performance evaluation parameters.

Figure 9

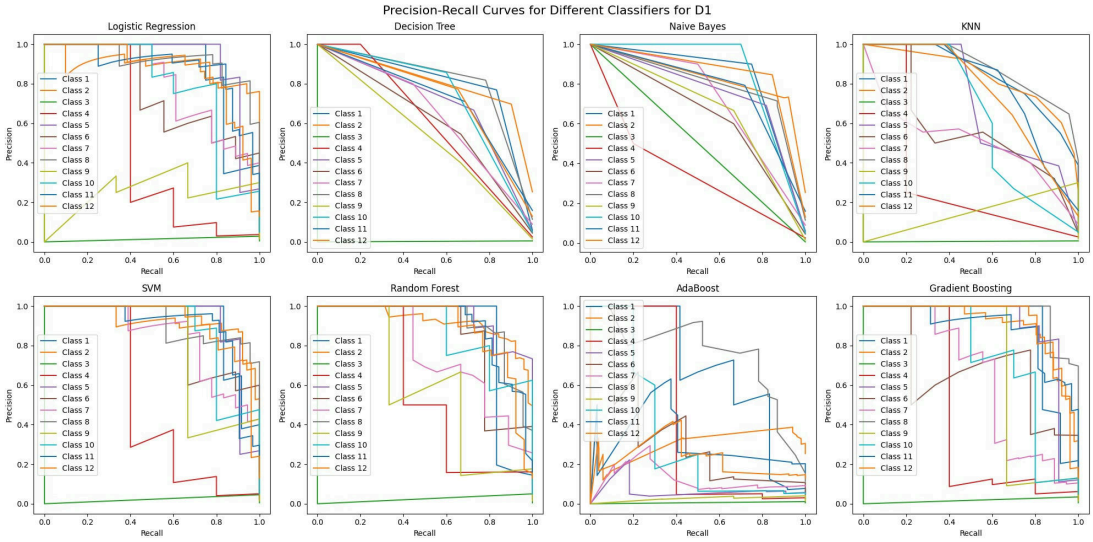
ROC Curve For D_f



Figures 10, 11, and 12 show the precision recall curves for D_1, D_2 and D_f using all the algorithms discussed in this study. A precision recall curve represents the trade-off between precision and recall i.e. sensitivity under varying thresholds. A good precision recall curve has higher precision at higher recall scores, which also results in a higher area under the curve i.e. AUC. The precision-recall curves in Figure 10 demonstrate the varying performance among the classifiers. Here, RF, LR, SVC, and GB exhibit the most consistent performance, and AB continues to perform the worst across classes. Class 3 has proved to be difficult to predict and exhibits extremely poor performance across multiple classifiers for which the AUC for class 3 is between 0.00 and 0.02. Most of the other classes continued to perform well, with an AUC greater than 0.80.

Figure 10

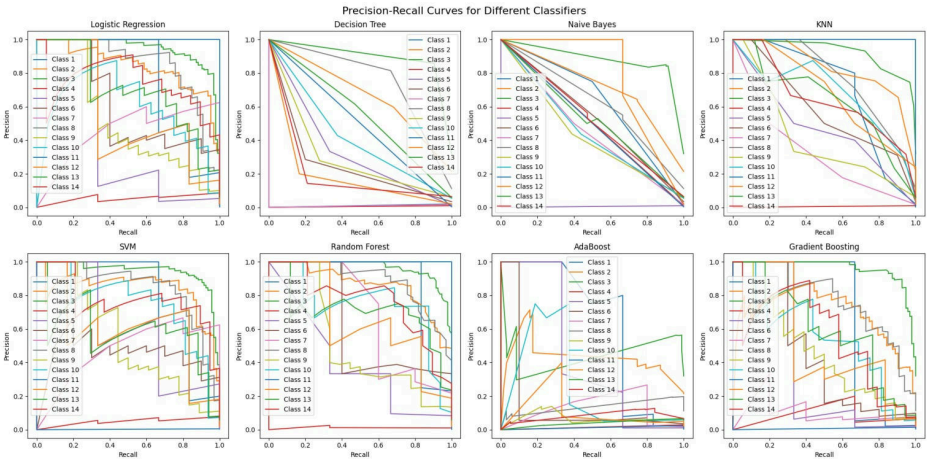
Precision recall curve for D_1



Charts in Figure 11 shows the precision-recall curve for D_2. Based on the graphs, it shows a degradation in model performance compared to D_1. The RF and GB algorithms maintained relatively better performance than the other algorithms on this dataset. For D_2, different classes, such as 4, 5, and 11, showed lower AUC scores across all classifiers, some with an AUC score of 0.50. Overall, AB proved to be the worst performing algorithm.

Figure 11

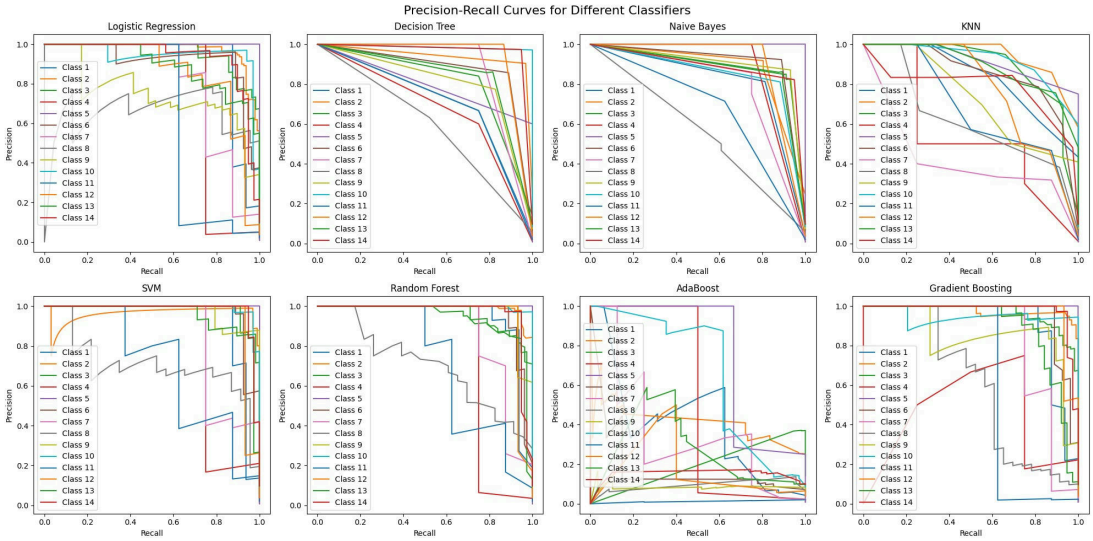
Precision recall curve for D_2



Precision-recall curves in Figure 12 shows superior performance to those of both D_1 and D_2. The curves in Figure 12 shows higher degree of performance across most classifiers, with particularly high AUC scores for GB, RF, LR, and SVC, with many classes achieving AUC values greater than 0.90. However, AB still demonstrated poor performance with several classes having very low AUC scores (less than 0.30.

Figure 12

Precision recall curve for D_f



The models evaluated in this study performed better than the highest recall scores documented literature from last fours years, it is shown in Table 3. The recall metric was selected as the main evaluation criterion due to its crucial role in reducing false negatives, which is a prerequisite for accurately classifying software requirements. To ensure a consistent basis for comparison, only machine learning models were used in the previous studies and the current assessment. The efficacy of the proposed strategy is demonstrated by the improved performance of the proposed models in both fusion and nonfusion settings. Under unfusion, our SVC, GBR, and RF outperformed the existing models with recall of 77.77%, 76.87%, and 74.88%, respectively, and under fusion settings, the SVC outperformed all the models by a huge margin with a recall score of 94.20%.

Table 3

Performance measure of different datasets with different algorithms

Study	Approach	Recall Scores	
(Dias Canedo & Cordeiro Mendes, 2020)	SVM	73%	(Dias Canedo & Cordeiro Mendes, 2020)
Multinomial Naive Bayes (MNB)	77%	(Binkhonain & Zhao, 2023)	Hierarchical Classification for Requirements Classification (HC4RC)
64%	This Study	SVC	77.77%
This Study	GRB	74.88%	This Study
RF	76.87%	This Study	SVC (Fused)
94.24%			

Conclusion

Software systems are based on requirements. The requirements phase of the software development process sets reasonable objectives that must be achieved. The most important and challenging phase of the SDLC (Software Development Life Cycle) is gathering requirements. In the age of automation, the process of requirement gathering need to automatics as well. One of the key features of automation is the recognition of the type of requirements. The data available for this task are very limited. Therefore, in this study, we perform data fusion to create a new dataset from existing one with more instances. Based on the results, data fusion proved to be effective. Separately D_1 and D_2 had an F1 score of 68.93% and 56.45%, respectively, but after fusion, the F1 score improved to 87.78%. In both the fusion and non-fusion scenarios, the models proposed in this study performed better than previous research. SVC achieved 94.20% recall in fusion and 77.77% in non-fusion, demonstrating that our methodology improved the classification of software requirements. The major limitation of this study was the lack of resources available for performing the classification task; thus, performing fusion would not scale. One potential future work of this study to combine more datasets and determine if fusion is as effective as shown in this study, or adding more datasets will alter the improvement.

Peer Review: Externally peer-reviewed.
Author Contributions: Conception/Design of Study- M.O.R., V.M., J.R.; Data Acquisition- M.O.R., V.M., J.R.; Data Analysis/Interpretation- M.O.R., V.M., J.R.; Drafting Manuscript- M.Y., J.R., M.O.R.; Critical Revision of Manuscript- M.Y., J.R., M.O.R., S.A.; Final Approval and Accountability- J.R., S.A.; Technical or Material Support- M.O.R., V.M., J.R.; Supervision- M.Y., J.R., S.A.
Conflict of Interest: The authors have no conflict of interest to declare.
Grant Support: The authors declared that this study has received no financial support.

Author Details

Muhammad Owais Raza
[1] İstanbul Sabahattin Zaim University, Department of Computer Engineering, İstanbul, Türkiye
0000-0002-3065-385X 6210024002@std.izu.edu.tr

Vajeeha Mir
[2] Mehran Univerisity of Engineering and Technology, Department of Software Engineering, Jamshoro, Pakistan
0009-0009-4935-9042

Jawad Rasheed
[1] İstanbul Sabahattin Zaim University, Department of Computer Engineering, İstanbul, Türkiye
[3] Applied Science Private University, Applied Science Research Center, Amman, Jordan
[4] İstanbul Nisantaşı University, Department of Software Engineering, İstanbul, Türkiye
0000-0003-3761-1641

Mirsat Yeşiltepe
[5] İstanbul University, Department of Intelligent Transportation System, İstanbul, Türkiye
0000-0003-4433-5606

Shtwai Alsubai

[6] Department of Computer Science, College of Computer Engineering and Sciences in Al-Kharj, Prince Sattam Bin Abdulaziz University, Al-Kharj, Saudi Arabia

0000-0002-6584-7400

References

Ali, Z. H., & Burhan, A. M. (2023). Hybrid machine learning approach for construction cost estimation: Evaluation of the extreme gradient boosting model. *Asian Journal of Civil Engineering*, 24(7), 2427–2442, 2017.

Binkhonain, M., & Zhao, L. (2023). A machine learning approach for hierarchical classification of software requirements. *Machine Learning with Applications*, 12, 100457. https://doi.org/10.1016/j.mlwa.2023.100457.

Binkhonain, M., & Zhao, L. (2019). A review of machine learning algorithms for the identification and classification of non-functional requirements. *Expert Systems with Applications: X*, 1, 100001.

Catal, C., & Diri, B. (2009). We investigate the effect of dataset size, metrics sets, and feature selection techniques on software fault prediction problems. *Information Sciences*, 179(8), 1040–1058.

Carreira-Perpiñán, M. Á., & Zharmagambetov, A. (2020). *The ensembles of bagged TAO trees consistently improved over random forests, AdaBoost, and gradient boosting.* In Proceedings of the 2020 ACM-IMS on Foundations of Data Science Conference (pp. 35–46).

Devlin, J., Chang, M.-W., & Toutanova, K. (2019). *BERT: Pre-training of deep bidirectional transformers for language understanding.* In Proceedings of NAACL-HLT (p. 2). Minneapolis, MN, USA.

Dias Canedo, E., & Cordeiro Mendes, B. (2020). Software requirements classification using machine learning algorithms. *Entropy*, 22(9), 1057. https://doi.org/10.3390/e22091057.

Eckhardt, J., Vogelsang, A., & Fern'andez, D. M. (2016). *Are non-functional requirements really non-functional? An investigation of non-functional requirements in practice.* In Proceedings of the 38th International Conference on Software Engineering (pp. 832–842). Austin, TX, USA.

García, S. M., Fernández-y-Fernández, C. A., & Pérez, E. R. (2023). Classification of non-functional requirements using convolutional neural networks. *Programming and Computer Software*, 49(8), 705–711.

Handa, N., Sharma, A., & Gupta, A. (2022). Framework for prediction and classification of non-functional requirements: A novel vision. *Cluster Computing*, 25(2), 1155–1173.

Haque, M. A., Rahman, M. A., & Siddik, M. S. (2019). *Non-functional requirements classification with feature extraction and machine learning: An empirical study.* In 2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT) (pp. 1–5). Dhaka, Bangladesh.

Hey, T., Keim, J., Koziolek, A., & Tichy, W. F. (2020). *Norbert: Transfer learning for requirements classification*. In 2020 IEEE 28th International Requirements Engineering Conference (RE) (pp. 169–179). Zurich, Switzerland.

Khan, M. A., Khan, M. S., Khan, I., Ahmad, S., & Huda, S. (2023). *Non-functional requirements identification and classification using transfer learning model.* IEEE Access.

Khatian, V. M., Arain, Q. A., Alenezi, M., Raza, M. O., & Shaikh, F. (2021). *Comparative analysis for predicting non-functional requirements using supervised machine learning*. In 2021 1st International Conference on Artificial Intelligence and Data Analytics (CAIDA) (pp. 7–12). Riyadh, Saudi Arabia.

Kaur, K., & Kaur, P. (2023). *Improving the BERT model for requirements classification using a bidirectional LSTM-CNN deep model.* Computers and Electrical Engineering.

Laplante, P. A., & Kassab, M. (2022). *Requirement engineering of software and systems.* Boca Raton, FL: Auerbach Publications.

Li, B., & Nong, X. (2022). Automatically classifying non-functional requirements using a deep neural network. *Pattern Recognition*, 132, 108948.

Luitel, D., Hassani, S., & Sabetzadeh, M. (2024). Improving requirements completeness: Automated assistance through large language models. *Requirements Engineering*, 29(1), 73–95.

Luo, X. (2021). Efficient English text classification using selected machine learning techniques. *Alexandria Engineering Journal*, 60(3), 3401–3409.

Mullis, J., Chen, C., Morkos, B., & Ferguson, S. (2024). Deep neural networks in natural language processing to classify requirements by origin and functionality: Application of BERT in system requirements. *Journal of Mechanical Design*, 146(4), 041401.

Okesola, O. J., Adebiyi, A. A., Owoade, A. A., Adeaga, O., Adeyemi, O., et al. (2020). Software requirement in iterative SDLC model. In Intelligent algorithms in software engineering: Proceedings of the 9th computer science on-line conference 2020 (pp. 26–34). Cham, Switzerland: Springer; 2020.

Raza, O. "Owais4321/software-requirement-dataset: Dataset for Software Requirement Engineering. dataset, *Zenodo*, Apr. 17, 2025, doi:10.5281/zenodo.15235904.

Rahman, K., Ghani, A., Misra, S., & Rahman, A. U. (2024). A deep learning framework for non-functional requirement classification. *Scientific Reports*, 14(1), 3216.

Rahimi, N., Eassa, F., & Elrefaei, L. (2021). One-and two-phase software requirement classification using ensemble deep learning. *Entropy*, 23(10), 1264.

Rashwan, A., Ormandjieva, O., & Witte, R. (2013). *Ontology-based classification of non-functional requirements in software specifications: A new corpus and SVM-based classifier.* In 2013 IEEE 37th Annual Computer Software and Applications Conference (pp. 381–386). Kyoto, Japan.

Sherif, E., Helmy, W., & Galal-Edeen, G. H. (2023). *Proposed framework to manage non-functional requirements in agile.* IEEE Access, 11, 53995–54005.

Syriopoulos, P. K., Kalampalikis, N. G., Kotsiantis, S. B., & Vrahatis, M. N. (2023). kNN classification: A review. *Annals of Mathematics and Artificial Intelligence*, 1–33.

Wang, H., Li, G., & Wang, Z. (2023). Fast SVM classifier for large-scale classification problems. *Information Sciences*, 642, 119136.

Yuvaraj, N., Chang, V., Gobinathan, B., Pinagapani, A., Kannan, S., Dhiman, G., & Rajan, A. R. (2021). Automatic detection of cyberbullying using multifeature artificial intelligence with deep decision tree classification. *Computers and Electrical Engineering*, 92, 107186.

Younas, M., Jawawi, D. N., Ghani, I., & Shah, M. A. (2020). Extraction of non-functional requirement using semantic similarity distance. *Neural Computing and Applications*, 32, 7383–7397.

İSTANBUL UNIVERSITY PRESS

Acta Infologica

Acta Infologica,
ACIN 2025, 9 (1): 293–313

https://doi.org/10.26650/acin.1626593
Submitted: 24.01.2025
Revision Requested: 01.05.2025
Last Revision Received: 25.05.2025
Accepted: 05.06.2025
Published Online 13.06.2025

Research Article

Open Access

Predicting Industry Maturity Index Using Machine Learning Methods

Ayhan Doğan [1] & Cihan Ünal [1]

[1] Hacettepe University, Başkent Organized Industrial Zone, Technical Sciences Vocational School, Ankara, Türkiye

Abstract Industry 4.0 has become a widely adopted concept in recent years. Maturity and readiness models are commonly used to assess the current state of industrial organizations in relation to Industry 4.0. Companies' maturity levels and index scores are typically determined through structured surveys. However, due to their complexity, time consumption, and high cost, many enterprises lack formally assessed maturity index (MI) scores. To address this limitation, this study initially employed survey data to evaluate the accuracy of the proposed machine learning (ML) framework. A 58-question survey was conducted to calculate the MI scores of the companies. These scores were then used as reference values to be predicted based on five easily accessible enterprise-level variables: company age, industry type, ownership structure, number of employees, and annual turnover. This approach tested whether MI could be accurately predicted without relying on lengthy survey processes, using only a minimal set of key enterprise attributes. The results of this study demonstrate that MI can be estimated successfully using ML techniques without the need for answering long and complex surveys. To reduce the burdens associated with conventional survey-based methods, this study employed multiple ML algorithms, including Support Vector Machines (SVM), Gaussian Process Regression (GPR), Linear Regression (LR), Regression Trees (RT), and Ensemble Tree-based models, and advanced boosting-based methods, such as extreme gradient boosting (XGB) and Light Gradient Boosting Machine (LGBM). The findings demonstrate that the proposed model predicts MI with high accuracy and offers a practical and scalable alternative for enterprises seeking to assess their Industry 4.0 readiness.

Keywords Machine learning · Industry 4.0 · Maturity model · Maturity index

Citation: Doğan, A. & Ünal, C. (2025). Predicting industry maturity index using machine learning methods. *Acta Infologica*, 9(1), 293-313. https://doi.org/10.26650/acin.1626593

This work is licensed under Creative Commons Attribution-NonCommercial 4.0 International License.

© 2025. Doğan, A. & Ünal, C.

Corresponding author: Ayhan Doğan aydogan@hacettepe.edu.tr

Introduction

Industry 4.0 was introduced as a concept at a fair in Germany in 2011. It was used as the name of the state policy used by the German government for digital transformation. According to this fiction, every revolutionary change in the industry is divided into phases: Industry 1.0, Industry 2.0, Industry 3.0, and Industry 4.0. Industry 1.0 represents the invention of steam engines, Industry 2.0 represents electricity and subsequent production, Industry 3.0 represents the beginning of the digital age, and Industry 4.0 represents and symbolizes the fully autonomous production approach.

Industry 4.0 has gained autonomous production capability thanks to many technologies. These include cloud computing, the Internet of Things, artificial intelligence, autonomous robots, sensors, big data, and 3D printers. This way, interaction based on the network, equipment, and data exchange that connect people and machines are established.

Industry 4.0 significantly contributes to businesses thanks to its technology-based production (Saad et al., 2021). However, companies must prepare for Industry 4.0 to adapt to the fourth industrial revolution (Hizam-Hanafiah et al., 2020). Competitiveness depends on efficiency, flexibility, and rapid production. Digitalization and technological transformation are essential (Çınar et al., 2021; Rahamaddulla et al., 2021). However, small businesses may not have the resources and equipment necessary for digital transformation, making competition more difficult (Jesus & Lima, 2020).

Industry 4.0 is a concept that proposes the use of new technologies to increase performance and efficiency (Chonsawat & Sopadang, 2020). Businesses should conduct internal evaluations, examine their competitors, and perform planning to determine Industry 4.0 levels and needs. They require maturity models (MMS) to realize all these processes (Carvalho et al., 2018). While applying for MMS, interviews are held with companies, and various surveys are conducted. Because of these surveys, companies' preparation level for Industry 4.0 can be determined. Appropriate studies should be carried out considering the maturity level of the companies under Industry 4.0. However, companies must be prepared and willing to allocate sufficient resources for this process.

The Industry 4.0 level of many companies is unknown due to reasons such as the time required to conduct surveys and high costs. In this study, the success of machine learning (ML) methods was investigated to make the industrial maturity index (MI) assessments of many enterprises located in industrial zones faster, more economical, and more accessible.

Background and Literature

Candanedo et al. (2018) conducted an extreme case study using the heating, ventilation, and air conditioning systems (HVAC) dataset. Kuo et al. (2017) explored using inexpensive triaxial sensors to monitor machines. To extract meaningful insights from the collected data, they developed a dimension reduction method with a low computational load and a neural network that allows the obtained data to be analyzed automatically (Kuo et al., 2017). Angelopoulos et al. (2019) presented ML solutions related to Industry 4.0 and classified them according to their learning processes. On the other hand, Rai et al. (2021) sought the work of a wide variety of researchers to understand the change that began with the use of ML techniques and to report current research on the fundamental theoretical and experimental aspects of ML and their applications in production and production systems. Brik et al. (2019) recommended a tool for Industry 4.0 to monitor system outages. They used an ML algorithm to develop a predictive model of resource localization,

considering the actual task timing in terms of resource localization. Lee and Lim (2021) analyzed numerous journal articles by text mining using unsupervised ML algorithms.

MMS is an evolutionary way to achieve a goal that takes shape in stages. This path guides the reorganization and restructuring of existing abilities to achieve perfection (Finance, 2015). Within the scope of this study, existing maturity and readiness models were examined. Accordingly, the number of dimensions of the existing models is between 3 and 13. Some measures are not preferred for a good analysis. However, in some sectors and exceptional cases, the small number of measurements is not a problem. Although some dimension names are used differently, the evaluation serves the same purpose. When the existing maturity and preparation models are examined, it can be seen that the maturity level (MTL) of the enterprise is between 4 and 6.

The Industry 4.0 revolution can contribute to small and medium-sized enterprises (SMEs), one of the cornerstones of growth and development (Chonsawat & Sopadang, 2020). The readiness of businesses is a measure of the extent to which they benefit from the new technologies that come with Industry 4.0 (Stentoft et al., 2021). In other words, Industry 4.0 preparation concerns companies preparing to use Industry 4.0 technologies (Rais, 2021; Vazire, 2018). Insufficient understanding of concepts and advantages by businesses constitutes a significant problem (Da Silva et al., 2020; Horváth & Szabó, 2019; Rauch et al., 2019).

During the digital transformation phase, organizations' software and hardware requirements also emerge (Haber et al., 2015; Wank et al., 2016). Achieving Industry 4.0 preparation is essential for businesses today (Schaupp et al., 2017). Assume digital transformation is not well managed. Perhaps, it may be faced with the fact that companies that do not attach sufficient importance to digitalization will disappear from the market (Canetta et al., 2018; Ivanov et al., 2019). MMS is used to measure and evaluate the capabilities required by businesses to reach the desired level (Schumacher et al., 2016). It is also used to see the company's progress over time and compare it with that of its competitors (Pessl et al., 2017).

Industry 4.0 Readiness and maturity models

As the concept of 'Industry 4.0' is new, readiness and maturity assessment tools continue to evolve. In this study, 30 different Industry 4.0 readiness and MMS approaches were examined. Experts in the industry developed nine (30%) of the models inspected, while academics developed the remaining 21 (70%). The list of readiness and MMS is presented in Table 1.

Table 1

Widely used readiness and MMS

No.	Model Name	Reference
1	"Industry 4.0 Readiness Evaluation for Manufacturing Enterprises"	(Basl and Doucek, 2019)
2	"Industry 4.0 Maturity Model"	(Bibby and Dehe, 2018)
3	"Future Readiness Level (FRL)/Industry 4.0 Future Readiness"	(Botha, 2018)
4	"E-Business Industry 4.0 Readiness Model"	(Városiné Demeter et al., 2018)
5	"Benchmarking Readiness I4.0"	"Fraunhofer Institute for Systems and Innovation"
6	"SMEs Maturity Model Assessment of IR4.0 Digital Transformation"	(Hamidi et al., 2018)
7	"Readiness for Industry 4.0"	(Horvat et al., 2018)
8	"SSCM Assessment for Industry 4.0"	(Manavalan and Jayakrishna, 2019)

No.	Model Name	Reference
9	“Industry 4.0 Business Model Innovations Tool”	(Müller and Voigt, 2018)
10	“Industry 4.0 Maturity Model”	(Geissbauer et al., 2016)
11	“Manufacturing Companies Industry 4.0 Adoption Model”	(Mittal et al., 2018)
12	“BMS Smart Industry Research Roadmap (Behavioral, Management, Social Sciences)- SIRM”	“University of Twente”
13	“ACATECH Industrie 4.0 Maturity Index”	“Acatech Academy”
14	“Enterprise 4.0 Assessment”	(Valentin, 2017)
15	“Industry 4.0 Maturity Model- SPICE (Software Process Improvement and Capability Determination)”	(Gökalp et al., 2017)
16	“Industry 4.0 Readiness Model for Tool Management”	(Schaupp et al., 2017)
17	“Three Stages Maturity Model in SMEs toward Industry 4.0”	(Erol et al., 2016)
18	“Design Business Modeling for Industry 4.0”	(Gerlitz, 2016)
19	“SIMMI 4.0–System Integration Maturity Model Industry 4.0”	(Leyh et al., 2016)
20	“Industry 4.0 Introduction Strategy”	“Merz Consulting”
21	“Roadmap Industry 4.0”	(Pessl et al., 2017)
22	“Assessment Model for Organizational Adoption of Industry 4.0 Based on Multi-criteria Decision Techniques”	“University of Warwick”
23	“Industry 4.0 Maturity Model”	(Akdil et al., 2018)
24	“Reference Architecture Model for the Industry 4.0 (RAMI4.0)”	(Kannan et al., 2017)
25	“Industry 4.0 Hindering Factors Model”	(Geissbauer et al., 2016)
26	“IMPULS—Industrie 4.0 Readiness”	“Verband Deutscher Maschinen- und Anlagenbau (VDMA)”
27	“Industry 4.0 Barometer”	“MHP Porsche Company”
28	“Roland Berger Industry 4.0 Readiness Index”	“Ronald Berger Consulting”
29	“Fraunhofer Industrie 4.0 Layer Model”	(Geissbauer et al., 2016)
30	“Industry 4.0 Readiness Model for Manufacturing”	(Methavitakul and Santiteerakul, 2018)
31	“Lean modified manufacturing maturity model for Industry 4.0 (LM4I4.0)”	(Sajjad et al., 2024)
32	“Maturity Model-ADIME 4.0”	(Skalli et al., 2023)
33	“Maturity SCAN 4.0”	(Muñoz et al., 2023)

Machine Learning

In a simple definition, ML involves learning from data methods developed for decision-making processes (Jalil et al., 2019). The goal is to find patterns in the data and make better decisions for the algorithm.

ML, a sub-branch of artificial intelligence, is an algorithm that produces predictions about the problem's solution using data about the given situation. Training models were created for these processes. Although various ML methods require various amounts of data, they often require large amounts of data to continually optimize models and make the best predictions (Wei et al., 2019). several algorithms have been developed for ML. Commonly used are linear regression (LR), ensemble trees (ET), gauss process regression (GPR), decision trees (DT), random forest (RF), support vector machines (SVM), regression trees (RT), and artificial neural networks (ANN). ML is used for different purposes, such as clustering, regression, and classification.

The process begins with the collection of the necessary data. The required data are converted into numerical values. Normalization and standardization processes were applied to the data. All data were used in two groups: training and testing.

This study used seven different methods: ET, SVM, GPR, LR, RT, XGB, and LGBM. Estimates were made on the test data using the models obtained from the training. The prediction accuracy can be measured using various metrics. In this study, root mean square error (RMSE), mean square error (MSE), mean absolute error (MAE), and coefficient of determination (R2) metrics were used.

The k-fold cross-validation method was used to evaluate the prediction results of ML. The reason for preferring this method is the relatively small amount of data. The proposed method is suitable for assessing model success because all data are used alternately in training and testing. However, this method requires relatively more computation because training and testing are performed multiple times. In the k-fold cross-validation method, all data are divided into k layers. A layer is used for testing, while the other is used for training. The most preferred number of layers is 5. Too few or too many layers can degrade the performance of the proposed method. In this study, the number of layers was taken as 5.

Linear regression

LR is used to determine the relationship between the dependent variable and the independent variable. This method is generally preferred when a linear relationship exists between inputs and outputs (Luu et al., 2021). Equation 1 defines the LR method.

$$y = a_0 + a_1 X + \varepsilon \tag{1}$$

In Equation 1, ε is the error term, X is the independent variable, and y is the dependent variable. In addition, a_0 is the constant term, and a1 is the regression coefficient showing the slope. If the value of a1 is positive, then the change is increasing, and if the value of a1 is negative, then the difference is decreasing (Bayazit & Oğuz, 1994; Doğan et al., 2023).

The LR method uses one or more arguments. If there is more than one independent variable, the method is called multiple linear regression (MLR) (Memnun & Kalaycı, 2006; Rong & Bao-Wen, 2018). The formula for MLR methods is given in Equation 2.

$$y = a_0 + a_1 X_1 + a_2 X_2 + \cdots + a_n X_n + \varepsilon \tag{2}$$

The values a1, a2, ... an in Equation 2 are multiple regression coefficients. X1, X2, ... Xn values are multiple independent variables (Aslan et al., 2011).

Support vector machines

In the early 1990s, Vapnik developed SVM (Cortes & Vapnik, 1995). The SVM algorithm does not require knowledge of the combined distribution function (Soman et al., 2009). It aims to find the optimal hyperplane to separate classes from each other (Ayhan & Erdoğmuş, 2014; Doğan et al., 2023). SVMs use neural networks and statistical methods together (Haykin, 1999; Tolun, 2008). It is used in various fields, such as identification and classification (Cristianini & Shawe-Taylor, 2000; Schölkopf et al., 1999).

SVM can also work with linear, nonlinear, and multi-class data (Kavzoğlu & Çölkesen, 2010). The equations for the optimal hyperplane are given in 3 and 4.

$$w.x_i + b \geq +1, y = +1 \tag{3}$$

$$w.x_i + b \leq +1, y = -1 \tag{4}$$

In these equations, x ∈ RN represents the N-dimensional space, y ∈ (-1,+1) class labels, b the trend value, and w the weight vector (Hearst et al., 1998; Huang et al., 2018). When the data are not linearly separated, the editing parameter and artificial variable are used to classify the data correctly (Doğan, 2024; Kavzoğlu & Çölkesen, 2010).

Gaussian process regression

The GPR method is used to identify unknown functions and to learn and optimize them effectively (Becker et al., 2018). A successful ML method is used primarily in probabilistic, non-parametric problems to produce predictions by solving nonlinear problems (Liu et al., 2019; Mehmet & Doğansoy, 2022). It can also produce successful results with little data (Ateş, 2020; Yesiloglu-Gultekin & Dogan, 2024). Different covariance functions can also be used (Heo & Zavala, 2012). Equation 5 gives the Gaussian process function f(x) (Boran, 2021; Rasmussen and Williams, 2006).

$$f(x) = GP(m(x), k(x.')) \tag{5}$$

In equation 5, m(x) is the mean function, and k(x. x') is the covariance function. The mean function and covariance function are given in Equations 6 and 7, respectively.

$$m(x) = E[f(x)] \tag{6}$$

$$k(x, x') = E[(f(x) - m(x))(f(x') - m(x'))] \tag{7}$$

Regression trees

The RT method is not parametric. RT aims to categorize members of the community into homogeneous subclasses. The basis of this approach is the fact that individuals who are similar are gathered in the same node (Yesiloglu-Gultekin & Dogan, 2024; Yücel, 2017). Nodes or branches with high error values are removed. This process is called pruning (Mendeş & Akkartal, 2009). Estimates for leaves correlate with weighted averages calculated at nodes (Maimon & Rokach, 2005).

Three separation rules apply to RT. The least squares (LS), Least Absolute Deviation (LAD), and Clark & Pregibon (CP) equations were used. The primary goal of homogeneous nodes is (Temel, 2004). The RT method divides arguments into separate rather than continuous intervals.

Ensemble of trees

ET is a step-by-step process that leverages the information from many trees. Each tree was grown using data from previous trees. Fitting to a regression tree was performed using the initial data. The model is constantly updated using the last residuals (Schiltz et al., 2018). Multiple classifiers are widely used. Many studies have demonstrated that multiple classifiers perform better than others (Breiman, 1996; Opitz & Maclin, 1999). Some classifiers may behave erratically. Ensemble classifiers may provide better results (Zhou et al., 2004). RF is an excellent example of an ensemble classifier (Breiman, 2001). It performs well in classification and regression tasks (Katuwal et al., 2018).

Ensemble methods improve prediction success in classification problems. Examples of these methods are RF and gradient boosting. DT, another successful way, are also widely used. The proposed method works faster than neural network classifiers and requires fewer parameters (Shoaran et al., 2018).

Extreme gradient boosting

The XGB method is an advanced ensemble learning method based on the gradient boosting algorithm that is optimized for both speed and performance. The model builds decision trees sequentially, where each new tree attempts to correct the errors of the previous tree. XGB integrates L1 and L2 regularization techniques to prevent overfitting and improve generalization (Chen & Guestrin, 2016). In this study, XGB was employed to evaluate whether MI could be accurately predicted using only five basic enterprise-level variables. The results demonstrate that the method is capable of effectively capturing nonlinear relationships even within relatively small datasets.

Light gradient boosting machine

LGBM is a modern gradient boosting algorithm developed by Microsoft that provides high efficiency and scalability, particularly when working with large datasets. Unlike traditional level-wise tree growth strategies, LGBM uses a leaf-wise growth approach, which allows the model to minimize loss more rapidly (Ke et al., 2017). It also supports parallel computation and is optimized for handling categorical variables. In this study, LGBM was applied to test its ability to predict MI scores with high accuracy and low computational cost.

Algorithm Design

The framework proposed in this study enables the prompt determination of enterprise maturity levels without the need for conducting a survey, resulting in significant time and cost savings.

In this study, a 58-question questionnaire was administered to enterprises, and the MI score was calculated using the data obtained from the questionnaire. The MI score was then calculated using five basic data (age, field of activity, capital size, number of employees and turnover) that can be obtained directly from enterprises without requiring a questionnaire. The MI score results obtained by these two methods are compared, and the success of the proposed method is demonstrated using performance metrics. The details of the implementation steps of this new framework are described in the following algorithm.

Step 1: Preliminary preparation

a) In the preparation phase, a questionnaire is given to gather information about the enterprises.
b) Next, a maturity model (MM) is selected, and the AHP method is used to determine the weights of its dimensions.
c) Finally, the results are used to calculate enterprises' maturity index (MI) scores.

Step 2: Machine learning model training

a) Data on enterprise age, field of activity, capital size, number of employees, and turnover are obtained from the survey and translated into numerical values for input data.
b) The data undergo standardization and normalization processes to be used in machine learning.
c) Training sessions are conducted, and models are generated using various machine learning methods.

Step 3) Machine learning test process

a) The data adapted for machine learning are used as input data. MI scores of each enterprise are used as output data. Model testing was performed using the input and output data. For this purpose, the k-fold cross-validation method was used.
b) The test results were evaluated using performance metrics.

Step 4) Predicting MI score for new business

a) Basic input data on the age of the enterprise, field of activity, capital size, number of employees, and turnover are obtained directly from the enterprise. Because enterprises already know this basic information, they do not need a survey.

b) The standardization and normalization processes were applied to the data for use in machine learning.

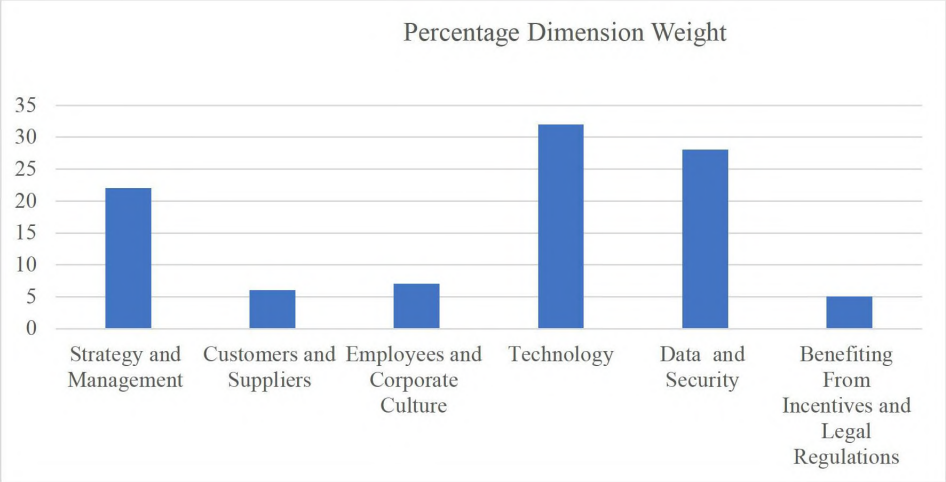
c) Using these data and models previously trained with machine learning methods, the MI score of the enterprise is estimated.

Implementation of the Proposed Method

The SANOL maturity model (SMM) was chosen in this study to test the proposed methodology on a sample SMM. The model measures maturity in six dimensions for Industry 4.0. These six dimensions and percentage weights are shown in Figure 1.

Figure 1

Percentage weight of dimensions



In the SMM, a questionnaire consisting of 58 questions was used as the evaluation method. The weights of the dimensions in the model are determined by the AHP method.

Since the study was initiated during the final stages of the COVID-19 pandemic, and considering the relative ease of both administration and data processing, the survey was conducted online. For this purpose, a dedicated digital form was developed to facilitate the online implementation of the questionnaire. The English version is accessible at https://form.jotform.com/251425426420953. However, no data were obtained from many businesses contacted, and the responses provided were incomplete. As a result, the study took longer than originally planned and was ultimately completed using questionnaires completed by 61 businesses.

The input and output data were prepared for use in machine learning studies. In this study, the machine learning models were trained using five easily accessible enterprise-level variables: company age, industry type, ownership structure, number of employees, and annual turnover. These variables were selected due to their wide availability, simplicity of collection, and conceptual relationship with organizational maturity. Table 2 provides a detailed definition of each variable along with its corresponding categorical groupings, which were used for model training and testing.

Table 2
Description and categorical breakdown of the input variables used in the model

Variable	Definition	Categories
Company Age	Number of years since the company officially began operations	1: 0–9 years 2: 10–19 years 3: 20 years or more
Industry Type	Primary economic sector in which the company operates	1: Electronics 2: Manufacturing 3: Services 4: Logistics 5: Machinery 6: Other
Ownership Structure	Ownership and partnership structure of the company	1: Domestic capital 2: Foreign capital 3: Joint domestic–foreign ownership
Number of Employees	Total number of personnel employed by the company	1: 1–9 (micro) 2: 10–50 (small) 3: 51–250 (medium) 4: 251–1000 5: More than 1000
Annual Turnover	Total annual revenue generated by the company	1: <1M (in the Turkish Lira, TRY) 2: 1–10 M 3: 10–50 M 4: 50–100 M 5: 100–250 M 6: 250–500 M 7: >500M

These variables served as the sole inputs to the machine learning models, replacing conventional survey-based indicators and enabling streamlined estimation of the Maturity Index (MI).

The output data is the MI value, which is calculated using SMM with the survey data. The MI score was obtained from the inputs of the six dimensions shown in Figure 1. Details are available in related publications (Ünal et al., 2022).

Of the 61 enterprises surveyed, 13 were electronics, 13 were manufacturing, 10 were service, nine were logistics, eight were machinery, and eight other enterprises. 26 enterprises were interviewed but could not be surveyed. Among the enterprises surveyed, 17 have more than 250 employees. When the annual turnover of enterprises was analyzed, 33% of enterprises had a turnover of less than 100 million TRY, 46% had a turnover between 100 million and 500 million TRY, and the rest had a turnover of more than 500 million TRY. Of the survey questions administered using the SMM, 12 are related to technology, 11 to strategy and management, ten to supports and incentives, seven to data and security, five to employees and corporate culture, three to customers and suppliers, and the remaining ten are concerned with general catalog information.

MI scores were calculated using enterprises' survey data. The MI score was calculated by adding the weighted dimension scores (Equation 8).

$$E = \Sigma_{i=0}^{n} Bi^{*}gi \tag{8}$$

E: MI score (between 1 and 5).

B: Dimension maturity score (between 1 and 5).

g: Overall dimension weight.

Older enterprises have higher MI scores than newer ones. Considering the establishment dates of the enterprises, approximately 33% are less than 10 years old, 23% are between 10 and 19 years old, and the rest are 20 years old or older.

Seven machine learning methods were used in the studies. Table 3 provides the technical details and parameters of each of these methods.

Table 3

Machine learning methods, subtypes and their parameters

ML Methods	ML Sub Methods	Parameters and technical details
LR	Linear (Selected)	Standard
	Interactions Linear	Standard
	Robust Linear	Standard
	Stepwise Linear	Maximum number of steps:1000
SVM	Linear SVM (Selected)	Manual box constarint:0.769 Manual epsilon: 0.077 Manual kernel scale: 1
	Quadratic SVM	Manual kernel scale: 1
	Cubic SVM	Manual kernel scale: 1
	Fine Gaussian SVM	Manual kernel scale: 0.56
	Medium Gaussian SVM	Manual kernel scale: 2.2
	Coarse Gaussian SVM	Manual kernel scale: 8.9
GPR	Rational Quadratic GPR, Squared Exponential GPR (Selected), Matern 5/2 GPR, Exponential GPR	Kernel scale:1.1690659
	Signal standard deviation: 0.4553253	Sigma: 0.4553253
RT	Fine Tree	Maximum surrogates per node:10 Minimum leaf size:4
	Medium Tree (Selected)	Minimum leaf size: 12
	Coarse Tree	Minimum leaf size:36
ET	Boosted Trees (Selected)	Minimum leaf size:4 Number of learners:30 Learning rate:0.1
	Bagged Trees	
XGB	Fine	n_estimators: 700 learning_rate: 0.02

ML Methods	ML Sub Methods	Parameters and technical details
		max_depth: 4
		subsample: 1.0
		colsample_bytree: 1.0
		gamma: 0.05
		min_child_weight:
		reg_alpha: 0.5
		reg_lambda: 1.5
LGBM	Fine	learning_rate=0.01
		n_estimators=300
		num_leaves=60
		max_depth=8
		subsample=0.8
		feature_fraction=0.9
		min_child_samples=15
		reg_alpha=0.0
		reg_lambda=0.0
		lambda_l1=0.0
		lambda_l2=0.05

Research Results and Discussion

Summary of Input Data and Model Objective

Within this study's scope, a survey was conducted to evaluate the industrial Maturity Index (MI). There are 58 questions in the survey. Because operating a study is very difficult and time-consuming, all the desired answers were obtained from 61 enterprises, and the industrial MI general score was calculated using SMM. In this study, the success of machine learning (ML) was investigated to make the industrial MI assessments of many enterprises located in industrial zones faster, cheaper, and more accessible. The enterprise age, field of activity, capital size, number of employees, and turnover information were used as inputs to the ML models, and the overall score was estimated as the output.

Evaluation Metrics

The metrics are the feedback of the designed model. The success of the model created according to these metrics was measured, and improvements are made when necessary (Özen et al., 2021).

The coefficient of determination, known as R^2, is a key metric used to evaluate the accuracy of regression models. The value ranges from 0 to 1 and indicates how well the predicted values match the actual data. A value closer to 1 implies a better fit.

The Root Mean Square Error (RMSE) measures the standard deviation of the prediction errors and reflects the magnitude of the model's errors in the same units as the output. The Mean Squared Error (MSE) was calculated as the average of the squared differences between actual and predicted values, emphasizing larger errors due to squaring. The Mean Absolute Error (MAE) computes the average of the absolute differences and provides a more intuitive interpretation of the model's average error per prediction. The coefficient of

determination (R^2) shows the performance of the predicted values in the model. R^2 determines how well the data fit a curve. The RMSE returns the standard deviation of the actual and predicted values and measures the magnitude of a model's error. The MSE provides the arithmetic mean of the squared difference between the actual and predicted values. The MAE reflects the average of the absolute distance between the actual and predicted values.

Model Performance and Algorithm Comparison

The predictive performance of all the machine learning models was assessed by comparing their outputs with the results obtained from the field survey. To validate the models on a limited dataset, a 5-fold cross-validation approach was used. As summarized in Table 4, the models demonstrated strong performance with R^2 values generally greater than 0.70. The RMSE and MAE values remained within acceptable ranges, supporting the consistency and stability of the predictions.

Among the seven algorithms, SVM achieved the highest R^2 value of 0.75, followed closely by GPR (0.74), ET (0.73), LGBM (0.72), and XGB (0.71). These results indicate that both traditional and gradient boosting-based models can provide reliable estimations and highlight the robustness and flexibility of the proposed framework across diverse algorithmic strategies.

Table 4 and Figures 2–8 present the detailed performance metrics and visual comparisons of measured versus predicted scores. The performance evaluation revealed that all seven machine learning algorithms demonstrated notable predictive capabilities in estimating the Maturity Index (MI). Among them, SVM achieved the highest R^2 value of 0.75, followed closely by GPR (0.74), ET (0.73), LGBM (0.72), and XGB (0.71). These results indicate that both traditional and gradient boosting-based models can provide reliable estimations and highlight the robustness and flexibility of the proposed framework across diverse algorithmic strategies.

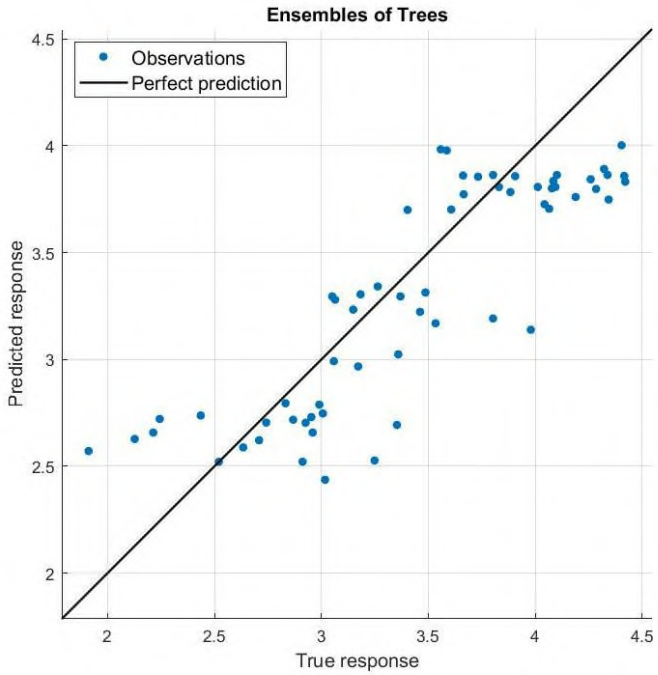
Table 4

Metrics of the test predictions

	LR	SVM	GPR	RT	ET	XGB	LGBM
R^2	0.74	0.75	0.74	0.69	0.70	0.71	0.72
RMSE	0.33	0.33	0.33	0.36	0.36	0.35	0.34
MSE	0.11	0.11	0.11	0.13	0.13	0.12	0.11
MAE	0.28	0.26	0.28	0.28	0.30	0.28	0.28

Figure 2

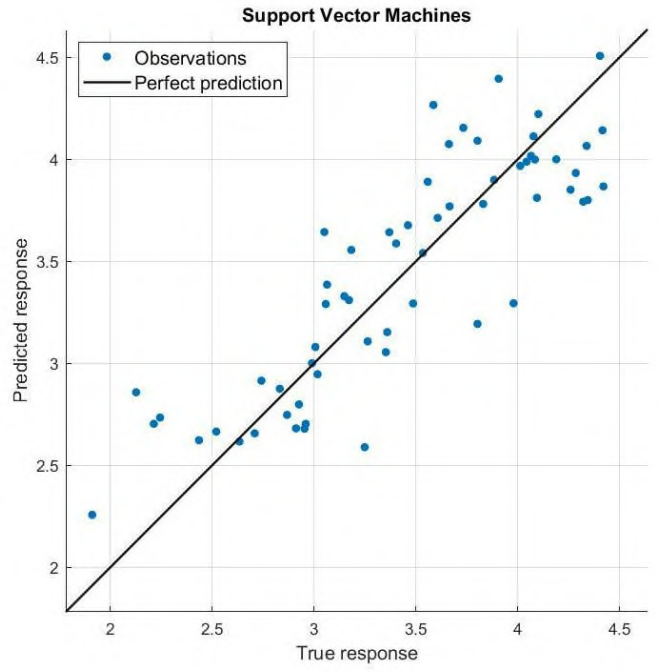
Index scores measured and predicted using the ET method.



The resulting graph of the tests obtained from the SVM studies is shown in Figure 3.

Figure 3

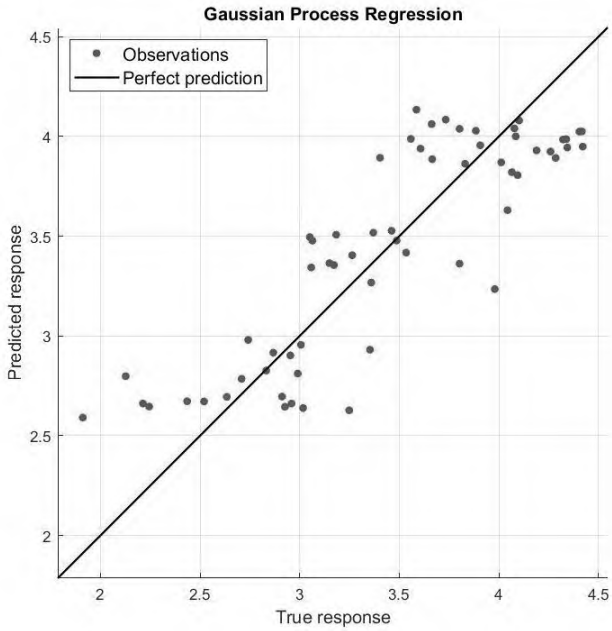
Index scores measured and predicted by the SVM method



The resulting graph of the tests obtained from the studies performed using the GPR is shown in Figure 4.

Figure 4

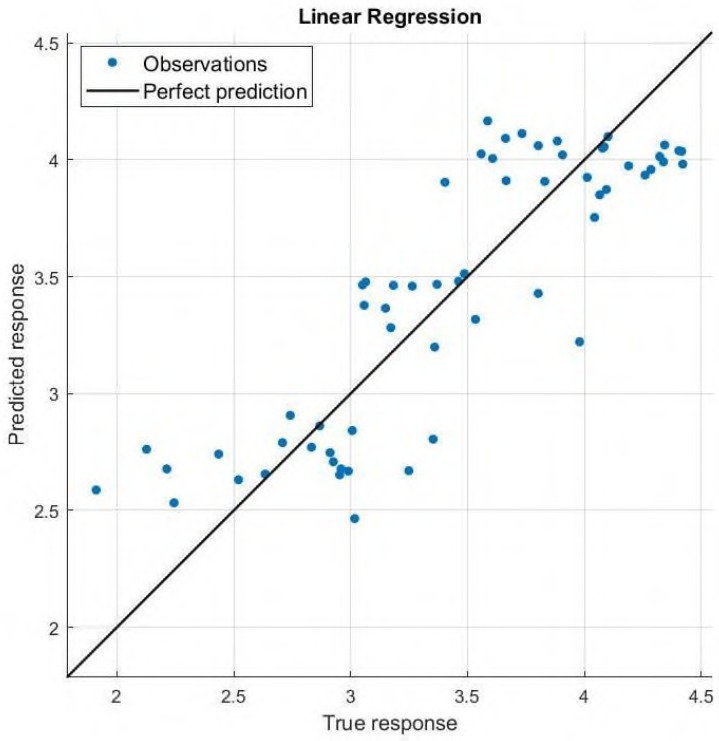
Index scores measured and predicted using the GPR method



The resulting graph of the tests obtained from the studies using LR is shown in Figure 5.

Figure 5

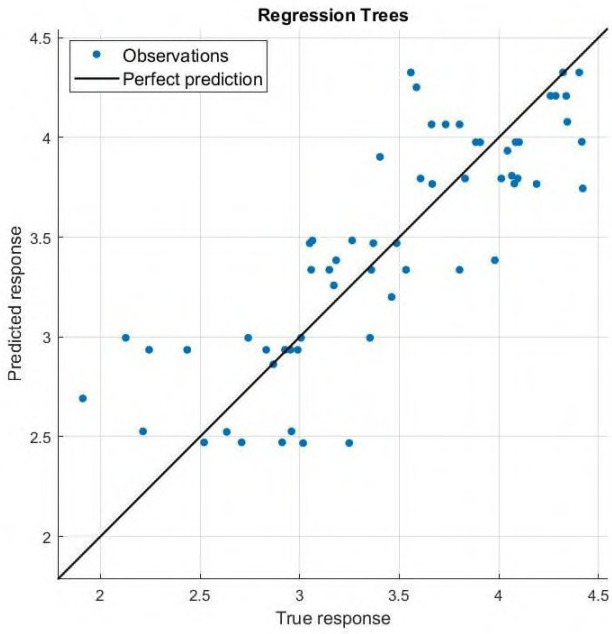
Index scores measured and predicted using the LR method



The resulting graph of the tests obtained from the studies performed with RT is shown in Figure 6.

Figure 6

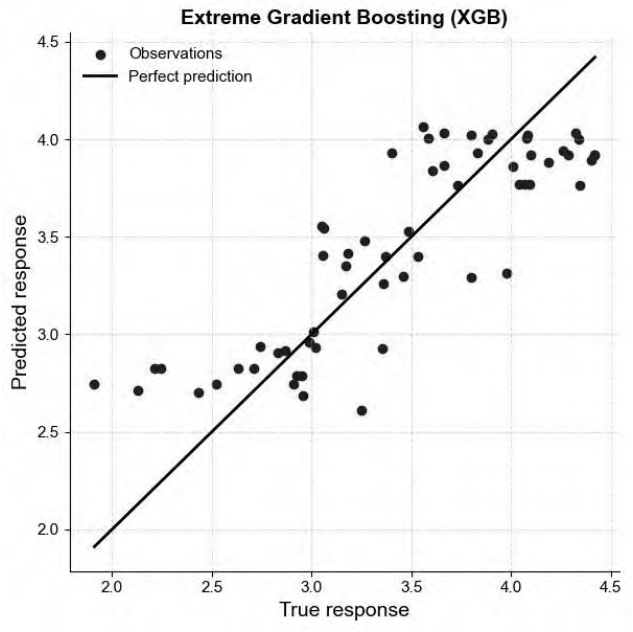
Index scores measured and predicted using the RT method



The resulting graph of the tests obtained from the studies performed with XGB is shown in Figure 7.

Figure 7

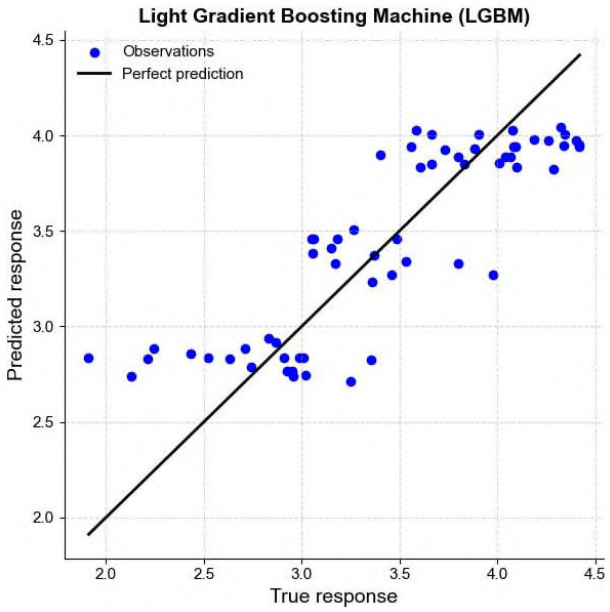
Index scores measured and predicted using the XGB method



The resulting graph of the tests obtained from the studies performed using LGBM is shown in Figure 8.

Figure 8

Index scores measured and predicted using the LGBM method



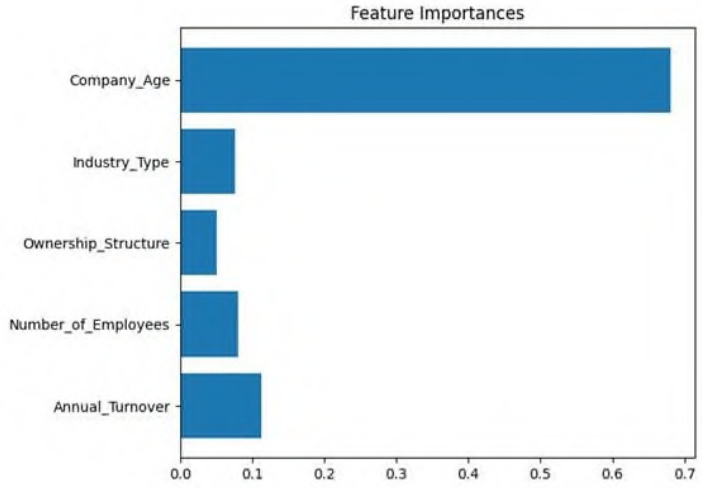
Feature Importance Analysis

To gain further insight into the internal dynamics of the prediction process, a feature importance analysis was conducted to evaluate the relative contributions of each input variable to the prediction of the Maturity Index (MI). As shown in Figure 9, company age is the most influential variable, followed by Annual Turnover and Number of Employees. These variables, associated with enterprise scale and longevity, have a stronger impact on MI predictions. In contrast, Industry Type and Ownership Structure exhibited comparatively lower levels of importance.

This analysis improves the interpretability of the model and supports practical decision-making by identifying the enterprise characteristics that are most influential in determining digital maturity.

Figure 9

Feature importance of input variables in predicting MI



Real-World Implementation and Future Research Directions:

Although the results presented in this study demonstrate the effectiveness of the proposed machine learning framework under controlled conditions, its integration into live industrial environments remains a critical area for future research. Real-time implementation will require not only technical adjustments, such as the automation of data flows and feedback loops, but also organizational readiness and stakeholder engagement. Collaborations with manufacturing enterprises are planned to deploy the model in operational settings and monitor its predictive accuracy and adaptability in dynamic real-world contexts. Such efforts are expected to validate the framework's practical value and guide necessary refinements for broader applicability. These real-world applications are also envisioned as key components of future research activities aimed at scaling up the model and verifying its impact in diverse industrial domains.

Conclusions

It is believed that Industry 4.0 will increase the success and quality of life of production. For this reason, the adaptation of enterprises to Industry 4.0 has gained importance. One measure of compliance is the MI score. Surveys are commonly used to determine MI scores; however, these procedures can be laborious and time-consuming.

This study presents an enhanced framework that integrates seven machine learning methods, including both traditional algorithms (e.g., SVM, GPR, LR, RT) and advanced boosting-based models (XGB, LGBM), to predict MI based on only five easily obtainable enterprise variables. The study demonstrates that the MI scores of enterprises can be predicted with high accuracy using machine learning, without the need for survey-based assessments.

The proposed framework was validated using a five-fold cross-validation technique, and performance metrics (R^2, RMSE, MSE, MAE) revealed that all models achieved strong predictive accuracy, with SVM yielding the highest R^2 value (0.75). Feature importance analysis further clarified the internal mechanics of the model, revealing that company age, turnover, and number of employees were the most influential factors in MI prediction. Thus, companies can reduce costs and accelerate their transition to Industry 4.0 using data-driven estimation tools. Although model performance may vary according to algorithm type, dataset size, and domain specificity, the proposed approach is scalable and practical solution.

In future research, this framework will be integrated into real-time industrial environments to evaluate its dynamic performance. Collaborations with manufacturing enterprises are envisaged to deploy the model in operational settings and monitor its predictive stability and utility across diverse industrial conditions.

Enterprises can rapidly determine Industry 4.0 readiness using the framework proposed in this study.

Peer Review: Externally peer-reviewed.
Author Contributions: Conception/Design of Study- A.D., C.Ü.; Data Acquisition- A.D., C.Ü.; Data Analysis/Interpretation- A.D., C.Ü.; Drafting Manuscript- A.D., C.Ü.; Critical Revision of Manuscript- A.D., C.Ü.; Final Approval and Accountability- A.D., C.Ü.; Technical or Material Support- A.D., C.Ü.; Supervision- A.D., C.Ü.
Conflict of Interest: The authors have no conflict of interest to declare.
Grant Support: The authors declared that this study has received no financial support.

Author Details

Ayhan Doğan

[1] Hacettepe University, Başkent Organized Industrial Zone, Technical Sciences Vocational School, Ankara, Türkiye

0000-0002-9872-8889 aydogan@hacettepe.edu.tr

Cihan Ünal

[1] Hacettepe University, Başkent Organized Industrial Zone, Technical Sciences Vocational School, Ankara, Türkiye

0000-0002-5255-4078

References

Akdil, K. Y., Ustundag, A., & Cevikcan, E. (2018). *Maturity and readiness model for industry 4.0 strategy. Industry 40: Managing the digital transformation*. Springer. pp 61-94

Angelopoulos, A., Michailidis, E. T., Nomikos, N., Trakadas, P., Hatziefremidis, A., Voliotis, S., & Zahariadis, T. (2019). Tackling faults in the industry 4.0 era—a survey of machine-learning solutions and key aspects. *Sensors*, 20, 109.

Aslan, Y., Yavasca, S., & Yasar, C. (2011). Long-term electric peak load forecasting of Kutahya using different approaches. *International Journal on Technical and Physical Problems of Engineering*, 3, 87-91.

Ateş, B. (2020). *Gemi yapılarında gerilme yığılması öngörülerinin kaba ağ yapısı ve makine öğrenmesi ile gerçekleştirilmesi*. Fen Bilimleri Enstitüsü

Ayhan, S., & Erdoğmuş, Ş. (2014). Destek vektör makineleriyle sınıflandırma problemlerinin çözümü için çekirdek fonksiyonu seçimi. *Eskişehir Osmangazi Üniversitesi İktisadi ve İdari Bilimler Dergisi*, 9, 175-201.

Basl, J., & Doucek, P. (2019). A metamodel for evaluating enterprise readiness in the context of Industry 4.0. *Information,* 10, 89.

Bayazit, M., & Oğuz, B. (1994). *Mühendisler için istatistik*. Birsen Yayınevi, İstanbul, 197s.

Becker, B. G., Klein, T., Wachinger, C., & Initiative AsDN (2018). Gaussian process uncertainty in age estimation as a measure of brain abnormality. *NeuroImage*, 175, 246-258.

Bibby, L., & Dehe, B. (2018). Defining and assessing industry 4.0 maturity levels–case of the defense sector. *Production Planning & Control*, 29, 1030-1043.

Boran, P. D. S. (2021). *Makine öğrenmesi yöntemleri kullanılarak kısa dönem rüzgar gücü tahmini*. Sakarya Üniversitesi

Botha, A. P. (2018). Rapidly arriving futures: future readiness for Industry 4.0. *South African Journal of Industrial Engineering*, 29, 148-160.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24, 123-140.

Breiman, L. (2001). Using iterated bagging to debian regressions. *Machine Learning*, 45, 261-277.

Brik, B., Bettayeb, B., Sahnoun, M. H., & Duval, F. (2019). Toward predicting system disruption in industry 4.0: Machine learning-based approach. *Procedia Computer Science*, 151, 667-674.

Candanedo, I. S., Nieves, E. H., González, S. R., Martín, M. T. S., & Briones, A. G. (2018). *Machine learning predictive model for industry 4.0*. Knowledge Management in Organizations: 13th International Conference, KMO 2018, Žilina, Slovakia, August 6–10, 2018, Proceedings 13. Springer. pp 501-510

Canetta, L., Barni, A., & Montini, E. (2018). *Development of a digitalization maturity model for the manufacturing sector*. 2018 ieee international conference on engineering, technology and innovation (ICE/ITMC). IEEE. pp 1-7

Carvalho, N., Chaim, O., Cazarini, E., & Gerolamo, M. (2018). Manufacturing in the fourth industrial revolution: A positive prospect in sustainable manufacturing. *Procedia Manufacturing*, 21, 671-678.

Chen, T., & Guestrin, C. (2016). *XGBoost: A scalable tree boosting system*. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. https://doi.org/10.1145/2939672.2939785

Chonsawat, N., & Sopadang, A. (2020). Defining SMEs' 4.0 readiness indicators. *Applied Sciences*, 10, 8998.

Cortes, C., & Vapnik, V. (1995). Support vector networks. *Machine Learning*, 20, 273-297.

Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge University Press,

Çınar, Z. M., Zeeshan, Q., & Korhan, O. (2021). A framework for industry 4.0 readiness and maturity of smart manufacturing enterprises: a case study. *Sustainability*, 13, 6659.

Da Silva, V. L., Kovaleski, J. L., Pagani, R. N., Silva, J. D. M., & Corsi, A. (2020). Implementation of Industry 4.0 concept in companies: Empirical evidence. *International Journal of Computer Integrated Manufacturing*, 33, 325-342.

Doğan, A. (2024). Prediction of major earthquakes that may affect İzmir using machine learning methods. *Yerbilimleri*, 45, 93-106. doi:10.17824/yerbilimleri.1402618

Doğan, A., Korkmaz, M., & Kirmaci, V. (2023). Estimation of Ranque-Hilsch vortex tube performance using machine learning techniques. *International Journal of Refrigeration*, 150, 77-88. https://doi.org/10.1016/j.ijrefrig.2023.01.021

Erol, S., Schumacher, A., & Sihn, W. (2016). *Strategic guidance toward Industry 4.0–a three-stage process model*. International conference on competitive manufacturing. pp 495-501

Finance, A. (2015). *Industry 4.0 Challenges and solutions for the digital transformation and use of exponential technologies*. Finance, audit tax consulting corporate: Zurich, Swiss:1-12.

Geissbauer, R., Vedso, J., & Schrauf, S. (2016). *Industry 4.0: Building the digital enterprise*.

Gerlitz, L. (2016). Design management as a domain of smart and sustainable enterprise: Business modeling for innovation and smart growth in Industry 4.0. *Entrepreneurship and Sustainability*, 3, 244.

Gökalp, E., Şener, U., & Eren, P. E. (2017). *Development of an assessment model for industry 4.0: industry 4.0-MM*. International Conference on Software Process Improvement and Capability Determination. Springer. pp 128-142

Haber, R. E., Juanes, C., del Toro, R., & Beruvides, G. (2015). Artificial cognitive control with self-x capabilities: A case study of a micro-manufacturing process. *Computers in Industry*, 74, 135-150.

Hamidi, S. R., Aziz, A. A., Shuhidan, S. M., Aziz, A. A., & Mokhsin, M. (2018). *SMEs maturity model assessment of IR4. 0 digital transformation*. International Conference on Kansei Engineering & Emotion Research. Springer. pp 721-732

Haykin, S. (1999). *Neural networks: A guided tour*. Soft computing and intelligent systems: theory and applications 71.

Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., & Scholkopf, B. (1998). Support vector machines. *IEEE Intelligent Systems and Their Applications*, 13, 18-28.

Heo, Y., & Zavala, V. M. (2012). Gaussian process modeling for measurement and verification of building energy savings. *Energy and Buildings*, 53, 7-18.

Hizam-Hanafiah, M., Soomro, M. A., & Abdullah, N. L. (2020). Industry 4.0 readiness models: a systematic literature review of model dimensions. *Information*, 11, 364.

Horvat, D., Stahlecker, T., Zenker, A., Lerch, C., & Mladineo, M. (2018). A conceptual approach to analyzing manufacturing companies' profiles concerning Industry 4.0 in emerging economies. *Procedia Manufacturing*, 17, 419-426.

Horváth, D., & Szabó, R. Z. (2019). Driving forces and barriers of Industry 4.0: Do multinational and small- and medium-sized companies have equal opportunities? *Technological Forecasting and Social Change*, 146, 119-132.

Huang, S., Cai, N., Pacheco, P. P., Narrandes, S., Wang, Y., & Xu, W. (2018). Applications of support vector machine (SVM) learning in cancer genomics. *Cancer Genomics & Proteomics*, 15, 41-51.

Ivanov, D., Dolgui, A., & Sokolov, B. (2019). *The impact of digital technology and Industry 4.0 on the ripple effect and supply chain risk analytics*. International Journal of Production Research, 57, 829-846.

Jalil, N. A., Hwang, H. J., & Dawi, N. M. (2019). *Machines learning trends, perspectives and prospects in education sector*. Proceedings of the 2019 3rd International Conference on Education and Multimedia Technology, 201-205.

Jesus, C. D., & Lima, R. M. (2020). Literature search of key factors for developing generic and specific maturity models for Industry 4.0. *Applied Sciences*, 10, 5825.

Kalayci, T. E. (2018) Kimlik hırsızı web sitelerinin sınıflandırılması için makine öğrenmesi yöntemlerinin karşılaştırılması. *Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi*, 24, 870-878.

Kannan, S. M., Suri, K., Cadavid, J., Barosan, I., Van Den Brand, M., Alferez, M., & Gerard, S. (2017). *Toward industry 4.0: Gap analysis between current automotive MES and industry standards using model-based requirement engineering*. 2017 IEEE International Conference on Software Architecture Workshops (ICSAW). IEEE. pp 29-35

Katuwal, G. J., Baum, S. A., & Michael, A. M. (2018). Early brain imaging can predict autism: Application of machine learning to a clinical imaging archive. *BioRxiv*:471169.

Kavzoğlu, T., & Çölkesen, İ. (2010). Destek vektör makineleri ile uydu görüntülerinin sınıflandırılmasında kernel fonksiyonlarının etkilerinin incelenmesi. *Harita Dergisi*, 144, 73-82.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T. Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 30.

Kuo, C. J., Ting, K. C., Chen, Y. C., Yang, D. L., & Chen, H. M. (2017). Automatic machine status prediction in the era of Industry 4.0: Case study of machines in a spring factory. *Journal of Systems Architecture*, 81, 44-53.

Lee, C., & Lim, C. (2021). From technological development to social advance: A review of Industry 4.0 through machine learning. *Technological Forecasting and Social Change*, 167, 120653.

Leyh, C., Bley, K., Schäffer, T., & Forstenhäusler, S. (2016). *SIMMI 4.0-a maturity model for classifying the enterprise-wide it and software landscape focusing on Industry 4.0*. 2016 federated conference on computer science and information systems (fedcsis). IEEE. pp 1297-1302

Liu, K., Hu, X., Wei, Z., Li, Y., & Jiang, Y. (2019). *Modified Gaussian process regression models for cyclic capacity prediction of lithium-ion batterie*s. IEEE Transactions on Transportation Electrification, 5, 1225-1236.

Luu, Q. H., Lau, M. F., Ng, S. P., & Chen, T. Y. (2021). Testing multiple linear regression systems with metamorphic testing. *Journal of Systems and Software*, 182, 111062. https://doi.org/10.1016/j.jss.2021.111062

Maimon, O., & Rokach, L. (2005). *Data mining and knowledge discovery handbook*.

Manavalan, E., & Jayakrishna, K. (2019). A review of Internet of Things (IoT) embedded sustainable supply chain for industry 4.0 requirements. *Computers and Industrial Engineering*, 127, 925-953.

Mehmet, A., & Doğansoy, G. A. (2022). Makine öğrenmesi ve derin öğrenme yöntemleri kullanılarak e-perakende sektörüne yönelik talep tahmini. *Gazi Üniversitesi Mühendislik Mimarlık Fakültesi Dergisi*, 37, 1325-1340.

Memnun, D., & Kalaycı, Ş. (2006). *SPSS Uygulamalı Çok Değişkenli İstatistik Teknikleri*. Ankara: Asil Yayın Dağıtım. İlköğretim Online 12.

Mendeş, M., & Akkartal, E. (2009). Regression tree analysis for predicting slaughter weight in broilers. *Italian Journal of Animal Science*, 8, 615-624.

Methavitakul, B., & Santiteerakul, S. (2018). *Analysis of key dimension and sub-dimension for Supply Chian of Industry to fourth Industry Performance Measurement*. 2018 IEEE International Conference on Service Operations and Logistics, and Informatics (SOLI). IEEE. pp 191-195

Mittal, S., Khan, M. A., Romero, D., & Wuest, T. (2018). A critical review of smart manufacturing & Industry 4.0 maturity models: Implications for small- and medium-sized enterprises (SMEs). *Journal of Manufacturing Systems*, 49, 194-214.

Muñoz, L. P., Hagemann, M., Palandella, L., & Cagliari, P. (2023). *Maturity SCAN 4.0: An analysis and discussion of results from scanning of digital capabilities adopted by Argentinian CPG companies*. ESSN: 2701-6277, 854-863.

Müller, J. M., Voigt, K. I. (2018). The impact of industry 4.0 on supply chains in engineer-to-order industries-an exploratory case study. *IFAC-PapersOnLine*, 51, 122-127.

Opitz, D., & Maclin, R. (1999). Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research*, 11, 169-198.

Özen, N. S., Saraç, S., & Koyuncu, M. (2021). COVID-19 Vakalarının Makine Öğrenmesi Algoritmaları ile Tahmini: Amerika Birleşik Devletleri Örneği. *Avrupa Bilim ve Teknoloji Dergisi*, 134-139.

Pessl, E., Sorko, S. R., & Mayer, B. (2017). Roadmap Industry 4.0–implementation guideline for enterprises. *International Journal of Science, Technology and Society*, 5, 193-202.

Rahamaddulla, S. R. B., Leman, Z., Baharudin, B. H. T. B., & Ahmad, S. A. (2021). Conceptualizing smart manufacturing readiness-maturity model for small and medium enterprise (sme) in malaysia. *Sustainability*, 13, 9793.

Rai, R., Tiwari, M. K., Ivanov, D., & Dolgui, A. (2021). *Machine learning in manufacturing and industry 4.0 applications*. Taylor and Francis. pp 4773-4778.

Rais, R. B. (2021). *Sharing the Pie The Fourth Industrial Revolution and the Sharing/Platform Economy: Distributive justice implications*. Beyond Free Market. Routledge. pp 125-140

Rasmussen, C. E., & Williams, C. (2006). *Gaussian processes for machine learning, vol. 1*. MIT press Cambridge MA.

Rauch, E., Dallasega, P., & Unterhofer, M. (2019). Requirements and barriers for introducing smart manufacturing in small- and medium-sized enterprises. *IEEE Engineering Management Review*, 47, 87-94.

Rong, S., & Bao-Wen, Z. (2018). The research of regression model in machine learning field. MATEC Web of Conferences. *EDP Sciences*. 01033, 2019.

Saad, S. M., Bahadori, R., & Jafarnejad, H. (2021). The smart SME technology readiness assessment methodology in the context of industry 4.0. *Journal of Manufacturing Technology Management*.

Sajjad, A., Ahmad, W., Hussain, S., Chuddher, B. A., Sajid, M., Jahanjaib, M., Ali, M. K., & Jawad, M. (2024). Assessment by Lean Modified Manufacturing Maturity Model for Industry 4.0: A Case Study of Pakistan's Manufacturing Sector. *IEEE Transactions on Engineering Management*, 71, 6420-6434. https://doi.org/10.1109/TEM.2023.3259005

Schaupp, E., Abele, E., & Metternich, J. (2017). Potentials of digitalization in tool management. *Procedia CIRP*, 63, 144-149.

Schiltz, F., Masci, C., Agasisti, T., & Horn, D. (2018). Using regression tree ensembles to model interaction effects: a graphical approach. *Applied Economics*, 50, 6341-6354.

Schölkopf, B., Williamson, R. C., Smola, A., Shawe-Taylor, J., & Platt, J. (1999). *Support vector method for novelty detection*. Advances in neural information processing systems 12.

Schumacher, A., Erol, S., & Sihn, W. (2016). A maturity model for assessing Industry 4.0 readiness and maturity of manufacturing enterprises. *Procedia Cirp*, 52, 161-166.

Shoaran, M., Haghi, B. A., Taghavi, M., Farivar, M., & Emami-Neyestanak, A. (2018). Energy-efficient classification for resource-constrained biomedical applications. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 8, 693-707.

Skalli, D., Charkaoui, A., & Cherrafi, A. (2023). *Design of Industry 4.0 Maturity Model and Readiness Assessment Under the Digital Transformation Toward the Implementation of LSS4.0*. In T. Masrour, I. El Hassani, and N. Barka, Artificial Intelligence and Industrial Applications Cham.

Soman, K., Loganathan, R., & Ajay, V. (2009). *Machine learning with SVM and other kernel methods*. PHI Learning Pvt. Ltd., New Delhi.

Stentoft, J., Adsbøll Wickstrøm, K., Philipsen, K., & Haug, A. (2021). Drivers and barriers for Industry 4.0 readiness and practice: empirical evidence from small and medium-sized manufacturers. *Production Planning & Control*, 32, 811-828.

Temel, G. O. (2004). *Sınıflama ve Regresyon Ağaçları*. Yayımlanmamış Yüksek Lisans Tezi, Mersin Üniversitesi, Sağlık Bilimleri Enstitüsü, Bioistatistik Anabilim Dalı, Mersin 11:23-24.

Tolun, S. (2008). *Destek vektör makineleri: Banka başarısızlığının tahmini üzerine bir uygulama*. İktisadî Araştırmalar Vakfı

Ünal, C., Sungur, C., & Yildirim, H. (2022). Application of the maturity model in industrial corporations. *Sustainability*, 14, 9478.

Valentin, B. A. (2017). *Methods Of Assessment And Training Of A Company Toward The Enterprise 4.0*. Annals of DAAAM & Proceedings 28.

Városiné Demeter, K., Losonci, D., Szász, L., & Rácz, B. G. (2018). *Assessing Industry 4.0 readiness: a multi-country industry level analysis*.

Vazire, S. (2018). Implications of the credibility revolution for productivity, creativity, and progress. *Perspectives on Psychological Science*, 13, 411-417.

Wank, A., Adolph, S., Anokhin, O., Arndt, A., Anderl, R., & Metternich, J. (2016). Using a learning factory approach to transfer Industrie 4.0 approaches to small-and medium-sized enterprises. *Procedia Cirp*, 54, 89-94.

Wei, J., Chu, X., Sun, X. Y., Xu, K., Deng, H. X., Chen, J., Wei, Z., & Lei, M. (2019). Machine learning in materials science. *InfoMat*, 1, 338-358.

Yesiloglu-Gultekin, N., & Dogan, A. (2024). Estimation of the elastic modulus of basaltic rocks using machine learning methods. *Earth Science Informatics*, doi:10.1007/s12145-024-01472-7

Yücel, Y. B. (2017). *Yaşam memnuniyetini etkileyen faktörlerin sınıflama ve regresyon ağacı ile belirlenmesi*. Fen Bilimleri Enstitüsü.

Zhou, G., Zhang, J., Su, J., Shen, D., & Tan, C. (2004). Recognizing names in biomedical texts: a machine learning approach. *Bioinformatics*, 20, 1178-1190.

İSTANBUL UNIVERSITY PRESS

Acta Infologica,
ACIN 2025, 9 (1): 314–338

https://doi.org/10.26650/acin.1515409
Submitted: 12.07.2024
Revision Requested: 22.02.2025
Last Revision Received: 14.04.2025
Accepted: 09.05.2025
Published Online 25.06.2025

Acta Infologica

Research Article

Open Access

Factors Influencing Consumer Adoption of Smart Home Systems: A Socio-Demographic Perspective

Onur Muço [1] & Üstün Özen [1]

[1] Atatürk Üniversitesi, Yönetim Bilişim Sistemleri, Erzurum, Türkiye

Abstract The purpose of this research is to determine the factors influencing consumers' adoption of smart home systems. Within this context, the relationships between factors such as perceived benefit, perceived risk, functionality, and price and their effects on the adoption of smart home systems were investigated. Additionally, whether there were significant differences in the mentioned factors based on gender, age group, education level, marital status, and income level was investigated. The population of the study consists of individuals living in Turkey who use smart home systems, while the sample comprises 556 individuals selected from this population. The data for the study were collected through a survey and evaluated using quantitative analysis methods. The statistical analyses were conducted using the SPSS 25 software, applying descriptive statistical methods, Pearson correlation analysis, independent samples t-test, and one-way variance analysis. The analyses reveal that consumers' overall levels of adoption of smart home systems are slightly above average. However, it is observed that the first three factors significantly influencing the adoption of smart home systems are price, perceived benefit, and functionality, respectively. Furthermore, strong and very strong positive and significant relationships were found among all factors related to the adoption of smart homes. Furthermore, it is concluded that the approach to smart homes varies significantly according to gender, age group, education level, marital status, and income level.

Keywords Smart home · planned behavior · technology adoption · consumer · Technology Acceptance Model

Citation: Muço, O. & Özen, Ü. (2025). Factors influencing consumer adoption of smart home systems: A socio-demographic perspective. *Acta Infologica*, *9*(1), 314-338. https://doi.org/10.26650/acin.1515409

This work is licensed under Creative Commons Attribution-NonCommercial 4.0 International License.

© 2025. Muço, O. & Özen, Ü.

Corresponding author: Üstün Özen uozen@atauni.edu.tr

Introduction

The integration of smart home systems represents a prevalent trend as part of the Internet of Things (IoT), signaling larger shifts in the types of products and services consumers use. Smart homes are conceived as fully connected passages of devices, providing convenience and security, but are largely receiving attention for energy efficiency and quality of living. Adoption of IoT devices for smart homes is variable across populational demographics, with a significant influence attributable to socio-demographic factors. These socio-demographic factors inform the technology adoption decision-making process, creating value for stakeholders who consider research on socio-demographic differences in home smart technologies as applicable.

One of the primary considerations for home consumers adopting technologies is the perceived value of smart home systems. Existing research emphasizes that consumers are more inclined to purchase systems when they understand the benefits of smart home systems, including the benefits of and the convenience from systems adopted for energy use, safety, and home comfort (Sequeiros et al., 2021). If examined according to futures or aging populations, the conveniences provided by smart home technologies to older adults are highly adoptable (Lau et al., 2022). As the population ages, more technologies will be created facilitations independent living purposes, safety, and mitigating risk, which puts agendas in understanding the technology use and needs of older adults (Patskanick et al., 2024).

Examining socio-demographic factors such as age, gender, and income are still significant drivers of consumer attitudes toward smart home technologies. Studies show that older adults show conflicting attitudes toward home smart technologies due to either confusion of efficacy employing the technology alone or usage concerns at the household level, and hurdles of understanding privacy and security implications (Wei et al., 2023). Under certain volition of cognitive barriers and social situations, older consumers are developing a rapport of acceptance for smart home systems (Patskanick et al., 2024). Each demographic objective highlights the need to develop marketing strategies to address specific aspects of ambivalence toward home smart technology consumers.

Security and privacy are significant understanding consumers have surrounding technologies, and even further for smart home technologies. Home smart technology could possess a risk of data breaches and thieves gaining access to homes, which would not support overall comfort for consumers. Security becomes complacent if trust in technology is fostered, and particularly trust because of the amount of personal data currently shared, is paramount (Liu et al., 2021). Both findings are essential to overcome through oversights of complex user experiences deemed unavailable or concerns managing personal data.

Finally, due to the complex ecosystem of smart home devices, connections, multiple connectivity protocols, necessary object usage (smart hubs), or other barriers to adoption, can be overwhelming for consumers wishing to maintain the value proposition of the device. Consumers are probably confused about how devices will integrate into their home (Phan & Kim, 2020). Barriers will reduce as smart home devices become more universal, easier to use or integrate, and more interoperable with each iteration (Türkyılmaz & Altındağ, 2022).

Recent advancements in smart home technologies have been profoundly shaped by the integration of artificial intelligence (AI) and sustainable energy solutions, yielding homes that are more efficient, secure, and environmentally responsible (Gazieva et al., 2023; Șerban & Lytras, 2020). The advent of sophisticated AI techniques, including deep learning and machine learning models, has enabled smart residential systems

to optimize energy consumption and automate everyday tasks, thereby enhancing user comfort while simultaneously reducing operational costs (Camacho et al., 2024; Khan et al., 2020). These innovations are not merely confined to the incorporation of intelligent applications within isolated systems but also extend to a comprehensive integration of IoT-based devices, which serve as the cornerstone of modern smart home architectures (Sepasgozar et al., 2020; Tarhouni & Aloui, 2024).

Concurrently, the incorporation of renewable energy sources has underscored the transition toward sustainable energy practices within domestic environments. AI-driven energy management systems are increasingly being deployed to facilitate real-time monitoring, predictive analytics, and the optimal distribution of renewable resources, such as photovoltaic systems and wind power, within smart grid frameworks (Hassan et al., 2023). Such systems not only enhance energy efficiency but also contribute significantly to reducing the overall carbon footprint of households, aligning with the global impetus toward net-zero emissions and sustainable living (Camacho et al., 2024; Stecuła et al., 2023).

These technological advancements promote a dynamic and responsive home environment wherein AI models continuously learn from user behavior and external variables to deliver personalized and energy-efficient services (Almusaed et al., 2023; Lashkari et al., 2019). In essence, the convergence of AI, IoT, and sustainable energy solutions is fundamentally redefining the smart home paradigm, thereby fostering innovative living spaces that deliver both technological sophistication and environmental sustainability (Singh & Dhablia, 2023).

The adoption of smart home systems has emerged as a significant area of interest in contemporary consumer technology research. As households increasingly integrate advanced technologies into their daily lives, understanding the factors that influence consumer acceptance becomes essential. This study aims to explore the socio-demographic factors that affect consumer adoption of smart home systems, using the Technology Acceptance Model (TAM) as a theoretical framework.

Strzelecki et al. (2024) emphasized that the increasing affordability of smart home technologies is a critical factor driving their adoption, yet they also noted a lack of understanding regarding the elements that sustain continued use. This suggests that while initial adoption may be influenced by economic factors, ongoing usage is contingent on user satisfaction and perceived benefits.

The role of socio-demographic factors cannot be understated. Cannizzaro et al. (2020) conducted a nationally representative survey in the UK, revealing that trust in smart home devices, privacy concerns, and user satisfaction are significantly influenced by socio-demographic characteristics. Older adults, for example, may exhibit different attitudes toward technology compared to younger users, often due to varying levels of technological literacy and perceived utility (Tiersen et al., 2021). Wong and Leung (2016) further explored how the attributes of smart home devices can enhance the quality of life for elderly users, indicating that tailored solutions are necessary to meet the unique needs of this demographic.

In addition to perceived usefulness, perceived ease of use—defined as the degree to which a person believes that using a particular system would be free of effort—also significantly impacts adoption attitudes. Shih (2013) argues that a high perceived ease of use can positively influence consumer attitudes toward smart home services, although it does not directly correlate with adoption. This nuanced understanding suggests that while ease of use is important, it must be coupled with tangible benefits to encourage adoption. Furthermore, the interplay between user satisfaction and technology adoption was highlighted by Orlov (Orlov et al., 2024), who asserted that usability and reliability are critical determinants of user satisfaction, which in turn influences the likelihood of adopting smart home technologies.

The adoption of smart home systems is a multifaceted issue influenced by various factors, including perceived usefulness, ease of use, and socio-demographic characteristics. The integration of these elements within the Technology Acceptance Model provides a robust framework for understanding consumer behavior in this context. As the market for smart home technologies continues to grow, further research is essential to unravel the complexities of user acceptance and to develop strategies that enhance the adoption rates among diverse consumer segments.

In this study, the adoption of smart home systems is examined through the Technology Acceptance Model (TAM) framework, incorporating a comprehensive set of factors that influence consumer preferences. Key dimensions such as Adoption Intention, Perceived Value, Perceived Benefit, Perceived Risk, Functionality, Design, Brand Name, Price, Perceived Service Usability, Perceived Network Size, Perceived Complementarity, and Privacy Concern are explored to understand their roles in shaping user attitudes and decision-making processes. By integrating these variables within the TAM framework, this research aims to provide a deeper insight into the socio-demographic determinants of smart home adoption, offering valuable implications for both academia and industry stakeholders. As smart home technologies continue to evolve, identifying the factors that drive or hinder their adoption becomes increasingly crucial for enhancing consumer acceptance and fostering widespread adoption across diverse user groups.

In terms of contribution to the literature, this study provides a comprehensive examination of smart home adoption through the Technology Acceptance Model (TAM) framework, integrating a diverse set of factors. While existing studies have individually examined perceived usefulness, ease of use, and security concerns, this study extends the discussion by incorporating a broader socio-demographic perspective, considering how variables such as income, education level, age, gender, and marital status influence adoption decisions. Additionally, by analyzing factors like Perceived Benefit, Perceived Risk, Functionality, Brand Name, and Perceived Complementarity, this research offers a holistic understanding of consumer attitudes toward smart home technologies. The findings contribute to the growing body of knowledge on consumer technology adoption, providing practical insights for developers, policymakers, and marketers seeking to enhance the design, accessibility, and security of smart home systems, ultimately fostering higher adoption rates and user satisfaction.

Methodology

Population and Sample

The population of the study consisted of consumers aged 20 and above who had previously purchased and currently used any product within the scope of smart home systems in Türkiye. The sample consisted of 556 individuals selected from among these through convenient and snowball sampling methods. Snowball sampling is often considered appropriate because of the difficulty of accessing a well-defined target population and the need to reach individuals with specific experiences or awareness. Although the snowball sampling method was appropriate for reaching participants in this study, it is important to acknowledge that such non-probability sampling techniques may introduce certain biases. These methods can limit the generalizability of the findings due to the potential homogeneity among respondents and the lack of representation of the broader population.

There are no official statistics regarding the number of people who have purchased and are currently using any product within the scope of smart home systems in Türkiye. However, in a survey conducted in 2022 with the participation of 1046 individuals, the usage rate of home technologies such as security, lighting,

smoke detectors, etc., was determined to be in the range of 20-21% (Statista, 2022). Information regarding the region where the research was conducted is not provided, but this rate was considered to be quite high and the usage rate of smart home technology was accepted as 10% for the calculation of sample adequacy. Based on the 2022 Address-Based Population Registration System (ADNKS) results, the population aged 20 and over in Türkiye is 60.229.333 (TÜİK, 2022). Roughly 10% of this figure amounts to approximately 6,025,000 individuals. In cases where the number making up the population is known, one of the formulas commonly used to determine the number of units to be included in the sample is as follows:

$$n = \frac{Nt^2pq}{d^2(N-1) + t^2pq}$$

In this formula; n: number of individuals to be sampled, N: number of individuals in the population, t: theoretical value obtained according to the t table at a certain significance level, p: frequency of occurrence of the event under investigation, q: frequency of non-occurrence of the event under investigation, and d: sampling error (Yazıcıoğlu and Erdoğan, 2014). When applying this formula to our study with a confidence level of 95% and a sampling error of 5%, the calculated required sample size was 384 individuals. We included 556 individuals in our research, approximately 1,5 times this calculated number.

Participants were reached through convenient (easy) and snowball sampling methods. Convenient (easy) sampling is based on selecting units that are easily accessible from among those thought to represent the population. Snowball sampling is conducted by reaching a new individual exhibiting suitable characteristics for participation in the research through a participant who has already joined the research (Tuna, 2016).

Data Collection Method

The research data were gathered through a questionnaire comprising two sections. The initial part encompassed a "Demographic Information Form" crafted by the researcher, containing five multiple-choice queries covering gender, age group, marital status, income level, and education level.

The measurement scale was developed by adapting items from previously validated and widely used instruments in the literature. The scale was designed to measure users' approaches toward smart home systems. A total of 45 items were included in the scale, which was structured using a 7-point Likert format (1 = Strongly Disagree, 7 = Strongly Agree).

The items related to functionality were adapted from the studies of Ziamou and Ratneshwar (2002) and Yu et al. (2017). The items used for the measuring design were derived from the scales developed by Van der Heijden (2004) and Cyr et al. (2006). The items related to price were adapted from the studies of Teng and Lu (2009) and Yu et al. (2017), while the items related to brand name were based on the scales developed by Brucks et al. (2000) and Yu et al. (2017).

Among the variables related to platform features, perceived service availability was measured using the items developed by Hong and Tam (2006). The perceived network size and perceived complementarity were evaluated based on the scale proposed by Lin and Lu (2011). Privacy concerns were measured using the items developed by Nepomuceno et al. (2014). The items related to perceived benefit were taken from the study of Hsu & Lin (2016), while the items related to perceived risk were adapted from the scales developed by Yang et al. (2015) and Featherman & Pavlou (2003). Finally, the measurements related to perceived value and adoption intention were derived from the scales developed by Kim et al. (2007).

The sub-dimensions included adoption intention, perceived value, perceived benefit, perceived risk, functionality, design, brand name, price, perceived service usability, perceived network size, perceived

complementarity, and privacy concerns. Notably, within the scale, 7 items in the price and perceived risk sub-dimensions were reverse-coded before analysis.

To ascertain the validity of the Smart Home Systems Approach Questionnaire, an exploratory factor analysis (EFA) was conducted. Initially, the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy yielded a value of 0,934, deemed "excellent" for sampling adequacy according to Field (2009). Furthermore, Bartlett's Test of Sphericity produced a chi-square value of 34375.124 with $p < .05$, meeting the criteria for data suitability for factor analysis (Büyüköztürk, 2011). Based on these results, a factor analysis was pursued.

The validity, represented by the explained variance, of the Smart Home Systems Approach Questionnaire was determined to be 88.04%. Given that this value surpassed 50%, the questionnaire was deemed valid (Büyüköztürk, 2011). Moreover, to assess the reliability of the questionnaire, Cronbach's Alpha reliability analysis was conducted for both the overall scale and each sub-dimension separately (refer to Table 1). As shown in Table 1, the scale exhibited a reliability coefficient of 0,993, with values exceeding 0,90 observed across all sub-dimensions. This aligns with the classification by George & Mallery (2003), where values of 0,90 and above are considered highly reliable.

Table 1

Reliability Findings of the Approach to the Smart Home Systems Scale

	Adoption Intention	Perceived Value	Perceived Benefit	Perceived Risk	Functionality	Design	Brand Name Usability	Price	Perceived Service Usability	Perceived Network Size	Perceived Complementarity	Privacy Concern	**Scale General**
Cronbach Alpha	0,925	0,927	0,976	0,973	0,965	0,991	0,994	0,974	0,988	0,977	0,969	0,978	**0,993**

Hypotheses of the Study

The adoption of smart home systems is influenced by a multitude of factors, including adoption intention, perceived value, perceived benefit, perceived risk, functionality, design, brand name usability, price, perceived service usability, perceived network size, perceived complementarity, and privacy concerns.

The intention to adopt smart home systems is significantly influenced by perceived value, which encompasses the benefits that consumers expect to receive from using these technologies. Research indicates that perceived usefulness and satisfaction are pivotal in forming positive adoption intentions (Gu et al., 2019; Wei et al., 2019). For instance, a study highlighted that service quality and perceived usefulness directly enhance user satisfaction, which in turn fosters the habitual use of smart home services (Gu et al., 2019). Furthermore, perceived value mediates the relationship between service innovation and consumers' willingness to use smart home technologies, emphasizing its importance in the adoption process (Li, 2022).

Consumers weigh the perceived benefits of smart home systems against the potential risks. The perceived benefits include enhanced convenience, security, and energy efficiency, which are crucial for motivating adoption (Wong & Leung, 2016). Conversely, perceived risks, particularly concerning privacy and security, can deter users from embracing these technologies.

The functionality of smart home systems, including their ease of use and integration with existing home environments, is vital for consumer preference. Studies have shown that users favor systems that offer seamless interoperability and user-friendly interfaces (Yang et al., 2017). Additionally, design esthetics and

the overall user experience contribute to the perceived usability of smart home technologies, influencing consumer choices (Wong & Leung, 2016). A well-designed system that aligns with consumer expectations can enhance the perceived value and reduce the perceived risk.

Brand reputation plays a significant role in shaping consumer perceptions of smart home systems. Established brands are often associated with reliability and quality, which can enhance perceived trust and usability (Liu et al., 2021). Price sensitivity also affects adoption intentions; consumers are likely to evaluate the cost against the perceived benefits and functionalities offered by smart home systems (Pliatsikas & Economides, 2022). Thus, competitive pricing strategies can enhance market penetration and consumer acceptance.

The usability of the services provided by smart home systems is crucial for consumer satisfaction. Users expect high-quality service that is responsive and reliable (Gu et al., 2019). Additionally, the perceived network size, which refers to the extent of connectivity and compatibility with other devices, influences consumer preferences. A larger network size can enhance the perceived complementarity of smart home devices, making them more appealing to potential users (Wei et al., 2019).

Privacy concerns remain a significant barrier to the adoption of smart home technologies. Consumers are increasingly aware of the implications of data collection and surveillance associated with smart home systems. Research indicates that addressing these concerns through transparent data practices and robust security measures can mitigate perceived risks and enhance consumer trust (Liu et al., 2021). Thus, companies must prioritize privacy in their smart home offerings to foster greater acceptance among consumers. In line with the literature mentioned above, the H1 hypothesis has been described as follows:

H1: There is a statistically significant relationship among the factors influencing consumers' preferences for smart home systems.

The intention to adopt smart home systems is significantly influenced by perceived value, which encompasses the benefits that consumers expect to receive from using these technologies. Research indicates that perceived usefulness and satisfaction are pivotal in forming positive adoption intentions. For instance, a study highlighted that service quality and perceived usefulness directly enhance user satisfaction, which in turn fosters the habitual use of smart home services (Gu et al., 2019). Furthermore, perceived value mediates the relationship between service innovation and consumers' willingness to use smart home technologies, emphasizing its importance in the adoption process (Li, 2022).

Consumers weigh the perceived benefits of smart home systems against the potential risks. The perceived benefits include enhanced convenience, security, and energy efficiency, which are crucial for motivating adoption (Wong & Leung, 2016). Conversely, perceived risks, particularly concerning privacy and security, can deter users from embracing these technologies.

The functionality of smart home systems, including their ease of use and integration with existing home environments, is vital for consumer preference. Studies have shown that users favor systems that offer seamless interoperability and user-friendly interfaces (Yang et al., 2017). Additionally, design esthetics and the overall user experience contribute to the perceived usability of smart home technologies, influencing consumer choices (Wong & Leung, 2016). A well-designed system that aligns with consumer expectations can enhance the perceived value and reduce the perceived risk.

Brand and price sensitivity also affects adoption intentions; consumers are likely to evaluate the cost against the perceived benefits and functionalities offered by smart home systems (Wei et al., 2019). Thus, competitive pricing strategies can enhance market penetration and consumer acceptance.

The usability of the services provided by smart home systems is crucial for consumer satisfaction. Users expect high-quality service that is responsive and reliable (Gu et al., 2019). Additionally, the perceived network size, which refers to the extent of connectivity and compatibility with other devices, influences consumer preferences. A larger network size can enhance the perceived complementarity of smart home devices, making them more appealing to potential users (Wei et al., 2019)

Privacy concerns remain a significant barrier to the adoption of smart home technologies. Consumers are increasingly aware of the implications of data collection and surveillance associated with smart home systems. Research indicates that addressing these concerns through transparent data practices and robust security measures can mitigate perceived risks and enhance consumer trust (Liu et al., 2021). Thus, companies must prioritize privacy in their smart home offerings to foster greater acceptance among consumers. Building upon the existing literature, the H2 hypothesis has been established as follows:

H2: Consumers' preferences for smart home systems significantly differ by gender.

Younger consumers often perceive greater value in smart home technologies due to their familiarity with technology and the benefits these systems offer, such as convenience and enhanced lifestyle. Research indicates that younger individuals are more adept at utilizing smart home technologies, which correlates with their higher education levels and income, leading to a greater affinity for such innovations (Arar et al., 2021; He et al., 2022). In contrast, older adults may view these technologies as complex or intrusive, which can diminish their perceived value and benefits (Baek et al., 2022). The perceived benefits of smart home systems, such as improved healthcare, safety, and social involvement, are particularly significant for older adults, as they often seek solutions that promote independent living and enhance their quality of life (Jain et al., 2024; Wei et al., 2023).

Older adults may have heightened concerns regarding privacy and the reliability of smart home technologies, which can hinder their adoption (Sequeiros et al., 2021). In contrast, younger consumers typically exhibit lower perceived risks, viewing smart home functionalities as integral to modern living (He et al., 2022; Katuk et al., 2018). The functionality of smart home systems, including ease of use and integration with existing technologies, is paramount for all age groups, but particularly for older adults who may require more intuitive interfaces (Cho & Choi, 2020; Yaldaie et al., 2023).

Some researches emphasize that systems must be user-friendly to cater to older adults, who may struggle with complex interfaces (Cho & Choi, 2020). Conversely, younger users may prioritize esthetic design and advanced functionalities, which can enhance their overall experience with smart home technologies (Shih, 2013). Established brands often instill confidence in their products, particularly among older consumers who may be more risk-averse (Arar et al., 2021; Maswadi et al., 2022). Price perception also varies by age group, with younger consumers often more willing to invest in innovative technologies, while older adults may be more price-sensitive, focusing on the long-term value and cost-effectiveness of smart home solutions (Maswadi et al., 2022; Wei et al., 2023).

Older adults may require more support and simpler services to feel comfortable with technology (Baek et al., 2022; Uddin et al., 2010). Additionally, perceived network size, or the extent of connectivity and integration with other devices, can enhance the appeal of smart home systems, especially for younger users who are

accustomed to interconnected devices (Katuk et al., 2018; Qiu et al., 2022). Privacy concerns are particularly pronounced among older adults, who may fear the implications of data collection and surveillance inherent in smart home technologies. Younger consumers, while also concerned about privacy, may prioritize the benefits of connectivity and convenience over these risks (He et al., 2022; Maswadi et al., 2022). Addressing these privacy concerns through transparent practices and robust security measures is essential for fostering trust across all age groups. Drawing from the aforementioned studies, the H3 hypothesis has been defined as follows:

H3: Consumers' preferences for smart home systems significantly differ across different age groups.

Higher perceived benefits, such as energy savings, enhanced security, and increased comfort, lead to a greater willingness to invest in smart home technologies. Research indicates that when consumers recognize the value of platform services, they also perceive greater value in smart home products, enhancing their overall experience (Tang & Inoue, 2021). This aligns with findings that emphasize the importance of perceived usefulness and satisfaction in driving the continuous use of smart home services (Gu et al., 2019; Nascimento & Fettermann, 2023).

The functionality of smart home systems is another critical factor. Consumers prefer user-friendly systems and offer seamless integration with existing devices (Cho & Choi, 2020; Hubert et al., 2019). Research has shown that perceived ease of use significantly affects adoption intentions, particularly among older adults who may find complex systems daunting (Zhou et al., 2024). The design and usability of smart home interfaces are essential for ensuring that users can interact with the technology without frustration.

A well-designed product that aligns with consumer expectations can enhance perceived value and usability (Hubert et al., 2020; Yaldaie et al., 2023). Brand name recognition can further bolster consumer confidence, as established brands are often associated with reliability and quality (Shih, 2013).

Price remains a significant barrier to adoption, particularly for consumers with lower educational backgrounds who may have limited financial resources (Nascimento & Fettermann, 2023). Studies suggest that consumers weigh the cost against the perceived benefits and functionality of smart home systems, making affordability a crucial consideration in their decision-making process (Sequeiros et al., 2021; Yang et al., 2017).

Consumers are more inclined to adopt smart home technologies that can integrate with various devices and services, enhancing their overall functionality (Tang & Inoue, 2021; Türkyılmaz & Altındağ, 2022). This interconnectedness can lead to a more cohesive and satisfying user experience, further encouraging adoption (Nascimento & Fettermann, 2023). Based on the literature reviewed above, the H4 hypothesis is expressed as follows:

H4: Consumers' preferences for smart home systems significantly differ by education level.

Research indicates that married individuals may prioritize functionality and security features more than single individuals, who might focus on cost and ease of use (Li et al., 2021). For instance, married couples often seek systems that enhance family safety and convenience, while single consumers may be more inclined to adopt technologies that offer entertainment and personal convenience (Mamonov & Benbunan-Fich, 2020). Married consumers may perceive higher value in systems that support family dynamics and household management, while singles may evaluate value based on personal lifestyle enhancements (Al-Husamiyah & Al-Bashayreh, 2022; Marikyan et al., 2019).

Married consumers often report a higher frequency of perceived benefits related to comfort and security, such as routine activity automation and remote home management (Nascimento et al., 2022). In contrast,

single consumers might prioritize benefits that enhance their individual lifestyle, such as energy savings and entertainment options (Maznah et al., 2021).

Studies have shown that married individuals may exhibit greater concern over the privacy risks associated with smart home systems, as these technologies often involve the collection and sharing of personal data (Wang et al., 2023). This concern can lead to hesitance in adopting smart home technologies, particularly if the perceived risks outweigh the anticipated benefits. On the other hand, single consumers may be less concerned about privacy issues, focusing instead on the immediate benefits and functionalities offered by smart home systems (Liu et al., 2021).

Married couples often look for systems that can integrate seamlessly into their daily routines and enhance household management, emphasizing the importance of user-friendly interfaces and reliable performance (Wright & Shank, 2019; Zhang et al., 2020). In contrast, single consumers may prioritize esthetic design and innovative features that align with their personal style and lifestyle preferences (Luor et al., 2015).

Price sensitivity is another factor that can differ between married and single consumers. Research suggests that married individuals may be more willing to invest in higher-priced smart home systems if they perceive significant long-term benefits, such as energy savings and enhanced security (Al-Husamiyah & Al-Bashayreh, 2022; Li et al., 2021). Conversely, single consumers may be more price-sensitive and inclined to opt for budget-friendly options that still meet their basic needs (Jacobsson et al., 2016). Brand reputation also plays a crucial role; married consumers might prefer established brands known for reliability and security, while singles may be more open to exploring newer brands that offer innovative features at competitive prices (Mennicken & Huang, 2012; Tural et al., 2021). In light of previous research, the H5 hypothesis has been proposed as follows:

H5: Consumers' preferences for smart home systems significantly differ by marital status.

Consumers with higher income levels often perceive greater value in smart home technologies due to their ability to enhance convenience, security, and energy efficiency. Studies indicate that perceived value is a critical determinant of consumer acceptance, particularly in relation to the modularity and inter-consumer connectivity offered by smart home platforms (Tang & Inoue, 2021). Furthermore, the perceived benefits, such as improved quality of life and enhanced safety, significantly influence adoption intentions among affluent consumers (Hubert et al., 2019; 2020). In contrast, lower-income consumers may prioritize immediate cost savings and practicality over advanced functionalities, which can affect their perceived value of smart home systems.

The perception of risk associated with smart home technologies varies with income levels. Higher-income consumers may be more willing to embrace new technologies despite potential risks, such as privacy concerns, because they can afford to mitigate these risks through enhanced security measures (Liu et al., 2021). Conversely, lower-income consumers often exhibit heightened sensitivity to perceived risks, which can deter them from adopting smart home systems (Hubert et al., 2019; Paetz et al., 2011).

Functionality is a paramount consideration for consumers across all income levels, but the emphasis on design and ease of use may differ. Higher-income consumers tend to favor esthetically pleasing designs and advanced functionalities that align with their lifestyle preferences (Chin et al., 2023). In contrast, lower-income consumers may prioritize basic functionalities that offer immediate utility, such as energy management and home security (Pliatsikas & Economides, 2022).

Brand reputation plays a significant role in the adoption of smart home technologies, particularly among higher-income consumers who may associate established brands with quality and reliability (Chin et al., 2023; Gu et al., 2019). Price sensitivity is more pronounced among lower-income consumers, who often view smart home systems as luxury items rather than necessities (Tural et al., 2021).

The usability of smart home systems is crucial for adoption, especially for older adults or less tech-savvy individuals (Maswadi et al., 2020; Zhang et al., 2024). Higher-income consumers may have greater access to support services and resources that enhance usability, while lower-income consumers may struggle with the complexity of these systems (Maswadi et al., 2020; Tural et al., 2021). Additionally, perceived network size, which refers to the interconnectedness of devices within a smart home ecosystem, can influence adoption intentions. Consumers with higher income levels may be more inclined to invest in comprehensive systems that offer extensive connectivity (Gu et al., 2019; Hubert et al., 2020).

Privacy remains a significant barrier to the adoption of smart home technologies, particularly for lower-income consumers who may lack trust in the security of their personal data (Liu et al., 2021). Addressing these concerns through transparent communication and robust security measures is essential for fostering trust and encouraging adoption across all income levels. Grounded in the discussed literature, the H6 hypothesis has been articulated as follows:

H6: Consumers' preferences for smart home systems significantly differ by income level.

Data Analysis

Quantitative analysis methods were employed to evaluate the data of the study. For this purpose, SPSS 25 software was utilized, and descriptive statistical methods, Pearson correlation analysis, independent samples t-test, one-way analysis of variance (ANOVA), and Hochberg GT2 post-hoc test were applied. The reason for using the Hochberg GT2 is that the compared group sizes are not similar (Field, 2009). Furthermore, all analyses were conducted at a 95% confidence interval and evaluated at a significance level of $p < .05$. Additionally, the skewness and kurtosis values of the relevant scores were examined to determine whether the normal distribution assumption was met (refer to Table 2).

Table 2

Skewness and Kurtosis Values for Subscale Scores

	Adoption Intention	Perceived Value	Perceived Benefit	Perceived Risk	Functionality	Design	Brand Name	Price	Perceived Service Usability	Perceived Network Size	Perceived Complementarity	Privacy Concern
Skewness	−0.567	−0.387	−0.456	−0.521	−0.33	−0.549	−0.14	−0.28	−0.552	−0.51	−0.462	−0.416
Kurtosis	−0.499	−0.796	−0.919	−0.662	−1.038	−0.671	−1.307	−1.21	−0.697	−0.742	−0.782	−0.851

In social sciences, especially when the sample size is large, it is often recommended to use skewness and kurtosis values to assess normality instead of relying solely on the Kolmogorov-Smirnov or Shapiro-Wilk tests, which can be overly sensitive in detecting minor deviations (Ghasemi & Zahediasl, 2012). Acceptable ranges for skewness and kurtosis are typically between −1 and +1 (George & Mallery, 2010). Additionally, the coefficient of variation can be used to evaluate the relative variability of the data (Rumsey, 2016). Based on these indicators, the data in this study can be considered to be normally distributed.

Relationships Among Factors Affecting the Adoption of Smart Homes

The relationships between subscale scores representing factors affecting the adoption of smart homes were examined using Pearson correlation analysis, and the findings are presented in Table 3.

The correlation analysis results indicate that there are significant relationships among the factors influencing consumers' preferences for smart home systems. The strongest positive correlations were observed between the intention to adopt and key factors such as perceived value, perceived complementarity, and perceived network size, suggesting that consumers' willingness to adopt smart home systems is shaped by how they perceive the system's overall value and its integration within their existing technological ecosystem. Additionally, price, perceived benefit, functionality, and brand name were moderately correlated with adoption intention, indicating that while these factors are important, their influence may vary based on individual consumer priorities.

Moreover, price exhibited a strong positive correlation with perceived benefit, highlighting that consumers associate higher-priced smart home systems with greater advantages. However, price had only a moderate correlation with most other subscales, implying that while cost is a significant factor, it is not the sole determinant of consumer preferences. Across all dimensions, most factors were strongly interrelated, reinforcing the idea that smart home adoption is influenced by a combination of perceived benefits, risks, usability, and financial considerations rather than a single dominant factor. These findings confirm the hypothesis that various interdependent factors collectively shape consumer decisions regarding smart home technologies.

Based on these findings, the hypothesis "H1: There is a statistically significant relationship among the factors influencing consumers' preferences for smart home systems" is supported.

Table 3

Correlation Findings for Relationships Among Subscale Scores

Dimension		1	2	3	4	5	6	7	8	9	10	11	12
1. Adoption Intention	r	1	.790	.684	.760	.677	.764	.648	.389	.767	.764	.717	.702
	p		**.000**	**.000**	**.000**	**.000**	**.000**	**.000**	**.000**	**.000**	**.000**	**.000**	**.000**
2. Perceived Value	r		1	.776	.903	.787	.899	.745	.506	.894	.905	.850	.812
	p			**.000**	**.000**	**.000**	**.000**	**.000**	**.000**	**.000**	**.000**	**.000**	**.000**
3. Perceived Benefit	r			1	.830	.911	.822	.835	.723	.831	.838	.793	.751
	p				**.000**	**.000**	**.000**	**.000**	**.000**	**.000**	**.000**	**.000**	**.000**
4. Perceived Risk	r				1	.858	.970	.806	.534	.969	.978	.914	.917
	p					**.000**	**.000**	**.000**	**.000**	**.000**	**.000**	**.000**	**.000**
5. Functionality	r					1	.873	.799	.667	.856	.847	.786	.772
	p						**.000**	**.000**	**.000**	**.000**	**.000**	**.000**	**.000**
6. Design	r						1	.802	.545	.959	.975	.919	.893
	p							**.000**	**.000**	**.000**	**.000**	**.000**	**.000**
7. Brand Name	r							1	.636	.798	.814	.754	.727
	p								**.000**	**.000**	**.000**	**.000**	**.000**
8. Price	r								1	.527	.547	.512	.481
	p									**.000**	**.000**	**.000**	**.000**
	r									1	.964	.912	.885

Dimension		1	2	3	4	5	6	7	8	9	10	11	12
9. Perceived Service Usability	p										**.000**	**.000**	**.000**
10. Perceived Network Size	r										1	.940	.909
	p											**.000**	**.000**
11. Perceived Complementarity	r											1	.837
	p												**.000**
12. Privacy Concern	r												1
	p												

Examination of Attitudes Toward Smart Homes Based on Sociodemographic Characteristics

The results of the detailed analyses conducted based on the participants' sociodemographic characteristics are presented below in detail under relevant subheadings.

Gender

An independent samples t-test was conducted to determine whether there were significant differences between men and women regarding their attitudes toward smart homes, and the findings are presented in Table 4. All types of scores examined on the adoption of smart homes displayed significant differences based on gender ($p < .05$). In other words, men and women exhibit distinct approaches toward smart homes across all features.

Table 4

Comparison of Subscale Scores by Gender

Dimension	Female (Mean)	Male (Mean)	t	p
Adoption Intention	5.17	4.24	6.90	0.000
Perceived Value	5.34	4.13	9.28	0.000
Perceived Benefit	5.84	4.24	14.03	0.000
Perceived Risk	5.48	4.28	8.87	0.000
Functionality	5.72	4.24	12.33	0.000
Design	5.66	4.27	10.26	0.000
Brand Name	5.58	4.30	10.11	0.000
Price	6.00	5.32	8.16	0.000
Service Usability	5.57	4.27	9.35	0.000
Perceived Network Size	5.54	4.26	9.56	0.000
Perceived Complementarity	5.47	4.20	9.64	0.000
Privacy Concern	5.37	4.29	7.71	0.000

When reviewing the means, it was observed that the means for all subscales were higher for women than for men. Thus, it can be concluded that women's preferences are significantly more dominant than men's across all subscales. Women tend to adopt smart homes more and perceive more value, benefit, risk, functionality, and usability. Additionally, design, brand name, price, network size, complementarity, and privacy concerns are more prominent reasons for preference among women compared to men.

Based on these findings, the hypothesis "H2: Consumers' preferences for smart home systems significantly differ by gender" is supported.

Age

One-way analysis of variance (ANOVA) was conducted to examine the approach of participants toward smart homes across different age groups and to determine if there were any significant differences among the groups. The findings are presented in Table 5. The age distribution of participants was as follows: 90 individuals in the 20-29 age group, 159 in the 30-39 age group, 193 in the 40-49 age group, and 114 in the 50+ age group, making a total of 556 participants. The letters (a, b, c, d) represent different age groups in the post-hoc analysis, where a corresponds to the 20-29 age group, b to the 30-39 age group, **c** to the 40-49 age group, and d to the 50+ age group. These letters indicate the statistical groupings derived from the post hoc comparisons, showing which age groups are significantly different from each other.

Table 5

Comparison of Subscale Scores by Age

Dimension	20-29 (Mean)	30-39 (Mean)	40-49 (Mean)	50+ (Mean)	F	p	Difference
Adoption Intention	4.01	4.62	5.11	5.14	12.56	.000	c,d > a,b
Perceived Value	4.00	4.74	5.04	5.37	13.98	.000	c,d > a,b
Perceived Benefit	4.26	5.07	5.24	6.04	27.21	.000	d > a,b,c
Perceived Risk	3.91	5.03	5.18	5.50	18.37	.000	b,c,d > a
Functionality	4.23	5.14	5.14	5.80	19.33	.000	b,c,d > a
Design	4.08	5.18	5.22	5.58	15.38	.000	b,c,d > a
Brand Name	4.19	5.09	4.92	5.98	24.94	.000	d > a,b,c
Price	5.44	5.40	5.69	6.45	27.82	.000	d > a,b,c
Service Usability	4.08	5.09	5.14	5.58	14.30	.000	b,c,d > a
Perceived Network Size	4.04	5.04	5.18	5.52	15.26	.000	b,c,d > a
Perceived Complementarity	4.04	4.89	5.20	5.37	12.90	.000	b,c,d > a
Privacy Concern	4.11	5.10	5.02	5.21	9.05	.000	b,c,d > a

The findings revealed significant differences in participants' attitudes toward smart home systems across all sub-dimensions based on age groups ($p < .05$). Analyzing the mean scores suggests that attitudes toward smart home adoption tend to become more positive with increasing age. To further explore these differences, the Hochberg GT2 post hoc test was conducted, identifying specific variations among the age groups.

The results indicate that participants aged 40 years exhibit significantly higher adoption intention and perceived value compared to younger individuals, suggesting that older consumers recognize greater benefits and are more inclined to adopt smart home technologies. Furthermore, those aged 50 and above demonstrated stronger perceptions of benefits, placed greater importance on brand reputation, and prioritized price more than younger participants, implying that older consumers considered both functional and financial aspects when evaluating smart home systems. Additionally, individuals aged 30 and above reported higher perceptions of risk, functionality, design, perceived service usability, perceived network size, perceived complementarity, and privacy concerns than those under 30, indicating that as consumers age, they develop a more comprehensive and cautious approach toward smart home technologies. The results suggest that age plays a crucial role in shaping consumer perceptions and decision-making processes, with

older individuals generally showing stronger preferences and a more thoughtful evaluation of smart home technologies.

These findings support the hypothesis "H3: Consumers' preferences for smart home systems significantly differ across different age groups" is supported.

Education Level

One-way analysis of variance (ANOVA) was conducted to examine the participants' attitudes toward smart homes based on education level and determine if there were any significant differences among the groups. The results are presented in Table 6. The education level distribution of participants was 45 individuals in the Primary/Middle School group, 359 in the High School group, 124 in the Undergraduate group, and 28 in the Graduate group, making a total of 556 participants. In the post hoc analysis, a represents Primary/Middle School, b represents High School, c represents Undergraduate, and d represents Graduate. Groups sharing the same letter are not significantly different from each other, whereas those with different letters indicate statistically significant differences.

Based on the findings, notable differences were observed in participants' attitudes toward smart homes across all aspects, contingent on their educational background ($p < 0,05$). When reviewing the means, a trend emerged indicating that scores generally increased alongside higher levels of education. Postgraduate degree holders showed significantly greater intent to adopt compared with other participants. Moreover, this intent was notably higher among high school and undergraduate degree holders than among those with primary and secondary education.

Table 6

Comparison of Subscale Scores by Education Level

Dimension	Primary/ Middle (Mean)	High School (Mean)	Undergraduate (Mean)	Graduate (Mean)	F	p	Difference
Adoption Intention	2.36	4.83	5.24	6.29	58.35	.000	d > a,b,c
Perceived Value	2.31	4.72	5.74	6.69	92.14	.000	d > a,b,c; c > a,b; b > a
Perceived Benefit	2.32	4.84	6.85	7.00	367.78	.000	d > a,b,c; c > a,b; b > a
Perceived Risk	2.29	4.82	6.10	6.78	109.10	.000	c,d > a,b; b > a
Functionality	2.47	4.86	6.68	5.93	191.81	.000	c > a,b,d; d > a,b; b > a
Design	2.47	4.91	6.17	7.00	100.99	.000	d > a,b,c; c > a,b; b > a
Brand Name	2.76	4.79	6.33	6.68	112.75	.000	c,d > a,b; b > a
Price	5.75	5.19	6.96	7.00	201.14	.000	c,d > a,b; a > b
Service Usability	2.29	4.86	6.21	6.64	104.54	.000	c,d > a,b; b > a
Perceived Network Size	2.37	4.83	6.10	7.00	110.01	.000	d > a,b,c; c > a,b; b > a
Perceived Complementarity	2.53	4.71	6.11	7.00	100.05	.000	d > a,b,c; c > a,b; b > a
Privacy Concern	2.50	4.84	5.72	6.55	65.11	.000	d > a,b,c; c > a,b; b > a

The post hoc analysis using the Hochberg GT2 test revealed significant contrasts among educational groups regarding their perceptions of smart home systems. Postgraduate degree holders exhibited the highest perceptions of value, benefits, design, perceived network size, perceived complementarity, and privacy concerns compared to other groups. These perceptions were also higher among undergraduate

degree holders than among high school graduates and those with lower education levels, while high school graduates scored higher than individuals with lower education. Similarly, both undergraduate and postgraduate degree holders perceived greater risks, brand significance, and perceived service usability compared to other groups, with high school graduates also scoring higher than those with lower education levels. These findings suggest that higher education levels are associated with a more analytical and comprehensive evaluation of smart home technologies, where individuals recognize both the potential advantages and risks.

Additionally, functionality was a critical preference factor, with significantly higher ratings among the undergraduate and postgraduate participants compared to the others. Moreover, postgraduates demonstrated stronger functionality preferences than high school graduates and those with lower education levels, while high school graduates also preferred functionality more than the lowest education group. A similar pattern was observed for price preferences, where undergraduate and postgraduate participants displayed significantly higher price sensitivity, yet individuals with lower education levels exhibited stronger price preferences than high school graduates. This suggests that while higher-educated consumers evaluate smart home systems based on their technological benefits, lower-educated consumers may focus more on cost-effectiveness and affordability when making purchasing decisions. These findings highlight that education level plays a crucial role in shaping consumer perceptions and adoption tendencies regarding smart home systems.

Based on these findings, since notable differences were observed in participants' attitudes toward smart homes across all aspects, contingent on their educational background ($p < 0,05$), the hypothesis "H4: Consumers' preferences for smart home systems significantly differ by education level" is supported.

Marital Status

An independent samples t-test was conducted to examine whether there was a significant difference between these groups in their approach to smart homes. The marital status groups consisted of 248 married and 308 single participants. The findings are presented in Table 7. The findings are presented in Table 7.

Table 7

Comparison of Scale Scores by Marital Status

Dimension	Married (Mean)	Single (Mean)	t	p
Adoption Intention	5.64	4.12	12.38	.000
Perceived Value	5.78	4.11	14.04	.000
Perceived Benefit	6.55	4.11	33.95	.000
Perceived Risk	6.03	4.17	15.61	.000
Functionality	6.24	4.22	21.25	.000
Design	6.15	4.25	15.78	.000
Brand Name	6.25	4.11	21.86	.000
Price	6.60	5.01	27.07	.000
Service Usability	6.17	4.14	17.01	.000
Perceived Network Size	6.12	4.15	17.02	.000
Perceived Complementarity	6.11	4.03	17.87	.000
Privacy Concern	5.87	4.18	13.45	.000

As shown in Table 7, all the score types examined for consumers' adoption of smart homes significantly differ by marital status (p < 0,05). In other words, married and single individuals exhibit different approaches to all the features of smart homes. When examining the means, it is seen that the averages for married individuals are higher than for singles in all sub-dimensions. Accordingly, it can be said that the preferences of married individuals are significantly more pronounced than those of singles in all sub-dimensions. Married individuals adopt smart homes more, perceiving more value, benefit, risk, functionality, and usability. Additionally, design, brand name, price, network size, complementarity, and privacy concerns are more dominant preference factors for married individuals compared to singles.

Based on these findings, the hypothesis "H5: Consumers' preferences for smart home systems significantly differ by marital status" has been accepted.

Income Level

A one-way ANOVA was conducted to examine whether there was a significant difference between participants' approaches to smart homes by income level. The findings are presented in Table 8. In this table, a represents individuals with an income of less than 10.000 (n=166), b represents those earning between 10.000 and 19.999 (n=212), and c represents those with an income of 20.000 or more (n=178), with these groups used in post-hoc comparisons to indicate statistically significant differences. Accordingly, participants' approaches to smart homes significantly differed by income level in all sub-dimensions (p < .05). When examining the means, it is generally observed that the scores increase with the increase in income level. However, to determine between which income levels the differences exist, Hochberg's GT2 post hoc test was used for comparisons.

Table 8

Comparison of Scale Scores by Income Level (TL)

Dimension	<10.000 (Mean)	10.000-19.999 (Mean)	20.000+ (Mean)	F	p	Difference
Adoption Intention	3.31	5.37	5.50	154.33	0.000	b,c > a
Perceived Value	3.32	5.21	5.86	190.57	0.000	b,c > a
Perceived Benefit	3.28	5.38	6.76	1335.90	0.000	b,c > a
Perceived Risk	3.29	5.38	6.15	254.32	0.000	b,c > a
Functionality	3.40	5.46	6.34	418.69	0.000	b,c > a
Design	3.41	5.49	6.21	230.49	0.000	b,c > a
Brand Name	3.50	5.22	6.35	276.90	0.000	b,c > a
Price	4.75	5.55	6.84	420.77	0.000	b,c > a
Service Usability	3.33	5.39	6.23	242.82	0.000	b,c > a
Perceived Network Size	3.30	5.40	6.20	267.34	0.000	b,c > a
Perceived Complementarity	3.39	5.15	6.20	206.30	0.000	b,c > a
Privacy Concern	3.30	5.45	5.85	187.17	0.000	b,c > a

The results indicate that income level has a significant impact on consumers' perceptions and preferences regarding smart home systems, as all dimensions show statistically significant differences (p = 0,000) across the income groups.

The findings indicate that income level plays a significant role in shaping consumers' perceptions and adoption intentions regarding smart home systems. Consumers in the highest income group (20.000+)

consistently reported the highest mean scores across all dimensions, suggesting that they perceive smart home systems as more valuable, beneficial, and reliable. The most pronounced differences between income groups were observed in Perceived Benefit (F = 1335,90), Functionality (F = 418,69), and Price (F = 420,77), highlighting that individuals with higher incomes not only recognize greater advantages in these technologies but also find them more functional and are more willing to invest in them despite their cost. This pattern suggests that financial capacity significantly influences adoption decisions, as higher-income consumers are better positioned to assess both the functional and financial aspects of smart home systems.

In contrast, consumers in the middle-income group (10.000-19.999) scored significantly higher than those in the lowest income group (<10.000) across all dimensions but still remained below the highest income group. This finding suggests that while middle-income consumers recognize the benefits of smart home systems, their perceptions are not as strong or developed as those in the highest income category. Meanwhile, the lowest income group consistently reported the lowest mean scores across all dimensions, indicating that financial constraints negatively impact the adoption intention, perceived value, and perceived usability of smart home technologies. These individuals may perceive smart home systems as less accessible or unnecessary luxuries, making affordability a primary barrier to adoption. Overall, the results confirm that income level is a key determinant of consumer preferences, with higher-income individuals perceiving smart home technologies as more valuable, usable, and trustworthy, while lower-income individuals exhibit significantly lower interest and confidence in adopting these innovations.

Based on the findings, the hypothesis "H6: Factors influencing consumers' preferences for smart home systems significantly differ by income level" has been accepted.

Discussion and Conclusion

The purpose of this study was to identify the essential aspects that strongly shape consumers' acceptance of smart home systems in Türkiye, a developing and vibrant digitally evolving market. Recently, there has been a surge in the use of smart technology in daily life and shifting consumer expectations for technologies that increase convenience, safety, and efficiency, among others, leading to increased interest in smart homes. However, understanding the processes shaping consumer adoption behavior remains a relevant issue, particularly in developing economies where socio-economic and cultural contexts are likely to shape technology usage behaviors. In this domain, the study provides valuable insights into how consumers adopt smart home systems in Türkiye based on an amalgamation of various technological and socio-demographic factors while yielding important implications for both theory and practice in good faith.

The investigation illustrated that the general adoption of smart home systems was somewhat above the mean. This indicates a favorable disposition toward smart home technologies, but still demonstrates an area for growth. The investigation identified three factors: price, perceived benefit, and functionality as significantly important. Specifically focusing on price, it related positively toward adoption, which indicated that the learner has sensitivity to the subjective financial cost associated with smart home systems. Similarly, the perceived benefits to the user, particularly the convenience, and limiting access to the home were beneficial to the user, and energy efficiency; functionality were other significant dimensions extending to shaping consumer behaviors. Considering all these elements, the study suggested that consumers valued the practical, life-improving, and value-unitizing aspects of smart home systems, specifically concerning decent brands providing the practical parts. In addition to this, the study was also able to identify practical functionality and ease of use as significant determinants of adoption; consumers generally preferred

systems providing user-friendly, reliable, and accessible features that assist with daily life tasks and reduce technology use complexity. Another key finding was that consumers generally possess a positive perception of privacy aspects related to smart home systems; perceiving these technologies as safe with respect to their personal data, opening to assist with routine, and valuable life improvements.

Another significant contribution of this study was with respect to the socio-demographic variables, including marital status and income level, which previous studies have generally avoided. The findings revealed that these aspects are key components of consumers' access to the technology, and more importantly, whether to purchase them. This observation was consistent with the view of Van Dijk (2020), who indicated that socio-economic inequalities—the digital divide—could limit consumers' access to technological solutions. In this sense, the current study implied more inclusion strategies to reach broader or alternative segments of consumers to use smart home technologies when purchasing based on socio-demographic elements.

Another area of interest was the relationships between adoption-related factors, which were not solitary but strongly mutually connected. All adoption-related factors were found to have a strong positive correlation, indicating that consumers evaluating or perceiving smart home systems was a complete package with multiple dimensions simultaneously. Moreover, the study suggested that an improvement in one area, i.e., design or functionality, was strongly likely to reflect a positive indication with other areas, i.e., usability, privacy, or perceived value. Finally, the study provides compelling evidence that consumers' adoption tendencies toward smart home systems vary significantly based on gender, age, education level, marital status, and income level. These demographic variables were found to meaningfully differentiate consumer perceptions and adoption behaviors across all examined factors. This result reinforces the importance of developing differentiated marketing and design strategies that consider the diverse needs, preferences, and expectations of various consumer segments.

The outcomes of this research offer valuable perspectives regarding the variables that impact consumer adoption of smart home systems and yield significant contributions to the monitoring of relevant systems in theory and practice. Consistent with previous literature (Gu et al., 2019; Hsu & Lin, 2016; Kim et al., 2007) and the Technology Acceptance Model (TAM), perceived benefit, perceived risk, functionality, and price appeared to be the most significant variables influencing adoption behavior. Among these four, price was the most significant indicator and supported previous findings regarding consumers' price sensitivity (Teng & Lu, 2009; Yu et al., 2017), particularly in emerging markets.

The significant role impacts of perceived benefit also provides further support for Wong & Leung (2016) and Hsu & Lin (2016), in that consumers' perceptions of increased convenience, efficiency, and security contribute considerably to their evaluation of the adoption of smart home systems. Furthermore, functionality, ease-of-use, and usability were also important predictors of adoption and were consistent with Ziamou & Ratneshwar, (2002) and Van der Heijden (2004) studies.

Significantly, all the variables showed strong and statistically significant relationships. This finding is important because it highlights that consumers are likely to make holistic consumer evaluations regarding smart home systems, which is consistent with Yang et al. (2017) and the reported synergistic nature of system usability, trust, and perceived value.

A further theoretical contribution of this study is that it illustrated that socio-demographic characteristics are also important moderator variables that shape consumer attitudes toward smart home systems. The socio-demographic differences that were found, for gender, age, education, marital status, and income

differences, extend previous literature (Shin et al., 2018; Wei et al., 2019), which has largely disregarded the moderator role of these variables in the smart technology context. Female consumers had significantly higher adoption intention and rated all dimensions of adoption more positively than males. These results align with findings by Shin et al. (2018) and Nikou (2019), who found that females show more sensitivity toward emotional and experiential values when adopting new technologies.

Older adults demonstrated increased perceived benefit, brand sensitivity, and price sensitivity, supporting prior research (Baek et al., 2022; Demiris et al., 2004). Moreover, older adults' higher perceived risks were aligned with assertions by Pal et al., (2018), who observed that privacy and security are particularly important to elderly users.

Educational level had a positive and statistically significant effect on all adoption factors, corroborating Baudier & Ammi (2020) and Shin et al. (2018) that education heightens users' ability to assess opportunities and risks associated with new technologies.

An important finding from this study indicates that married participants reported significantly greater adoption intentions and perceptions across all factors compared with single participants. In part, this supports the literature (Al-Husamiyah & Al-Bashayreh, 2022), suggesting that family-oriented consumers probably viewed functionality, security, and usability as more important as they considered the wider implications of smart home systems for all family members.

Income level significantly differed in adoption behavior. Higher income consumers rated smart home systems as more beneficial, valuable, and functional, despite the greater expense, which aligns with Hubert et al. (2020). Although income did not completely align with previous research and findings. Yang et al. (2017) and Shin et al. (2018)) that state income may influence user perceptions indirectly in various contexts.

While privacy was an important concern among participants, it was not the most important factor in their intention to adopt. However, concern for privacy did emerge as a statistically significant factor in their intention to adopt, a finding that is consistent with the concerns expressed by Nepomuceno et al. (2014) and Yang et al. (2015), indicating that privacy/security is a concern when adopting technology on platforms. Moreover, perceived service usability, perceived network size, and perceived complementarity appear to support the conclusions of Hong and Tam (2006) and Lin and Lu (2011) that platform-related factors are important in the decision-making process of consumers.

Although these findings provide valuable insights into the adoption of smart home systems, there are some limitations. First, this study was conducted in a geographically defined country (Türkiye), limiting the study sample to consumers in that country. Thus, the findings may not be generalizable to other geographic or cultural contexts. Future studies might expand the scope of this study, including consumers from different countries or regions, to understand the implementation and adoption of smart home systems in comparative contexts, including possible cultural influences on adoption.

Second, the study used a cross-sectional research design where consumer perceptions were obtained at one point in time; however, technology adoption is a potential evolving process; therefore, consumer perceptions and behaviors may change. Longitudinal research designs can further examine how consumer attitudes, perceptions of service usability, perceived network size, and perceived complementarity evolve as technologies become more accessible and frequently adopted.

Third, the data were obtained through a self-reported questionnaire. There is a potential for common method bias and social desirability bias, meaning that participants may not have responded in the way

they intended and rather provided socially desirable answers to the questions posed rather than answering based on their own behaviors or beliefs. Using other methods of data collection can help address this shortcoming.

Finally, while the study focused on some aspects of the Technology Acceptance Model (TAM) in assessing smart home systems, there are other variables that have been studied together with TAM factors, for example, trust as a protective factor, technology anxiety, perceived innovativeness, or environmental concern —one or all of which could provide a fuller understanding of consumer adoption behaviors. In particular, examining the psychological or contextual factors that impact consumers in their decision-making may provide deeper insight into consumers' processes of adopting smart home technologies.

Peer Review: Externally peer-reviewed.
Author Contributions: Conception/Design of Study- Ü.Ö., O.M.; Data Acquisition- O.M.; Data Analysis/Interpretation- Ü.Ö., O.M.; Drafting Manuscript- Ü.Ö., O.M.; Critical Revision of Manuscript- Ü.Ö.; Final Approval and Accountability- Ü.Ö., O.M.; Technical or Material Support- Ü.Ö., O.M.; Supervision- Ü.Ö.
Conflict of Interest: The authors have no conflict of interest to declare.
Grant Support: The authors declared that this study has received no financial support.

Author Details
Onur Muço
[1] Atatürk Üniversitesi, Yönetim Bilişim Sistemleri, Erzurum, Türkiye
0000-0001-6830-9897

Üstün Özen
[1] Atatürk Üniversitesi, Yönetim Bilişim Sistemleri, Erzurum, Türkiye
0000-0002-7595-4306 uozen@atauni.edu.tr

References

Al-Husamiyah, A. & Al-Bashayreh, M. (2022). A comprehensive acceptance model for smart home services. *International Journal of Data and Network Science*, *6*(1), 45-58. https://doi.org/10.5267/j.ijdns.2021.10.005

Almusaed, A., Yitmen, İ., & Almssad, A. (2023). Enhancing smart home design with ai models: a case study of living spaces implementation review. *Energies*, *16*(6), 2636. doi:10.3390/en16062636

Arar, M., Jung, C., Awad, J., & Chohan, A. (2021). Analysis of smart home technology acceptance and preference for the elderly in Dubai, UAE. *Designs*, *5*(4), 70. doi:10.3390/designs5040070

Baek, J., Na, S., Lee, H., Jung, H., Lee, E., Jo, M., ... & Jang, I. (2022). Implementation of an integrated home internet of things system for vulnerable older adults using a frailty-centered approach. *Scientific Reports*, *12*(1). doi:10.1038/s41598-022-05963-9

Baudier, P., Ammi, C., & Deboeuf-Rouchon, M. (2020). Smart home: Highly-educated students' acceptance. *Technological Forecasting and Social Change*, 153. https://doi.org/10.1016/j.techfore.2018.06.043.

Brucks, M., Zeithaml, V. A. & Naylor, G. (2000). Price and brand name as indicators of quality dimensions for durable goods. *Journal of the Academy of Marketing Science*, *28*(3), 359–374.

Büyüköztürk, Ş. (2011). Sosyal Bilimler için Veri Analizi El Kitabı. Pegem Akademi.

Camacho, J., Aguirre, B., Ponce, P., Anthony, B., & Molina, A. (2024). Leveraging artificial intelligence to bolster the energy sector in smart cities: a literature review. *Energies*, *17*(2), 353. doi:10.3390/en17020353

Cannizzaro, S., Procter, R., Ma, S., & Maple, C. (2020). Trust in the smart home: findings from a nationally representative survey in the uk. *Plos One*, *15*(5), e0231615. https://doi.org/10.1371/journal.pone.0231615

Chin, C., Wong, W., Cham, T., Thong, J., & Ling, J. (2023). Exploring the usage intention of ai-powered devices in smart homes among millennials and zillennials: the moderating role of trust. *Young Consumers Insight and Ideas for Responsible Marketers*, *25*(1), 1-27. https://doi.org/10.1108/yc-05-2023-1752

Cho, Y., & Choi, A. (2020). Application of affordance factors for user-centered smart homes: a case study approach. *Sustainability*, *12*(7), 3053. https://doi.org/10.3390/su12073053

Cyr, D., Head, M., & Ivanov, A. (2006). Design esthetics leading to m-loyalty in mobile commerce. *Information & Management*, *43*(8), 950–963.

Demiris, G., Rantz, M., Aud, M., Marek, K., Tyrer, H., Skubic, M., & Hussam A. (2004). Older adults' attitudes toward and perceptions of "smart home" technologies: a pilot study. *Med Inform Internet Med*,;29(2), 87-94.

Featherman, M. S., & Pavlou, P. A. (2003). Predicting e-service adoption: A perceived risk facets perspective. *International Journal of Human-Computer Studies*, *59*(4), 451–474.

Field, A. (2009). *Discovering Statistics Using SPSS*. SAGE.

Gazieva, L., Aygumov, T., & Zaripova, R. (2023). The role of smart technologies in the development of cost-effective and sustainable energy. *E3s Web of Conferences*, *451*, 01007. doi:10.1051/e3sconf/202345101007

George, D., & Mallery, P. (2003). *SPSS for Windows step by step: A simple guide and reference. 11.0 update (4th ed.)*. Boston: Allyn and Bacon.

George, D., & Mallery, P. (2010). *SPSS for Windows step by step: A simple guide and reference (10th ed.)*. Pearson.

Ghasemi, A., & Zahediasl, S. (2012). Normality tests for statistical analysis: A guide for non-statisticians. *International Journal of Endocrinology and Metabolism*, *10*(2), 486–489. https://doi.org/10.5812/ijem.3505

Gu, W., Bao, P., Hao, W., & Kim, J. (2019). Empirical examination of the intention to continue using smart home services. *Sustainability*, *11*(19), 5213. https://doi.org/10.3390/su11195213.

Hassan, Q., Sameen, A., Salman, H., Al-Jiboory, A., & Jaszczur, M. (2023). The role of renewable energy and artificial intelligence toward environmental sustainability and net zero. https://doi.org/10.21203/rs.3.rs-2970234/v1

He, F., Wu, Y., Jiao, Y., Chen, K., Xie, J., Tuersun, Y., ... & Chen, J. (2022). Chinese adult segmentation according to health skills and analysis of their use for smart home: a cross-sectional national survey. *BMC Health Services Research*, *22*(1). https://doi.org/10.1186/s12913-022-08126-8

Hong, S.-J., & Tam, K. Y. (2006). Understanding the adoption of multipurpose information appliances: The case of mobile data services. *Information Systems Research*, *17*(2), 162–179.

Hsu, C.-L., & Lin, J. C.-C. (2016). Empirical examination of consumer adoption of Internet of Things services: Network externalities and concern for information privacy perspectives. *Computers in Human Behavior*, *62*, 516–527.

Hubert, M., Blut, M., Brock, C., Zhang, R., Koch, V., & Riedl, R. (2019). The influence of acceptance and adoption drivers on smart home usage. *European Journal of Marketing*, *53*(6), 1073-1098. https://doi.org/10.1108/ejm-12-2016-0794

Hubert, M., Carugati, A., Brock, C., & Obel, B. (2020). Take it personally—the role of consumers' perceived value of personalization on cross-category use in a smart home ecosystem. *Proceedings of the 53rd Hawaii International Conference on System Sciences*, 2020. doi:10.24251/hicss.2020.144.

Jacobsson, A., Boldt, M., & Carlsson, B. (2016). A risk analysis of a smart home automation system. *Future Generation Computer Systems*, *56*, 719-733. https://doi.org/10.1016/j.future.2015.09.003

Jain, P., Pendyala, S., Etu, E., Zhang, Z., Shah, M., Larot, J., ... & Huang, G. (2024). Exploring attitudes toward smart home technology through focus groups: comparing older adults with and without health conditions. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *68*(1), 3-8. doi: 10.1177/10711813241276479

Katuk, N., Ku-Mahamud, K., Zakaria, N., & Maarof, M. (2018). Implementation and recent progress in cloud-based smart home automation systems., 71-77. https://doi.org/10.1109/iscaie.2018.8405447

Khan, M., Seo, J., & Kim, D. (2020). Toward energy efficient home automation: a deep learning approach. *Sensors*, *20*(24), 7187. https://doi.org/10.3390/s20247187

Kim, H.-W., Chan, H. C., & Gupta, S. (2007). Value-based adoption of mobile internet: An empirical investigation. *Decision Support Systems*, *43*(1), 111–126.

Lashkari, B., Chen, Y., & Musílek, P. (2019). Energy management for smart homes—state of the art. *Applied Sciences*, *9*(17), 3459. doi:10.3390/app9173459

Lau, J., Lau, J., & Harun, Z. (2022). Factors influencing the intention to adopt smart home technology among households in johor bahru, malaysia. *International Journal of Academic Research in Business and Social Sciences*, *12*(12). https://doi.org/10.6007/ijarbss/v12-i12/16016

Li, J. (2022). Research on the impact of smart home service innovation on consumers' willingness to use from the perspective of privacy concerns. *BCP Business & Management*, *29*, 204-217. doi:10.54691/bcpbm.v29i.2272

Li, W., Yiğitcanlar, T., Erol, I., & Liu, A. (2021). Motivations, barriers and risks of smart home adoption: from systematic literature review to conceptual framework. *Energy Research & Social Science*, *80*, 102211. doi:10.1016/j.erss.2021.102211

Lin, K.-Y., & Lu, H.-P. (2011). Why people use social networking sites: An empirical study integrating network externalities and motivation theory. *Computers in Human Behavior*, *27*(3), 1152–1161.

Liu, Y., Gan, Y., Song, Y., & Liu, J. (2021). What influences the perceived trust of a voice-enabled smart home system: an empirical study. *Sensors*, *21*(6), 2037. https://doi.org/10.3390/s21062037

Luor, T., Lu, H., Yu, H., & Lu, Y. (2015). Exploring the critical quality attributes and models of smart homes. *Maturitas*, *82*(4), 377-386. https://doi.org/10.1016/j.maturitas.2015.07.025

Mamonov, S., & Benbunan-Fich, R. (2020). Unlocking the smart home: exploring key factors affecting the smart lock adoption intention. *Information Technology and People*, *34*(2), 835-861. https://doi.org/10.1108/itp-07-2019-0357

Marikyan, D., Papagiannidis, S., & Alamanos, E. (2019). Smart home technology acceptance: an empirical investigation., 305-315. https://doi.org/10.1007/978-3-030-29374-1_25

Maswadi, K., Ghani, N., & Hamid, S. (2022). Factors influencing the elderly's behavioral intention to use smart home technologies in Saudi Arabia. *Plos One*, *17*(8), e0272525. https://doi.org/10.1371/journal.pone.0272525

Maznah, M., Najwa, M., Kamaliah, M., Jeffery, L., Sahithi, A. & Preece, C. (2021). Sustainable townships and sustainable homes: public perceptions. *Journal of the Society of Automotive Engineers Malaysia*, *5*(3), 331-347. https://doi.org/10.56381/jsaem.v5i3.176

Mennicken, S. & Huang, E. (2012). Hacking the natural habitat: an in-the-wild study of smart homes, their development, and the people who live in them., 143-160. doi:10.1007/978-3-642-31205-2_10

Nascimento, D., Tortorella, G., & Fettermann, D. (2022). Association between the benefits and barriers perceived by the users in smart home services implementation. *Kybernetes*, *52*(12), 6179-6202. https://doi.org/10.1108/k-02-2022-0232

Nepomuceno, M. V., Laroche, M., & Richard, M.-O. (2014). How to reduce perceived risk when buying online: The interactions among intangibility, product knowledge, brand familiarity, privacy and security concerns. *Journal of Retailing and Consumer Services*, *21*(4), 619–629.

Nikou, S. (2019). Factors driving the adoption of smart home technology: an empirical assessment. *Telematics and Informatics*, *45*, 101283. https://doi.org/10.1016/j.tele.2019.101283

Orlov, A., Saxena, A., Mittal, A., Ranjan, R., Singh, B., & Yellanki, V. (2024). User satisfaction and technology adoption in smart homes: a user experience test. *Bio Web of Conferences*, *86*, 01087. https://doi.org/10.1051/bioconf/20248601087

Paetz, A., Dütschke, E., & Fïchtner, W. (2011). Smart homes as a means to sustainable energy consumption: a study of consumer perceptions. *Journal of Consumer Policy*, *35*(1), 23-41. doi:10.1007/s10603-011-9177-2

Pal, D., Funilkul, S., Vanijja, V., & Papasratorn, B. (2018). Analyzing the Elderly Users' Adoption of Smart-Home Services. *in IEEE Access*, *6*, 51238-51252. DOI: 10.1109/ACCESS.2018.2869599

Patskanick, T., Cerino, L., Ashebir, S., FakhrHosseini, S., D'Ambrosio, L. & Coughlin, J. (2024). Aging in a smart home? the oldest olds' attitudes toward technology for aging-in-place. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *68*(1), 9-12. https://doi.org/10.1177/10711813241260397

Phan, L. & Kim, T. (2020). Breaking down the compatibility problem in smart homes: a dynamically updatable gateway platform. *Sensors*, *20*(10), 2783. https://doi.org/10.3390/s20102783

Pliatsikas, P. & Economides, A. (2022). Factors influencing the intention of Greek consumers to use smart home technology. *Applied System Innovation*, *5*(1), 26. https://doi.org/10.3390/asi5010026

Qiu, M., Najm, E., Sharrock, R., & Traverson, B. (2022). Reinforcement learning-based architectures for dynamic generation of smart home services., 7-14. https://doi.org/10.1109/icmla55696.2022.00010

Rumsey, D. J. (2016). *Statistics for dummies (2nd ed.)*. Wiley.

Sepasgozar, S., Karimi, R., Farahzadi, L., Moezzi, F., Shirowzhan, S., Ebrahimzadeh, S., ... & Aye, L. (2020). A systematic content review of artificial intelligence and the internet of thing applications in the smart home. *Applied Sciences*, *10*(9), 3074. https://doi.org/10.3390/app10093074

Sequeiros, H., Oliveira, T., & Thomas, M. (2021). The impact of iot smart home services on psychological well-being. *Information Systems Frontiers*, *24*(3), 1009-1026. doi:10.1007/s10796-021-10118-8

Șerban, A. & Lytras, M. (2020). Artificial intelligence for the smart renewable energy sector in europe—smart energy infrastructures for next-generation smart cities. *IEEE Access*, *8*, 77364-77377. https://doi.org/10.1109/access.2020.2990123

Shih, T. (2013). Determinates of consumer adoption attitudes. *International Journal of E-Adoption*, *5*(2), 40-56. https://doi.org/10.4018/jea.2013040104

Shin, J., Park, Y., & Lee, D. (2018). Who will be the smart home users? An analysis of adoption and diffusion of smart homes. *Technological Forecasting and Social Change*, *134*, 246-253, https://doi.org/10.1016/j.techfore.2018.06.029.

Singh, M. & Dhablia, A. (2023). Smart home automation using iot: prototyping and integration of home devices. *RJCSE*, *4*(2), 130-143. https://doi.org/10.52710/rjcse.83

Statista (2022). *Smart home device ownership in Turkey in 2022*. https://www.statista.com/forecasts/1003015/smart-home-device-ownership-in-Turkey. AD: 25.02.2023.

Stecuła, K., Wolniak, R., & Grebski, W. (2023). Ai-driven urban energy solutions—from individuals to society: a review. *Energies*, *16*(24), 7988. https://doi.org/10.3390/en16247988

Strzelecki, A., Kolny, B., & Kucia, M. (2024). Smart homes as catalysts for sustainable consumption: a digital economy perspective. *Sustainability*, *16*(11), 4676. doi:10.3390/su16114676

Tang, R. & Inoue, Y. (2021). Services on platform ecosystems in the smart home 2.0 era: elements influencing consumers' value perception of smart home products. *Sensors*, *21*(21), 7391. https://doi.org/10.3390/s21217391

Tarhouni, M. & Aloui, I. (2024). Toward smart home automation based on containerization. https://doi.org/10.21203/rs.3.rs-4761820/v1

Teng, W. & Lu, H.-P. (2009). Consumer adoption of PDA phones in Taiwan. *International Journal of Mobile Communications*, *8*(1), 1–20.

Tiersen, F., Batey, P., Harrison, M., Naar, L., Serban, A., Daniels, S., ... & Calvo, R. (2021). Smart home sensing and monitoring in households with dementia: a user-centered design approach. *Jmir Aging*, *4*(3), e27047. https://doi.org/10.2196/27047

Tuna, F. (2016). *Sosyal Bilimler İçin İstatistik*. Pegem Akademi,

Tural, E., Lu, D., & Cole, D. (2021). Safely and actively aging in place: older adults' attitudes and intentions toward smart home technologies. *Gerontology and Geriatric Medicine*, 7. https://doi.org/10.1177/23337214211017340

TÜİK. (2022). *TÜİK Veri Portalı*. https://data.tuik.gov.tr. Last Access: 01.01.2025.

Türkyılmaz, S. & Altındağ, E. (2022). Analysis of smart home systems in the context of the internet of things in terms of consumer experience. *International Review of Management and Marketing*, *12*(1), 19-31. doi:10.32479/irmm.12709

Uddin, M., Kim, T., & Kim, J. (2010). Video-based indoor human gait recognition using depth imaging and hidden markov model: a smart system for smart home. *Indoor and Built Environment*, *20*(1), 120-128. https://doi.org/10.1177/1420326x10391140

Van der Heijden, H. (2004). User acceptance of the hedonic information systems. *MIS Quarterly*, *28*(4), 695–704.

Van Dijk, J. (2020). *The Digital Divide: Bridging the Gap*. Cambridge University Press.

Wang, Y., Zhang, R., Zhang, X., & Zhang, Y. (2023). Privacy risk assessment of a smart home system based on the stpa–fmea method. *Sensors*, *23*(10), 4664. https://doi.org/10.3390/s23104664

Wei, N., Baharudin, A., Hussein, L., & Hilmi, M. (2019). Factors affecting user's intention to adopt smart home in malaysia. *International Journal of Interactive Mobile Technologies (IJIM)*, *13*(12), 39. https://doi.org/10.3991/ijim.v13i12.11083

Wei, W., Gong, X., Li, J., Tian, K., & Xing, K. (2023). A study on community older people's willingness to use the smart home—an extended technology acceptance model with intergenerational relationships. *Frontiers in Public Health*, 11. https://doi.org/10.3389/fpubh.2023.1139667

Wong, J. & Leung, J. (2016). Modeling factors influencing the adoption of smart-home technologies. *Facilities*, *34*(13/14), 906-923. https://doi.org/10.1108/f-05-2016-0048

Wright, D. & Shank, D. (2019). Smart home technology diffusion in a living laboratory. *Journal of Technical Writing and Communication*, *50*(1), 56-90. https://doi.org/10.1177/0047281619847205

Yaldaie, A., Porras, J., & Drögehorn, O. (2023). This state of home automation: a systematic literature review. *International Journal of Hybrid Innovation Technologies*, *3*(1), 23-46. https://doi.org/10.21742/ijhit.2653-309x.2022.2.1.03

Yang, H., Lee, H. & Zo, H. (2017). User acceptance of smart home services: an extension of the theory of planned behavior. In*dustrial Management & Data Systems*, *117*(1), 68-89. https://doi.org/10.1108/IMDS-01-2016-0017

Yang, Y., Liu, Y., Li, H. & Yu, B. (2015). Understanding the perceived risks in mobile payment acceptance. *Industrial Management & Data Systems*, *115*(2), 253–269.

Yazıcıoğlu, Y., & Erdoğan, S. (2014). *SPSS Uygulamalı Bilimsel Araştırma Yöntemleri*. Detay Yayıncılık.

Yu, J., Lee, H., Ha, I. & Zo, H. (2017). User acceptance of media tablets: An empirical examination of perceived value. *Telematics and Informatics*, *34*(4), 206–223.

Zhang, N., Wang, S., & Li, H. (2024). Improving user satisfaction by analyzing users' subjective cognitive types in smart home systems. *Universal Access in the Information Society*. doi:10.1007/s10209-024-01105-2

Zhang, Q., Li, M., & Wu, Y. (2020). Smart home for elderly care: development and challenges in China. *BMC Geriatrics*, *20*(1). https://doi.org/10.1186/s12877-020-01737-y

Zhou, C., Qian, Y., & Kaner, J. (2024). A study on the smart home use intention of elderly consumers based on technology acceptance models. *Plos One*, *19*(3), e0300574. https://doi.org/10.1371/journal.pone.0300574

Ziamou, P. L. & Ratneshwar, S. (2002). Promoting consumer adoption of high-technology products: Is more information always better? *Journal of Consumer Psychology*, *12*(4), 341–351.

İSTANBUL UNIVERSITY PRESS

Acta Infologica,
ACIN 2025, 9 (1): 339–358

https://doi.org/10.26650/acin.1540759
Submitted: 29.08.2024
Revision Requested: 19.02.2025
Last Revision Received: 02.04.2025
Accepted: 09.05.2025
Published Online 23.06.2025

Acta Infologica

Research Article

Open Access

Data Governance for Businesses: Challenges, Recommendations, and Critical Success Factors

Vildan Ateş [1] & Atacan Garip [2]

[1] Ankara Yıldırım Beyazıt University, School of Business, Department of Management Information Systems, Ankara, Türkiye
[2] DATAMIND, İstanbul, Türkiye

Abstract Data governance is the set of roles, processes, policies, and tools that ensure data quality throughout its lifecycle and appropriate usage across an organization. Data governance (DG) empowers users to locate, prepare, utilize, and disseminate dependable datasets independently. This study aims to reveal the challenges that businesses may encounter in the data governance process, propose recommendations, and highlight the critical success factors (CSFs) for data governance. The study's method is qualitative research. Interviews were conducted with 12 experienced volunteer participants in data governance. The data collected from the interviews were analyzed using descriptive and content analysis. The study revealed that the main challenges encountered were a lack of understanding of the impact of the data governance process on other business processes, the complexity of the process, and the role of the process manager. The study revealed that these challenges can be mitigated with effective communication, a clear demonstration of the contribution of the data governance process to the business, and leadership support. Furthermore, the study identified 20 critical success factors for DG success. These CSFs are classified into four categories: organization, people, data and technology, and regulations. As a result of the study, suggestions for businesses, managers, and individuals regarding data governance are presented.

Keywords Data Governance · Data Governance Challenges · Recommendations for Data Governance · Data Governance Success' Critical Success Factors · Qualitative Research Method

Author Note This article is based on the master's thesis titled "Identification of challenges and critical success factors encountered by businesses in the data governance process," written by Atacan Garip in 2024 under the supervision of Assoc. Prof. Dr. Vildan Ateş at the Department of Management Information Systems, Ankara Yıldırım Beyazıt University.

Citation: Ateş, V. & Garip, A. (2025). Data governance for businesses: Challenges, recommendations, and critical success factors. *Acta Infologica*, *9*(1), 339-358. https://doi.org/10.26650/acin.1540759
This work is licensed under Creative Commons Attribution-NonCommercial 4.0 International License.
© 2025. Ateş, V. & Garip, A.
Corresponding author: Vildan Ateş vates@aybu.edu.tr

Introduction

Data governance (DG) involves making decisions and exercising authority over data-related matters. It includes establishing and applying policies, procedures, and standards to ensure proper management and use of an organization's data assets. This includes defining roles and responsibilities for managing data, setting benchmarks for data quality, and complying with regulatory obligations (Khan et al., 2018; McMahon & Howe, 2018). DG is the set of decisions, processes, standards, policies, technologies, and other resources necessary to effectively manage, maintain, and utilize information as an organizational asset (Cheong, & Chang, 2007). In another definition, data governance is the process by which a company manages the quantity, consistency, availability, security, and accessibility of its data (Newman & Logan, 2006). Effective DG requires a holistic approach that encompasses both information technology (IT) and DG. The governance of IT provides the decision-making framework that links IT strategy and operations with organizational goals and objectives. On the other hand, DG ensures that data assets are effectively managed and utilized, focusing on meeting the needs and expectations of data consumers. Organizations must prioritize data quality management and consider data quality as a strategic asset. This includes establishing policies, processes, and tools to manage the data quality, concerning meeting the needs and expectations of data consumers. By doing so, organizations can ensure that their data is accurate, complete, consistent, and timely, leading to improved decision-making, increased productivity, and better business outcomes.

Effective DG is crucial for organizations to realize their strategic targets and maintain a competitive advantage in a data-driven business environment. The absence of effective data governance can result in the isolation of data in silos, data inconsistencies, and compliance issues (Oshiro, 2020). Therefore, organizations must develop and implement a comprehensive data governance framework that addresses critical success factors, like quality, ownership, and data security (Khan et al., 2018). Organizations must prioritize data governance to guarantee that data assets are effectively managed and used in support of business objectives. Achieving this requires a thorough data governance framework that addresses the key factors for success. It is crucial to recognize that data governance is a continuous process requiring ongoing monitoring and enhancement to adapt to evolving regulatory demands and business needs.

A well-designed data governance structure is crucial for organizations to manage and use their data assets effectively, and this requires delineating clear roles and responsibilities, emphasizing collaboration, and communication. To achieve this, organizations should establish a comprehensive data governance structure, consisting of a council for strategic decision-making, stewards for data quality management and meeting data consumer expectations, custodians for managing technical aspects such as storage and security, and an office for overseeing the entire data governance program in alignment with organizational objectives. By implementing such a structure and ensuring that all stakeholders comprehend their obligations, organizations can optimize data management to enhance decision-making, heightened efficiency, and better business outcomes. In the evolving landscape of organizational data management, the effective implementation of DG programs has emerged as a critical concern (Alhassan et al., 2019).

The literature review reveals a limited number of studies on data governance in Turkey. The first study was conducted by Kayrak in 2013, titled "Information Technology Governance." The study's findings indicated that organizations with effective IT governance, able to respond to corporate needs, will gain a competitive advantage over other institutions within their sector. Another study is a master's thesis, in which Karındaş (2019) proposed a data governance framework for the banking sector. The study aims to inform those who want to learn how to follow a path and what to do in the data governance creation and implementation

processes. A recent master's thesis on the banking sector used a special "Data Management Scale" survey to assess a bank's data governance practices and its employees' perceptions. The study found that the Kuveyt Turk Participation Bank has a solid foundation in data management, but that it needs to keep developing and updating these practices in response to the changing technological and regulatory environment (Çiçek, 2024). It is also observed that there is a lack of research in the literature on the key factors that ensure the success of data governance in Turkey. This study aims to identify the potential challenges and CSFs for implementing data governance programs in businesses. By identifying CSFs and barriers, this research advances the academic understanding and provides practical insights for organizations seeking to succeed in the complex world of data governance. The findings of this study are significant for businesses seeking to develop their data governance strategies and align them with ever-evolving technological and organizational demands.

The study has four main research questions, which are as follows:

1. What are the participants' views on data governance and its constituent elements?
2. What challenges are encountered in the data governance process?
3. What are the CSFs in implementing a data governance program?
4. What are the recommendations for successfully implementing a data governance program?

The scarcity of studies on CSFs of data governance in the foreign literature and the significant gap in Turkish research in this area highlight the novelty of this study. This research bridges this gap by combining information collected from interviews with professionals working in this sector with the researcher's own experience and interpreting it. This study is expected to be a valuable resource for businesses or individuals planning to implement data governance in the future. It is also intended to help organizations anticipate potential challenges and implement proactive measures to address them effectively. The study also aims to contribute significantly to the literature in this area by providing valuable suggestions for institutions, individuals, and managers.

Literature Review

Critical success factors are essential elements in nearly every business line, facilitating success. In other words, the critical success factor (CSF) can be defined as the key strategies that must be considered to achieve success. The determination of CSFs is the responsibility of the organization's managers or the initiative. Managers help determine the factors organizations need to achieve their goals and strategic plans. Managers consider the most important factors in the organization's operation when determining the critical success factors. Similarly, it is essential to understand the critical success factors of the data governance process, which is a crucial aspect of business operations in the modern era.

As a result of the literature review, it was seen that the critical success factors for the successful implementation of data governance studies included in the literature are as follows: roles and responsibilities (Alhassan et al., 2019a; Mahanti, 2018; Panian, 2010), competencies (Alhassan et al., 2019a; Mahanti, 2018), partnership (Dember, 2006; Ladley, 2012; Luftman, 2004; Mahanti, 2018; Thomas, 2005; Wong et al., 2022), change management (Ladley, 2012; Mahanti, 2018), sponsorship (Mahanti, 2018; Panian, 2010), approach (Mahanti, 2018), KPI and tracking (Mahanti, 2018), business strategy (Mahanti & Mahanti, 2021), and technology (Alhassan et al., 2019a; CDI Institute, 2006; Khatri & Brown, 2010; Larkin, 2008; Panian, 2010; Reeves & Bowen, 2013; Tallon et al., 2013; Tallon et al., 2013; Watson et al., 2004). As a result, when the related studies were evaluated, five basic critical success factors for data governance emerged. They are senior

management support, roles and responsibilities, sponsorship, partnership, and technology. One of the most important factors is senior management support and involvement, as without it, the program is unlikely to succeed. As Sun and Zhao (2015) noted, "*Senior management support is essential to the success of data governance. Without it, the data governance program is unlikely to have the necessary resources, funding, and visibility to make a significant impact*" (p. 56). Another critical factor is the establishment of clear policies and procedures for data management, which must be communicated effectively to all stakeholders. As Kwon and Johnson (2013) pointed out, "*The development of clear policies and procedures is important to ensure that data is managed effectively and efficiently and that all stakeholders understand their roles and responsibilities in the data governance process*" (p. 23). In addition, the availability of high-quality data is also crucial to the success of data governance initiatives. As stated by Strong et al. (2014), "*Data quality is the foundation of data governance. Without it, decisions will be based on inaccurate or incomplete information, which can lead to serious consequences*" (p. 19). The comprehension and acknowledgment of data governance's business value by senior-level management is imperative for the success of such initiatives. Data governance programs must originate and receive backing from the upper echelons of management within business units, as these units typically bear the responsibility for data governance. Senior executives must perceive data not just as a resource but as a strategic asset, warranting their sponsorship and active support for data governance programs (Panian, 2010; Soini, 2024).

Sponsorship is a critical component of DG as another CSF, as companies and government organizations strive to achieve their data governance goals. The sponsor plays a vital role in the board of directors for any project, program, or portfolio. The sponsor is responsible for ensuring that the work is managed effectively and that the objectives are delivered in a manner that meets the identified needs. Additionally, the sponsor plays a critical role in DG, specifically in achieving the data governance objectives of the business. Their primary responsibility within the DG process is to align it with the business goals, strategy, and objectives and to ensure that it is initiated and executed in a manner that aligns with these goals. They also manage risks and changes while maintaining the security of the DG process (Mahanti, 2018).

Another critical success factor for data governance is clear roles and responsibilities. To enhance accountability and operational efficiency in DG endeavors, it is advisable to establish and communicate roles and responsibilities from the outset, as DG activities encompass decision-making and design processes before technical operations (Alhassan et al., 2019a). Wende and Otto (2007) pointed out that defined roles and responsibilities for data governance are at the heart of successful data governance. It is essential to delineate roles, responsibilities, and decision-making authority across various levels, from operational to executive, concerning business, organizational, and technical facets, ensuring proper segregation of duties is established (Mahanti, 2018). Data governance impacts organizational operations, responsibilities, and activities beyond technical aspects; therefore, the human factor is paramount. It is crucial to define procedures, roles, and responsibilities related to data at all levels within the organization and ensure informed participation. Identifying individuals responsible for day-to-day data quality issues or granting strategic authorization for system acquisitions is critical in both strategic and operational data governance activities (Alhassan et al., 2019a). Data stewards and owners emerge as catalysts for driving change, assuming pivotal roles encompassing data quality, data dictionary management, change management, and more. Given that data governance includes a range of roles and responsibilities, it requires a diverse mix of skills, expertise, and knowledge. This diversity is crucial for the successful implementation of data governance initiatives. Essential skills include soft skills such as communication, persuasion, and organizational abilities, as well as technical skills like understanding of data, data quality, and data lineage (Mahanti, 2018). Employee data

competencies are indeed regarded as CSF because of the direct role employees play throughout the data lifecycle, which begins with data entry and encompasses various data governance activities.

Since data governance spans multiple areas within the organization, partnerships between different teams can be another critical success factor. Business units have the responsibility for data, making the management of data governance programs primarily the responsibility of business units and not IT units themselves (Thomas, 2005). IT units, on the other hand, are responsible for governing the IT infrastructure that supports the business's objectives (Luftman, 2004). Thus, a partnership between these sides is necessary to achieve the organization's goals (Dember, 2006). Data governance applies to all business units within the enterprise that interact with data, including those involved in inputting, managing, storing, and analyzing it. To achieve success, stakeholders from all these areas must be engaged and persuaded of the benefits that data governance can offer, which will help minimize resistance to change (Mahanti, 2018). For successful data governance activities, the involvement of teams from different backgrounds, their collaborative efforts, and progress aligned with set targets are crucial, in addition to the framework of roles and policies (Wong et al., 2022). Data stewardship plays a vital role in data governance implementation, as it brings together team members from diverse backgrounds and must be established to go beyond the framework. Therefore, fostering partnerships is crucial for effective data governance (Wong et al., 2022).

Another critical success factor is technology, which plays a key role in implementing data governance programs, particularly in areas such as data integration, reporting, and data quality (Cohen, 2006). These subparts require a technological perspective in their execution. Data governance encompasses technology and human aspects, as each component is interconnected. The implementation of data governance aims to enhance decision-making mechanisms, improve the efficiency of data-related operations, ensure compliance with data regulations, and promote data literacy. To achieve these goals, technology should be widely utilized, as it is inherently linked to the data objectives. Organizations, as noted by Alhassan et al. (2019a), invest in suitable IT infrastructure for data governance implementation. Indeed, there is a wide array of data governance tools available, and the selection of these tools should be based on their inherent capabilities, their ability to integrate with existing systems, and their suitability for meeting specific business needs and cases. Notable examples of companies offering diverse data governance tools include Collibra, Informatica, IBM, Oracle, SAP, and Microsoft, each with their unique offerings in this domain (Mahanti, 2018).

Apart from implementing technology in the implementation of data governance, tracking the measurement of the whole program plays a crucial role in its success. Measurement is significant for assessing the effectiveness of data governance implementation and its management. Although capturing and reporting metrics may seem challenging, it is important to define and establish a process for this purpose (Mahanti, 2018). The application approach is crucial. Data governance, as a living process and approach, should have a long-term structure that changes and updates over time. Thus, the data governance application becomes a repetitive structure and easily escapes the "one-time project" structure. This increases the success of data governance implementation (Mahanti, 2018). Therefore, these CSFs should be carefully considered and addressed in any data governance program to ensure its success.

A review of the literature shows a lack of studies on data governance in Turkey, with a few exceptions (Çiçek, 2024; Karındaş, 2019; Kayrak, 2013). There is also a lack of research on the CSFs of data governance in Turkey. The literature review revealed a gap in the research on data governance in the Turkish business context. This study addresses this gap by investigating the challenges, solution suggestions, and critical

success factors faced by businesses in Turkey in the data governance process. This study will guide businesses that either have implemented or will implement data governance programs in Turkey.

Methods

In this section of the study, the research method, participants, the data collection tool, the data collection process, the data analysis, and the validity and reliability of the study are presented.

Research Method

The qualitative research method was used in this research. The objective of utilizing this methodology is to obtain insights into the experiences and challenges encountered by individuals during their involvement in data governance processes and to identify the critical success factors that contribute to effective data governance practices. This study, which tries to understand the perspectives of the employees working in the different sectors to determine their experiences and challenges, is phenomenological research, one of the qualitative research methods designed and aims to reveal the essence of the person's perception of their experiences.

Participants

The participants in the study were selected using criterion sampling, one of the purposeful sampling methods. Purposive sampling is a random sampling method in which the sample group is targeted to have specific characteristics. This method can be used in many populations, but is more effective with a smaller sample size and a more homogeneous population (Patton, 2002). The purposive sampling method was preferred because researchers had a clear idea about the characteristics they wanted to examine and because they wanted to select a sample that represented these characteristics. The participants in the study were individuals with professional experience who had taken part in data management programs/projects. The criterion sampling method was preferred to determine a specific importance criterion. The study's criteria were that the participants were required to have participated in a data governance program's implementation and to reside in Istanbul. The study's eligible participants are 12 employees of eight firms in Istanbul, namely automotive, insurance, consulting, banking, and energy production, who have been involved in the data governance program implementation.

The participants in this study were voluntarily involved in the research. Before the interviews, the participants were provided with comprehensive information regarding the study's objectives and potential benefits. Each participant was asked to sign an informed consent form, indicating that they had chosen to participate voluntarily and without any external influence. Additionally, the form stated that their involvement in the study was entirely optional and that they could withdraw from the study at any time without consequence. No personal information was requested from the participants. Each individual was assigned a code, ranging from P1 to P12.

Data Collection Instrument

To gather data for this study, a semi-structured interview form was developed as the primary data collection instrument. Interviews are the most commonly used data collection tool in qualitative research. Patton (1997) emphasizes that the purpose of the interview is to explore an individual's inner world and understand their perspective (Patton, 1997. p. 108). Bailey (2007) outlined the following strengths of the

interview method: flexibility, high response rate, non-verbal behaviors, control of the environment, question order, immediate responses, completeness of information, and in-depth insights.

The draft version of the interview form was created based on the review of relevant literature and the researcher's own experience. The draft form was updated based on feedback from five different experts to create the final version. To make the interview form more efficient and consistent, expert opinions were sought from professionals who have worked in the field of data governance and gained experience in this domain across different companies and industries. These experts are highly educated in their respective fields, hold high-level positions in their organizations, and have multiple project experiences in the DG field. The experts' demographic information has been presented in Table 1.

Table 1

Demographic information of the experts

Expert Code	Gender	Age	Education Level	Sector	Working Experience (years)	Data Governance Experience (years)
E1	Male	33	Master's degree	Automotive, Banking	11	4
E2	Female	30	Master's degree	Automotive, Consultancy, and Banking	8	5
E3	Female	26	Bachelor's degree	Consultancy	3	3
E4	Male	32	Bachelor's degree	Consultancy, Insurance, Telecom	10	8
E5	Male	40	Doctoral Degree	Consultancy, Banking, Stock Market, University	18	9

Table 1 indicates that the experts have an average of 10 years of working experience and 5.8 years of experience in data governance.

The draft interview form consisted of 22 questions, including six questions about the participant's demographic information and 16 questions about the data governance domain. Table 2 shows the number of questions before and after the expert opinions.

Table 2

Number of questions before/after expert opinions

Section No	Section name	# of Questions before expert opinions	# of Questions after expert opinions
1	Demographic Information	6	7
2	Basics of DG	7	4
3	Challenges Encountered	5	3
4	Critical Success Factors and Suggestions	4	2

Upon examination of Table 2, a new question was added to the participant information section. However, it was also observed that questions were withdrawn from the other three sections. Furthermore, the experts were requested to provide an overall assessment of the interview form questions. The experts offered feedback that the interview form is "valuable, impressive, scope-expanding, and enriching." Before the expert opinions, the total number of questions was 22. Following the expert opinions, the total number of

questions decreased to 16. Seven questions sought demographic information about the participants, while 9 questions were about the data governance area (Table 2).

The questions in the interview form have three main sources: literature, experts, and researchers. Table 3 illustrates the distribution of questions according to their sources.

Table 3
Source distribution of the questions

Section No	Section name	# of Questions	# of Questions from the Literature	# of Questions from Experts	# of Questions from Researchers
Section 1	Demographic Information	7	4	1	2
Section 2	Basics of DG	4	2	1	1
Section 3	Challenges Encountered	3	1	1	1
Section 4	Critical Success Factors and Suggestions	2	1	1	-
Total number of questions: 16					

The interview form comprises four sections (Table 3). The first section of the interview form consists of seven questions about the participants' demographic information (age, gender, education level, area of expertise, employment sector, work experience, and experience in data governance). The second section contains four questions about the basics of DG. The third section has three questions regarding the challenges encountered with DG. Finally, the fourth and last section has two questions about the critical success factors and suggestions.

Data Collection Process

The data collection process for this study was approved by Ankara Yıldırım Beyazıt University Ethics Committee Approval (03-46), which was obtained from the Ethics Committee to conduct the research. The decision of the ethical committee was obtained on March 15, 2023/, with reference number (03-46).

The researcher conducted interviews from April 14, 2023, to May 30, 2023. Face-to-face interviews were organized and conducted with participants at prearranged locations, following the prior scheduling of appointments. Half of the interviews were held in online environments such as Zoom, Google Meet, and Microsoft Teams sessions because of the distance between the participant and the researcher. The interviews lasted an average of one and a half hours and were conducted at the participants' request. The participants were interviewed in a face-to-face setting in their respective offices in Istanbul, and online interviews were conducted via Zoom, Google Meet, and Microsoft Teams sessions.

Data Analysis

This study used descriptive and content analysis techniques to analyze the data collected from the interviews. Descriptive analysis is the summarization and interpretation of data within the framework of the determined themes (Patton, 2018; Yıldırım & Şimşek, 2016). A descriptive analysis comprises a summary of the data in question. It is based on predetermined themes, supported by direct quotes.

Content analysis consists of examining research data, generating codes from the data, determining themes, creating thematic networks, ensuring the integrity of the themes and interpreting them (Miles &

Huberman, 2016; Robson, 2017). The conceptual framework obtained from the literature review was used in the content analysis. Responses were grouped according to the framework, and various themes and categories were determined. Similar data were clustered to form specific concepts, which were then interpreted accordingly. In general, this analytical process facilitated the researcher to grasp the interview data more deeply and draw meaningful conclusions.

Validity and Reliability

To ensure the validity and reliability of the study, the following approaches were adopted:

- The preparation of the interview questions was based on individual experiences, which served as a foundation. Subsequently, the questions were enriched with an academic and literature-based perspective by examining commonly used frameworks in DG.
- To enhance the reliability and diversity of the interview questions, experts with varying levels of expertise from different sectors and backgrounds were consulted. Their feedback was incorporated, and necessary adjustments were made to the data collection process.
- The participants were randomly selected from different sectors and companies. The study prioritized validity and accuracy by including participants with various levels and disciplines, thus obtaining a representative and generalizable working group.
- The interviews were conducted online or in person based on the participant's preferences and convenience.
- Before the interviews, the participants were given information to ensure they understood the study. They also exhibited adherence to the principles of voluntariness and confidentiality.
- After the interviews, the interview transcripts, a document comprising their responses to each question, were shared with the participants to obtain confirmations to ensure accuracy and validation.

The participants in the study were thoroughly informed about the researcher's identity, the purpose of the research, how and where the information gathered will be used, and the voluntary nature of their participation. They were also informed that they had the right to withdraw from the study at any stage. This comprehensive approach ensures ethical rigor and informed consent in the research process.

Results

This section consists of two subsections. In the first subsection, the findings regarding the demographic characteristics of the participants are presented. In the second subsection, the participants' responses to the questions in the interview form were analyzed and presented.

Findings of the Demographic Information of the Participants

The demographic information (age, gender, education level, expertise area, sector of employment, working experience, and DG working experience) of the participants is presented in Table 4.

Table 4

Source distribution of the questions

ID	Age	Gender	Education Level	Area of Expertise	Employment	Working Experience (years)	DG Working Experience (years)
P1	33	Male	Master's degree	Math	Automotive	11	2
P2	38	Female	Bachelor's degree	Informatics	Finance	14	5
P3	25	Female	Bachelor's degree	Computer Eng.	Consultancy	2	2
P4	39	Female	Bachelor's degree	Math	Consultancy	15	4
P5	30	Female	Master's degree	Computer Eng.	Consultancy	8	5
P6	32	Male	Bachelor's degree	Computer Eng.	Consultancy	10	8
P7	32	Female	Master's degree	Math	Banking	8	2
P8	26	Male	Bachelor's degree	Computer Eng.	Consultancy	2	1
P9	40	Male	Doctoral Degree	Statistics	Consultancy	18	9
P10	28	Male	Bachelor's degree	Computer Eng.	Insurance	5	3
P11	44	Male	Master's degree	Data Analytics	Banking	20	5
P12	31	Female	Bachelor's degree	Computer Eng.	Energy	7	1

When analyzing the participants' demographic information (Table 4), it was observed that there was an equal distribution across the "gender" category among the participants. This indicates a balanced representation in terms of gender demographics. It is seen that most participants fall within the 30-39 age range, with a total of seven participants. This indicates a relatively higher representation of individuals in their thirties. Table 4 indicates that the participants' working experience is spread across different ranges. 33% of participants fall within the 5–10 years of experience category. This means a significant representation of individuals with mid-level experience. This comprehensive analysis highlights the multidimensional perspectives and insights provided by participants with varying combinations of working and data governance experience. When Table 4. is examined, it can be observed that most participants, accounting for 50% of the total, have 1-3 years of data governance experience. This indicates a significant representation of individuals with relatively little experience in the field. Additionally, the 4-6 years of experience category comprised 33% of the total participants, demonstrating a significant number of participants with mid-level experience. Lastly, the 7+ years of experience category is represented by 17% of the participants.

Findings of the Interview Questions

The second part of the interview form contains nine questions in three sections. These sections discuss the basics of data governance, challenges encountered, CSFs, and recommendations. The data collected from the participants will be classified according to these three sections, and the analysis results will be presented accordingly.

The second section of the interview form regarding the basics of data governance contains four questions. Participants were first asked for the definition of data governance. Three of the participants' definitions of data governance are as follows:

P1: *"The function carried out to manage, control, and ensure the correct use and sharing of an organization's data assets, thus optimizing business processes, is called Data Governance."*

P3: *"It determines the custodians of the data assets within an institution, ensures that the orphaned assets are appropriated by addressing them to the right people and points, determines the processes, standards, policies, and procedures used in the institution, ensures their traceability and sustainability, and also ensures decision-making on the relevant processes. It is the process cycle in which activities are carried out to ensure that business units can speak the same language."*

P10: *"Data governance encompasses the principles, procedures, and strategies to ensure that organizations can discover and track data, place data accurately in appropriate business processes, and maximize data security, quality, and value."*

Participants' definitions of data governance were examined, and the words used in these definitions and their frequency of use are presented in Table 5. Participants' responses were analyzed based on keywords.

Table 5

Data governance definition's word frequency

Keywords	Participants	Frequency
Management	P1, P4, P5, P7, P8, P10	6
Business Process	P1, P2, P6, P11, P12	5
Classification	P2, P9, P11, P6	4
Data asset	P1, P3, P7	3
Policy	P2, P3, P8	3
Procedure	P3, P9, P10	3
Efficiency	P6, P7	2
Quality	P7, P5	2

Table 5 illustrates that the three most frequently referenced terms in defining data governance among participants were: management, business process, and classification. On the other hand, the definitions were only referenced once in the terms "asset management," "strategy," "common understanding," and "framework." Upon evaluation of the definitions proposed by the participants, it was observed that they were largely compatible with one another. However, the degree of compatibility varied according to the participants' experiences and focus points.

In the second question of this section, participants were asked about the data governance subcomponents they were working on. Thus, this study aims to evaluate the different experiences and perspectives of the participants in this field. In this question, participants were allowed to highlight more than one data governance subcomponent. As a result of examining the subcomponents that the participants worked on in the field of data governance, it was seen that 11 participants worked on the sub-areas of metadata, data quality, and data ownership. These sub-components were followed by data compliance with nine participants and master data with five participants. Notably, the other four areas (content management, data security, data lineage, and data architecture) were addressed by the four participants.

In the third question, participants were asked to rank the subcomponents of data governance according to their priorities. The priority value ranges from 1 to 5 (1: highest priority, 5: lowest priority). Participants were asked to rank the options in this question.

Table 6

DG subcomponents rankings by priority

Subcomponent	1	2	3	4	5
Data architecture	5	3	-	-	4
Data compliance	2	3	4	-	3
Data lineage	2	2	3	4	1
Data ownership	6	1	1	1	3
Data quality	4	3	1	1	3
Data security	4	3	2	-	3
Document and content management	-	2	6	3	1
Metadata	3	4	1	-	4
Reference and master data	3	3	3	1	2

As seen in Table 6, data ownership ranked first, and six participants gave data ownership the highest priority. Data architecture ranked second, while data quality and data security ranked third.

In the fourth question of this section, participants were asked to rank the subcomponents of data governance according to their difficulty (1: highest difficulty level, 5: lowest difficulty level)

Table 7

DG subcomponents rankings by difficulty

Subcomponent	1	2	3	4	5
Data architecture	6	-	2	2	2
Data compliance	2	2	5	2	1
Data lineage	1	2	3	5	1
Data ownership	2	3	2	3	2
Data quality	-	6	2	2	2
Data security	3	6	-	1	2
Document and content management	1	2	3	5	1
Metadata	2	3	4	2	1
Reference and master data	2	2	6	2	-

When Table 7 data governance sub-dimensions are examined in terms of difficulty, it is seen that only data architecture comes to the fore. This is followed by data quality and data security. On the other hand, data quality was not mentioned by any participant at the highest difficulty level.

The third subsection of the interview form regarding data governance contains questions related to the challenges encountered. In this section, the participants were asked three questions. The first question of this section was asked the participants in an open-ended manner. Participants were asked to describe any challenges they encountered in implementing the data governance program and, if so, to explain them. Three participants (P2, P4, P6) stated they did not experience any challenges. Participant 1 also emphasized that no difficulties were encountered due to the support of senior management. Similarly, P11 stated that

as follows: "*Due to the comprehensive IT evaluation consultancy output, it was not difficult to convince the relevant people of the necessity of data governance studies.*" The challenges noted by the other participants are summarized in Table 8.

Table 8

Challenges encountered

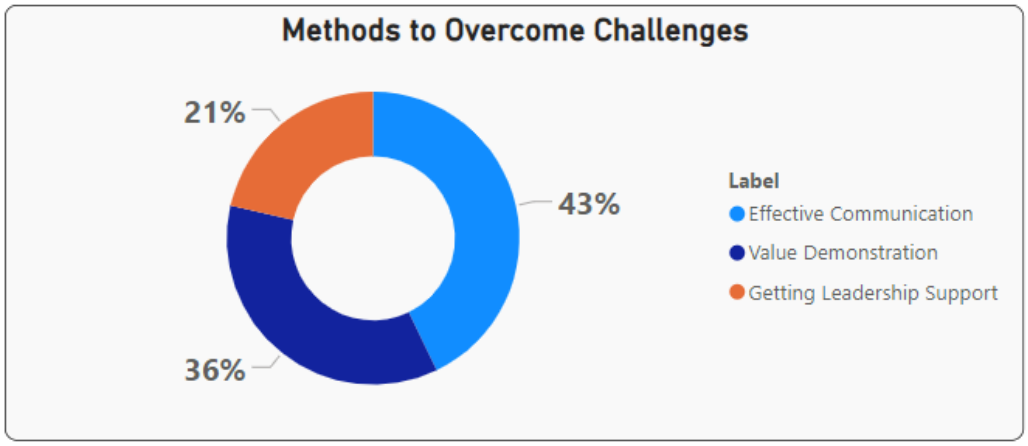
Challenges	Participants	Frequency
Business involvement	P3, P5, P6, P8, P10	5
Complexity	P8, P9, P10, P12	4
Ownership	P9, P12	3
Convince managers	P5, P7	2
Leadership support	P3, P8	2
Technology	P7, P10	2
Continuity	P8	1
DG Framework	P9	1

According to Table 8, the involvement of business units and complexity stand out as the most frequently encountered challenges by the participants.

In the second question of this section, participants were asked how they overcame the challenges they encountered. The participants' responses were analyzed and presented in Figure 1, along with the strategies employed to address these challenges.

Figure 1

Methods to overcome challenges



Upon examination of Figure 1, three methods emerged as effective communication, value demonstration, and leadership support. Of the participants, 43% indicated effective communication, 36% stated value, and 21% stated receiving leadership support as a method.

In the third question of this section, participants were asked to identify any potential challenges they may encounter in the data governance process in the future. They were also asked to provide solutions to these challenges. The response of P4 to this question is as follows: "*It usually starts as a program and stays in focus for a while. However, as time goes by, the flame of the program fades away, and it is observed that there is a sustainability problem. The data is defined and owned, but as time passes, it becomes difficult to keep this*

information alive because it is treated as a finite program." In response to this question, P10 provided the following answer: "*With the constant change of people within the data governance structure in business units, difficulties may arise in ensuring the adaptation of new colleagues.*" Participants indicated possible future challenges such as loss of motivation (P2, P6, P8, P12), security and control (P3, P5, P6, P11), sustainability (P1, P4, P7), staff turnover and shortage (P7, P9, P10), infrastructure and technological requirements (P7, P11), and compatibility (P2, P12). Participants also suggested the following solutions to these future challenges. The solutions suggested by the participants for the challenges that may be encountered in the future are presented in Table 9.

Table 9

The suggested solutions to future challenges

Solutions	Participants	Frequency
Benefit demonstration	P1, P3, P6, P10, P11, P12	6
Senior management support	P1, P2, P5, P9, P10, P12	6
Having good standards	P7, P8, P11, P12	4
Automate the processes	P4, P9, P11	3
Preventing personnel loss	P2, P5, P8	3
Detailed planning	P6, P7, P10	3
Dedicated team	P7, P8, P9	3
Training	P3, P4, P9	3

When the solution suggestions in Table 9 are examined, the first solution is to explain and demonstrate the benefits to be obtained from the DG process. The second solution suggestion expressed by the same number of participants is to ensure the support of the senior management. Standards are in third place, followed by automated processes, detailed planning, training, dedicated teams, and measures to prevent personnel losses.

The fourth section of the interview form regarding the data governance field contains questions related to critical success factors and suggestions. In this section, the participants were asked two questions. In the first question of this section, participants were asked to identify the critical success factors necessary for the successful execution of a data governance program. The following is a list of the CSFs cited by the participants:

- Awareness
- Budget
- Clear DG Processes
- Continuity
- Currentness
- Data management
- Data security
- Data Policies
- Linked DG with Business Results
- Training

- Technology
- Teamwork and collaboration
- Prioritization
- Regulations
- Responsibilities
- Roles
- Scope
- Senior Management Support
- Standardization
- Stewardship

The second question of this section is related to the recommendations. Participants asked for recommendations for a business that would start implementing a data governance program.

Table 10

Recommendations for businesses that will implement a data governance program

Recommendations	Participants	Frequency
Senior management support	All participants	12
Business requirements determination	P1-P10, P12	11
Technology investment	P1-P7, P10-P12	10
Data ownership	P1-P7, P9, P10	9
Key person identification	P1-P3, P5, P8-P12	9
Vision	P4-P7, P9, P10, P11	7
Determination of the KPIs	P6, P7, P9, P10, P11	5
Pilot group selection	P1, P2, P7, P10	4

All responses given by the participants were examined and presented in Table 10. When Table 10 is examined, it can be seen that senior management support, determination of business needs, and technology investment are the most frequently expressed suggestions. Pilot group selection was the least recommended suggestion by the participants.

In the third question of this section, participants were asked to score to what extent the importance of data management is understood by institutions in Turkey, with 1 being the lowest and 10 being the highest.

Table 11

Participants' scores regarding the importance of DG in Turkey

Score	Participants	Frequency
1	-	-
2	-	-
3	P1, P5, P7, P12	4
4	-	-
5	P2, P8, P11	3
6	P3, P4, P6, P10	4

Score	Participants	Frequency
7	P9	1
8	-	-
9	-	-
10	-	-

The scores given by the participants on how important the data governance process is for institutions are shown in Table 11. According to Table 11, none of the participants gave a score of 8 or above. On the other hand, it is seen that the participants did not give a score of two or below. When the average of the scores is taken, it is seen that it is 4.8.

Discussion and Conclusion

In this study, interviews were conducted with participants working in different sectors in Turkey and involved in the data governance process. Questions were asked about the four research questions of the study in the interviews, and the responses of the participants were analyzed.

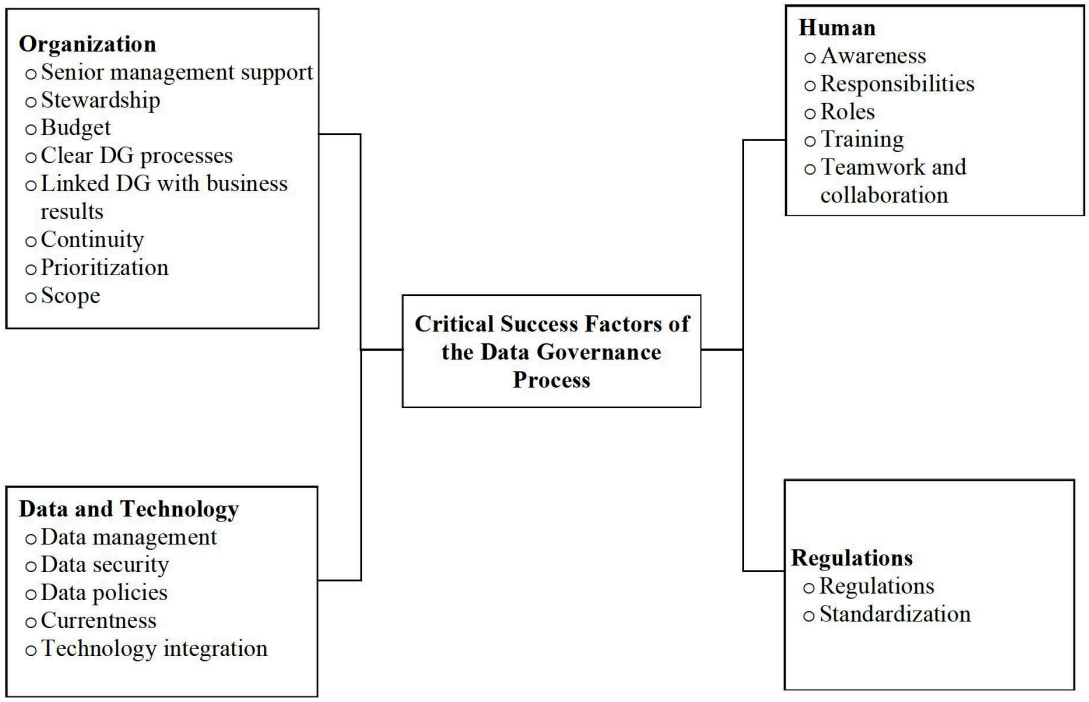
As a result of the study, it was seen that the data governance definitions of the participants were compatible with each other. The difficulties encountered in the data governance process were the participation of the DG process in other business processes, the complexity of the process, and the DG process manager. In addition to these difficulties, the persuasion of managers, senior management support, and technology-related problems were emphasized. It was revealed that these difficulties can be overcome with three methods: effective communication, demonstration of the contribution of the DG process, and leadership support.

Another study's finding is that the average score of the question of understanding the importance of data governance by institutions in Turkey is 4.8. This can be interpreted as the importance of DG being understood by 48%, which is close to the medium level. In other words, it can be said that the importance of data management is not fully understood by institutions in Turkey.

Because of the study, it was seen that there are 20 critical success factors affecting the DG process. These 20 CSFs were grouped under four categories: institutional, data, people, and regulations, and are presented in Fig. 2. When Fig. 2 is examined, it is seen that the categories are institution, data, regulations, and human. It is seen that the most critical success factors are gathered under the institutional category. Therefore, it is important for the institution to be ready for the data governance process and to provide full support. Another important category is humans. Only the managers' requests will not be enough in any process. All employees in the business must be ready for the DG process both mentally and competently. Another category is data and technology, and for the DG process to be successful, technology, data management, data security, data policies, and data to be up-to-date are also required. Another category is human, and no business process can be carried out without humans. Therefore, roles and responsibilities should be clearly defined and, if necessary, employees should be trained. In addition to all these, the process should be carried out by laws and regulations.

Figure 2

Categorization of the CSFs of DG



In parallel with the results of this study, roles, responsibilities, and competencies have been reported in the literature as critical success factors for data governance success (Alhassan et al., 2019a; Mahanti, 2018; Panian, 2010). Partnerships and collaborations between different teams, which are similar to the results of this study, have also been widely mentioned as a CSF for data governance practices and have been expressed in studies as partnership and collaboration (Dember, 2006; Ladley, 2012; Luftman, 2004; Mahanti, 2018; Thomas, 2005; Wong et al., 2022). In a related study, Alhassan, Sammon, and Daly (2019b) conducted a single case study with a major telecommunications service provider in Saudi Arabia to identify CSFs for data governance. The results demonstrated that a business-driven data integration strategy is the most critical success factor.

Senior management support has been reported as CSF by other researchers in the literature (Kwon & Johnson, 2013; Strong et al., 2014; Sun & Zhao, 2015). Senior management support was a critical success factor that was acknowledged by most participants. Senior management support is among the most important success factors in both initiating data governance efforts and successfully transforming the efforts into a sustainable format and disseminating them throughout the organization. The study results also showed that senior management support is the most common obstacle encountered in the DG process. A study with similar results was conducted by interviewing five data experts in the Finnish banking sector. The objective was to identify the critical success factors of DG in the banking sector. The CSFs identified were managing regulatory and financial risk, customer and stakeholder value and business value, senior management support, data ownership, data-centric culture, and business strategy (Soini, 2024). In a separate study, researchers examined the CSFs that affect the success of cloud data governance. The study identified 10 critical success factors for data governance, including strategic accountability, standards, administrative blind spots, embracing the complexities of data, cross-departmental issues, data quality metrics, partnership, strategic control points, education and awareness of data stakeholders, and compliance monitoring.

Although the study investigated cloud data governance CSFs, some CSFs were common to our study (Al-Ruithe, & Benkhelifa, 2017).

This study identified technology, data management, data security, data policies, and data updates as CSFs in the DG process. A literature review revealed that technology or data-related topics, including data quality, integration, data dictionaries, data life cycles, reporting, and root cause analysis, are also critical success factors. Additionally, the integration of technology with the institution's existing systems has been identified as a significant factor (Khatri & Brown, 2010; Panian, 2010; Reeves & Bowen, 2013; Tallon et al., 2013).

Another CSF category seen in this study was human. It is seen that there is no human category in the studies in the literature and this study brings a new perspective to the literature with this result. The human factor has always been a critical factor for the success of information systems. The results of this study also support this. The competencies of employees involved in the DG process are regarded as critical success factors (CSFs) due to the direct role they play in the data lifecycle, which begins with data entry and encompasses various data management activities. Managing tasks such as data entry and data access requires employees to have a basic skill set and a certain degree of awareness to properly handle the organization's data (Alhassan et al., 2019a).

The study participants also provided recommendations for businesses that will implement data governance in the future. These recommendations encompass the following: support from top management, identification of business needs, and investment in technology. It can be reasonably concluded that businesses that want a successful data governance process in the future should consider these three recommendations.

Considering the findings of the study, a series of recommendations can be put forth for consideration by those in managerial roles. It is recommended that senior management provide active support for data governance initiatives. This may entail providing the necessary resources and showing visible support for these endeavors. Defining roles and responsibilities and communicating them to employees is essential to prevent ambiguity and overlap. DG strategies should be aligned with business results, with tasks that provide business impact given the highest priority. A dedicated task force can be established for the DG. It is not advisable to attempt to implement all DG activities simultaneously; instead, a pilot study can be initiated. If necessary, the scope can be expanded to other areas with support. Since DG is a long-term process, it is essential to determine key performance indicators and to closely monitor progress to ensure long-term success.

The following recommendations are intended for organizations that are either implementing or planning to implement a data governance program. It is essential to allocate sufficient budgets and invest in appropriate technology to streamline data management processes. It is essential to develop and implement a long-term vision for success and business engagement within a framework that aligns with the organization's current business strategy. A culture of continuous improvement must be encouraged, and the DG process must be updated based on trends and evolving business needs. It is crucial to invest in the right team or people before embarking on data management initiatives. If necessary, outsource services to an experienced service provider with successful practices in a similar industry.

Those working in the field of data governance can benefit from several suggestions. First, they should be aware of the importance of this issue. Second, they should receive training to maintain their knowledge of the latest practices. Third, they should develop effective collaboration, teamwork and communication skills. Fourth, they should demonstrate the beneficial effects of the DG process to business units and senior

management regularly. Finally, they should receive senior management support to initiate and sustain the process over time.

This study has several limitations. It relies on a relatively small working group and focuses on a limited number of firms that may not fully represent the variety of challenges and practices in DG. To gain a more comprehensive understanding of the field, future researchers could conduct research with a larger, more diverse sample and examine additional aspects such as the impact of cultural differences on DG practices. They could also investigate the long-term effects of implemented DG strategies and their adaptation to rapidly changing technological environments.

Peer Review: Externally peer-reviewed.
Author Contributions: Conception/Design of Study- V.A., A.G.; Data Acquisition- A.G.; Data Analysis/Interpretation- V.A., A.G.; Drafting Manuscript- V.A., A.G.; Critical Revision of Manuscript- V.A.; Final Approval and Accountability- V.A.; Technical or Material Support- A.G.; Supervision- V.A.
Conflict of Interest: The authors have no conflict of interest to declare.
Grant Support: The authors declared that this study has received no financial support.

Author Details

Vildan Ateş
[1] Ankara Yıldırım Beyazıt University, School of Business, Department of Management Information Systems, Ankara, Türkiye
0000-0002-8855-8556 vates@aybu.edu.tr

Atacan Garip
[2] DATAMIND, İstanbul, Türkiye
0000-0001-8414-536X

References

Alhassan,I., Sammon, D., & Daly, M. (April, 2019a). Critical Success Factors for Data Governance: A Theory Building Approach. *Infomation System Management*, 1-15.

Alhassan, I., Sammon, D., & Daly, M. (2019b). Critical success factors for data governance: a telecommunications case study. *Journal of Decision Systems*, *28*(1), 41-61.

Al-Ruithe, M., & Benkhelifa, E. (2017). *Analysis and classification of barriers and critical success factors for implementing a cloud data governance strategy*. *Procedia Computer Science*, 113, 223-232.

Bailey, K. (2007). *Methods of Social Research* (Fourth Edition). New York Free Press, 68.

CDI Institute (2006). *Corporate Data Governance Best Practices – 2006-07 Scorecards for Data Governance in the Global 5000*, CDI Institute, Retrieved from http://www.tcdii.com/#.

Cheong, L. K., & Chang, V. (2007). *The need for data governance: a case study.* ACIS 2007 Proceedings, 100.

Dember, M. (2006). The 7 Stages of Highly Effective Data Governance: Advanced Methodology for Implementation. CIBER Whitepaper.

Khan, S., Khan, M. A., Asghar, M. J., & Ahmad, W. (2018). Critical success factors for implementing data governance. *IEEE Access, 6*, 22951-22968.

Khatri, V.& Brown, C. V. (2010). Designing data governance. *Communications of the ACM, 53*(1), 148–152.

Kwon, Y., & Johnson, M. E. (2013). The critical success factors for effective data governance. *Journal of Data and Information Quality (JDIQ),* 4(3), 12:1-12:22. https://doi.org/10.1145/2500423.2500425

Ladley, J. (2012). *Data governance: How to design, deploy, and sustain an effective data governance program*. Morgan Kaufmann Publishers.

Larkin, B. S. (2008). Increasing information integrity: Cultural impacts of changing the way we manage data. *International Journal of Organization Theory & Behavior (Pracademics Press), 11*(4), 558–578.

Luftman, J. (2004). *Managing the Information Technology Resource*. Pearson Prentice Hall, Upper Saddle River, New Jersey, USA.

Mahanti, R. (2018). Data Governance Implementation: Critical Success Factors. *Software Quality Professional*, *20*(4), 4-26.

Mahanti, R., & Mahanti, R. (2021). Strategy and Data Governance. *Data Governance Success: Growing and Sustaining Data Governance*, 67-98.

McMahon, P., & Howe, D. (2018). *Data governance: Creating value from information assets*. Apress.

Miles, M.B., & Huberman, A.M. (2016). *Nitel veri analizi* (S. Akbaba Altun ve A. Ersoy, Çev. Ed). Pegem Akademi.

Newman, D., & Logan, D. 2006. Governance Is an Essential Building Block for Enterprise Information Systems. Gartner Research, May. https://www.gartner.com/en/documents/492444

Oshiro, T. (2020). Why data governance is critical for successful digital transformation. Forbes. https://www.forbes.com/sites/tedoshiro/2020/05/04/why-data-governance-is-critical-for-successful-digital-transformation/?sh=1c8d78b518df

Panian, Z. (2010). Some practical experiences in data governance. World Academy of Science. *Engineering and Technology Management, 62*, 468–475.

Patton, M.Q. (1987). *How to use qualitative methods in evaluation* (Second edition). California: Newbury Park Sage Publication, 108.

Patton, M. Q. (2018). *Nitel araştırma ve değerlendirme yöntemleri*. (M. Bürün ve S.B. Demir, Çev. Ed.). Ankara: Pegem Akademi.

Patton, M. Q. (2002). *Qualitative research & evaluation methods*. (Third edition). Thousand Oaks. Sage Publications.

Reeves, M. G., & Bowen, R. (2013). Developing a data governance model in health care: although the term may be unfamiliar, data governance is a longstanding obligation of the healthcare industry. *Healthcare Financial Management*, *67*(2), 82-87.

Robson, C. (2017). *Bilimsel araştırma yöntemleri gerçek dünya araştırmaları* (Ş. Çımkır ve N. Demirkasımoğlu, Çev. Ed.). Ankara: Anı Yayıncılık.

Soini, T. (2024). Critical Success Factors of Data Governance in Finnish Banking Organizations. Master's Thesis in Information Systems Åbo Akademi University, Finland. https://www.doria.fi/handle/10024/189274

Strong, D. M., Lee, Y. W., Wang, R. Y., & Yang, S. B. (2014). Data quality in context. *ACM Journal of Data and Information Quality, 5*(4), 15:1-15:22.

Sun, Y.& Zhao, Y. (2015). A framework for data governance. *Journal of Industrial Engineering and Management, 8*(1), 55-71.

Tallon, P. P., Ramirez, R. V., & Short, J. E. (2013). The information artifact in IT governance: Toward a theory of information governance. *Journal of Management Information Systems, 30*(3), 141–178.

Tallon, P. P., Short, J. E., & Harkins, M. W. (2013). The evolution of information governance at Intel. *MIS Quarterly Executive, 12*(4), 189–198.

Thomas, G. (2005). *Alpha Males and Data Disaster*. Published by Brass Cannon Press, USA.

Watson, H. J., Fuller, C., & Ariyachandra, T. (2004). Data warehouse governance: Best practices at the Blue Cross and Blue Shield of North Carolina. *Decision Support Systems, 38*(3), 435–450.

Wende, K., & Otto, B. (2007). *A contingency approach to data governance*, 12th International Conference on Information Quality, Cambridge, USA, 163-176.

Wong, D., Maarop, N., & Samy, G. (8 May, 2022). Data Governance and Data Stewardship: A Success Procedure. 8th International Conference on Information Technology and Multimedia. 54-61.

Yıldırım, A., & Şimşek, H. (2016). *Sosyal bilimlerde nitel araştırma yöntemleri*. (10. baskı). Seçkin Yayıncılık.

İSTANBUL UNIVERSITY PRESS

Acta Infologica

Acta Infologica,
ACIN 2025, 9 (1): 359–384

https://doi.org/10.26650/acin.1612225
Submitted: 03.01.2025
Accepted: 13.02.2025
Published Online 24.02.2025

Review Article

Open Access

Engineering Education Meets the Metaverse

Veselina Nedeva [1], Snejana Dineva [1] & Svetoslav Atanasov [1]

[1] Trakia University, Faculty of Engineering and Technologies, Department of Electrical Engineering, Automation, Computer Systems and Communications, Yambol, Bulgaria

Abstract The metaverse, a collective virtual shared space created by the convergence of virtually enhanced physical reality and persistent virtual space, encompassing all virtual worlds, augmented realities, and the Internet, offers a unique platform for experiential learning and collaborative education. This study used the Web of Science, Google Scholar, and Scopus via Publish or Perish for source searching, with VOSviewer employed for bibliometric visualization. Following the PRISMA protocol, the study systematically reviewed relevant literature to address the following research questions: (1) What specific fields within engineering education incorporate Metaverse-based learning? (2) Which tools, technologies, or software platforms are employed in these practices? (3) What teaching strategies or educational frameworks are utilized in this context? (4) How do collaboration and interaction between humans and computers take place within the Metaverse? (5) What is the role of digital twins in Metaverse education? This review investigates the metaverse's role in engineering education, highlighting both its positive impacts and potential drawbacks.

Keywords Metaverse · Engineering Education · Virtual Reality · E-Learning

Author Note This paper is part of the research project *Metaverse Approach for a New Generation of Engineering Education and Collaborative Learning (MAGURA)*, conducted under Project No. BG-RRP-2.004-0006-C02, titled *Development of Research and Innovation at Trakia University in Service of Health and Sustainable Well-being*.

Citation: Nedeva, V., Dineva, S. & Atanasov, S. (2025). Engineering education meets the metaverse. *Acta Infologica*, *9*(1), 359-384. https://doi.org/10.26650/acin.1612225

This work is licensed under Creative Commons Attribution-NonCommercial 4.0 International License.

© 2025. Nedeva, V., Dineva, S. & Atanasov, S.

Corresponding author: Svetoslav Atanasov svetoslav.atanasov@trakia-uni.bg

Introduction

The concept of the Metaverse gained significant attention in 2021, with companies like Microsoft, Samsung, and ByteDance introducing Metaverse platforms. By bridging the gap between the physical and virtual worlds, the Metaverse overcomes the limitations of traditional 2D e-learning environments. It integrates technologies such as Virtual Reality (VR) and Augmented Reality (AR) to create immersive experiences (Wu & Gao, 2022; Zhonggen, 2023; Zonaphan et al., 2022). In particular, VR allows users to interact with digital environments and engage with others in real-time, fostering a connected social ecosystem across various platforms (Mystakidis, 2022). Education has emerged as a key area of innovation within the Metaverse, significantly enhancing students' cognitive engagement (Aşıksoy, 2023; Bradbury, 2022).

The Metaverse comprises a network of technology devices and platforms intertwined with IoT, Blockchain, Artificial Intelligence, Big Data Analytics, and other advanced industries, including education (Zhou, 2022). It offers numerous benefits, including improved user experiences, seamless collaboration and communication, virtual commerce opportunities, support for training and education, and avenues for entertainment, social interaction, and research innovation. However, the Metaverse remains a nascent concept (Deng, 2023).

In education, the Metaverse can serve as a platform for hosting virtual classrooms, delivering interactive learning experiences, and facilitating training simulations in fields such as healthcare, aviation, and engineering. These features make it particularly valuable for distance education and skills training. Despite this potential, questions about the Metaverse's adoption and impact on education remain underexplored (De Felice et al., 2023). For widespread adoption, technological advancements, legal frameworks, and regulatory oversight are necessary (Deng, 2023). Nevertheless, utilizing the Metaverse to train engineers capable of meeting international standards and embracing digital transformation represents a forward-looking strategy (Thongprasit & Piriyasurawong, 2022). Among the tools used to validate various Metaverse models, SmartPLS (PLS-SEM) is the most prominent (Flores-Castañeda et al., 2024).

The metaverse, a hypothetical virtual space where humans interact with digital environments, has gained attention as a transformative concept in industries as diverse as education, healthcare, and business. Therefore, it is worth exploring the evolution of the Metaverse in education, the way it has been designed and the trends in its research over the years, as well as its place in engineering education.

With the research and review, we answered the following research questions:

- RQ1. In what areas of engineering education is Metaverse learning practiced?
- RQ2. Which tools, technologies, or software platforms are employed in these practices?
- RQ3. What teaching strategies or educational frameworks are used in this context?
- RQ4. How do collaboration and interaction between humans and computers occur within the Metaverse?
- RQ5. What is the role of digital twins in Metaverse education?

Scientific research and analysis are essential for selecting appropriate software and designing teaching materials to experiment with engineering education in the Metaverse. This initiative is set to be implemented at the Faculty of Engineering and Technology - Yambol as part of the *Metaverse Approach for a New Generation of Engineering Education and Collaborative Learning (MAGURA)* project, under Project No. BG-RRP-2.004-0006-C02, titled Development of research and innovation at Trakia University in service of health and sustainable well-being.

The study aims to explore the Metaverse's potential as a tool in engineering education (EE), employing content and bibliometric analyses to identify research trends, focal points, and gaps within this domain. Notably, few studies have addressed mobile learning, hybrid learning, or microlearning scenarios, and no studies to date have specifically examined the application of the Metaverse in engineering education.

The findings offer a roadmap for future research to enhance Metaverse adoption in education globally and improve teaching and learning experiences within this virtual environment. By conducting a comprehensive bibliometric analysis of the field, this paper sheds light on research trends, academic contributions, and the conceptual themes of scholarly work related to the Metaverse in education. It represents a significant step toward understanding the evolving research landscape and advancing the integration of the metaverse into educational practices.

Research Methods

Sources of Data and Search Methods

This study employs a mixed-methods approach, combining both quantitative and qualitative synthesis, to review educational research related to the Metaverse. Traditional systematic reviews, while valuable, often risk introducing reporting biases and subjective interpretations of findings (He et al., 2017). To address this, the study adopts a systematic review framework incorporating bibliometric and content analysis, which allows for a scientific evaluation of the knowledge base and the evolution of the topic (Tlili et al., 2022). This review follows the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines for structure and methodology (Moher et al., 2010).

This systematic review analyzes Metaverse-related educational research sourced from databases such as Web of Science, Google Scholar, and Scopus. Search queries utilized keywords including "(metaverse OR virtual reality OR VR OR extended reality OR XR) AND (engineering education AND immersive education OR immersive learning OR immersive teaching) NOT (artificial intelligence OR deep learning OR machine learning)." These keywords were chosen based on the authors' expertise and experience. The search produced a list of 4,386 publications from the Web of Science database.

To broaden the scope, Harzing's Publish or Perish tool was used with the same keywords to extract an additional 1,200 publications from Google Scholar (1,000 records) and Scopus (200 records), covering the period between 2013 and February 2024. Although data from 2024 is incomplete and limited to publications up to March, it is included to highlight a clear trend of increasing research activity in the field.

The analysis focuses on citation networks, authors' research areas, and the content of the selected papers. This bibliometric approach results in the development of a comprehensive model detailing key elements such as authors, article titles, keywords, and abstracts. By aligning these elements with studies on engineering education in the Metaverse, the study establishes stronger connections within the research taxonomy and provides valuable insights into this emerging field.

Data cleaning

Certain publications were excluded based on the following criteria: (1) failure to address the research questions related to the Metaverse in education, particularly in engineering education; (2) lack of relevance to the topics associated with the research questions; (3) absence from peer-reviewed conferences; and (4) unavailability online. After applying these criteria, 85 studies from the Web of Science, Scopus, and Google Scholar databases were included in the analysis.

The study employed both content analysis and bibliometric analysis to interpret the data, ensuring a comprehensive perspective and enhancing the validity of the findings through data triangulation. VOSviewer software was utilized to generate distance-based co-occurrence maps, where terms derived from keywords, titles, and abstracts were grouped and visualized based on their relatedness within a similarity matrix (Van Eck & Waltman, 2010).

Results

Features of the Selected Literature

Figure 1 illustrates the literature identification procedure. Initially, 4386 records were retrieved—1000 from Google Scholar and 200 from Scopus through the WoSCC database. After removing ineligible entries, the titles and abstracts of 2271 records were manually evaluated, and 653 additional records were excluded. Based on the predefined study questions, 150 records were extracted. The final review included 85 scientific publications after applying the exclusion criteria. To understand publication trends, important authors, organizations, and study fields, the data were thoroughly examined.

Filtering of the results

Only open-access articles were included in the further analysis because they are peer-reviewed and can be accessed in scientific databases. For this reason, only Article 2243 (Table 1) and Review Article 251 were included in the following analyses.

Web of Science articles were filtered according to the following criteria:

- by years for the period from 2013 to m.02.2024.
- by types of documents – Articles; Review Articles; and some Proceeding Paper articles that were reviewed; Early Access and Book chapters were scanned according to titles and abstracts to assess the inclusion of some of them.
- Research area: Computer Science, Education Educational Research, Engineering, Science Technology, and Automation Control Systems;
- Web of Science Categories.

Table 1

Type of Publications

Document Types	Record Count	% of 4386
Article	2243	51.140
Book Chapters	36	0.798
Early Access	221	5.039
Proceeding Paper	1920	43.776
Review Article	251	5.723

A search was also conducted through Google Scholar and Scopus and the results were combined. A total of 2271 articles were included in the final set. They were analyzed by VOSviewer. VOS creates a map based on text data to create a term co-occurrence map.

The search for "Engineering Education in Metaverse" across various academic databases yielded 4,386 results, with a total of 39,139 citations and an average of 8.92 citations per item. The h-index, which measures both the productivity and citation impact of the publications, is 76 (Table 2).

Table 2
Publication by Years

Publication Years	Web of Science		Google Scholar	
	Record Count	% of 4386	Record Count	% of 997
2024	307	7.000	52	5,22
2023	1324	30.187	158	15,85
2022	846	19.289	172	17,25
2021	524	11.947	160	16,05
2020	352	8.026	145	14,54
2019	327	7.456	119	11,94
2018	233	5.312	60	6,02
2017	180	4.104	37	3,71
2016	107	2.440	29	2,91
2015	91	2.075	31	3,11
2014	47	1.072	19	1,91
2013	48	1.094	15	1,5

There has been a noticeable trend of increasing publication volume recently in both Web of Science and Google Scholar, peaking in 2023 and 2022, respectively. The highest record count was observed in 2023 for Web of Science (1324 records, 30.187%) and in 2022 for Google Scholar (172 records, 17.25%). This indicates significant recent interest and activity in the field. Publications before 2016 are relatively sparse, indicating that the intersection of engineering education and the metaverse is a relatively new field of study.

Comparative Analysis of Data Sources-Web of Science vs. Google Scholar:

Coverage: The Web of Science consistently has higher record counts compared to Google Scholar for the same years. This might be due to the more stringent inclusion criteria and indexing practices of the Web of Science.

Percentages: The percentages of total records vary, with Google Scholar generally showing a higher percentage in recent years (e.g., 2022 and 2021), suggesting broader indexing or more diverse sources included in Google Scholar.

Implications for Engineering Education and the Metaverse:

Emerging Interest: The data indicates a growing interest in integrating metaverse concepts with engineering education, especially in the last five years.

Research Focus: The peak recently suggests active research and possibly significant advancements or findings being made in this intersection.

Future Trends: The slight dip in 2024 could suggest either a stabilization in the publication rate or a shift in research focus. Continuous monitoring of publication trends will be necessary to understand the evolving landscape.

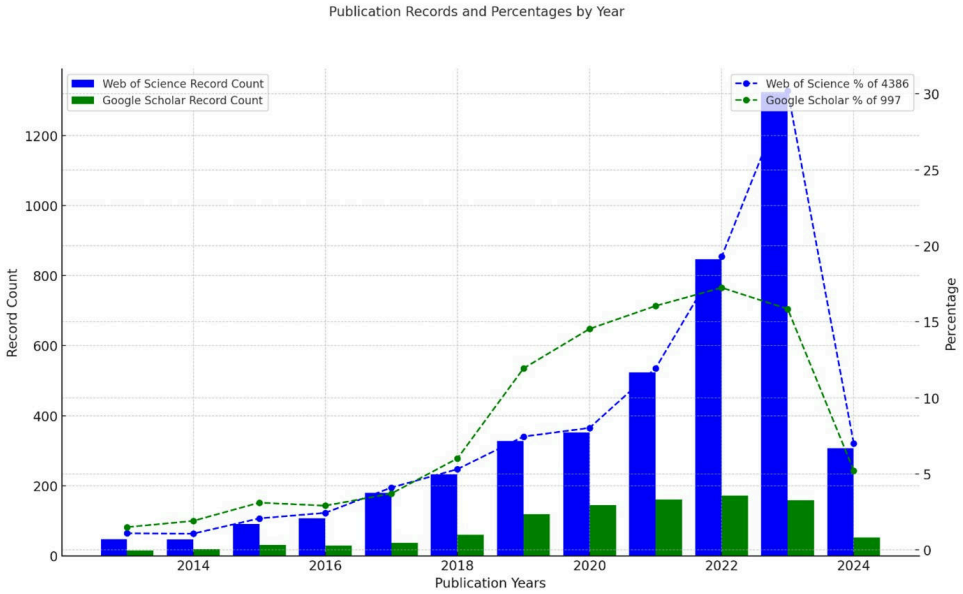
Table 2 provides a clear picture of the growing academic interest in the intersection of engineering education and the metaverse, particularly in the last few years. The analysis indicates that the topic has gained significant traction, likely due to technological advancements and the increasing relevance of immersive

and virtual learning environments. Researchers and educators should consider these trends when exploring new studies or educational initiatives in this dynamic field.

This chart (Figure 1) helps to visualize how the share of publications varies between the two databases over time, providing a clearer comparison based on relative proportions rather than absolute numbers.

Figure 1

Publication Records and Percentages by Year



PRISMA protocol results

The study is based on a systematic review approach performed by means of the Preferred Reporting Items for Systematic Review and Meta-Analyses (PRISMA) protocol, to reveal the effect of the Metaverse in the field of engineering education (Page et al., 2021. n71; Page et al., 2021. n160). A PRISMA's rigorous protocol is used to systematically review academic articles indexed in different databases Web of Science, Scopus, and Google Scholar.

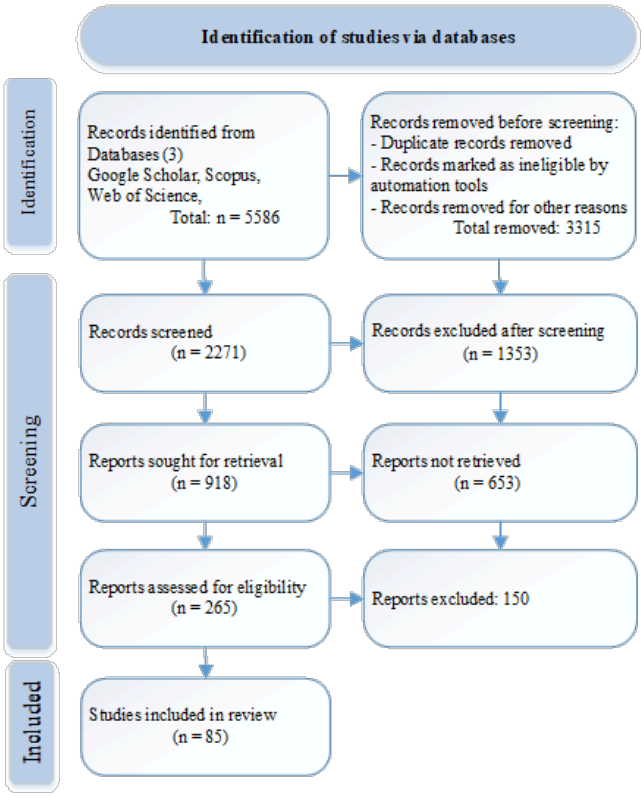
After conducting our initial searches for relevant studies, you can use the PRISMA protocol to organize and report your findings. Here is a step-by-step guide on how to use the PRISMA protocol after searching.

Figure 2 is a flow diagram explaining the results of the search protocols, keywords and research questions.

We completed the PRISMA flow diagram to visually represent the flow of studies through the screening and selection process. After this, we include the number of records identified, screened, excluded, and included, along with the reasons for the exclusions.

Figure 2

Flow diagram of the PRISMA protocol



Findings from the VOSviewer Analysis

VOSviewer is a widely used software for visualizing and analyzing bibliometric networks, enabling researchers to uncover patterns and connections within extensive collections of scientific publications. This information typically includes details about the authors, publications, citations, and keywords. The software supports various file formats such as BibTeX, RIS, and plain text. For this article review, data were sourced from platforms such as Google Scholar, Scopus, and Web of Science.

The construction of bibliometric maps in VOSviewer relies on the VOS mapping technique (Van Eck & Waltman, 2010), where "VOS" signifies the visualization of similarities. While VOSviewer itself does not perform searches, users need to collect and import their search results into the software. Nonetheless, the techniques employed by VOSviewer are also used in other tools, and the insights it provides are invaluable for gaining a deeper understanding of research domains. This is particularly beneficial for researchers exploring new topics, such as Explorer. Within the program, users can select from various similarity-based options tailored to the dataset and the specific analysis required.

The keywords identified by VOSviewer are presented in Table 3.

Table 3
Keywords that VOSviewer identifies

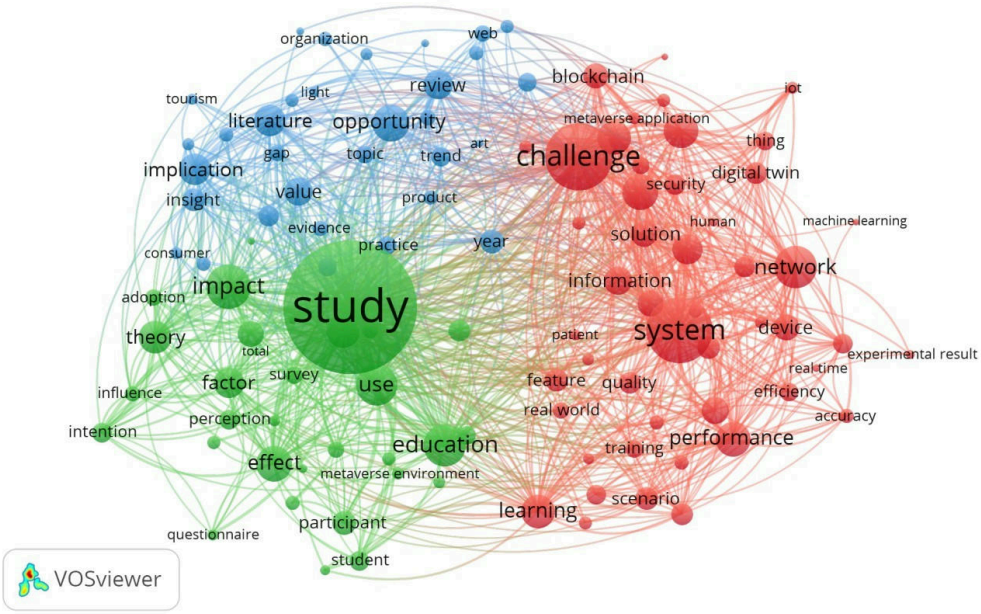
Id	Term	Occurrences	Relevance score
1	learning	26	1,6997
2	student	32	1,6225
3	education	32	1,3664
4	virtual reality	24	1,2557
5	experience	26	1,0869
6	study	27	1,0681
7	teaching	19	1,0298
8	article	10	0,9168
9	effectiveness	11	0,8526
10	technology	28	0,8277
11	immersive virtual reality	18	0,8245
12	use	16	0,7507
13	application	20	0,6396
14	research	17	0,6052
15	paper	21	0,4539

VOS aims to project all documents onto a two-dimensional chart, where the distance between them is inversely proportional to their similarity (i.e., more similar documents appear closer together).

The analysis of keywords identified by VOSviewer reveals that the literature on "Engineering Education Meets the Metaverse".is rich and varied, with a strong focus on student engagement, learning outcomes, and the integration of immersive technologies. The three clusters formed are shown by the visualization–Figure 3). The frequent occurrence of terms related to research and publications indicates an active and evolving field. This analysis can guide researchers and educators in identifying key areas for further study and in understanding the current trends and priorities in this interdisciplinary domain.

Overall, the keywords identified by VOSviewer provide a comprehensive view of the prominent themes and topics within the literature on engineering education and the metaverse. The analysis underscores the importance of learning methodologies, VR technologies, and educational research in shaping the future of engineering education within virtual environments.

Figure 3
Clusters represented with different colors



The Implications for Engineering Education and the Metaverse are as follows:

Student-Centric Approaches: The high occurrence of terms like 'student' and 'learning' suggests that the literature emphasizes student engagement and the impact of metaverse technologies on learning outcomes.

Technological Integration: The frequent mention of “virtual reality,” “immersive virtual reality,” and “technology” underscores the integration of advanced technologies in education and their role in transforming learning experiences.

Evaluating Impact: The terms “effectiveness” and “teaching” imply a critical examination of how these technologies affect teaching methodologies and their effectiveness in improving educational outcomes.

Research and Development: The presence of “study,” “research,” “article,” and “paper” highlights the research-intensive nature of this field, with ongoing studies contributing to a growing body of knowledge.

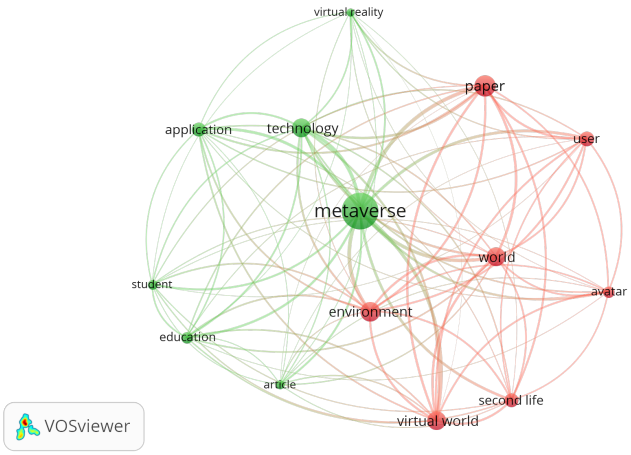
Practical Applications: The term “application” suggests a focus on how theoretical insights are being translated into practical educational tools and methods.

After new filtering, the terms are 5199, of which 23 meet the threshold. In the mapping, which was done according to the title, abstract and keywords in the publications, only those terms with a minimum of 25 occurrences were included. From them, 2 clusters were formed, represented in the two different colors in Figure 4.

After the analysis of titles and abstracts, 85 articles were analyzed according to their content.

Figure 4

Results from the last filtering



Discussion

RQ1: In what areas of engineering education is Metaverse learning practiced?

The Metaverse offers opportunities for immersive experiences, collaboration, and interaction, fostering the development of social experiences and enabling the emergence of "parallel worlds" (Schlemmer & Backes, 2015). According to Lee et al., 2021, developing the Metaverse involves three key stages: (1) Digital Twins: The creation of digital models and representations of the physical world, serving as virtual replicas of physical environments that operate synchronously; (2) Individuals with High Digital Competencies: The necessity for individuals to possess advanced technological skills to effectively manage and operate in digital environments; (3) Co-existence of Physical and Virtual Realities: Integrating and connecting the virtual and physical worlds into a seamless ecosystem.

Additionally, Davis et al. (2009) proposed a research model for the Metaverse, incorporating Virtual Reality elements, which includes the following: (1) the Metaverse itself, (2) users/avatars, (3) technological capabilities of the Metaverse, (4) behaviors, and (5) outcomes (Tlili et al., 2022).

In the realm of engineering education, Metaverse learning is applied across various fields, enhancing VR/AR-based content for three-dimensional object explanations and process visualizations. It is also employed in software engineering education (SEE) to improve learning outcomes and deliver immersive learning experiences.

However, Fernandes & Werner, 2022 highlighted a gap in the literature regarding the characterization of Metaverse applications in SEE. Figure 5 (De Felice et al., 2023) outlines the major educational subject areas benefiting from the Metaverse integration.

Research in WoS on the selected keywords indicates research interest in the following areas. The data in Table 4 cover the top 25 of all analyzed scientific fields.

Table 4

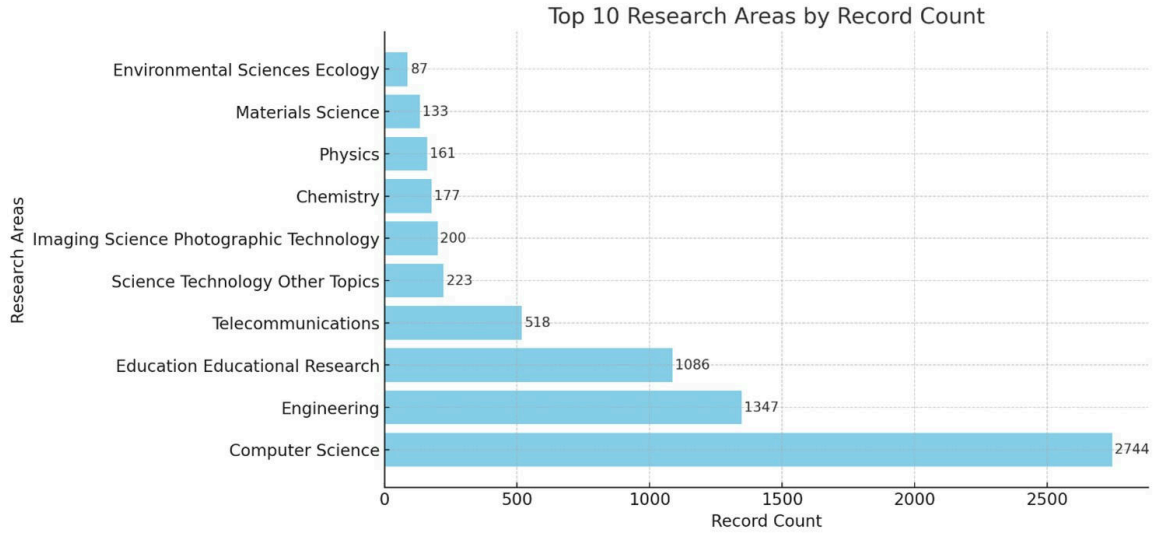
Publications by Research Areas

Research Areas	Record Count	% of 4386
Computer Science	2744	62.563
Engineering	1347	30.711
Education Educational Research	1086	24.761
Telecommunications	518	11.810
Science Technology Other Topics	223	5.084
Imaging Science and Photographic Technology	200	4.560
Chemistry	177	4.036
Physics	161	3.671
Materials Science	133	3.032
Environmental Sciences and Ecology	87	1.984

Showing 10 of 70 entries

Figure 5

Research Areas in Engineering Education



The integration of the Metaverse in education is still limited, but its use in training engineers holds significant promise for preparing them for future digital transformations (Thongprasit & Piriyasurawong, 2022). To evaluate the impact and effectiveness of Metaverse-based training, it is essential to develop reliable instruments for assessing the educational experiences it offers (De Felice et al., 2023; Lopez-Belmonte et al., 2023). One notable limitation of adopting the Metaverse is the unequal access to high-speed internet and advanced hardware, potentially excluding certain populations from participation (De Felice et al., 2023).

Compared to traditional e-learning platforms like Moodle, Zoom, Google Classroom, and Meet, the Metaverse offers a distinct advantage by creating a sense of reality in virtual classes (Kanematsu et al., 2014). For engineering students, it provides more realistic simulation scenarios (Zaga, 2023). As an educational technology, the Metaverse has the potential to enhance student motivation and improve learning outcomes (Camilleri, 2023; Daz et al., 2020). Currently, the Metaverse is increasingly associated with education (Flores-Castañeda et al., 2024), although its applications extend successfully to fields like industrial automation, agriculture, healthcare, and more, with promising benefits for humanity in the years ahead (Zhou, 2022).

In engineering education, Metaverse learning spans several areas. One application involves using Metaverse platforms to enhance VR/AR-based content for explaining three-dimensional objects and visualizing object transformations (Lee & Lee, 2022). Another study focuses on integrating Metaverse platforms into engineering courses to boost motivation, encourage active learning, and deepen topic comprehension (Alvarez et al., 2023). In software engineering education, virtual worlds are employed to improve learning outcomes and offer immersive learning environments (Fernandes & Werner, 2023). Additionally, in computer science education, Metaverse technologies utilize VR and AI to create video recordings, personalized class notes, and virtual workspaces for students (Ho et al., 2023).

According to Daz et al., 2020, the potential impacts of incorporating the Metaverse in education include the following:

- **Enrichment and Transformation of Education:** Enhanced learning outcomes and increased student engagement and motivation (Camilleri, 2023; Flores-Castañeda et al., 2024; Zahra et al., 2021).
- **Creation of New Learning Environments:** Providing learners with access to parallel, safe, and personalized realities for education.
- **Elimination of Distance and Time Barriers:** Enabling remote and asynchronous learning opportunities.
- **Immersive and Interactive Experiences:** Improving the understanding and retention of knowledge through engaging experiences.
- **Facilitation of Collaborative Learning:** Promoting interaction among students and with digital objects.
- **Development of Digital Literacy:** Equipping students with the technological skills required for the future workforce.
- **Potential for Inclusive Education:** Accommodating various learning styles and needs through experiential, collaborative, cooperative, significant, explicit, and emotional learning approaches (Flores-Castañeda et al., 2024).
- **Opportunities for Interdisciplinary and Experiential Learning:** Bridging the gap between theoretical knowledge and practical application.
- **Cultural Exchange and Global Collaboration:** Connecting students from diverse countries and cultural backgrounds to foster mutual understanding.

The Metaverse can revolutionize the educational landscape (Locurcio, 2022).

Short summaries of the analyzed papers, including their focus, results, and conclusions, are presented in Table 5.

Table 5

A summary of some analyzed papers

Authors	Study focus	Results	Conclusion
Aşıksoy, 2023	This systematic review offers important insights for researchers and educators seeking to leverage the Metaverse in education.	The review emphasizes the need for additional empirical research on the use of mixed reality technology in educational metaverse environments, encouraging educators to explore its	The scarcity of experimental studies addressing students' behavioral engagement in metaverse-based education highlights a promising avenue for future research. By designing and implementing such studies, educators can

Authors	Study focus	Results	Conclusion
		potential for enriching student learning experiences.	explore how immersive virtual environment influence student engagement and develop targeted strategies to optimize learning outcomes.
Bradbury, 2022	This paper seeks to serve as a reference and source of inspiration for advancing the quality of vocational education in the Metaverse era.	The paper suggests that Metaverse technology can enhance vocational education by driving innovation and fostering the reform of teaching methods in vocational schools.	The paper emphasizes the importance of analyzing the potential risks associated with the Metaverse, including privacy, ethical, and addiction risks, and calls for the implementation of effective measures to mitigate these challenges.
Braguez et al., 2023	This review aims to define the Metaverse, outline the roles of augmented reality (AR), mixed reality (MR), and virtual reality (VR), and explore the concepts of digital twins and lifelogging.	The paper suggests incorporating the Metaverse into classroom settings as a complementary tool, rather than adopting it for full-time learning, ensuring its use is purposeful and enhances educational outcomes.	The significant potential of Metaverse technologies lies in their ability to provide immersive experiences, enriching both content engagement and social interactions.
Camilleri, 2023	This paper reviews how, why, where, and when Metaverse technologies can be utilized to enhance educational experiences and outcomes.	The Metaverse can facilitate the real-time learning of academic subjects through immersive virtual reality (VR) environments.	The Metaverse can serve as an educational technology that enhances student motivation and improves learning outcomes.
Chen & Liu, 2020	This study investigates how augmented reality (AR) can be applied in a chemistry course to enhance student engagement and improve learning outcomes.	The results indicate that integrating augmented reality (AR) technology into chemistry education can significantly enhance student engagement and improve learning outcomes.	The findings suggest that augmented reality (AR) can help students understand abstract concepts more effectively in chemistry education.
Chua & Yu, 2023	This study seeks to map trends in the educational use of the Metaverse and evaluate its perceived usefulness and ease of use from 2008 to 2022.	Perceived usefulness and perceived ease of use are crucial factors influencing the adoption of Metaverse technologies in education.	The study recommends future research to investigate the perceived usefulness and ease of use across diverse educational fields, incorporating various Metaverse platforms and devices into the research design.
De Felice et al., 2023	This review investigates the role of the Metaverse as an educational tool, exploring its potential to enhance learning experiences and engagement.	The research findings provide insights into the benefits, potential, and risks of the Metaverse as a tool for creating immersive and innovative learning experiences.	One major limitation identified is the absence of standardized assessment tools and metrics to effectively evaluate the impact and effectiveness of Metaverse-based education

Authors	Study focus	Results	Conclusion
			compared to traditional educational methods.
Mazzuco et al., 2022	A systematic literature review was conducted to understand the specific aspects of this context, particularly focusing on the main chemistry topics where augmented reality (AR) is applied.	Augmented reality enhances the understanding of complex chemistry concepts across different educational levels, making abstract and intricate topics more accessible and engaging for students.	The paper emphasizes the necessity for further research to bridge existing gaps and offers suggestions and directions to guide future studies in the field.
Roy et al., 2023	The study focuses on educational settings, emphasizing the significant impact that the metaverse system will have on teaching and learning environments.	This study outlines metaverse applications in education, highlights challenges, identifies adoption factors and themes, and suggests directions for future research.	The metaverse system benefits teaching and learning by enhancing engagement, enabling immersive experiences, facilitating real-time collaboration, and improving the understanding of complex concepts through interactive simulations.
Thongprasit & Piriyasurawong, 2022	The research focuses on developing a metaverse framework aimed at enhancing engineering competency through immersive, interactive, and collaborative virtual environments.	The metaverse framework for engineering competency development consists of three key components: Metaverse: The virtual environment where immersive and interactive learning takes place; Metaverse Platform: The technological infrastructure and tools that facilitate interaction, collaboration, and simulation; Engineering Competency for Future Readiness – The specific skills, knowledge, and abilities that the framework aims to develop to prepare students for future engineering challenges.	Training engineers through the metaverse is an effective approach to prepare them for the digital transformation of the future. The immersive, interactive, and collaborative nature of the metaverse allows engineers to develop practical skills, engage in virtual simulations, and adapt to emerging technologies, ensuring they are equipped to navigate the evolving digital landscape.
Zhonggen, 2023	This paper performed a systematic literature review to assess the acceptance of Metaverse technology in education over 16 years.	The study identified perceived usefulness and ease of use as the key factors influencing the acceptance or rejection of Metaverse technology in education.	Future research should investigate the perceived usefulness and ease of use across diverse educational fields, while also examining various Metaverse platforms and devices in their design.

RQ2. Which tools, technologies, or software platforms are employed in these practices?

A systematic review of 77 research articles (Flores-Castañeda et al., 2024) identified 14 emerging technologies shaping metaverse education, including artificial intelligence (AI), virtual reality (VR), 5G, EON-XR, digital twins (DT), 3D virtual reality, and immersive VR. These technologies contribute to real-world simulations

within virtual environments (immersion), foster interaction with others (interactivity), enhance educational settings through innovative content delivery, and boost learning motivation by capturing student attention.

Recent advancements in metaverse-based education highlight the integration of VR, augmented reality (AR), and 3D simulations (Meena et al., 2023). The metaverse is expected to combine AR, VR, AI, and blockchain to create accurate and accessible virtual environments (De Felice et al., 2023). AI and IoT technologies underpin the development of digital virtual worlds in the metaverse, allowing students to engage in educational and social activities securely and dynamically (Mozumder et al., 2023).

The study underscored the role of 14 key technologies, including AI, VR, 5G, and DT, in promoting immersion, interactivity, and innovative content presentation, ultimately enhancing the educational landscape.

Virtual reality (VR)

Virtual reality (VR) aims to simulate realistic human environments by providing immersive, computer-generated 3D simulations that users can explore and interact with. This technology enables systems to detect user input and modify virtual worlds in real-time, fostering dynamic, interactive experiences. In education, VR helps address challenges in understanding complex objects, such as chemical compounds or anatomical structures, enhancing the learning process and improving efficiency (Petkov et al., 2019).

Platforms like Google Expeditions offer students with learning difficulties unrestricted participation in educational activities (Mitsea et al., 2023). VR environments and applications are specifically designed to enhance cognitive skills and provide immersive learning experiences for students with disabilities (Elfakki et al., 2023). By visualizing and interacting with material in a virtual setting, these students process information more effectively. VR games (VRGs) also support individuals with special educational needs and disabilities (SEND) by boosting motivation, fostering metacognitive skills, and promoting emotional intelligence (Prakash & Rajendran, 2023).

VR typically involves headsets or goggles that display virtual environments, alongside controllers or gloves for interaction, creating a sense of presence and immersion. Applications span entertainment, gaming, education, training, and simulation. Recent advancements in graphics, motion tracking, and haptic feedback have significantly enhanced VR experiences. Collaborative VR environments promote teamwork and communication by enabling students to interact and work on projects together (Petkov et al., 2019).

Due to their cost-effectiveness, VR technologies are widely adopted by universities, educational institutions, and industrial companies, offering practical alternatives to real-world scenarios (Novak-Marcincin et al., 2014).

Extended Reality (XR)

Extended reality (XR) is an umbrella term encompassing virtual environments, human-machine interactions, and wearable technologies. XR integrates virtual reality (VR) and augmented reality (AR), offering immersive experiences that merge the digital and physical worlds.

Augmented reality overlays computer-generated elements onto the real world, enhancing user perception and engagement across sectors such as gaming, retail, and education (Hayes, 2024). Mixed reality (MR) further blends physical and virtual environments, allowing users to interact with both simultaneously.

Table 6 outlines the key distinctions between AR, VR, MR, and XR, highlighting their unique applications and technological features.

Table 6

Comparison between Augmented Reality (AR) vs. Virtual Reality (VR) vs. Mixed Reality (MR) vs. Extended reality (XR)

Comparison	Virtual Reality	Augmented Reality	Mixed Reality	Extended reality (XR)
Definition:	Virtual Reality (VR), users are fully immersed in a digital environment, blocking out the real world and preventing direct interaction with their physical surroundings.	Augmented Reality (AR) overlays digital content or virtual objects onto the real world, enriching the user's experience by enhancing their perception and interaction with the physical environment.	Mixed Reality (MR) blends real and virtual environments, allowing physical and digital objects to co-exist and interact in real time. It integrates aspects of both Virtual Reality (VR) and Augmented Reality (AR) to create immersive and interactive experiences.	Extended Reality (XR) is an overarching term that includes all immersive technologies, such as Virtual Reality (VR), Augmented Reality (AR), and Mixed Reality (MR). XR represents any technology that enhances or extends our perception of reality by blending digital and physical environments.
Scope:	Virtual Reality (VR) seeks to generate a feeling of presence and immersion, giving users the impression that they are physically situated within the virtual environment.	Unlike Virtual Reality (VR), Augmented Reality (AR) enables users to interact with and perceive both the real world and virtual elements simultaneously.	Mixed Reality (MR) integrates virtual and physical objects, enabling users to interact with both simultaneously. It offers a spectrum of experiences that range from fully virtual environments to entirely physical ones, blending digital content seamlessly with the real world.	Extended Reality (XR) covers a broad spectrum of immersive technologies, including Virtual Reality (VR), Augmented Reality (AR), and Mixed Reality (MR). XR refers to any technology that modifies or enhances our perception of reality, regardless of how much the physical and virtual worlds are integrated.
Technology:	Virtual Reality (VR) is a computer-generated simulation of a 3D environment, allowing users to interact with it through specialized equipment like headsets, gloves, or controllers.	Augmented Reality (AR) is commonly experienced through devices such as smartphones, fitness trackers, tablets, or smart glasses. These devices use cameras and sensors to detect the real-world environment and overlay digital content onto it. AR technology can gather data on user health, activity levels, and surroundings, while also interacting with other devices to provide notifications, track fitness goals, or enable immersive AR experiences.	Mixed Reality (MR) relies on advanced sensors, cameras, and tracking systems to map the physical environment and overlay virtual objects onto it. Users are immersed in VR through technologies like VR head-mounted devices (HMDs) and high-degree-of-freedom tracking controllers. This technology demands a high level of spatial understanding and tracking capabilities to seamlessly integrate the real and virtual worlds.	Extended Reality (XR) encompasses various technologies. Virtual Reality (VR) generates a fully immersive digital environment, Augmented Reality (AR) overlays digital content onto the real world, and Mixed Reality (MR) blends elements of both to create an interactive mixed environment.
Applications:	Virtual Reality (VR) often demands significant computing power and is widely used in gaming, training, education, simulations, and virtual tours.	Augmented Reality (AR) is applied in diverse fields, including gaming, entertainment, education, and industrial training.	Mixed Reality (MR) is used across multiple sectors such as gaming, entertainment, education, and training. It enables users to engage with virtual objects within real-world environments, creating immersive and interactive experiences.	Extended Reality (XR) encompasses a broad spectrum of applications. Virtual Reality (VR) is commonly used in gaming, simulations, and virtual tours, while Augmented Reality (AR) is applied in education, healthcare, and industrial training. Mixed Reality (MR) merges elements of both VR and AR, enabling its use across diverse fields.

These technologies have been effectively integrated into manufacturing engineering education, offering students virtual environments to develop and refine their skills (Novak-Marcincin et al., 2014). Unlike virtual reality (VR), which constructs an entirely digital environment, augmented reality (AR) enhances the real world by overlaying additional content and information (Hayes, 2024). VR enables users to interact with digital objects and other participants in ways that feel lifelike, while AR uses devices such as smartphones, tablets, or headsets to introduce virtual elements into real-world scenarios.

Mixed reality (MR) merges physical and virtual environments by leveraging advanced computing, imaging, and input technologies. Devices like Microsoft HoloLens incorporate sophisticated sensors and cameras to

track user movements and place virtual objects accurately within the real world. This interaction allows users to engage with virtual objects as if they existed physically, fostering highly immersive experiences. MR finds applications in fields such as architecture, design, and engineering, where it is used to visualize and manipulate 3D models. The term extended reality (XR) encompasses all immersive technologies, including VR, AR, and MR, and is frequently referenced in academic and research settings.

In XR, VR immerses users in a fully digital environment, AR overlays virtual components on the physical world, and MR blends the two seamlessly, allowing users to interact with virtual objects in real-time environments.

Augmented reality (AR) is increasingly utilized in chemistry education, serving as a valuable pedagogical tool for understanding complex scientific concepts (Mazzuco et al., 2022; Saidin et al., 2016). Hands-on AR applications have demonstrated potential in motivating students and sustaining long-term engagement. Research indicates that students using AR for chemistry experiments achieved greater conceptual understanding and higher engagement levels compared to those taught through conventional methods (Chen & Liu, 2020). AR enhanced the learning experience by enabling students to visualize and manipulate chemical elements in virtual environments, facilitating the comprehension of abstract ideas. This approach resulted in improved learning outcomes and increased student interest (Chen & Liu, 2020).

AR technologies also boosted student motivation, promoted teamwork, and contributed to lasting knowledge retention (Muhammad et al., 2021). Both AR and VR are applied across STEAM disciplines (Science, Technology, Engineering, Arts, and Mathematics) without differentiation between subjects in metaverse-based education (Braguez et al., 2023; Kanematsu et al., 2014). AR-supported flipped learning strategies not only enhanced students' project performance but also fostered greater learning motivation, critical thinking, and collective self-efficacy within groups (Chang & Hwang, 2018). Findings revealed that students expressed satisfaction with AR experiences and expressed interest in continuing to use such applications in the future (Sahin & Yilmaz, 2020).

Internet of Things

The Internet of Things (IoT) refers to a network connecting physical objects, known as "things," embedded with sensors, communication tools, and networking technologies. These objects exchange data, information, and knowledge with other devices and systems through existing network infrastructures (Zhou, 2022). IoT devices facilitate data transfer between computers and humans across networks. IoT is seen as a transformative force in education, with the potential to enhance learning outcomes, boost operational efficiency, and improve student performance (Sailesh, 2023).

Digital technologies in education aim not only to expand accessibility but also to improve the efficiency and inclusivity of traditional educational systems. IoT is reshaping education by fostering interactivity, collaboration, and inclusivity (Ingole, 2022; Kalid et al., 2022; Rodrigues, 2023). Through IoT devices, teachers can track real-time data on student progress, enabling more accurate assessments of academic performance (Ingole, 2022).

Smart boards function similarly to traditional blackboards but can also display visuals and images relevant to the topic. IoT innovations, such as GPS-equipped school buses, smart security cameras, and tablets with educational applications, are transforming how educational institutions operate. Daily concerns like student attendance can also be managed more effectively with IoT.

The IoT will enhance operational efficiency across diverse learning environments. IoT in classrooms can refine lesson planning, improve learning resources, enrich teaching methods, optimize administrative functions, and reduce operational costs (Sailesh, 2023). In academic contexts, IoT sensors gather data and suggest relevant academic topics for students, guiding their future learning paths.

RQ3. What teaching strategies or educational frameworks are used in this context?

A metaverse framework for developing engineering competencies includes three elements: the metaverse itself, the metaverse platform, and the competencies required for future readiness (Thongprasit & Piriyasurawong, 2022). In the metaverse, students engage with digital content and their peers in immersive and interactive environments. They participate in virtual simulations, role-playing scenarios, and collaborative global projects (Abraham et al., 2023).

The proposed metaverse education system integrates virtual reality (VR) and blockchain technologies to foster social learning. Learners collaborate to deepen their knowledge, while the blockchain ensures secure and equitable tracking of progress and grading (Zheng et al., 2023).

The pedagogical framework for metaverse education is grounded in constructivist principles, emphasizing active participation, experimentation, and exploration. The essential elements include immersive simulations, XR-based student engagement, and social learning facilitated through avatars.

Principles for developing the framework of training in edu-metaverse

The framework for training in the edu-metaverse, as outlined in (Sin et al., 2023) and supported by references (Honebein, 1996) and (Murphy, 1997), is based on several core principles:

- P1: Facilitate Knowledge Construction Experiences–Learners must recognize that knowledge is a constructed representation of the world.
- P2: Integrate Relevance into Learning–New learning should be connected to prior knowledge to enhance understanding.
- P3: Promote Exposure to Multiple Solutions–Providing learners with diverse perspectives fosters a deeper and more comprehensive grasp of concepts.
- P4: Embed Learning in Realistic Contexts: Associating knowledge with real-world applications enhances engagement and reinforces relevance.
- P5: Foster Ownership of Learning–Encouraging students to independently seek out knowledge broadens their learning experiences.
- P6: Use Multiple Representations–Presenting information through different formats helps learners form a well-rounded understanding of concepts.
- P7: Promote Exploration: Learners should be encouraged to pursue knowledge independently, fostering curiosity and discovery.
- P8: Encourage Social Learning–Collaborative learning environments enable learners to refine their understanding by engaging with peers.

These principles align with the constructivist framework, which posits that knowledge is internally constructed through active engagement with new concepts. Learning occurs not by passively receiving information but through exploration and interaction (Sin et al., 2023).

Traditional teaching methods often emphasize memorization and passive learning, whereas constructivist approaches prioritize active participation, experimentation, and the discovery of cause-and-effect

relationships. Constructivist teaching cultivates students' cognitive, emotional, psychomotor, and experiential abilities (Mijanović, 2023).

Metaverse Learning Principles:

The metaverse learning principles center around crafting immersive, interactive educational experiences through virtual environments. These principles emphasize collaborative and project-based learning to deliver instructional content and achieve educational goals (Lopez-Belmonte et al., 2023). The design of metaverse spaces is pivotal in fostering interaction among students and enhancing their overall learning capabilities (Ahn & Heo, 2023). Similar to online learning, metaverse-based education promotes self-directed learning, collaboration, and experiential learning; however, it assumes a fundamentally altered sense of reality within the metaverse (Dreamson & Park, 2023). This approach allows learners to engage with digital content, participate in virtual simulations, role-playing exercises, and collaborative projects, while also accessing diverse resources within a single platform (Sabarinath, 2023). Multiple models and frameworks have been developed to facilitate the integration of the metaverse into educational settings.

A defining feature of metaverse learning is Immersive Learning (IL), which takes place within extended reality (XR). XR's primary draw is its ability to create a heightened sense of presence, or the feeling of "being" in a virtual environment (Gibson, 2014). Additionally, immersion through IL helps students concentrate on mastering the educational material rather than navigating the user interfaces (Sin et al., 2023).

The key metaverse learning models include:

- (M1) Visualizing Knowledge as Constructs: Knowledge is seen as an internal construct requiring mental organization. Students must understand knowledge to actively engage with its formation (P1).
- (M2) Knowledge Association: New concepts are grasped through connections to the learner's prior knowledge (P2).
- (M3) Immersive Simulations: Knowledge serves to model reality. Students benefit from exploring problems from various perspectives to develop comprehensive internal models (P3).
- (M4) Real-World Problem Reenactment: Daily experiences form the basis for new learning. IL supports situated learning by reenacting real-life problems, allowing students to relate problem-solving processes to real-world contexts (P4).
- (M5) XR Student Engagement: Students take responsibility for their learning journeys (P5).
- (M6) Multimodal Content Delivery: Presenting content through different modalities enhances the construction of robust mental models (P6).
- (M7) Knowledge Exploration: Interactive experiences play a crucial role in synthesizing new knowledge (P7).
- (M8) Social Learning with Avatars: Collaborative learning environments enable students with diverse experiences to guide and challenge each other (P8).

A pedagogical framework for metaverse-based training has been proposed by (Sin et al., 2023), highlighting the following:

1. A ubiquitous platform for exploring interconnected educational concepts.
2. Intuitive interaction grounded in physical metaphors.
3. Integration with real-world learning through the advantages of digital content.

The central idea behind edu-metaverse frameworks is to merge the structure of knowledge graphs (KG) with VR immersion, enabling deeper exploration of educational content.

RQ4. How do collaboration and interaction between humans and computers occur within the Metaverse?

Metaverse collaboration allows multiple users to engage, communicate, and interact within virtual environments. This can involve activities such as virtual meetings, shared workspaces, gaming, and social interactions. The metaverse acts as a digital platform that connects individuals regardless of their physical location.

Version 2 (V2) of metaverse collaboration emphasizes virtual environments that facilitate communication and interaction, enhancing user engagement and fostering teamwork. Human-computer interaction (HCI) in the metaverse focuses on designing immersive and intuitive interfaces to improve user experience.

HCI is a multidisciplinary field dedicated to understanding, designing, and evaluating the interactions between humans and computers (Haux, 2023). It seeks to create user-friendly interfaces, efficient, and intuitive, drawing from fields such as anthropology and sociology (Wardhani, 2023). Methods within HCI include ethnography, participatory design, controlled experiments, and usability testing (Martínez-Miranda & Pérez-Espinosa, 2023). HCI research extends into various aspects of life, including spirituality, accessibility for individuals with disabilities, and emerging technologies like virtual reality and wearable devices (Jyoti & Kaur, 2023). The discipline has evolved from developing basic user interfaces to creating complex interactive systems capable of recognizing and simulating human emotions (Meen Teen-Hang & Tijus, 2023). HCI aims to enhance human-computer interaction by making technology more usable and accessible, integrating knowledge from design, psychology, ergonomics, and computer science.

The convergence of metaverse collaboration and HCI focuses on developing and studying interaction techniques and user interfaces within virtual environments. This involves analyzing how users engage with virtual worlds and one another, with an emphasis on creating interfaces that enable seamless communication and teamwork.

The field explores methods to improve usability, elevate user experience, and craft immersive digital spaces. Social and psychological dimensions—such as presence, identity, and social norms—are also considered when developing HCI for the metaverse.

Fang, 2022 outlines four key HCI frameworks:

1. A generic computing model for HCI;
2. A touchscreen display control model for the HCI;
3. An immersive virtual reality HCI model;
4. A metaverse-specific HCI model.

RQ5: What is the role of digital twins in Metaverse education?

Digital twins are virtual representations of physical objects or systems, created using data and simulations to enable real-time monitoring, optimization, control, and decision-making. They have garnered significant interest across industries for their capacity to enhance efficiency and inform better decisions (Ferrigno, 2023; Fuller, Fan, & Day, 2020). The concept of "digital twins" originated from NASA's Apollo program, where two identical spacecraft were constructed to replicate conditions during space missions. The vehicle that remained on Earth was referred to as the "twin" (Cellina et al., 2023). Michael Grieves was the first to formally introduce the idea of a Digital Twin (DT) (Babbar, 2023).

By replicating physical systems, digital twins facilitate real-time simulations, optimizations, and predictive insights, contributing to more informed decision-making processes. These models are increasingly vital in education, offering realistic simulations that help simplify complex subjects and improve learning outcomes. Throughout a product's lifecycle, digital twins assist in simulating, predicting, and refining both the product and its manufacturing systems before committing to real prototypes and physical assets.

In sectors such as oil and gas, digital twins enhance asset management, operational efficiency, and simulate the performance of new infrastructure (Kortelainen, Minav, & Tammi, 2023). In healthcare, digital human twins (DHTs) replicate patient data, advancing personalized and data-driven treatments (Babbar, 2023). In educational contexts, DTs can provide teachers with insights into students' academic and behavioral patterns, allowing for tailored support to address learning difficulties. Students with special needs can engage in simulated scenarios, preparing them for real-world challenges. Integrating digital twins with virtual reality (VR) offers further support for overcoming learning barriers.

Digital twins also play a role in vehicular networks by improving resource allocation and edge computing efficiency (Losovsky, 2023). The evolution of digital twins has accelerated with advancements in artificial intelligence, data analytics, and the Internet of Things (IoT) (Cali et al., 2023; El-Din, El-Shafai, & El Sayed, 2023; Ferrigno, 2023; Jadhav & Sarnikar, 2023). Their application spans industries such as construction, healthcare, manufacturing, oil and gas, and smart cities (Saha, Banik, & Banik, 2023). For example, Singapore has adopted digital twin models to predict and simulate disaster scenarios, using IoT data and satellite imagery to monitor infrastructure and enhance response strategies (Maiti & Kayal, 2024).

Despite their promise, the widespread adoption of digital twins faces challenges, including concerns related to data security, quality, and interoperability. Addressing these issues will be critical for their continued growth and integration.

Conclusion

Education in the Metaverse remains in its infancy, with efforts focused on concept development, experimental testing, and discussions among educators, researchers, and technology stakeholders. The concept development stage involves developing and refining frameworks to deliver and experience education in virtual reality. This includes exploring potential applications, understanding associated challenges, and envisioning how such environments can complement or transform traditional education systems.

Pilot studies and experiments play a critical role in evaluating VR-based learning. These initiatives involve designing specific educational interventions, assessing learner engagement, and analyzing outcomes to identify strengths and limitations. Insights gained from these experiments inform best practices and guide the optimization of Metaverse-based education.

The current findings emphasize the Metaverse's potential to create immersive, interactive, and collaborative learning experiences, enabling personalized and engaging education. However, challenges such as technological barriers, access inequality, and the need for effective pedagogical strategies must be addressed.

Future implications for the field include the following:

- Improving Accessibility: Developing cost-effective technologies and addressing infrastructure limitations to ensure equitable access.

- Refining Pedagogical Models: Creating adaptable and interdisciplinary teaching frameworks to leverage the Metaverse's unique capabilities.
- Fostering Collaboration: Encouraging partnerships among educators, technologists, and policymakers to design inclusive and scalable solutions.
- Evaluating Long-term Impact: Conducting longitudinal studies to measure the effectiveness of Metaverse education on learning outcomes, skill development, and workforce readiness.

While the Metaverse holds promise to revolutionize education, its successful integration requires further research, innovation, and collaboration to overcome current limitations and unlock its full potential.

Peer Review: Externally peer-reviewed
Author Contributions: Conception/Design of Study- V.N., S.D., S.A.; Data Acquisition- V.N., S.D., S.A.; Data Analysis/Interpretation- V.N., S.D., S.A.; Drafting Manuscript- V.N., S.D., S.A.; Critical Revision of Manuscript- V.N., S.D., S.A.; Final Approval and Accountability- V.N., S.D., S.A.; Technical or Material Support- V.N., S.D., S.A.; Supervision- V.N., S.D., S.A.
Conflict of Interest: The authors have no conflict of interest to declare.
Grant Support: The authors declared that this study has received no financial support.

Author Details

Veselina Nedeva
[1] Trakia University, Faculty of Engineering and Technologies, Department of Electrical Engineering, Automation, Computer Systems and Communications, Yambol, Bulgaria
0000-0002-4082-1896

Snejana Dineva
[1] Trakia University, Faculty of Engineering and Technologies, Department of Electrical Engineering, Automation, Computer Systems and Communications, Yambol, Bulgaria
0000-0002-4655-6640

Svetoslav Atanasov
[1] Trakia University, Faculty of Engineering and Technologies, Department of Electrical Engineering, Automation, Computer Systems and Communications, Yambol, Bulgaria
0000-0002-2658-1611 svetoslav.atanasov@trakia-uni.bg

References

Abraham, A., Suseelan, B., Mathew, J., & Sabarinath, P. (2023). A study on Metaverse in education. *2023 7th International Conference on Computing Methodologies and Communication (ICCMC)*, 1570–1573. https://doi.org/10.1109/ICCMC56507.2023.1008391

Ahn, S., & Heo, Y. (2023). The principles and practices of college writing education using the Metaverse. *Korean Journal of General Education*. https://doi.org/10.46392/kjge.2023.17.1.6

Alvarez, J., Deyanira, D., Angel, M. J., & Martínez, M. (2023). Tec Virtual Campus, a metaverse for engineering learning. *Conference Name (if applicable)*. https://doi.org/10.1109/EDUNINE57531.2023.1010285

Aşıksoy, G. (2023). Empirical studies on metaverse-based education: A systematic review. *International Journal of Engineering Pedagogy, 13*(3), 120–133. https://doi.org/10.3991/ijep.v13i3.36227, doi:10.3991/ijep.v13i3.3622

Babbar, H. (2023). Digital twin for edge computing in smart vehicular systems. https://doi.org/10.1109/InCACCT57535.2023.10141784

Bradbury, L. (2022). Research on the application and risk prevention of the Metaverse in vocational education. *Lecture Notes in Computer Science,* 41–54. https://doi.org/10.1007/978-3-031-23518-4_4, doi:10.1007/978-3-031-23518-4_

Braguez, J. M., Braguez, S., Moreira, C., & Filipe, T. (2023). The possibilities of changes in learning experiences with the Metaverse. *Procedia Computer Science, 219*, 504–511

Cali, U., Berhane, D., Hajialigol, P., Moazami, A., Gupta, S. N., Lobaccaro, G., & Papa, S. (2023). Digital twins: Shaping the future of energy systems and smart cities through cybersecurity, efficiency, and sustainability. https://doi.org/10.1109/FES57669.2023.10182868, doi: 10.1109/FES57669.2023.10182868

Camilleri, M. A. (2023). Metaverse applications in education: A systematic review and a cost–benefit analysis. *Interactive Technology and Smart Education*. https://doi.org/10.1108/ITSE-01-2023-0017, doi: 10.1108/ITSE-01-2023-001

Cellina, M., Cè, M., Alì, G., Irmici, E., Caloro, D., Fazzini, D., Oliva, G., & Papa, S. (2023). Digital twins: The new frontier for personalized medicine? *Applied Sciences*. https://doi.org/10.3390/app13137940 doi:10.3390/app13137940

Chang, S.-C., & Hwang, G.-J. (2018). Impacts of an augmented reality-based flipped learning guiding approach on students' scientific project performance and perceptions. *Computers & Education*, *125*(1), 226–239. https://doi.org/10.1016/j.compedu.2018.06.007 DOI: 10.1016/j.compedu.2018.06.00

Chen, S. Y., & Liu, Y. (2020). Using augmented reality to experiment with elements in a chemistry course. *Computers in Human Behavior, 111*, 106418. https://doi.org/10.1016/j.chb.2020.106418, doi:10.1016/j.chb.2020.10641

Chua, H. W., & Yu, Z. (2023). A systematic literature review of the acceptability of the use of Metaverse in education over 16 years. *Journal of Computers in Education*. https://doi.org/10.1007/s40692-023-00273-z, doi:10.1007/s40692-023-00273-

Davis, A., Murphy, J. D., Owens, D., Khazanchi, D., & Zigurs, I. (2009). Avatars, people, and virtual worlds: Foundations for research in metaverses. *Journal of the Association for Information Systems, 10*(2), 90

Daz, T. B., Karagölge, Z., & Ceyhun, İ. (2020). Üstün yetenekli öğrencilerin kimya dersine yönelik görüşlerinin incelenmesi: Erzurum Bilsem örneği. *Atatürk Üniversitesi Kazım Karabekir Eğitim Fakültesi Dergisi, 41*, 159–179

De Felice, F., Petrillo, A., Iovine, G., Salzano, C., & Baffo, I. (2023). How does the Metaverse shape education? A systematic literature review. *Applied Sciences, 13*, 5682. https://doi.org/10.3390/app13095682 doi:10.3390/app1309568

Deng, B. (2023). The impact of the Metaverse on the future development of enterprises. *BCP Business & Management, 36*, 151–157. https://doi.org/10.54691/bcpbm.v36i.340

Dreamson, N., & Park, G. (2023). Metaverse-based learning through children's school space design. https://doi.org/10.1111/jade.12449 doi: 10.1111/jade.12449;

El-Din, W. H., El-Shafai, W., & El Sayed, A. T. (2023). Integrating digital twins with the IoT-based blockchain: Concept, architecture, challenges, and future scope. *Wireless Personal Communications*. https://doi.org/10.1007/s11277-023-10538-6, doi:10.1007/s11277-023-10538-6

Elfakki, A. O., Sghaier, S., & Alotaibi, A. (2023). An efficient system based on experimental laboratory in 3D virtual environment for students with learning disabilities. *Electronics*. https://doi.org/10.3390/electronics12040989, doi:10.3390/electronics1204098

Fang, Y. M. (2022). Exploring generic and immersive frameworks for human-computer interactions. In C. Stephanidis, M. Antona, and S. Ntoa (Eds.), *HCI International 2022 Posters* (Vol. 1580, pp. 116-120). Springer. https://doi.org/10.1007/978-3-031-06417-3_8, doi:10.1007/978-3-031-06417-3_8

Fernandes, F. A., & Werner, C. (2023). A scoping review of the Metaverse for software engineering education: Overview, challenges, and opportunities. *Presence: Teleoperators & Virtual Environments*. https://doi.org/10.1162/pres_a_0037

Fernandes, F. A., & Werner, C. M. L. (2022). A scoping review of the Metaverse for software engineering education: Overview, challenges, and opportunities. *Presence: Virtual and Augmented Reality, 31*, 107–146. https://doi.org/10.1162/pres_a_0037

Ferrigno, E. (2023). 3D real-time digital twin. https://doi.org/10.2118/213115-ms doi:10.2118/213115-ms

Flores-Castañeda, R. O., Olaya-Cotera, S., & Iparraguirre-Villanueva, O. (2024). Benefits of Metaverse application in education: A systematic review. *International Journal of Engineering Pedagogy, 14*(1), 61–81. https://doi.org/10.3991/ijep.v14i1.42421, doi:10.3991/ijep.v14i1.4242

Fuller, A. Z., Fan, C. R., & Day, C. Barlow. (2020). Digital twin: Enabling technologies, challenges, and open research. *IEEE Access, 8*, 108952-108971. https://doi.org/10.1109/ACCESS.2020.2998358

Gibson, J. J. (2014). *The ecological approach to visual perception: Classic edition*. Psychology Press

Haux, R. (2023). Human–computer interaction. *Oxford Research Encyclopedia of Psychology*. https://doi.org/10.1093/acrefore/9780190236557.013.4

Hayes, A. (2024). Augmented Reality (AR): Definition, Examples, and Uses. *Investopedia*. https://www.investopedia.com/terms/a/augmented-reality.asp [Updated February 4, 2024]

He, Q., Wang, G., Luo, L., Shi, Q., Xie, J., & Meng, X. (2017). Mapping the managerial areas of building information modeling (BIM) using scientometric analysis. *International Journal of Project Management, 35*(4), 670–685. https://doi.org/10.1016/j.ijproman.2016.08.001, doi:10.1016/j.ijproman.2016.08.00

Ho, K.-H., Hou, Y., Chu, C. F. C., Pan, H., & Chan, T.-T. (2023). Work in progress: An AI-assisted metaverse for computer science education. *Conference Name (if applicable)*. https://doi.org/10.1109/EDUNINE57531.2023.1010281

Honebein, P. C. (1996). Seven goals for the design of constructivist learning environments. In *Constructivist Learning Environments: Case Studies in Instructional Design* (pp. 11–24). Educational Technology Publications

Ingole, J. G. (2022). Impact of IoT (Internet of Things) on today's education system. *International Journal of Advanced Research in Science, Communication and Technology*. https://doi.org/10.48175/ijarsct-5883 doi:10.48175/ijarsct-5883

Jadhav, S. G., & Sarnikar, S. (2023). Digital twin of a digital world: Process, data, and experience perspectives. *IT Professional, 25*(3), 68-73. https://doi.org/10.1109/MITP.2023.3264209

Jyoti, & Kaur, G. (2023). Research paper on human–computer interaction (HCI). *International Journal for Multidisciplinary Research (IJFMR)*, 5(2), 1-9

Kalid, K. S., Wan, F., Wan, A., Muhammad, S., Shahrul, A., Sulisworo, D., Fitrianawati, M., & Arshad, C. S. (2022). A proposed implementation of the Internet of Things as a teaching aid for learning science collaboratively. *2022 IEEE AIIC*, 9914591. https://doi.org/10.1109/AiIC54368.2022.9914591, doi: 10.1109/AiIC54368.2022.991459

Kanematsu, H., Kobayashi, T., Barry, D. M., Fukumura, Y., Dharmawansa, A., & Ogawa, N. (2014). Virtual STEM class for nuclear safety education in metaverse. *Procedia Computer Science, 29*, 1255–1261. https://doi.org/10.1016/j.procs.2014.08.224, doi:10.1016/j.procs.2014.08.22

Kortelainen, J. T., Minav, T., & Tammi, K. (2023). Digital twin-The dream and the reality. https://doi.org/10.3389/friot.2023.1108777;

Lee, L. H., Braud, T., Zhou, P., Wang, L., Xu, D., Lin, Z., Kumar, A., Bermejo, C., & Hui, P. (2021). All one needs to know about Metaverse: A complete survey on technological singularity, virtual ecosystem, and research agenda. *arXiv preprint arXiv:2110.05352*

Lee, Y.-S., & Lee, T.-Y. (2022). Applying the Metaverse Platform and Content in Practical Engineering Education. *Journal Name*. https://doi.org/10.15701/kcgs.2022.28.3.3

Locurcio, L. L. (2022). Dental education in the metaverse. *British Dental Journal, 232*(4), 191–191. https://doi.org/10.1038/s41415-022-3990-7, DOI: 10.1038/s41415-022-3990-

Lopez-Belmonte, J., Pozo-Sánchez, S., Moreno-Guerrero, A.-J., & Lampropoulos, G. K. (2023). Metaverse in education: A systematic review. *RED. Revista de Educación a Distancia, 23*(73). https://doi.org/10.6018/red.51142

Lopez-Belmonte, J., Pozo-Sánchez, S., Moreno-Guerrero, A.-J., & Marín-Marín, J.-A. (2023). We have reached the GOAL: Teaching methodology for transforming learning in the Metaverse. *Metaverse Research*, *30*, 2023. https:doi.org/10.56294/mr20233

Losovsky, B. Ya. (2023). Digital twin for spacecraft concepts. "https://doi.org/10.1109/aero55745.2023.10115665;

Maiti, M., & Kayal, P. (2024). Exploring innovative techniques for damage control during natural disasters. *Journal of Safety Science and Resilience*. https://doi.org/10.1016/j.jnlssr.2024.02.004;

Martínez-Miranda, J., & Pérez-Espinosa, H. (2023). 104 The use of language in affective computing for human-computer interaction. In G. L. Schiewer, J. Altarriba, & B. C. Ng (Eds.), *Volume 3* (pp. 2127-2158). De Gruyter Mouton. https://doi.org/10.1515/9783110795486-04

Mazzuco, A. E., Rocha, A. L., Krassmann, E. B., Reategui, R. S., & Gomes, M. (2022). A systematic review of augmented reality in chemistry education. *Review of Education, 10*(1). https://doi.org/10.1002/rev3.332

Meen Teen-Hang, C., & Tijus, S. (2023). Special issue on human–computer interactions 2.0. *Applied Sciences*. https://doi.org/10.3390/app1307426

Meena, S. D., Mithesh, G. S. S., Panyam, R., Chowdary, M. S., Sadhu, V. S., & Sheela, J. (2023). Advancing education through Metaverse: Components, applications, challenges, case studies, and open issues. *2023 International Conference on Sustainable Computing and Smart Systems (ICSCSS)*, Coimbatore, India, 880–889. https://doi.org/10.1109/ICSCSS57650.2023.10169535, doi: 10.1109/ICSCSS57650.2023.1016953

Mijanović, N. (2023). Constructivism as a contemporary teaching paradigm. *Inovacije u nastavi*, *36*(1), 21–32. https://doi.org/10.5937/inovacije2301021m, doi:10.5937/inovacije2301021

Mitsea, E., Drigas, A., & Skianis, C. (2023). VR gaming for meta-skills training in special education: The role of metacognition, motivations, and emotional intelligence. *Education Sciences*. https://doi.org/10.3390/educsci13070639 doi:10.3390/educsci1307063

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & PRISMA Group. (2010). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *International Journal of Surgery, 8*(5), 336–341

Mozumder, M. A. I., Athar, A., Armand, T. P. T., Sheeraz, M. M., Uddin, S. M. I., & Kim, H.-C. (2023). Technological roadmap of the future trend of Metaverse based on IoT, blockchain, and AI techniques in Metaverse education. *2023 25th International Conference on Advanced Communication Technology (ICACT)*, Pyeongchang, Korea, 1414–1423. https://doi.org/10.23919/ICACT56868.2023.1007946

Muhammad, A. K., Khan, N., Lee, M. Y., Imran, A. S., & Sajjad, M. (2021). School of the future: A comprehensive study on the effectiveness of augmented reality as a tool for primary school children's education. *Applied Sciences*, *11*(11), 5277. https://doi.org/10.3390/app11115277 doi:10.3390/app1111527

Murphy, E. (1997). Constructivism: From philosophy to practice. *Educational Resources Information Center (ERIC)*, U.S. Department of Education. ED 444 966

Mystakidis, S. (2022). Metaverse. *Encyclopedia, 2*(1), 486–497. https://doi.org/10.3390/encyclopedia201003

Novak-Marcincin, J., Janak, M., Barna, J., & Novakova-Marcincinova, L. (2014). Application of virtual and augmented reality technology in the education of manufacturing engineers. *Lecture Notes in Computer Science*, 439–446. https://doi.org/10.1007/978-3-319-05948-8_42, doi:10.1007/978-3-319-05948-8_4

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., & others. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ, 372*, n71. https://doi.org/10.1136/bmj.n71 doi: 10.1136/bmj.n7

Page, M. J., Moher, D., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., et al. (2021). PRISMA 2020 explanation and elaboration: Updated guidance and exemplars for reporting systematic reviews. *BMJ*, *372*, n160. https://doi.org/10.1136/bmj.n16

Petkov, T., Mitkova, M., Surchev, S., Popov, S., Todorov, M., Sotirova, E., Sotirov, S., & Bozov, H. (2019). An application of virtual reality technology in education. https://doi.org/10.1109/EAEEIE46886.2019.900041

Prakash, A., & Rajendran, R. (2023). Investigating the interaction behaviors of learners in the VR learning environment. *Conference Name (if applicable)*. https://doi.org/10.1109/VRW58643.2023.0030

Rodrigues, I. S. (2023). IoT as assistive technology: Applications in education as a tool for inclusion. *International Journal of Technology in Education*. https://doi.org/10.46328/ijte.357 doi:10.46328/ijte.35

Roy, R., Babakerkhell, M. D., Mukherjee, S., Pal, D., & Funilkul, S. (2023). Development of a framework for Metaverse in education: A systematic literature review approach. *IEEE Access, 11*, 57717–57734. https://doi.org/10.1109/ACCESS.2023.3283273, doi: 10.1109/ACCESS.2023.328327

Sabarinath, P. (2023). A study on Metaverse in education. https://doi.org/10.1109/ICCMC56507.2023.10083910

Saha, S. K., Banik, S. T., & Banik, T. (2023). A new era of mobility: Exploring digital twin applications in autonomous vehicular systems. *arXiv.org*. https://doi.org/10.48550/arXiv.2305.16158

Sahin, D. R., & Yilmaz, M. (2020). The effect of augmented reality technology on middle school students' achievements and attitudes toward science education. *Computers & Education*, *144*, 103710. https://doi.org/10.1016/j.compedu.2019.103710, doi:10.1016/j.compedu.2019.10371

Saidin, N. F., Halim, N. D. A., & Yahaya, N. (2016). Designing Mobile Augmented Reality (MAR) for Learning Chemical Bonds. https://doi.org/10.1007/978-981-10-0237-3_3

Sailesh, B. S. (2023). Transformation of higher education and research using the Internet of Things. *International Journal of Innovative Technology and Exploring Engineering*. https://doi.org/10.35940/ijitee.g9601.0612723, doi:10.35940/ijitee.g9601.061272

Schlemmer, E. and Backes, L. (2015). Learning in the metaverses: Coexisting in real virtuality. *IGI Global*. https://doi.org/10.4018/978-1-4666-6351-0", doi:10.4018/978-1-4666-6351-

Sin, Z. P. T., Ye, J., Wu, A. C. H., Zhao, I. D., Li, R. C., Ng, P. H. F., Huang, X., Baciu, G., Cao, J., & Li, Q. (2023). Toward an Edu-Metaverse of knowledge: Immersive exploration of university courses. *IEEE Transactions on Learning Technologies*, *16*(6), 1096–1110. https://doi.org/10.1109/TLT.2023.3290814", doi: 10.1109/TLT.2023.329081

Thongprasit, J., & Piriyasurawong, P. (2022). Metaverse for developing engineering competency. *2022 Research, Invention, and Innovation Congress (RI2C),* 119–123. https://doi.org/10.1109/RI2C56397.2022.991029

Tlili, A., Altinay, F., Huang, R., Altinay, Z., Olivier, J., Mishra, S., Jemni, M., & Burgos, D. (2022). Are we there yet? A systematic literature review of open educational resources in Africa: A combined content and bibliometric analysis. *PLOS ONE, 17*(1), e0262615. https://doi.org/10.1371/journal.pone.0262615, doi:10.1371/journal.pone.026261

Tlili, A., Huang, R., Shehata, D., Liu, J., Zhao, J., Metwally, A. H. S., Wang, H., Denden, M., Bozkurt, A., Lee, L. H., Beyoglu, D., Altinay, F., Sharma, R. C., Altinay, Z., Li, Z., Liu, J., Ahmad, F., Hu, Y., Salha, S., Abed, M., & Burgos, D. (2022). Is the metaverse in education a blessing or a curse: A combined content and bibliometric analysis. *Smart Learning Environments, 9*, 24. https://doi.org/10.1186/s40561-022-00205-x, doi: 10.1186/s40561-022-00205-

Van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics, 84*(2), 523–538. https://doi.org/10.1007/s11192-009-0146-3, doi:10.1007/s11192-009-0146-

Wardhani, I. P. (2023). Sistem human computer interaction (HCI) keamanan rumah pintar berbasis IoT. *Jurnal Ilmu Komputer dan Informatika*. https://doi.org/10.54082/jiki.29 doi:10.54082/jiki.29;

Wu, J., & Gao, G. (2022). Edu-Metaverse: Internet education form with fusion of virtual and reality. *Advances in Social Science, Education and Humanities Research, 664*, 1082–1085

Zaga, F. A. Ramos. (2023). Enhancing interactive learning in engineering education: Exploring the Metaverse's virtual horizons. *2023 World Engineering Education Forum-Global Engineering Deans Council (WEEF-GEDC)*, 1-6. https://doi.org/10.1109/WEEF-GEDC59520.2023.10343748, doi: 10.1109/WEEF-GEDC59520.2023.1034374

Zahra, S., Silvianita, A., Pradana, M., & Utami, F. N. (2021). Analysis of factors affecting the work motivation of teachers at the state private vocational school 08 Kab. In Sleman. *Proceedings of the International Conference on Industrial Engineering and Operations Management*, April 5–8, 2021

Zheng, W., Yan, L., Zhang, W., Ouyang, L., & Wen, D. (2023). DKI: Data-knowledge-driven group intelligence framework for smart service in education Metaverse. *IEEE Transactions on Systems, Man, and Cybernetics*. https://doi.org/10.1109/TSMC.2022.3228849, doi: 10.1109/TSMC.2022.322884

Zhonggen, Y. (2023). A systematic literature review of the acceptability of the use of Metaverse in education over 16 years. *Journal of Computers in Education*. https://doi.org/10.1007/s40692-023-00273-z, doi:10.1007/s40692-023-00273-

Zhou, M. (2022). Editorial: Evolution from AI, IoT and big data analytics to Metaverse. *IEEE/CAA Journal of Automatica Sinica, 9*(12), 2041–2042. https://doi.org/10.1109/JAS.2022.10009854, doi: 10.1109/JAS.2022.1000985

Zonaphan, L., Northus, K., Wijaya, J., Achmad, S. S., & Sutoyo, R. (2022). Metaverse as a future of education: A systematic review. *2022 CHIuXiD*, 77–81. https://doi.org/10.1109/CHIuXiD57244.2022.10009854