



ACTA INFOLOGICA (ACIN)

DECEMBER, 2023
Volume: 7 | Issue: 2

ISTANBUL UNIVERSITY, INFORMATICS DEPARTMENT

acin.istanbul.edu.tr
<https://dergipark.org.tr/tr/pub/acin>



Indexing and Abstracting

TÜBİTAK-ULAKBİM TR Dizin
EBSCO Applied Sciences Source Ultimate
Erih Plus
DOAJ
Bielefeld Academic Search Engine (BASE)
OpenAIRE
ResearchBib
ASOS Index



Owner

Prof. Dr. Sevinç GÜLSEÇEN
Istanbul University, Department of Informatics, Istanbul, Türkiye

Responsible Manager

Prof. Dr. Sevinç GÜLSEÇEN
Istanbul University, Department of Informatics, Istanbul, Türkiye

Correspondence Address

Istanbul University, Department of Informatics
Kalenderhane Mahallesi, 16 Mart Şehitleri Caddesi, No: 8
Vezneciler, Fatih, İstanbul, Türkiye
Phone: +90 212 440 00 00/10037
E-mail: acin@istanbul.edu.tr
<http://iupress.istanbul.edu.tr/tr/journal/acin/home>
<https://dergipark.org.tr/tr/pub/acin>

Publisher

Istanbul University Press
İstanbul Üniversitesi Merkez Kampüsü, 34452 Beyazıt,
Fatih, İstanbul, Türkiye
Phone: +90 212 440 00 00

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.

Publication Type: Periodical

EDITORIAL MANAGEMENT BOARD

Editors-in-Chief

Prof. Dr. Sevinç GÜLSEÇEN – İstanbul University, Department of Informatics, İstanbul, Türkiye – gulsecen@istanbul.edu.tr

Co-Editors-in-Chief

Assoc. Prof. Dr. Çiğdem EROL – İstanbul University, Department of Informatics, İstanbul, Türkiye – cigdems@istanbul.edu.tr

Section Editors

Dr. Serra ÇELİK – İstanbul University, Department of Informatics, İstanbul, Türkiye – serra.celik@istanbul.edu.tr

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

Assoc. Prof. Fatma Önay KOÇOĞLU – Muğla Sıtkı Koçman University, Faculty of Engineering, Software Engineering Department, Muğla, Türkiye – fonaykocoglu@mu.edu.tr

Publicity Manager

Şüheda ŞENYUVA – İstanbul University, Faculty of Communication, İstanbul, Türkiye – suhedasenyuva@gmail.com

Language Editor

Elizabeth Mary EARL – İstanbul University, Department of Foreign Languages, İstanbul, Türkiye – elizabeth.earl@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, Türkiye – timorm@istanbul.edu.tr

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

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

Prof. Dr. Selim YAZICI – İstanbul University, Faculty of Political Sciences, Department of Business Administration, İstanbul, Türkiye – 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, Türkiye – t.bensghir@hbv.edu.tr

Prof. Dr. Üstün ÖZEN – Atatürk University, Faculty of Economics and Administrative Sciences, Department of Management Informations Systems, Erzurum, Türkiye – 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, Cezayir – laffi.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, Türkiye – csutcu@marmara.edu.tr

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

Assist. Prof. Dr. Enis KARAARSLAN – Muğla Sıtkı Koçman University, Faculty of Engineering, Department of Computer Engineering, Muğla, Türkiye – enis.karaarslan@mu.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 Reis – İstanbul University-Cerrahpaşa, Hasan Ali Yücel Faculty of Education, Department of Computer Education and Instructional Technology, İstanbul, Türkiye – ayvazzer@iuc.edu.tr

Dr. Luis Miguel CARDOSO – University of Lisbon, Polytechnic Institute of Portalegre, Centre for Comparative Studies, Portugal – lmc Cardoso@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

CONTENTS / İÇİNDEKİLER

RESEARCH ARTICLES / ARAŞTIRMA MAKALELERİ

- A Shorter Form of the Game User Experience Satisfaction Scale in Turkish: GUESS-20-TR
Oyun Kullanıcı Deneyimi Tatmin Ölçeği Türkçe Kısa Versiyonu: GUESS-20-TR
Mehmet İlker Berkman, Çakır Aker 229
- Determining the Happiness Class of Countries with Tree-Based Algorithms in Machine Learning
Makine Öğrenmesinde Ağaç Tabanlı Algoritmalarla Ülkelerin Mutluluk Sınıfının Belirlenmesi
Merve Doğruel, Selin Soner Kara 243
- Converting Image Files to LaTeX Format Using Computer Vision, Natural Language Processing, and Machine Learning
Resim Formatındaki Dokümanların Bilgisayarlı Görü, Doğal Dil İşleme ve Makine Öğrenmesi Kullanılarak Latex Formatına Dönüştürülmesi
Murat Kazanç, Tolga Ensari, Mustafa Dağtekin 253
- Correlation Analysis of the Relationship between Demographic Variables, Computer Self-Efficacy, and Information-Seeking Behavior of Nigerian University Students
Tunde T. Oyedokun, Medinant D. Laaro, Zainab O. Ambali, Olabisi F. Adesina 267
- A Deep Learning-Based Classification Study for Diagnosing Corneal Ulcers on Ocular Staining Images
Oküler Boyama Görüntülerinde Kornea Ülserinin Teşhisi İçin Derin Öğrenmeye Dayalı Bir Sınıflandırma Çalışması
Onur Seveli 281
- Performance Evaluation of Magnitude-Based Fuzzy Analytic Hierarchy Process (MFAHP) Method
Magnitüde Bağlı Bulanık Analitik Hiyerarşi Süreci (MBAHS) Yöntemi Performans Değerlendirmesi
Barış Tekin Tezel, Ayşe Övgü Kımay 293
- Vision-Based Amateur Drone Detection: Performance Analysis of New Approaches in Deep Learning
Görüntü Tabanlı Amatör Drone Tespiti: Derin Öğrenmede Yeni Yaklaşımların Performans Analizi
Ahmet Aydın, Tarık Talan, Cemal Aktürk 308
- Digital Transformation and Innovation in Health for Future Health Services: Turkey Global Innovation Index Time Series Analysis Between 2018 and 2022
Geleceğin Sağlık Hizmetleri için Sağlıkta Dijital Dönüşüm ve Inovasyon: 2018-2022 Yılları Arası Türkiye Inovasyon İndeksi Zaman Serileri Analizi
Ayça Asena Özdemir, Zehra Alakoç Burma 317

CONTENTS / İÇİNDEKİLER

RESEARCH ARTICLES / ARAŞTIRMA MAKALELERİ

Comparison of Outlier Detection Methods in Linear Regression: A Multiple-Criteria Decision-Making Approach

Doğrusal Regresyonda Uç Değer Tespit Yöntemlerinin Karşılaştırılması: Çok Kriterli Karar Verme Yaklaşımı
Mehmet Hakan Satman 333

Query by Image Examination: Classification of Digital Image-Based Forensics Using Deep Learning Methods
Görüntü İncelemesine Göre Sorgulama: Dijital Görüntü Tabanlı Adli Görüntülerin Derin Öğrenme Yöntemleri Kullanılarak Sınıflandırılması

İlker Kara 348

The Efficiency of Regularization Method on Model Success in Issue Type Prediction Problem

Sorun Türü Tahmini Probleminde Düzenleştirme Yönteminin Model Başarısı Üzerindeki Etkisi

Ali Alsaç, Mehmet Mutlu Yenisey, Murat Can Ganiz, Mustafa Dağtekin, Taner Ulusinan 360

The Future of Smart Campuses: Combining Digital Twin and Green Metrics

Akıllı Kampüslerin Geleceği: Dijital İkiz ve Yeşil Metriklerin Birleştirilmesi

İlknur Teke, Orkun Teke, Murat Kılınc 384

REVIEW / DERLEME


A Systematic Literature Review for New Technologies in IT Audit

Bilgi Teknolojileri Denetiminde Yeni Teknolojiler Üzerine Bir Sistemik Literatür Taraması

Nur Sena Tanrıverdi, Nazım Taşkın 396

A Shorter Form of the Game User Experience Satisfaction Scale in Turkish: GUESS-20-TR

Oyun Kullanıcı Deneyimi Tatmin Ölçeği Türkçe Kısa Versiyonu: GUESS-20-TR

Mehmet İlker Berkman¹ , Çakır Aker² 

¹(Assoc. Prof.) Bahcesehir University, Faculty of Communication, Department of Communication Design, Istanbul, Türkiye

²(Assist Prof.) Bahcesehir University, Faculty of Communication Department of Digital Game Design, Istanbul, Türkiye

Corresponding author : Çakır AKER

E-mail : cakir.aker@bau.edu.tr

ABSTRACT

Games User Research (GUR) has become an important subdomain of the Human-Computer Interaction field within the last few years, as gaming has become a daily entertainment for many people rather than being in the interest of a few game enthusiasts. Researchers require specific tools to measure the users' responses and attitudes towards the games. Game User Experience Satisfaction Scale is one of the recent additions to GUR tools, which had already been adapted into Turkish as GUESS-TR. Through this study, we aimed to verify a shorter form of GUESS-TR which is compatible with a currently existing English shorter version with 18 items that measure game user experience through 9 factors. Data revealed that a 20-item version resembling 10 factors, namely GUESS-20-TR, is a valid and reliable measure of game user experience. We provided evidence for construct validity through confirmatory factor analysis. Spearman-Brown prophecy coefficients indicate that the 2-item subscales are reliable. Heterotrait - monotrait ratios show that items indicate different constructs, i.e. discriminant validity. Based on Pearson correlation, the mean scores obtained with the short form GUESS-20-TR are highly consistent with the 51-item Turkish version.

Keywords: Games user research, games user experience, scale adaptation

ÖZ

Oyunların sadece tutkunlarına özel bir alan olmaktan çıkıp pek çok insan için gündelik bir eğlence haline gelmesi ile, Oyun Kullanıcı Araştırmaları (OKA), İnsan-Bilgisayar Etkileşimi alanının önemli bir alt dalı haline gelmiştir. Araştırmacılar, oyuncuların oyunlara verdikleri tepkileri ve oyunlara karşı tutumlarını ölçmek için özelleştirilmiş araçlara ihtiyaç duymaktadırlar. Oyun Kullanıcı Deneyimi Ölçeği (GUESS) de son yıllarda OKA araçlarına eklenmiş olup, GUESS-TR adı ile Türkçe'ye de çevrilmiştir. Bu çalışmada, 51 sorudan oluşan GUESS-TR'nin kısa bir versiyonu geliştirilmiştir. Halihazırda 9 faktörü ölçen 18 soruluk kısaltılmış GUESS-18 versiyonu da göz önüne alınarak, 20 sorudan oluşan ve 10 faktörü ölçen bir varyant ortaya konmuştur. Önaylayıcı faktör analizi ölçüm modelinin yapısal geçerliliğinin sağlanması yapılmıştır. İki sorudan oluşan ölçekler için önerilen Spearman-Brown korelasyon katsayısı alt ölçeklerin güvenilirliğine delalet etmektedir. Heterotrait- monotrait oranları, her bir ölçeğin farklı yapıları ölçtüğünü, yani ayırıcı geçerliliği ortaya koymaktadır. Pearson korelasyon katsayıları göstermektedir ki, her bir faktör için 2 soru olmak üzere 20 soru ile elde edilen ortalama skorlar, 51 sorudan oluşan uzun ölçek ve alt ölçeklerinin ortalama skorları ile yüksek seviyede tutarlılık göstermektedir.

Anahtar Kelimeler: Oyun kullanıcı araştırması, oyuncu deneyimi, ölçek adaptasyonu

Submitted : 07.10.2022

Revision Requested : 31.05.2023

Last Revision Received : 03.06.2023

Accepted : 22.08.2023

Published Online : 17.10.2023



This article is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0)

1. INTRODUCTION

The field of Game User Research is a rapidly developing direction within the scope of human-computer interaction. Many different researchers have attempted to evaluate game user experience and proposed methods to measure player experience from different perspectives. The methods used to evaluate the user experience with the conventional approaches are effective for evaluating the applications with the aim of productivity in the most basic sense. However, they are insufficient in examining the user experience in games, since games are deliberately designed to be challenging, unlike productivity applications. The techniques and approaches in the field of user experience basically aim to minimize possible problems or obstacles in the application. On the contrary, obstacles and difficulties in digital games appear as elements of fun. (Aker, Rizvanoglu, & Bostan, 2017).

As given in the Related Studies section, researchers from different disciplines have different approaches to user experience in games and aim to analyze experience with heuristic methods through expert evaluation, assessment based on players' performance, ad hoc surveys, usability questionnaires or game-specific scales such as GUESS (Phan, Keebler & Chaparro, 2016). Undoubtedly, understanding the player experience is also essential for the gaming industry. In the face of increasing competition, many game companies are trying to make inferences from the field of player experience. An approach that is preferred today and which is increasingly common in game studies is based on analyzing the player experience through mixed methods, especially with the methods using game heuristics in expert evaluations. However, contrary to this qualitative perspective in the field, quantitative approaches are very few. The Game User Experience Satisfaction Scale (GUESS), which was introduced to fill this gap, is a relatively recent addition to the game user research toolkit. It has become one of the most used player experience evaluation methods in recent years. In this study, a short version of the GUESS was collated and its suitability was examined, using the Turkish GUESS-TR items and data obtained by Berkman, Bostan & Senyer (2022a; 2022a), considering the 18-item short-form version suggested by Keebler et al. (2020). GUESS-18 is validated for its construct validity, convergent validity, discriminant validity and reliability.

Phan et al. (2016), suggested the following definition for the term playability/usability: "The ease with which the game can be played with clear goals and objectives in the user's mind, without being hindered by the user interface and controls, with minimal cognitive interference". As for the scale suggested by Keebler et al. (2020), only items related to the game interface were included, and items related to controls and clarity of game objectives were not. On the other hand, Berkman et al. (2022a) suggested that some of the items of "usability" are about the game interface and controls. On the other hand, they also examined the items related to game objectives as a separate dimension under the title of "playability". It was emphasized that the scale, named GUESS-TR after different validation analyses, is a valid and reliable tool for the field of game user research. This study also overcame the language barrier and produced usable results. However, the adaptation of the GUESS scale to other languages has not gained much popularity, as it is still considered new and is not among the classical scales like GEQ (Game Experience Questionnaire) (Ijsselstein, de Kort and Poels, 2007) or System Usability Scale (SUS) (Brooke, 1996). For that reason, we aimed to present a short and localized version of the GUESS scale based on the recently translated GUESS-TR scale and the short-form GUESS-18. With this scale, which can be put forward in light of the data obtained from our previous research, a useful and up-to-date method can be suggested that would benefit both the industry in the country and academia. This will be presented through the Turkish version of the short GUESS scale, which is considered much more practical, fast, and effective. In order to achieve this, the Turkish translation study and the 18-item version of GUESS will be used.

We decided to create a short version of GUESS in Turkish for several reasons. First of all, it had already been adapted into Turkish with 51 items measuring 10 factors, and the dataset was available as open data. In addition, the original scale was developed by analyzing a large spectrum of prior research. Thirteen surveys on gaming experience, 15 game heuristic lists and, 3 user satisfaction surveys of human-computer interaction were explored to determine the original GUESS items, ensuring the content validity of the measurement tool. Furthermore, there is already a short-form version based on the English version, suggesting that a short-form version is viable, as well as providing a starting point for selecting the smaller set of items.

The impact of technological developments in the gaming industry should also be considered. Due to the rapidly developing technology, the questionnaires used in the assessment of human-computer interaction tend to become outdated. A recently developed measurement tool is more likely to gratify state-of-the-art requirements. Localization is also necessary in the implementation of the scale in order to ensure widespread use and overcome the language barrier.

2. METHODS OF EVALUATING GAMES FOR USER EXPERIENCE

Many of the methods used in Games User Research are inherited from Human-Computer Interaction and usability studies and these methods mainly focus on either users' behavior or users' attitude (Medlock, 2018). Some methods can be applied early in the lifecycle of the game design such as focus groups, interviews, or ethnographic field studies as well as card sorting, personas, and online surveys to envision what the game is going to be like. During the design and development phase, usability tests, physiological measurements, expert reviews, and heuristic evaluations can be used to assess users' behavior, whereas surveys and playtests with interview sessions help to understand users' attitudes. When the game is being released, telemetry analysis on the gameplay data, benchmark tests, and unmoderated usability tests can be used. After the release, A/B tests are employed to explore further updates. Some of these methods are for understanding the target user group of the end product, and others are for determining the product's attributes. Furthermore, these methods may help to identify the business model and provide valuable information about how the product is supposed to be built. Since standardized questionnaires such as GUESS are used in the design and development phase along with other methods, below we provide a review of methods that are mainly used in the design phase.

2.1. HEURISTIC APPROACHES

Malone (1982) proposed a set of heuristics for instructional games and suggested that there were three main heuristics for achieving entertainable interfaces. Three empirical tests were employed to understand what gamers like with a total number of 81 participants. As a result, he proposed the three heuristics categories; challenge, fantasy, and curiosity. Desurvire et al. (2004), proposed the Heuristics of Playability (HEP) framework and prepared a heuristics set of 43 items, based on literature and reviewed by several experts. During the study, the researchers conducted a user-testing method for validating and comparing the results from the heuristic evaluation. Federoff (2002) explored existing game heuristics and collated them to analyze the 'fun' aspect of the games. Five people from a game development team were observed and interviewed to suggest a set of heuristics for the evaluation of video games. On the other hand, Korhonen & Koivisto (2007;2006) were the first to publish playability heuristics for mobile games. They proposed a modular structure for their playability heuristics. There were two phases of the study; the first part involved the use of the three categories of heuristics with different mobile games. Four experts analyzed five mobile games in total. In the second phase, the set was iteratively improved and the experts conducted the test for the second time, but with different games. In their latter study, they included another module for the multiplayer aspect of mobile games. They prepared the heuristics for the multiplayer category by examining three multiplayer mobile games and by conducting a literature study. Schaffer (2008) suggested evaluating usability in video games. The aim of his study was to suggest a guideline for evaluating games. It was indicated that with both the utilization of user tests and expert evaluation methods, it would be possible to analyze the usability of games. 21 heuristics were suggested with five categories: general, graphical user interface, gameplay, control mapping and level design.

2.2. PERFORMANCE-BASED EVALUATIONS

The status of the player and game -world, which can be further analyzed through gameplay videos are widely used indicators of players' performance, where the player achievements and failures can be clearly observed. The in-game scores, which are obtained in many games, are also very easy to acquire for research purposes (Desurvire and Wixon, 2018). Both the gameplay status and the scores can be acquired via telemetry, where game software records the events in the game and reports them to the researchers (Drachen & Connor, 2018). Using game analytics, researchers can acquire performance measures, e.g. time spent on challenges and tasks, in-game objects that were mostly interacted with, or the number of failures in a gameplay session until the player quits the game.

2.3. PHYSIOLOGICAL MEASURES

Physiological measures are usually considered as an indicator of users' emotional state, whereas some measures also indicate the cognitive effort of users (Akan & Berkman, 2020). Although they may include noise and can be difficult to interpret, they are suggested to be immune to researcher or participant bias (Kivikangas et al., 2011). Cardiovascular measures such as heart rate, heart rate variability, and blood pressure can be acquired non-invasively during gameplay, and can be used as an indicator of valence and arousal. Brain-computer technologies such as EEG (electroencephalography) and fMRI (functional magnetic resonance imaging) have been used in game user studies. EEG can be employed to assess the emotional changes and cognitive activities of the players during gameplay (Hafeez et al.,

2021). fMRI mainly depicts cognitive aspects such as immersion, flow, challenge but also reveals data on affection (Ju & Wallraven, 2019). The facial expressions, which can be detected either via electromyography or image processing are assessed as indicators of emotions in several studies (e.g. Isman, Prasasti & Nugrahaeni, 2021). Furthermore, measures of electrodermal activity are employed in game user studies as indicators of emotional changes, however, obtained results are mixed, as some studies did not detect any changes in measurements due to gameplay conditions or could not identify a correspondence between subjective measures. Eye-Related measures, such as fixations and movements of the eye were found to be related to cognitive activity (e.g. Jennet et al., 2008; Mueller, Jackson & Skelton., 2008; Alkan & Çağiltay, 2007) and pupil size indicated the emotional changes (Mojzisch et al., 2006).

2.4. STANDARDIZED QUESTIONNAIRES

Some widely known examples of questionnaires used in the evaluation of gameplay experiences are “Game Experience Questionnaire (GEQ)” (IJsselsteijn et al., 2007), “Gameplay Experience Questionnaire” (Ermi & Mäyrä, 2005), “Player Experience of Need Satisfaction (PENS)” (Ryan, Rigby & Przybylski., 2006) and Immersion questionnaire (Jennett et al., 2008). Phan et al. (2016) criticized the lack of psychometric validation for some of the GEQ (Game Experience Questionnaire (IJsselsteijn et al., 2007) and PENS (Ryan et al., 2006). Some subsequent studies provided partial evidence of the psychometric qualities of the GEQ and PENS (Johnson, Gardner & Perry, 2018; Law, Brühlmann & Mekler, 2018; Berkman & Bostan, 2017) and it is suggested that the GEQ scale is not fully effective in measuring experience (Aker et al., 2017).

Some of the other questionnaires focus on only a single aspect of gaming or are designed to assess specific types of game genres such as interplayer interactions in serious games (Gorsic et al., 2019), attitudes toward game narrative (Qin, Patrick-Rau & Salvendy, 2009) or social presence (de Kort, IJsselsteijn & Poels, 2007). The PLEXQ (Playful Experiences Questionnaire) (Boberg et al., 2015) is not intended only for assessing game user experience, but it is a measure of playfulness, which can also be applied to games.

PXI (Player Experience Inventory) is another recent addition to the GUR toolbox (Abelee et al., 2020), which is highly similar to GUESS in its content. Aiming to assess player experience based on game design choices, there are 10 sub-dimensions with three items each; “meaning,” “mastery,” “immersion,” “autonomy,” “curiosity,” “ease of control,” “challenge,” “progress feedback,” “audio-visual appeal,” and “goals and rules.” Thus, it has similar characteristics with GUESS, in terms of its focus on the relationship between game design elements and player experience. It should be noted that GUESS has more dimensions querying about the game design elements such as narratives and social interaction.

3. STUDIES RELATED WITH THE GUESS

The original GUESS is a 55-item tool that defines video game satisfaction/satisfaction and evaluates nine different constructs. Phan et al. (2016) introduced this psychometrically validated scale for the purpose of comprehensive measurement of video game satisfaction. More than 450 different games with more than 1,300 participants were evaluated in their study. As a result, they produced a 55-item satisfaction scale with nine constructs: Usability/playability, Narrative (NA), Player Engrossment (PE), Enjoyment (EN), Creative Freedom (CF), Audio Aesthetics (AA), Personal Gratification (PG), Social Connectivity (SC), and Visual Aesthetics (VA). The Usability/Playability dimension refers to the ability to play the game without any hindrance due to the game’s interface or controls, as well as the ease of setting and determining goals. The Narrative dimension includes storytelling elements such as characters, events, and fictional elements. The Player Engrossment dimension deals with how much value and dedication the player gives to the game. The Enjoyment, as its name suggests, indicates how much delight the player has and/or enjoys playing. The Creative Freedom dimension asks the players about the extent that they can express themselves as individuals in the game, as well as their curiosity motive. The Audio Aesthetics dimension examines the effect of the sound and music used in the game on the experience, and the Visual Aesthetic dimension expresses the contribution of the visual elements with the same regard. The Personal Gratification dimension relates to the motivations that support the player’s sense of achievement that the game offers. Finally, the Social Connection dimension was introduced to query the players’ thoughts on playing games with other people.

The GUESS has been used in many studies and has succeeded in producing useful results. For example, it has been used previously to evaluate the application of procedural content generation for video games (Wijaya, Hansun & Kristanda, 2019). Xu et al. (2019) used GUESS in a study in Japan to evaluate the effect of "Player Domination Adjustment" on the gaming experience. The same scale was also assessed in the context of various VR games (e.g. Shelstad, Smith, and Chaparro., 2017; Yildirim et al., 2018; Aksayim & Berkman, 2020). Ali, Arumugam & Kumaran

(2021) used GUESS to assess the gamification elements in a medical rehabilitation program. A serious game within the novel concept of “audience participation game with a purpose” had also been evaluated through GUESS dimensions (Nguyen et al., 2020) as well as it was used for assessing a board game (Thevin et al., 2021). Studies show that GUESS is being used not only in GUR studies but also gaining demand in gamification and serious game research.

Findings of these studies also provide evidence for the sensitivity of GUESS, i.e. it is capable of producing significantly different scores regarding the attributes of the games evaluated. Although it is possible to use each dimension of GUESS as an independent measurement tool, responding to 55 items is cumbersome when repeated assessments are required (Keebler et al., 2020). On the other hand, one of the important advantages of GUESS is that each subscale can be used independently from the other, based on the requirements of the researcher. However, a 55-item scale is not practical to obtain an overall game user experience score. Hence, the necessity of introducing a much shorter version of the GUESS emerged, and Keebler et al. (2020) suggested an 18-item version, in which each of the 9 dimensions of the original GUESS is assessed through two items. A configurable 55-item model and an initial 18-item model of the GUESS were assessed based on the data obtained from 419 valid surveys from participants aged between 18 to 72 ($M = 35.11$, $SD = 11.63$). The final 18-item model was assessed through a total of 197 valid responses from participants ages ranging from 18 to 68 ($M = 33.21$, $SD = 10.90$). The final scale, which is called GUESS-18, is a valid and reliable measure and it is stated that it can be acquired as a short, practical, and comprehensive measure of game satisfaction for practitioners and researchers.

4. TRANSLATING THE GUESS INTO TURKISH

As mentioned before, removing the language barrier is also an important factor in practice. Non-English survey respondents can belikely to experience a foreign language as confusing, even if they are bilingual. For this reason, it should be considered an important criterion for the aforementioned scale’s practicality which should be adapted to Turkish, to be presented to Turkish gamers. Although heuristics and in-depth interviews meet some of the industry’s needs, the growing game industry in Turkey and the increase in game user research initiatives need reliable and valid measurement tools such as GUESS. However, employees and teams in the industry have difficulties in implementing this scale, which is not available in their users’ mother tongue. The GUESS, which is used much more frequently today compared to other scales, was translated into Turkish by Berkman et al., (2022a) for this very reason. In the study, whereas the GUESS was translated into Turkish through a series of expert reviews and back-translation processes, the translated set was tested in the laboratory environment and in the participants’ homes, which can be considered their natural environment for playing games. During the translation, they received feedback from three different English language experts and ensured that the language of the scale was translated properly. In the study, it was underlined that it is insufficient to simply translate materials for adaptation research since the translators should have an in-depth understanding of the subject matter and culture. Hence, in the abovementioned study, two of the authors took on the role of translators and translated GUESS items in Turkish, paying close attention to the original statements and Turkish gamers’ understanding and vocabulary. The translated version was then presented to three English language specialists who were either licensed translators or foreign language instructors. They gave each translation a grade ranging from 1(proper) to 3 (improper), and if they gave the translation of an item a grade other than “1”, they were invited to give an alternative translation (Berkman et. al., 2022a). Subsequently, six different games were examined in their study. The Turkish version of the entire 55-item set of GUESS was tested with Turkish participants. As a result of the measurements made with the data set consisting of 449 questionnaires, in which 121 participants evaluated 6 games in total, 51 items were retained, but it was pointed out that the ‘Playability (PL) and the Usability (US) factors should be kept separately, both structurally and conceptually.

5. METHODOLOGY

5.1. DATA SET & PARTICIPANTS

The data set used by Berkman, et al. (2022a) and published openly (Berkman, Bostan & Şenyer, 2022b) was employed in our analysis. As mentioned above, 6 games were evaluated by 121 participants in the study. One game was played on a game console platform whereas one was played on a mobile platform, in the laboratory prepared for play testing. The other four games were played on the participants’ personal computers at their homes. As for the selection of the games, similar to studies of Phan et al. (2016) and Keebler et al. (2020), it is not left to the player’s own choice, and the games chosen by the researchers are requested to be played. The game design students forming the sample are relatively more qualified in terms of domain knowledge than other players. The list of games as given by Berkman et al. (2022a) is summarized below in Table 1.

Table 1. The list of games and their attributes, Berkman et al. (2022a)

Game	Company/ Year	Platform	Genre	Environment	Narrative
Super Mario Odyssey	Nintendo, 2017	Nintendo Switch	Platformer and action game	Cartoonish and vivid colors	Level-based narrative with environmental cues
Sniper Elite 4	Rebellion, 2017	PlayStation 4	Third-person tactical shooter stealth game	Realistic	Fast-paced, action-oriented storytelling with fast beats
Contrast	Compulsion Games, 2013	PC	Indie puzzle and adventure game	Gloomy and cartoonish	Slow-paced storytelling relying on the puzzles and the environment
Control	Remedy Entertainment, 2019	PlayStation 4	Supernatural action-adventure game	Realistic, dark, and red colors are highly used to give a mysterious and dangerous atmosphere	Mixed-paced (slow and fast) storytelling around the challenges
Hellblade: Senua's Sacrifice	Ninja Theory, 2017	PC	Action-adventure game	Realistic, gloomy, and mysterious atmosphere	Fast-paced, action-oriented storytelling as the player fights through the world
The Council	Big Bad Wolf, 2018	PC	Adventure role-playing game	Gloomy and semi-realistic	Complex, slow-paced storytelling enriched by deep characters

As a part of a course they took, the participants were asked to play the above games and write a report on the narratives of these games, and they were also asked to answer the GUESS items given in Turkish. As participants, there were 24 females and 97 males between the ages between 19 and 32 ($M = 21.01$, $SD = 1.86$). Approximately one-third of the participants ($N=35$) evaluated all the games (210 evaluations), which corresponds to half of the evaluations. 20 participants evaluated five of the games (100 evaluations), 5 participants evaluated four, 15 participants evaluated three, 28 participants evaluated two, and 18 participants evaluated only one of the games.

6. ANALYSIS

Our analysis included both the 10-factor solution (Berkman et al., 2022a) through 20 items and the 9-factor solution (Keebler et al., 2020; Phan et al., 2016) through 18 items comparatively. Indicators are reported also in comparison with the results obtained for long-form scales.

Cronbach's alpha is reported as an indicator of reliability for the 18 and 20 items as an overall measure. As the short version has 2 items per dimension, the Spearman-Brown Prophecy Formula was used instead of Cronbach's alpha, as it is suggested for two-item scales (Eisinga, Grotenhuis & Pelzer, 2013). Cronbach's alpha values are reported only to be compared with previous short-form studies, but they were not intended as a valid indicator of reliability for two-item subscales.

Standardized Root Mean Squared Residual (SRMR), Normed Fit Index (NFI), Tucker-Lewis Index (TLI), Comparative Fit Index (CFI), and Root-mean-square error of Approximation (RMSEA) values are reported as indicators of model fit for assessment of construct validity, using the widely accepted threshold values (see Hooper, Coughlan & Mullen, 2008).

For discriminant validity, the Fornell-Larcker criterion regarding the comparison of Average Variance Extracted (AVE) and Maximum Shared Variance (MSV) is employed to provide results that can be compared with the findings of Keebler et al. (2020). However, we have taken the Heterotrait Monotrait (HTMT) ratio as the main method of assessing discriminant validity, as it is suggested to be superior to the Fornell-Larcker criterion (Henseler, Ringle & Sarstedt, 2015). In order to determine the convergent validity mean scores obtained from the 51-item GUESS-TR and 20-item GUESS-20-TR are compared through Pearson correlations. Item E03 is reverse coded before the analysis since it has a

negative statement. The overall score and the scores for the subscales are calculated as mean values of the corresponding items.

7. RESULTS AND DISCUSSION

7.1. CONSTRUCT VALIDITY

Both the 9-dimension (18 items) Phan-Keebler model and the 10-dimension (20 items) Berkman et al. (2022a) model lead to a Heywood case on the Social Connectivity dimension, with a negative variance estimate on the SOC2 item. Since having a negative variance is impossible, the model should be corrected. Reasons for a Heywood case could be the small-sized sample for adequate estimation of parameters, a dataset that is not normally distributed or has many outliers, or a misspecified model. Furthermore, very high and very low correlations are also suggested as a reason for the Heywood cases (Rindskopf, 1984). Since the model is known to be working with other datasets (Keebler et al, 2020) and different measurement items on the same dataset (Berkman et al, 2022b), we explored the data quality of the Social Connectivity items. As given in Table 2, the Shapiro-Wilk tests greater than 0.05 indicate a normal distribution of data. However, skewness is greater than $|2|$ for the SC01 item. The relatively lower standard deviation (SD) of 1.65 also indicates a lesser variance compared to other items. Furthermore, the item is very close in content to the item SC02. Besides, there is a higher correlation between SC02 and 03 ($r=.657$, $p<0.001$) compared to the correlation between SC01 and SC03 ($r=.412$, $p<0.001$). Considering these issues, we decided to replace item SC01 with item SC03 and keep the SC02 item.

Table 2. Descriptives for Social Connectivity Items

	SC01	SC02	SC03
N	292	272	327
Missing	157	177	122
Mean	3.27	3.99	3.94
Median	3.00	4.00	4
Standard deviation	1.65	1.75	2.05
Skewness	0.355	-0.109	-0.00343
Std. error skewness	0.143	0.148	0.135
Kurtosis	-0.704	-0.776	-1.27
Std. error kurtosis	0.284	0.294	0.269
Shapiro-Wilk W	0.923	0.930	0.905
Shapiro-Wilk p	< .001	< .001	< .001

This helped us to achieve validity for both the 9-dimension and 10-dimension measurement models, without any parameter estimate having an impossible value. Many of the model fit indicators given in Table 3 provide evidence that both measurement models with Turkish items have construct validity except the TLI value. However, it is also close to the suggested threshold of 0.95, but lower than the values observed in Keebler et al. (2020) findings with initial and final short models. Since other studies did not report their findings on TLI, it is not possible to make a comparison. However, we observe a minor improvement in TLI values and SRMR indicator when the model is based on 18 Turkish GUESS items measuring 9 dimensions, compared to the 20-item model measuring 10 dimensions. When the CFI and RMSEA are compared to the previous studies, our findings are very similar to the observation of Phan et al. (2016) based on 55 items measuring 9 dimensions but slightly poorer than Keebler et al. (2020) findings.

Table 3. Model Fit Indicators

	SRMR (.05< <.08)	NFI closer to 1	TLI (>.95)	CFI (>.95)	RMSEA (<.08/<.05)	χ^2 , df, p
GUESS-20-TR	0.035	0.932	0.94	0.96	0.052	278, 125, p<0.001
GUESS-18-TR	0.032	0.938	0.942	0.962	0.055	234, 99, p<0.001
GUESS-TR 10 dimension	0.089	0.705				4821.087, NA, p<0.001
Original GUESS				0.82	0.053	4428.63, 1394, p<0.001
Kebbler's configural model			0.866	0.875	0.050	2827.186, 1391, p<0.001
Kebbler's initial short model			0.96	0.975	0.042	171.966, 99, p<0.001
Kebbler's final short model			0.961	0.974	0.043	137.015, 100, p<0.001
GUESS-GA-18			0.932	0.957	0.041	179.077, 108, p<0.001

7.2. RELIABILITY

Although the previous studies reported Cronbach's alpha as the indicator of reliability, we embraced the Spearman-Brown Prophecy Formula which is recommended for 2-item scales as a better indicator of reliability compared to Cronbach's alpha (Eisinga et al., 2013), as the Brown formula assumes that the split-halves are parallel measures. Whereas its first row reports the Spearman-Brown correlations for each subdimension, Table 4 also depicts the Cronbach's alpha values of 18 and 20-item versions in comparison with previous studies.

Results show that all of the values are above the generic threshold of .70 for reliability for the Spearman-Brown correlation values, except the Usability dimension. However, it is at an acceptable level to consider the two item measurement as a reliable scale

Table 4. Reliability indicators for each dimension

	PL	US	NA	VA	AA	CF	EN	PE	PG	SC
GUESS-20-TR / Sp.Br. r	0.765	0.680	0.889	0.915	0.81	0.888	0.884	0.923	0.802	0.881
GUESS-20-TR / Cr0. α	0.620	0.515	0.8	0.843	0.68	0.799	0.792	0.857	0.670	0.793
GUESS-TR 10 dimension	0.744	0.823	0.869	0.844	0.832	0.84	0.923	0.92	0.801	0.786
Original GUESS		0.83	0.85	0.79	0.89	0.86	0.8	0.81	0.72	0.86
Kebbler's final short model		0.769	0.809	0.818	0.890	0.819	0.8	0.722	0.771	0.794
GUESS-GA-18 / Sp.Br. r		0.665	0.807	0.722	0.777	0.61	0.56	0.575	.645	0.545

When the overall items' reliability is assessed through Cronbach's alpha, the GUESS-20-TR reveals a value of 0.902, slightly lower than the 18-item version with 0.907. The original 55-item GUESS is reported to have a Cronbach's alpha value of 0.93 (Phan et al, 2016) and the Keebler et al. (2020) final 18-item model revealed a score of 0.785.

The following Figure 1 shows the factor loadings for the 20-item version with 10 dimensions, revealing that items have strongly loaded on their intended latent variable, which is also a common indicator of construct validity.

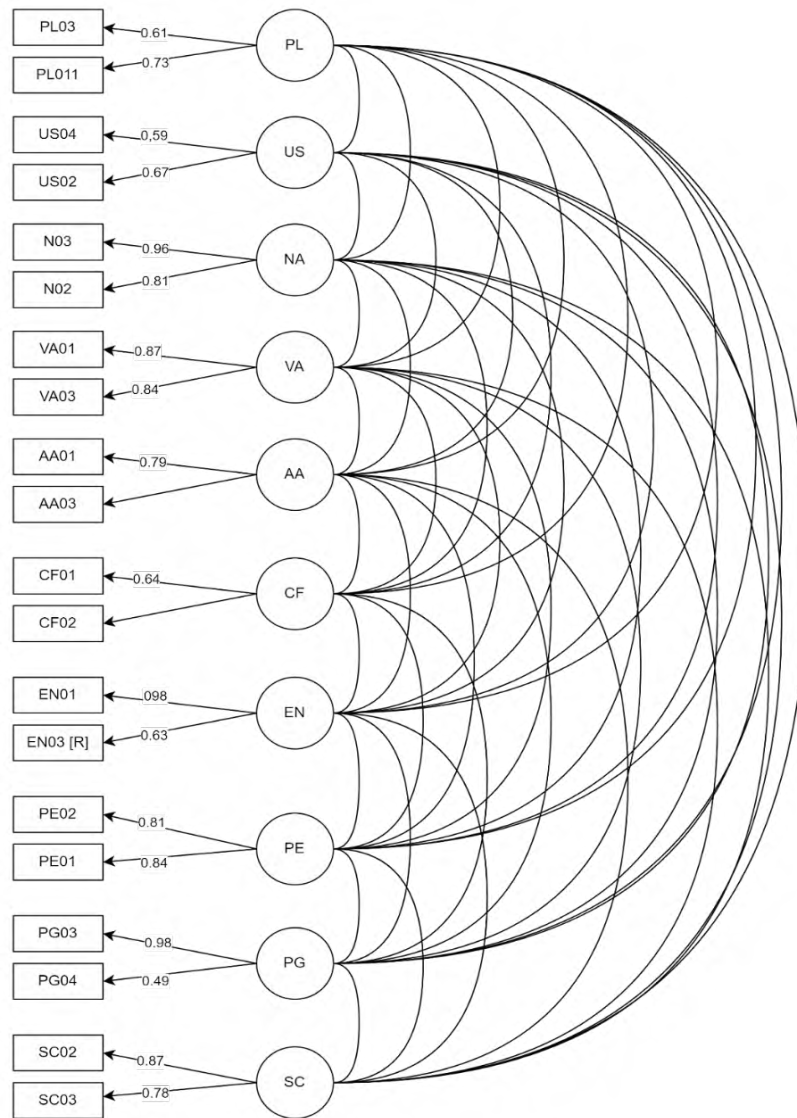


Figure 1. Measurement Model and Factor Loadings. See Table 5 for Factor Correlations

7.3. DISCRIMINANT VALIDITY

We evaluated both the 18-item and 20-item versions of GUESS-TR through the Fornell-Larcker criterion to provide results that are comparable to Keebler et al. (2020) and Berkman et al. (2022a). However, the HTMT ratio is shown to be a superior method of assessing discriminant validity, compared to the Fornell-Larcker criterion (Henseler et al., 2015). We also evaluated GUESS-18-TR and GUESS-20-TR through a conservative HTMT ratio criterion of 0.85.

According to the Fornell-Larcker criterion, the $\sqrt{\text{AVE}}$ values of each dimension should exceed its correlation to any of the other dimensions. As given in Table 5, our two-item short version GUESS-TR meets this criteria, except the $\sqrt{\text{AVE}}$ of Playability dimension being lower than the correlation of Playability with Usability and Enjoyment. It should be remembered that Playability is suggested by Berkman et al. (2022a), by separating the united Usability/Playability dimension suggested by Phan et al (2016) and embraced by Keebler et al. (2020). According to the Fornell-Larcker criterion, the two-item Playability dimension is not a unique measure, although it appeared to be discriminated from other dimensions in Berkman et al (2022a) study, with a $\sqrt{\text{AVE}}$ of .748 value which is higher than its correlations with other dimensions.

On the other hand, the HTMT ratio values we observed in the 10-dimension model confirm the discriminant validity

Table 5. AVE and Factor Correlations. $\sqrt{\text{AVE}}$ values are given in bold and italic.

$\sqrt{\text{AVE}}$	Playability	Usability	Narratives	VisualAesth	AudioAesth	CreativeFree	Enjoyment	PlayerEngross	PersonalGrat	SocialConnect
Playability	<i>0.567</i>	0.836	0.448	0.285	0.433	0.424	0.638	0.299	0.514	0.162
Usability		<i>0.848</i>	0.437	0.555	0.455	0.398	0.567	0.373	0.358	0.171
Narratives			<i>0.844</i>	0.575	0.645	0.618	0.732	0.624	0.544	0.201
VisualAesth				<i>0.732</i>	0.696	0.393	0.525	0.492	0.455	0.208
AudioAesth					<i>0.803</i>	0.423	0.686	0.533	0.572	0.303
CreativeFree						<i>0.824</i>	0.536	0.642	0.542	0.128
Enjoyment							<i>0.867</i>	0.607	0.608	0.302
PlayerEngross								<i>0.736</i>	0.554	0.163
PersonalGrat									<i>0.792</i>	0.191
SocialConnect										<i>0.661</i>

for all dimensions. As shown in Figure 2, none of the HTMT ratio values exceed the conservative threshold of 0.85. In Berkman et al. (2022a) study, the HTMT ratio of Narratives to Creative Freedom was reported to be 0.89, but still in line with the more liberal 0.90 HTMT ratio threshold.

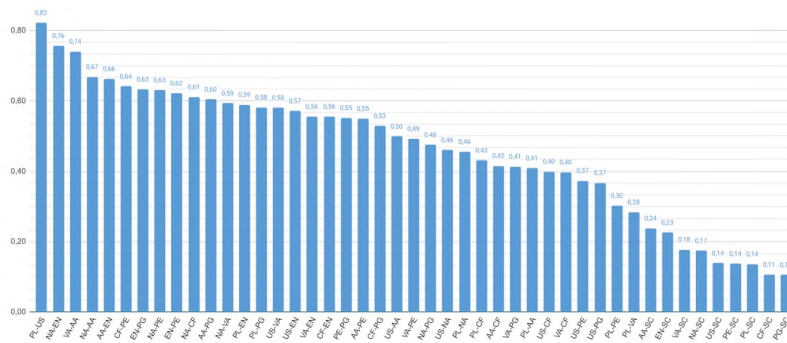


Figure 2. Heterotrait - Monotrait ratios of subscales from highest to lowest

7.4. CONVERGENT VALIDITY

Mean scores obtained with the 51-item GUESS-TR and the short form 20-item GUESS-20-TR scores are compared to assess the convergent validity of the short version. The overall mean scores are strongly correlated at a significant level; $r=0.975$; $p<0.001$. The weakest correlation is observed on the Creative Freedom dimension; $r=0.856$, $p<0.001$, but still shows a strong correlation between measurement made with 7 items in GUESS-TR and the two-item short version. Another relatively weaker correlation is observed on Personal Gratification; $r=0.884$, $p<0.001$, which is measured via 5 items in the long-form scale. The Usability mean scores also correlate at 0.887 ($p<0.001$). The other 7 dimensions have strong correlations between 0.91 to 0.965 at a significant level ($p<0.001$). These results show that the 20-item GUESS-20-TR is capable of producing results that are very similar to the 51-item version.

8. CONCLUSION

Our results show that the 20-item short version of GUESS-TR, given in Appendix-I, is a valid and reliable measure of game user experience with two items per dimension. The Turkish dimension names are given as suggested by Bostan (2022). Although GUESS-TR is shown to be applicable with 18 items indicating a 9-factor model, the additional Playability dimension is determined to be different from the other 9 dimensions. Both models are verified for their construct validity according to several model fit criteria. For the 10-dimension model with 20 items, Playability fails to be distinguished from Usability according to the Fornell- Larcker criterion, but the novel HTMT approach provides evidence that the two-item measure of Playability is different from the two-item measure of Usability. As previously discussed in Berkman et al. (2022a), these item sets are conceptually different. The Usability items query the user about the ergonomics of the interface elements and the controls, whereas the Playability items ask about the clarity of goals and self-confidence of the player about the actions that need to be taken to achieve these goals. In other words; Usability is about the user interface but Playability is about the game mechanics. Based on our findings, we suggest using the

20-item measure, namely GUESS-20-TR for assessing the game user experience and reporting the measurements in 10 dimensions including Playability. Please note again that the items in the Usability dimension of GUESS-20-TR are the same as the so-called Playability/usability dimension of GUESS-18 (Keebler et al., 2020). However, our additional two items indicate another dimension, Playability; which is conceptually different and can be statistically discriminated from Usability.

Although the reliability of the dimensions seems to be decreasing when the number of items per dimension is limited to two, indicators obtained through the Spearman-Brown Prophecy Formula suggest an acceptable level of reliability for the dimensions based on Turkish items.

Pearson correlations obtained between the short and long versions of GUESS-TR reveal that similar mean scores can be achieved with either of the scales. However, further studies are required in order to compare their sensitivity, i.e. the scale's capability of producing significantly different scores for games that provide different qualities in terms of user experience. The GUESS dimensions are reported to be sensitive to the differences between games (Aksayim & Berkman, 2020), yet there is no evidence on the sensitivity of the GUESS-18, the GUESS-TR, and the GUESS-20-TR. We suggest researchers use the subscales of the 51-item version GUESS-TR to obtain a higher sensitivity when their study does not require the assessment of all dimensions.

This study shows that the 10-factor model of Berkman et al. (2022a) measured via 51 items is applicable in a shorter form of GUESS. However, the model is only tested on the data collected with Turkish-translated items. Future studies employing GUESS in English should consider assessing the 10-factor model. Furthermore, future adaptation studies into other languages should also consider the 10-factor model. However, it should be noted that the factor model is assessed using a data set that is collected through a study in which the participants played the games they were asked to play. They are different from 9-factor model studies (Phan et al., 2016, Keebler et al, 2020) where participants were evaluating a game of their choice that they have recently played. Since most of the participants are likely to evaluate a game that they love to play, the indicators may reveal a lesser amount of variance that affects the granularity of the data.

Furthermore, both shorter form and full-length GUESS versions should be assessed against other criteria, such as PXI (Abelee et al., 2020), measures obtained regarding the player performance such as time on task, tasks completed and the score achieved in the game, as well as the measures obtained through biometric methods used in game studies (Akan & Berkman, 2020).

Peer Review: Externally peer-reviewed.

Author Contributions: Conception/Design of Study- İ.B., Ç.A.; Data Acquisition- İ.B.; Data Analysis/Interpretation- İ.B., Ç.A.; Drafting Manuscript- Ç.A.; Critical Revision of Manuscript- Ç.A., İ.B.; Final Approval and Accountability- Ç.A., İ.B.; Material and Technical Support- Ç.A., İ.B.; Supervision- İ.B., Ç.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.

ORCID IDs of the authors / Yazarların ORCID ID'leri

Mehmet İlker Berkman 0000-0002-2340-9373
Çakır Aker 0000-0002-0945-9251

REFERENCES

- Abelee, V. V., Spiel, K., Nacke, L., Johnson, D., & Gerling, K. (2020). Development and validation of the player experience inventory: A scale to measure player experiences at the level of functional and psychosocial consequences. *International Journal of Human-Computer Studies*, 135, 102370. <https://doi.org/10.1016/j.ijhcs.2019.102370>
- Akan, E., Berkman, M.İ. (2020). Physiological Measures in Game User Research. In: Bostan, B. (eds) *Game User Experience And Player-Centered Design. International Series on Computer Entertainment and Media Technology*. Springer, Cham. https://doi.org/10.1007/978-3-030-37643-7_10
- Aker, Ç., Rızvanoğlu, K., & Bostan, B. (2017). Methodological review of playability heuristics. *Proc. Eurasia Graph. Istanbul. Turk*, 405.
- Aksayim, A., & Berkman, M. İ. (2020). Effect of physical activity on VR experience: an experimental study. In Richir, S. (Ed.) *Laval Virtual ConVRgence (VRIC) Virtual Reality International Conference-VRIC*, pp. 19-27. 22-24 April 2020, Laval, France.
- Ali, A. S., & Arumugam, A. & Kumaran, S. (2021). Effectiveness of an intensive, functional, gamified Rehabilitation program in improving upper limb motor function in people with stroke: A protocol of the EnterTain randomized clinical trial. *Contemporary Clinical Trials*, 105, 106381. <https://doi.org/10.1016/j.cct.2021.106381>

- Alkan S., Çağıltay K. (2007) Studying computer game learning experience through eye tracking. *British Journal of Educational Technology*, 38(3), pp 538–542. <https://doi.org/10.1111/j.1467-8535.2007.00721.x>
- Berkman, M. İ., Bostan, B., & Şenyer, S. (2022a). Turkish Adaptation Study of the Game User Experience Satisfaction Scale: GUESS-TR. *International Journal of Human–Computer Interaction*, 38(11), 1081-1093. <https://doi.org/10.1080/10447318.2021.1987679>
- Berkman, M. İ., Bostan, B., & Şenyer, S. (2022b), “Game User Experience Satisfaction Scale Applied in Turkish”, *Mendeley Data*, V2, <https://doi.org/10.17632/pxnzgdz287.2>
- Boberg, M., Karapanos, E., Holopainen, J., & Lucero, A. (2015, October). PLEXQ: Towards a playful experiences questionnaire. In Proceedings of the CHI PLAY '15: Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play, October 2015, pp. 381–391 <https://doi.org/10.1145/2793107.2793124>
- Bostan, B. (2022) *Dijital Oyunlar ve İnteraktif Anlatı*. Istanbul, Turkey: The Kitap, 2022.
- Brooke, J. (1996). SUS: A quick and dirty usability scale. In P.W. Jordan, B. Thomas, B. A. Weerdmeester & I. L. McClelland (Eds.), *Usability Evaluation in Industry* (pp. 189-194). London: Taylor & Francis
- Desurvire, H., & Wixon, D. (2018). Heuristics uncovered for games user researchers and game designers. In Anders Drachen, Pejman Mirza-Babaei, Lennart E. Nacke (Eds) *Games User Research*, 217-256. <https://doi.org/10.1093/oso/9780198794844.003.0014>
- Desurvire, H., Caplan, M., & Toth, J. A. (2004, April). Using heuristics to evaluate the playability of games. In *CHI'04 extended abstracts on Human factors in computing systems* (pp. 1509-1512).
- Drachen, A., & Connor, S. (2018). Game analytics for games user research. In Anders Drachen, Pejman Mirza-Babaei, Lennart E. Nacke (Eds) *Games User Research* (pp. 333-353). Oxford University Press. <https://doi.org/10.1093/oso/9780198794844.003.0019>
- Eisinga, R., Grotenhuis, M. T., & Pelzer, B. (2013). The reliability of a two-item scale: Pearson, Cronbach, or Spearman-Brown? *International Journal of Public Health*, 58(4), 637-642. <https://doi.org/10.1007/s00038-012-0416-3>
- Ermi, L., & Mäyrä, F. (2005). Fundamental Components of the Gameplay Experience: Analysing Immersion. *DiGRA conference Changing Views: Worlds in Play*, Vancouver, Canada.
- Federoff, M. A. (2002). *Heuristics and usability guidelines for the creation and evaluation of fun in video games* (Doctoral dissertation, Indiana University).
- Gorsic, M., Clapp, J. D., Darzi, A., & Novak, D. (2019). A brief measure of interpersonal interaction for 2-player serious games: questionnaire validation. *JMIR serious games*, 7(3), e12788. <https://doi.org/10.2196/12788>
- Hafeez, T., Umar Saeed, S. M., Arsalan, A., Anwar, S. M., Ashraf, M. U., & Alsubhi, K. (2021). EEG in game user analysis: A framework for expertise classification during gameplay. *Plos one*, 16(6), e0246913. <https://doi.org/10.1371/journal.pone.0246913>
- Henseler, J., Ringle, C.M. & Sarstedt, M. (2015) A new criterion for assessing discriminant validity in variance-based structural equation modeling. *J. of the Acad. Mark. Sci.* 43, 115–135. <https://doi.org/10.1007/s11747-014-0403-8>
- Hooper, D., Coughlan, J., & Mullen, M. R. (2008). Structural equation modelling: Guidelines for determining model fit. *Electronic journal of business research methods*, 6(1), pp53-60.
- IJsselsteijn, W. A., De Kort, Y. A. W., & Poels, K. (2007). The Game Experience Questionnaire: Development of a self-report measure to assess the psychological impact of digital games. *Manuscript in Preparation* (pp. 9241-210). FUGA technical report Deliverable 3.3.
- Isman, F. A., Prasasti, A. L., & Nugrahaeni, R. A. (2021, April). Expression Classification For User Experience Testing Using Convolutional Neural Network. In *2021 International Conference on Artificial Intelligence and Mechatronics Systems (AIMS)* (pp. 1-6). IEEE. <https://doi.org/10.1109/AIMS52415.2021.9466088>
- Jennett, C., Cox, A. L., Cairns, P., Dhoparee, S., Epps, A., Tijs, T., & Walton, A. (2008). Measuring and defining the experience of immersion in games. *International Journal of Human-Computer Studies*, 66(9), 641–661. <https://doi.org/10.1016/j.ijhcs.2008.04.004>
- Johnson, D., Gardner, M. J., & Perry, R. (2018). Validation of two game experience scales: the player experience of need satisfaction (PENS) and game experience questionnaire (GEQ). *International Journal of Human-Computer Studies*, 118, 38-46. <https://doi.org/10.1145/1152215.1152218>
- Ju, U., & Wallraven, C. (2019). Manipulating and decoding subjective gaming experience during active gameplay: a multivariate, whole-brain analysis. *NeuroImage*, 188, 1-13. <https://doi.org/10.1016/j.neuroimage.2018.11.061>
- Keebler, J. R., Shelstad, W. J., Smith, D. C., Chaparro, B. S., & Phan, M. H. (2020). Validation of the GUESS-18: a short version of the Game User Experience Satisfaction Scale (GUESS). *Journal of Usability Studies*, 16(1), 49. <https://uxpajournal.org/validation-game-user-experience-satisfaction-scale-guess/>
- Kivikangas, J. M., Chanel, G., Cowley, B., Ekman, I., Salminen, M., Järvelä, S., & Ravaja, N. (2011). A review of the use of psychophysiological methods in game research. *Journal of Gaming & Virtual Worlds*, 3(3), 181–199. https://doi.org/10.1386/jgvw.3.3.181_1
- Korhonen, H., & Koivisto, E. M. (2006, September). Playability heuristics for mobile games. In *Proceedings of the 8th conference on Human-computer interaction with mobile devices and services* (pp. 9-16). <https://doi.org/10.1145/1152215.1152218>
- Korhonen, H., & Koivisto, E. M. (2007, September). Playability heuristics for mobile multi-player games. In *Proceedings of the 2nd international conference on Digital interactive media in entertainment and arts* (pp. 28-35).
- de Kort, Y. A., IJsselsteijn, W. A., & Poels, K. (2007). Digital games as social presence technology: Development of the Social Presence in Gaming Questionnaire (SPGQ). *Proceedings of PRESENCE*, 195203, 1-9.
- Law, E. L. C., Brühlmann, F., & Mekler, E. D. (2018, October). Systematic review and validation of the game experience questionnaire (GEQ)-implications for citation and reporting practice. In *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play* (pp. 257-270).
- Malone, T. W. (1982, March). Heuristics for designing enjoyable user interfaces: Lessons from computer games. In *Proceedings of the 1982 conference on Human factors in computing systems* (pp. 63-68). <https://doi.org/10.1145/800049.801756>

- Medlock, M.C. (2018) "An overview of GUR methods." In Anders Drachen, Pejman Mirza-Babaei, Lennart E. Nacke (Eds). *Games User Research*. Oxford University Press. <https://doi.org/10.1093/oso/9780198794844.003.0007>
- Mittmann, G., Zehetner, V., Krammer, I., & Schrank, B. (2023). Translation, adaptation and validation of the German version of the Game User Experience Satisfaction Scale (GUESS-GA-18) for adolescents. *Behaviour & Information Technology*, 1-15. <https://doi.org/10.1080/0144929X.2023.2212807>
- Mojzisch, A., Schilbach, L., Helmert, J.R., Pannasch, S., Velichkovsky, B.M., Vogeley K. (2006) The effects of self-involvement on attention, arousal, and facial expression during social interaction with virtual others: A psychophysiological study. *Social Neuroscience*, 1(3-4), pp 184–195. <https://doi.org/10.1080/17470910600985621>
- Mueller S.C., Jackson C.P., Skelton R.W. (2008) Sex differences in a virtual water maze: An eye tracking and pupillometry study. *Behavioural Brain Research*, 193(2), pp 209–215. <https://doi.org/10.1016/j.bbr.2008.05.017>
- Schaffer, N. (2008). Heuristic UX Evaluation of Games. In Isbister K., Hodent, C. (Eds) *Game Usability* (pp. 105-114). CRC Press.
- Nguyen, N. C., Thawonmas, R., Paliyawan, P., & Pham, H. V. (2020, August). JUSTIN: An audience participation game with a purpose for collecting descriptions for artwork images. In *2020 IEEE Conference on Games (CoG)* (pp. 344-350). IEEE. <https://doi.org/10.1109/CoG47356.2020.9231771>
- Phan, M. H., Keebler, J. R., & Chaparro, B. S. (2016). The development and validation of the Game User Experience Satisfaction Scale (GUESS). *Human Factors*, 58(8), 1217–1247. <https://doi.org/10.1177/0018720816669646>
- Qin, H., Patrick Rau, P. L., & Salvendy, G. (2009). Measuring player immersion in the computer game narrative. *Intl. Journal of Human-Computer Interaction*, 25(2), 107-133. <https://doi.org/10.1080/10447310802546732>
- Ryan, R. M., Rigby, C. S., & Przybylski, A. (2006). The motivational pull of video games: A self-determination theory approach. *Motivation and Emotion*, 30(4), 344–360. <https://doi.org/10.1007/s11031-006-9051-8>
- Rindskopf, D. (1984). Structural equation models: Empirical identification, Heywood cases, and related problems. *Sociological Methods & Research*, 13(1), 109-119. <https://doi.org/10.1177/0049124184013001004>
- Shelstad, W. J., Smith, D. C., & Chaparro, B. S. (2017). Gaming on the rift: How virtual reality affects game user satisfaction. *Proceedings of the Human Factors and Ergonomics Society*, 2017-October, 2072–2076. Texas, Austin: HFES.
- Thevin, L., Rodier, N., Oriola, B., Hachet, M., Jouffrais, C., & Brock, A. M. (2021). Inclusive adaptation of existing board games for gamers with and without visual impairments using a spatial augmented reality framework for touch detection and audio feedback. *Proceedings of the ACM on Human-Computer Interaction*, 5(ISS), 1-33. <https://doi.org/10.1145/3488550>
- Wijaya, I. G. N. T., Hansun, S., & Kristanda, M. B. (2019). DISDAIN: An auto content generation VR game. *Indian Journal of Science and Technology*, 12(7), 1–7. <https://doi.org/10.17485/ijst/2019/v12i7/141370>
- Xu, J., Paliyawan, P., Thawonmas, R., & Harada, T. (2019). Player dominance adjustment: Promoting self-efficacy and experience of game players by adjusting dominant power. *2019 IEEE 8th Global Conference on Consumer Electronics, GCCE 2019*, 487–488. Osaka, Japan: IEEE. <https://doi.org/10.1109/GCCE46687.2019.9015408>
- Yildirim, C., Carroll, M., Hufnal, D., Johnson, T., & Pericles, S. (2018, August). Video game user experience: to VR, or not to VR? In *2018 IEEE Games, Entertainment, Media Conference (GEM)* (pp. 1-9). IEEE. <https://doi.org/10.1109/GEM.2018.8516542>

How cite this article

Berkman, M. İ., Aker, C. (2023). A shorter form of the game user experience satisfaction scale in Turkish: GUESS-20-TR. *Acta Infologica*, 7(2), 229-242. <https://doi.org/10.26650/acin.1185840>

Appendix I - GUESS-20-TR items

Audio Aesthetics / Ses Estetiği

A01 Bu oyundaki ses efektlerinden keyif alıyorum.

A03 Bu oyunun seslerinin (örneğin ses efektleri, müzik) oyun deneyimimi arttırdığını hissediyorum.

Creative Freedom / Yaratıcılı Özgürlük

CF01 Bu oyunun hayal gücümü kullanmama olanak sağladığını düşünüyorum.

CF02 Bu oyunu oynarken kendimi yaratıcı hissediyorum.

Enjoyment / Eğlence

E01 Bu oyunun eğlenceli olduğunu düşünüyorum.

E03 Bu oyunu oynarken sıkılıyorum. [R]

Narratives / Anlatı

N02 Bu oyunun hikayesini başından itibaren çekici buluyorum.

N03 Bu oyunda sunulan fanteziden veya hikayedenden keyif alıyorum.

Player Engrossment / Oyun Meşguliyeti

PE01 Bu oyunu oynarken dış dünyadan kopmuş hissediyorum.

PE02 Bu oyun sırasında gerçek dünyada olan biteni umursamıyorum.

Personal Gratification / Kişisel Tatmin

PG03 Bu oyunu elimden geldiğince iyi oynamak istiyorum.

PG04 Bu oyunu oynarken kendi performansına çok odaklanırım.

Social Connectivity / Sosyal Bağlanırlık

SC03 Canım isterse bu oyunu diğer oyuncularla oynayabilirim.

SC02 Bu oyunu diğer oyuncularla birlikte oynamak hoşuma gidiyor.

Visual Aesthetics / Görsel Estetik

VA01 Bu oyunun grafiklerinden keyif alıyorum.

VA03 Bu oyunun görsel olarak çekici olduğunu düşünüyorum.

Usability / Kullanılabilirlik

US02 Bu oyunun kontrollerini açık anlaşılır buluyorum.

US04 Bu oyunun arayüzünü gezinmesini kolay buluyorum.

Playability / Oynanabilirlik

PL03 Bu oyunda hedeflerime/amaçlarıma nasıl ulaşacağımı her zaman bilirim.

PL11 Bu oyunu oynarken kendime çok güveniyorum.

Participants respond to the items through a Likert scale from “1 – Hiç katılmıyorum” (Strongly Disagree) to “7 – Tamamen katılıyorum” (Strongly Agree).

[R] marked item E03 should be reverse coded for scoring.

Item numbers/names are given according to the 51-item GUESS-TR (Berkman et al. 2022a; 2022b).

See Berkman et al. (2022a; 2022b) for the 51-item version, which can be used separately for each dimension regarding the requirement of the researchers.

Determining the Happiness Class of Countries with Tree-Based Algorithms in Machine Learning

Makine Öğrenmesinde Ağaç Tabanlı Algoritmalarla Ülkelerin Mutluluk Sınıfının Belirlenmesi

Merve Doğruel¹ , Selin Soner Kara² 

¹(Assist. Prof.), University of Istanbul Esenyurt,
Faculty of Business and Management Sciences,
Department of Management Information Systems,
Istanbul, Türkiye

²(Prof. Dr.), Yıldız Technical University, Faculty of
Mechanical Engineering, Department of Industrial
Engineering, Istanbul, Türkiye

Corresponding author : Merve DOĞRUDEL

E-mail : mervedogruel@esenyurt.edu.tr

ABSTRACT

Today, the concept of happiness is a frequently researched subject in the fields of economy, medicine, and social and political fields, as well as psychology. It has been an important research area for everyone, from policymakers to companies, to determine the factors affecting happiness. With machine learning algorithms, it is possible to make classifications with very high accuracy. The aim of this study is to use tree-based machine learning algorithms to classify the happiness scores of countries. In order to accomplish this, data from the World Happiness Index published in 2022 were used. On these data, tree-based algorithms CART, tree-based ensemble algorithms Bagging, and Random Forest were used. The test data of the model were obtained with 85% precision, recall, and F1 metrics, which were calculated using Bagging and Random Forest algorithms. The outcomes of the models obtained during the study were interpreted.

Keywords: Machine learning, World Happiness Index, Ensemble learning

ÖZ

Mutluluk kavramı günümüzde psikoloji alanı dışında ekonomi, tıp, sosyal ve politik alanlarda da sıklıkla araştırılan bir konu haline gelmiştir. Mutluluğu etkileyen faktörlerin belirlenmesi, politika yapıcılardan işletmelere kadar önemli bir araştırma alanı olmuştur. Makine öğrenmesi algoritmaları ile yüksek doğrulukta sınıflandırmalar çalışmaları yapmak mümkündür. Bu çalışmada, ağaç tabanlı makine öğrenmesi algoritmaları kullanılarak ülkelerin mutluluk puanlarının sınıflandırılması amaçlanmaktadır. Bu amaçla 2022 yılında yayınlanan Dünya Mutluluk Endeksi'nden alınan veriler kullanılmıştır. Bu veriler üzerinde ağaç tabanlı algoritmalar SRT, ağaç tabanlı topluluk algoritmaları torbalama ve rastgele orman kullanılmıştır. Torbalama ve rastgele orman algoritmaları ile elde edilen modelin test verilerinde %85 kesinlik, duyarlılık ve F1 metrikleri hesaplanmıştır. Çalışmada elde edilen bu modellerin sonuçları yorumlanmıştır.

Anahtar Kelimeler: Makine öğrenmesi, Dünya Mutluluk Endeksi, Topluluk öğrenmesi

Submitted : 15.02.2023

Revision Requested : 26.07.2023

Last Revision Received : 01.08.2023

Accepted : 01.08.2023

Published Online : 24.08.2023



This article is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0)

1. INTRODUCTION

Happiness was previously only discussed in the field of psychology, since it is a subjective concept that cannot be defined with an absolute definition. Over time, the concept of happiness, which is affected by many factors, has become an important research topic for economics under the heading economics of happiness (Öztürk & Suluk, 2020), which is separate from the fields of medicine, sociology, and politics.

International organizations have carried out various index studies on happiness, which is important in many fields. Examples of these are the World Happiness Report, Happy Planet Index, OECD Better Life Index, and Gallup Global Emotions.

The World Happiness Report (Helliwell et al., 2022), which was first published in 2013, has been published every year and its 10th edition was published in 2022. It has a readership of over 9 million and has been cited numerous times.

When the literature studies are examined, it is seen that there are many different results regarding the factors affecting happiness and the level of importance of these factors. It is vital to determine the effects of factors affecting happiness, including the mental states of individuals, the behavior of companies, and the policy determinations of countries. For this purpose, many statistical studies have been carried out in the literature in order to determine happiness.

Farooq and Shanmugam (2022) analyzed performance metrics using validated COVID-19 datasets of countries and happiness reports showing how free citizens are. Various machine learning techniques were used, such as linear regression, logistic regression, support vector machine (SVM), Naive Bayes (NB), and K-nearest neighbor (KNN) algorithms. In their study, Kiroğlu and Yıldırım (2022) examined the determinants of happiness in Turkey using multivariate logit models. In their research, Jannahi, Sael, and Benabbou (2021) aimed to predict quality of life using basic machine learning models using data from the 2015-2021 World Happiness Index reports.

In the research conducted by Ulkhaq and Adyatama (2021), countries were clustered based on distinct clustering algorithms using the World Happiness Report 2019 data. Ibnat, Gyalmo, Alom, Abdul Awal, and Azim (2021) utilized the 2019 World Happiness Report data to identify the happiest countries and regions through supervised machine learning techniques and to assess the life satisfaction of the country. Chaudhary, Dixit, and Sahni (2020) used the World Happiness Index data from 2016-18 to build models with Predictive Modeling and Bayesian Networks methods. They then evaluated them for 2019 data.

In the study conducted by Garces, Adriatico, and Timbal (2019), the data obtained from the World Happiness Index 2014 was utilized in the cluster analysis based on the quality of life of the countries, in conjunction with the various indicators determined. Carlsen's study (2018) used the partial ranking methodology to calculate the happiness index. The author proposed a distinct ranking from the country rankings as reported in the 2016 report, arguing that the ranking obtained through this calculation method is more nuanced. Dao's study (2017) examined the direct impact of government spending on happiness in 183 countries between 1990 and 2016 using pooled OLS, fixed effects, random effects models, and cross-sectional analysis.

This study aims to establish a model with the data from the 2022 World Happiness Index report using tree-based machine learning techniques and to determine the most important indicators in the appropriate model. Machine learning applications acquire knowledge through experience, similar to humans, without direct programming. The algorithm, which learns from training data and experience, can then detect the data it encounters and perform estimation and classification with a high accuracy rate. Within the scope of the purpose of the study, the main reason for using decision trees, one of the supervised machine learning methods, is both the ease of explanation and interpretation of the results obtained and the ability of decision trees to make variable selections. Additionally, ensemble learning algorithms, which enable the creation of more than one tree, in other words, forests, by adding the concept of randomness to the decision tree algorithms, were applied within the scope of the study. Decision trees were created using the CART algorithm, and decision forests were created using the Bagging and Random Forest algorithms. Decision forests are also known as ensemble learning algorithms.

When the test data metrics showing the classification success of the decision trees obtained were examined, it was determined that the precision, recall, and F1 values were equal in the ensemble learning algorithms and higher than the CART algorithm. As in many studies in the literature, success metrics obtained with ensemble learning were higher in this study. Based on results obtained with Bagging and Random Forest ensemble learning algorithms, according to the World Happiness Index 2022 report, GDP, social support, and health life expectancy are the most important factors in determining the happiness classes of countries. In this study, world happiness classes are investigated using tree-based machine learning algorithms. It is important that policymakers who want to increase the happiness level of countries evaluate this study first.

2. TREE-BASED ALGORITHMS IN MACHINE LEARNING

The most commonly preferred decision tree in data science is a predictive model that can be expressed as the recursive division of the covariate space into subspaces. Decision trees, which were previously subject to decision theory and statistics, have been developed with applications in other fields such as time and data mining, machine learning, and pattern recognition. A sub-branch of artificial intelligence, machine learning develops a model that allows making predictions for new data by learning from training data thanks to computer software (Okumuş, Ekmekçioğlu & Kara, 2021). Decision trees are algorithms based on supervised learning in machine learning. As in other learning algorithms, the decision tree learning algorithm chosen aims to generate the most appropriate model from the training data. Afterward, the validity of the model created with the test data is tested, and if the model is approved, it is used to make predictions (Doğruel & Fırat, 2021). Decision trees are algorithms that are easy to understand and have a high success rate because they imitate human thinking ability while making decisions (Efeoğlu, 2022).

Ensemble learning is the realization of a specific learning task by creating and combining multiple models in order to arrive at a better decision than the decisions made separately. Many studies in the literature have shown that ensemble learning increases the predictive power of a single model. Ensemble learning produces more effective results, especially in decision tree models. In some sources, it is seen that the trees obtained by combining ensemble learning and decision tree learning are also called decision forests. Random Forest algorithm is the most popular decision forest (Rokach, 2016).

In this study, the CART algorithm from decision trees, Bagging and Random Forest algorithms from decision trees, and ensemble learning algorithms are employed in conjunction (decision forests).

2.1. CART

The CART (Classification and Regression Tree) algorithm is a non-parametric and non-linear decision tree algorithm that makes predictions based on repeated binary separation, used to create both classification and regression trees. If the target variable is categorical, the tree is referred to as a classification tree (CT), whereas if the target variable is continuous, the tree is referred to as a regression tree (RA) (Doğruel & Fırat, 2021).

Using the CART algorithm, a classification tree can be grown by dividing the dataset into two sub-partitions (lower leaves) on a rule-based basis by binary recursion. Each split involves a single variable; some variables can be used multiple times, while others are not utilized at all. Each sub-leaf is then further divided according to independent rules. The rule-based approach used in CART relies on a binary iterative partitioning path that divides a subset of the dataset, known as leaves, into two subsets known as sub-leaves based on minimizing a computed heterogeneity criterion. The Gini index and cross-entropy are the two most preferred heterogeneity criteria (Bel, Allard, Laurent, Cheddadi, & Bar-Hen, 2009 and James, Witten, Hastie, Tibshirani, 2013).

The Gini index is a measure of the total variance among K classes:

$$G = \sum_{k=1}^k \hat{P}_{mk}(1 - \hat{P}_{mk}) \quad (1)$$

In the formula, \hat{P}_{mk} is the proportion of training observations in the mth region that are from the kth class.

The Gini index takes a small value if all P_{mk} 's are close to zero or one. Therefore, the Gini index is expressed as a measure of node purity, and a small Gini index indicates that a node generally contains observations from a single class. Cross-entropy is an alternative to the Gini index:

$$D = - \sum_{k=1}^K \hat{P}_{mk} \log \hat{P}_{mk} \quad (2)$$

Since $0 \leq \hat{P}_{mk} \leq 1$, it follows that $0 \leq \hat{P}_{mk} \log \hat{P}_{mk}$. If the \hat{P}_{mk} 's are all close to or close to zero, the cross-entropy will take a value close to zero. Therefore, the cross-entropy will take a small value like the Gini index, if the mth the node is pure.

Rules that include antecedents describing all nodes of the tree, from the root to the leaves, can be too complex. It may turn out that the initial decisions, especially in large trees, are not crucial for the classification of data vectors terminating in a leaf. Therefore, unnecessary rule predecessors should be removed (Grabczewski & Duch, 1999). Commonly employed pruning techniques, such as degree-based pruning of the separability of a split value (SSV) and

CART's cross-complexity pruning, rely on cross-validation. Some pre-cleaning algorithms, including cross-validation-based strategies and reduced error pruning (REP), have also been proposed as an alternative to post-pruning methods, but they are not as common as post-pruning because the results are not as good (Grabczewski, 2011).

2.2. Bagging

Working with high variance in standard decision trees can be a challenge. In this case, it is possible to obtain different results when working with different training data. In contrast, similar results will be obtained if a low variance procedure is repeatedly applied to different datasets. The process of bootstrap picking, commonly referred to as Bagging, is a general-purpose technique employed to reduce the variance of Bagging in a statistical learning method, as stated by James, Witten, Hastie, and Tibshirani (2013).

The Bagging algorithm is an ensemble learning method for creating a classifier ensemble. It involves combining basic learning algorithms trained on different samples of the training set. The basic point of the Bagging algorithm is based on the principle of providing diversity by training each basic learning algorithm that makes up the community on different training sets. Here, a simple random substitution sampling method is generally applied to generate different training sets from the data set. The results of the training sets obtained by the sampling method and the classification methods trained are merged through a majority vote (Onan, 2018)

The working principle of the Bagging algorithm is shown in Figure 1. In this principle, which is characterized by random sampling, there is no relationship between weak learning models. Randomly selected samples are subsequently returned to the data set subsequent to each extraction. This means that the previous sample can be collected continuously in the next sample (Li, 2022).

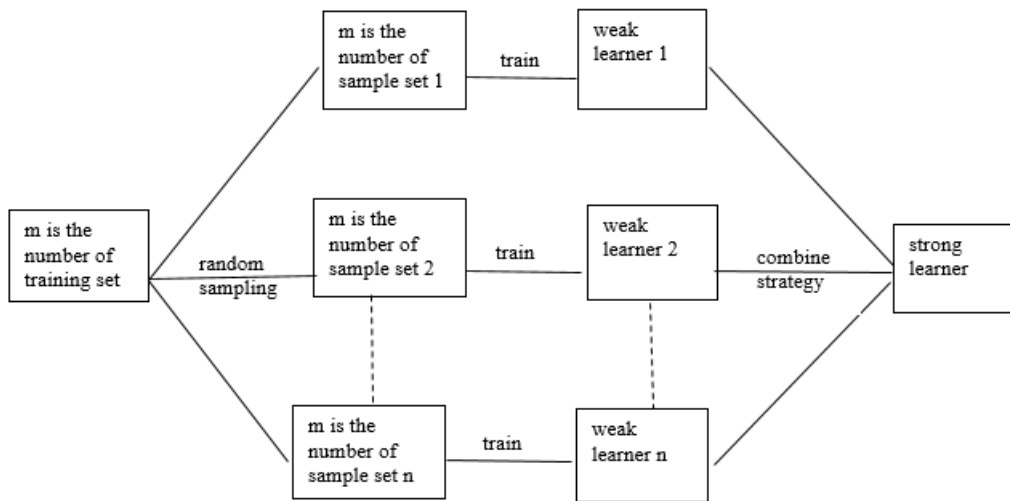


Figure 1. Bagging algorithm working principle (Li, 2022)

In this approach, B different bootstrapped training datasets are generated, then the method is trained on the b th bootstrapped training set to obtain $\hat{f}^b(x)$, and finally all predictions are averaged (James, Witten, Hastie, Tibshirani, 2013). This process is called Bagging:

$$\hat{f}_{bag}(x) = \frac{1}{B} \sum_{b=1}^B \hat{f}^{*b}(x) \quad (3)$$

2.3. Random Forest

The Random Forest algorithm developed by Leo Breiman in 2001 can be described as an evolutionary version of the Bagging algorithm (Li, 2022).

Using the Random Forest method, Bagging is used along with randomly selected features. Each new training set is drawn back from the original training set using the bootstrap method. Using random feature selection, a tree is then grown on the new training set. Trees that are grown are not pruned (Breiman, 2001).

There is a major difference between the Bagging and Random Forest methods in the selection of m-dimensional variables. During the construction of decision trees, a random sample of m estimators is selected as split candidates from all p estimators, considering each split in a tree (James, Witten, Hastie, Tibshirani, 2013). This random selection of predictors results in a reduction in correlation between trees in the forest, as well as a reduction in variance, which results in a higher accuracy in predictions (Suchetana, Rajagopalan & Silverstein, J., 2017). Also, Random Forests provide some measures of the importance of variables for the prediction of the outcome variable (Gregorutti, Michel, & Saint-Pierre, 2017). The Random Forest algorithm is used only for variable selection in many publications due to this feature.

In the original paper, Breiman (Breiman, 2001) proposed the size of the candidate feature set at each node as $m \approx \log_2(p+1)$. Later, many studies on Random Forest used the default size of the candidate feature set as $m \approx \sqrt{p}$ in classification problems and $m \approx p/3$ in regression problems.

Among the difficulties of the algorithm is that the image of the model obtained with the Random Forest algorithm cannot be obtained.

3. APPLICATION AND RESULTS

The happiness score published for 146 countries in the 2022 World Happiness Report and six variables used to explain this score were used in this study. The variables values of GDP levels, life expectancy, generosity, social support, freedom, and corruption in the dataset are not raw data in this study (World Happiness Report, 2022a). The values show the estimated extent to which each variable will contribute to making life assessments higher in each country than in Dystopia (an imaginary country with the most unhappy people in the world) (World Happiness Report, 2022b).

The goal of the study was to figure out how happy or unhappy countries are based on certain rules. We did this by looking at data from the 2022 Happiness Index Report. In order to determine the happiness class of the countries, the 37th country, corresponding to the 1st quartile of 146 countries, was accepted as the boundary. For this reason, countries with a happiness score of 6.3 or above were considered "happy," while other countries were considered "unhappy."

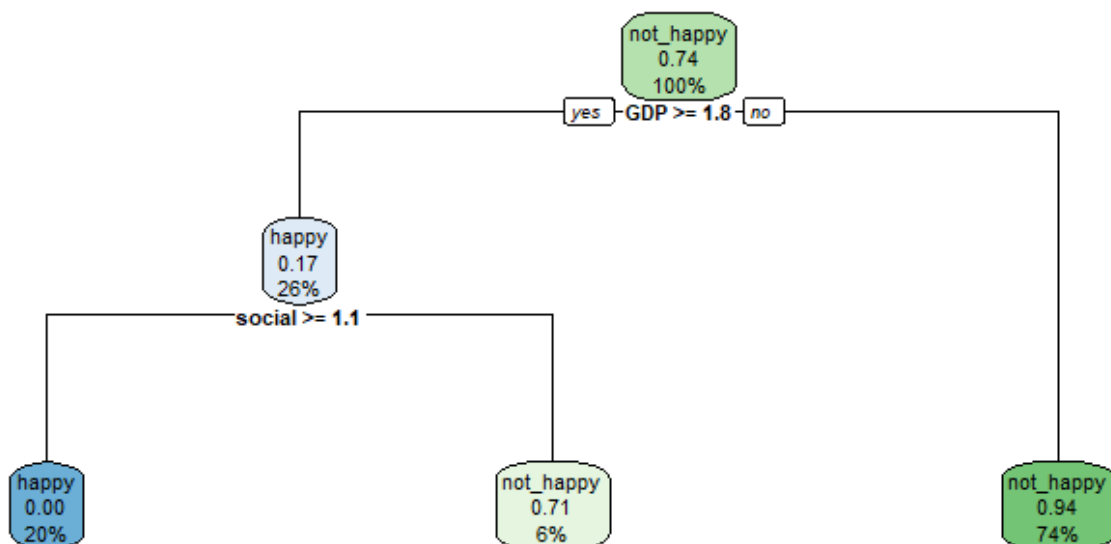


Figure 2. Tree with CART method

As the model must be initially trained and then tested in machine learning, 80% of the data set is randomly allocated

as training data and 20% as test data. In the field of machine learning, tree-based CART, Bagging, and Random Forest methods were obtained through the R program, and the outcomes of the methods were analyzed based on the accuracy parameters.

3.1. Model with CART method

In order to construct a classification tree using the CART algorithm, a model is constructed using default values such as minsplit=20 and complexity parameter=0.01.

The results are shown in Figure 2.

The resulting tree shows that the precision value obtained from the training data is 1.00, the recall value is 0.7667, and the F1 value is 0.8679, while the precision value obtained from the test data is 0.8333, the recall value is 0.7143, and the F1 value is 0.7692.

- A country is "happy" if its GDP per capita ≥ 1.8 and social support ≥ 1 .
- A country is "not happy" if its GDP per capita ≥ 1.8 and social support < 1.1
- A country is "not happy" if its GDP per capita < 1.8

3.2. Model with Bagging method

500 tree experiments were conducted using the Bagging method. The "mean decrease Gini" values indicate the accuracy of the model when the relevant variable is removed from the model. They are 28.5179 for GDP per capita, 7.4294 for social support, 2.7132 for freedom, 2.3950 for cheating, 1.7719 for corruption, and 1.3569 for generosity. The measurement of "mean decrease accuracy" is a measure of the contribution of the aforementioned variable to the homogeneity of the nodes and their leaves in the resulting Random Forest. In the study, it was found that GDP per capita was 0.1288, social support was 0.0858, freedom was 0.0142, corruption was 0.0008, and generosity was 0.0234. The visuals of these "mean decrease accuracy" and "mean decrease Gini" measures showing the importance of the variables in the model are presented in Figure 3.

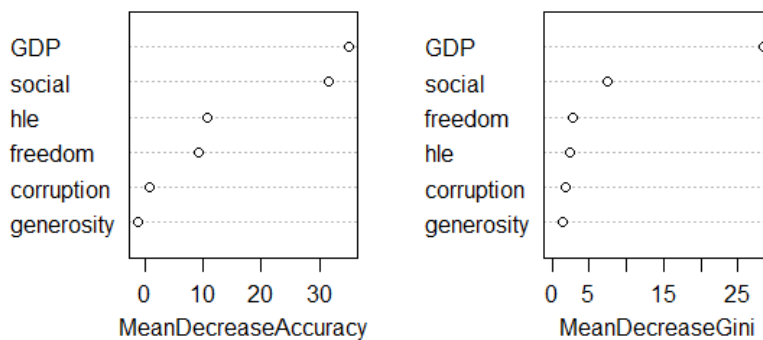


Figure 3. Variable importance plot

According to Figure 3 and the metrics obtained, the most important variables in the model can be stated as GDP per capita and social support, respectively.

The precision value, recall value, and F1 value were obtained in the model obtained with the training data as 1.00. The precision value, recall value, and F1 value of the model as determined by the test data are 0.8571.

3.3. Model with Random Forest method

To determine the most suitable model with this method, trials with a number of randomly sampled variables (mtry) between 2 and 6 as candidates in each compartment were made with the tune process. In the experiments, the number of random variables giving the highest accuracy value was suggested to be 3, as shown in Figure 4.

The "mean decrease Gini" values of the model generated by utilizing the training data and the proposed random three variables are 18.6584 for GDP per capita, 8.7225 for social support, 7.5078 for health, 3.8291 for freedom, 3.4187 for corruption, and 1.8424 for generosity. The "mean decrease accuracy" values of the model generated from the training

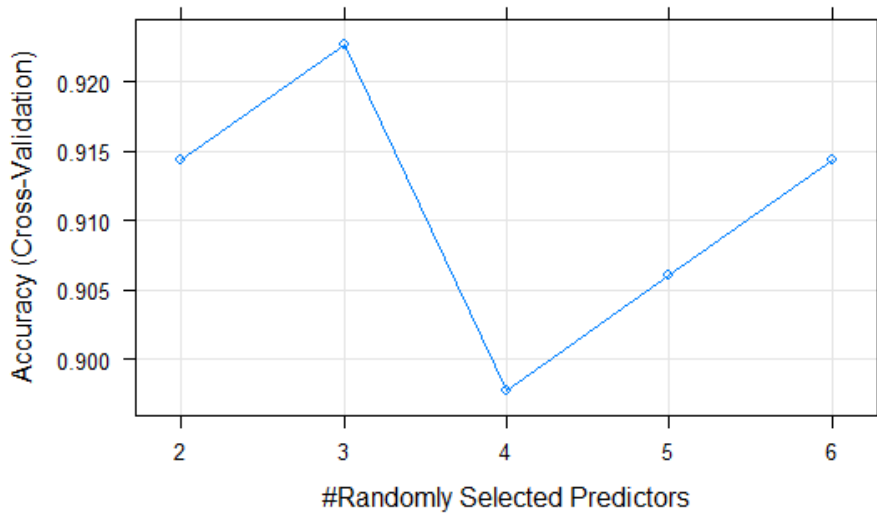


Figure 4. Tune model plot

data with the proposed random three variables are 0.1183 for GDP per capita, 0.0683 for social support, 0.0282 for hle, 0.0181 for freedom, 0.0025 for corruption, and -0.0000 for generosity. The visual representations of the "mean decrease accuracy" and "mean decrease accuracy" measures, which demonstrate the significance of the variables in the model, are depicted in Figure 5.

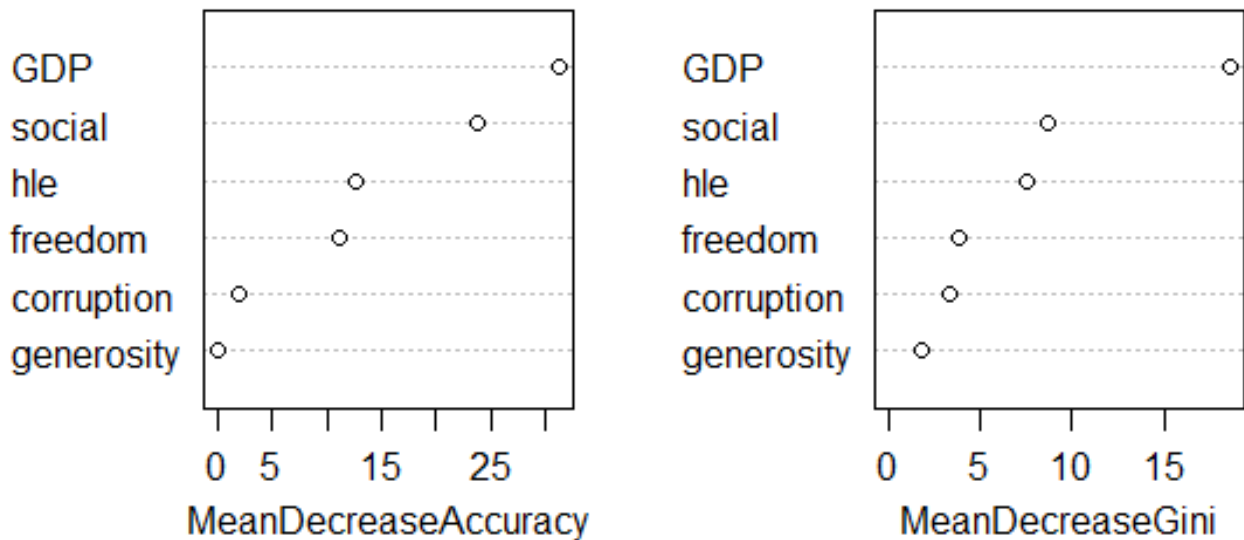


Figure 5. Variable importance plot

According to Figure 5 and the metrics obtained, the most important variables in the model are GDP per capita and social support, respectively. These two variables are the same as the results obtained with the CART and the Bagging method.

In the model constructed using the training data, the precision, recall, and F1 values were all equal to 1.00. The precision value, recall value, and F1 value of the model as determined by the test data are 0.8571.

4. CONCLUSION

The objective of this study was to determine the most significant variables used in the estimation of the World Happiness index of the countries and to determine their significance in the study. Therefore, decision trees, which are commonly used in machine learning, were used. The CART algorithm was employed to generate a single decision tree, while the Bagging and Random Forest algorithms were employed to generate decision trees (decision forests) with ensemble learning.

Upon examination of the model metrics presented in Table 1, it is evident that the test values for precision, recall, and F1 obtained from ensemble learning are superior to those obtained with a single decision tree. This result, as in many previous studies, supports the idea that learning in a community learning gives better performance.

Table 1. Metrics of tree-based models

	CART		Bagging		Random Forest	
	Train	Test	Train	Test	Train	Test
Precision	1.0000	0.8333	1.0000	0.8571	1.0000	0.8571
Recall	0.7667	0.7143	1.0000	0.8571	1.0000	0.8571
F1	0.8679	0.7692	1.0000	0.8571	1.0000	0.8571

Upon examination of Table 1, it is evident that all tree-based machine learning techniques yield satisfactory outcomes when attempting to classify nations based on their happiness scores. The precision, recall, and F1 values of the Bagging model, which has the advantage of preparing a lower variance model, and the Random Forest method, which has the advantage of reducing the risk of overfitting, are equivalent. According to the literature, this result supports the claim that ensemble-based learning machines have higher performance.

The various reasons for the performance increase in ensemble learning can be stated as follows (Erdem, Uslu & Firat, 2021):

- Combination of different models reduces overfitting.
- Better data representation and fitting when working with non-linear datasets.
- Reducing class imbalances.
- Increase in calculation performance.

According to all three methods, the two most important variables that determine the happiness classes of countries are GDP and social support. Upon examination of the outcomes obtained through the Random Forest algorithm (Figure 5), it is evident that the variables of healthy life expectancy (HLE) and freedom to make life choices (freedom) hold significant importance subsequent to GDP.

Similar to the findings of this study, Khder, Sayfi, and Fujo (2022) stated that they identified the most significant variables that impact the happiness score through machine learning techniques, namely GDP per capita and health life expectancy.

Jannani, Sael, and Benabbou (Jannani, Sael, & Benabbou, 2021) utilized the 7-year World Happiness Index Report data to generate predictions using diverse machine learning techniques. In this study, the most effective ensemble learning algorithm for prediction was Random Forest, which achieved an R2 value of 0.85. This result supports the conclusion that the Random Forest algorithm is also suitable for the estimation of the happiness index.

As a result of the application and supportive studies in the literature, it can be stated that the classification power of tree-based ensemble learning algorithms for happiness is high. The Bagging and Random Forest algorithms indicate that GDP, social support, and healthy life expectancy are the most important factors in determining happiness class. It is imperative to take into account the foremost priorities of policymakers who aim to enhance the welfare standard of nations.

We believe that dealing with the concept of happiness, which is an important research topic for many fields, will contribute to the literature. Tree-based applications are preferred due to their high accuracy in machine learning and interpretation possibilities. It would be possible to conduct comparative analyses with different happiness indices in future studies.

Peer Review: Externally peer-reviewed.

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.

ORCID IDs of the authors / Yazarların ORCID ID'leri

Merve Doğruel 0000-0003-2299-7182

Selin Soner Kara 0000-0002-0894-0772

REFERENCES

- Bel, L., Allard, D., Laurent, J. M., Cheddadi, R. & Bar-Hen, A. (2009). CART algorithm for spatial data: Application to environmental and ecological data. *Computational Statistics and Data Analysis*, 53, 3082-3093. <https://doi.org/10.1016/j.csda.2008.09.012>
- Breiman, L. (2001). Random Forests. *Machine Learning*, 45, 5–32. <https://doi.org/10.1023/A:1010933404324>
- Carlsen, L. (2018). Happiness as a sustainability factor. The World Happiness Index: A posetic-based data analysis. *Sustain Sci*, 13, 549-571. <https://doi.org/10.1007/s11625-017-0482-9>
- Chaudhary, M, Dixit, S. & Sahni, N. (2020). Network learning approaches to study world happiness. A Preprint. <https://arxiv.org/pdf/2007.09181.pdf>
- Dao, T. K. (2017). *Government expenditure and happiness: Direct and indirect effects*. Institute of Social Studies, Netherlands. Retrieved from <https://thesis.eur.nl/pub/41656/Dao-Tung-K.-.pdf>
- Doğruel, M & Fırat, S. Ü. (2021). Veri madenciliği karar ağaçları kullanarak ülkelerin inovasyon değerlerinin tahmini ve doğrusal regresyon modeli ile karşılaştırmalı bir uygulama. *Istanbul Business Research*, 50(2), 465-493. <https://www.doi.org/10.26650/ibr.2021.50.015019>
- Efeoğlu, E. (2022). Kablosuz sinyal gücünü kullanarak iç mekan kullanıcı lokalizasyonu için karar ağaçları algoritmalarının karşılaştırılması. *Acta Infologica*. <https://doi.org/10.26650/acin.1076352>
- Erdem, Z. U., Uslu, B. Ç. & Fırat, S. Ü. (2021). Customer churn prediction analysis in a telecommunication company with machine learning algorithms. *Journal of Industrial Engineering*, 32(3), 496-512.
- Farooq, S.A. & Shanmugam, S.K. (2022). A performance analysis of supervised machine learning techniques for COVID-19 and happiness report dataset. In: S. Shakya, V.E. Balas, S. Kamolphiwong, KL. Du (Eds) *Sentimental analysis and deep learning*. Singapore Springer.
- Garces, E. J., Adriatico C. & Timbal, L. R. E. (2019). Analysis on the relationships on the global distribution of the World Happiness Index and selected economic development indicators. *Open Access Library Journal*, 6, 1-16. <https://doi.org/10.4236/oalib.110545>
- Grabczewski, K., & Duch, W. (1999, June). *A general purpose separability criterion for classification systems*. Proceedings of the 4th Conference on Neural Networks and Their Application. Zakopane, Poland, 203-208.
- Grabczewski, K. (2011), Validated decision trees versus collective decisions. In P. Jedrzejowicz & N. T Nguyen (Eds.) *Computational Collective Intelligence Technologies and Applications Third International Conference* (pp.324-351). Verlag Berlin Heidelberg: Springer.
- Gregorutti, B., Michel, B. & Saint-Pierre, P. (2017). Correlation and variable importance in Random Forests. *Stat Comput*, 27, 659–67. <https://doi.org/10.1007/s11222-016-9646-1>
- Helliwell, J. F., Layard, R., Sachs, J. D., De Neve, J.-E., Akinin, L. B., & Wang, S. (Eds.). (2022). *World Happiness Report 2022*. New York: Sustainable Development Solutions Network.
- Ibnat, Gyalmo, Alom, Abdul Awal and Azim (2021, December). Understanding world happiness using machine learning techniques. International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2). Rajshahi, Bangladesh. <https://www.doi.org/10.1109/IC4ME253898.2021.9768407>
- James, G., Witten, D., Hastie, T. & Tibshirani, R. (2013). *An introduction to statistical learning: With applications in R*. New York: Springer.
- Jannani, A. Sael, N. & Benabbou, F. (2021, December). *Predicting quality of life using machine learning: Case of World Happiness Index*. 4th International Symposium on Advanced Electrical and Communication Technologies (ISAECT), Alkhobar, Saudi Arabia. <https://www.doi.org/10.1109/ISAECT53699.2021.9668429>
- Khder, M. A., Sayfi, M. A. & Fujo, S. W. (2022). Analysis of World Happiness Report dataset using machine learning approaches. *International Journal of Advances in Soft Computing and its Applications*, 14(1), 14-34. <https://doi.org/10.15849/IJASCA.220328.02>
- Kıroğlu, B. S. & Yıldırım, K. (2022). Mutluluk ve belirleyicileri: Türkiye için bir analiz. *Journal of Emerging Economies and Policy*, 7(2), 5070.
- Li, A. (2022). Stock forest model based on Random Forest. In X. Huang & F. Zhang (Eds.) *Economic and business management*. The Netherlands: CRP Press.
- Okumuş, F., Ekmekçioğlu, A. & Kara, S. S. (2021). Modelling ships main and auxiliary engine powers with regression-based machine learning algorithms. *Polish Maritime Research*, 28(1), 83-96. <https://www.doi.org/10.2478/pomr-2021-0008>
- Onan, A. (2018). A Clustering Based Classifier Ensemble Approach to Corporate Bankruptcy Prediction. *The Journal of Operations Research, Statistics, Econometrics and Management Information Systems*, 6(2), 365-376.
- Öztürk, S. & Suluk, S. (2020). Mutluluk Ekonomisi: G8 ülkeleri açısından ekonomik büyüme ve mutluluk arasındaki ilişkinin incelenmesi. *Süleyman Demirel Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, (37), 226-249.
- Rokach, L. (2016). Decision forest: Twenty years of research. *Information Fusion*, 27, 111-125. <http://dx.doi.org/10.1016/j.inffus.2015.06.005>




- Suchetana, B., Rajagopalan, B. & Silverstein, J. (2017). Assessment of wastewater treatment facility compliance with decreasing ammonia discharge limits using a regression tree model. *Science of the Total Environment*, 598, 249- 257. <https://doi.org/10.1016/j.scitotenv.2017.03.236>
- Ulkhaj, M. M. & Adyatama, A. (2021). Clustering countries according to the world happiness report 2019. *Engineering and Applied Science Research*, 48(2). 137-150.
- World Happiness Report, 2022a, “Data for table 2.1”, <https://worldhappiness.report/ed/2022/#appendices-and-data> , (29.07.2022)
- World Happiness Report, 2022b, “FAQ”, <https://worldhappiness.report/faq/> , (29.07.2022)

How cite this article

Dogruel, M., Soner Kara, S. (2023). Determining the happiness class of countries with tree-based algorithms in machine learning. *Acta Infologica*, 7(2), 243-252. <https://doi.org/10.26650/acin.1251650>

Converting Image Files to LaTeX Format Using Computer Vision, Natural Language Processing, and Machine Learning

Resim Formatındaki Dokümanların Bilgisayarlı Görü, Doğal Dil İşleme ve Makine Öğrenmesi Kullanılarak Latex Formatına Dönüştürülmesi

Murat Kazanç¹ , Tolga Ensari² , Mustafa Dağtekin³ 

¹(M.Sc.), Istanbul University-Cerrahpasa, Department of Computer Engineering, Istanbul, Türkiye

²(Assist. Prof.), Arkansas Tech University, College of Engineering & Applied Science, Department of Computer and Information Science, Arkansas, USA

³(Assist. Prof.), Istanbul University-Cerrahpasa, Department of Computer Engineering, Istanbul, Türkiye

Corresponding author : Mustafa DAĞTEKİN

E-mail : dagtekin@iuc.edu.tr

ABSTRACT

A few decades ago, people used printed resources such as books and magazines to learn. With the development of technology, digital documents have replaced printed resources. These documents can occur in the form of images or various text formats. Many different applications exist for preparing digital documents, one of these being LaTeX. LaTeX is a document preparation system and typesetting software that is used especially in the field of scientific publications and mathematics for preparing high quality documents. When preparing a document using LaTeX, the content is made ready using a markup language, which creates difficulties for some users. However, one of the main advantages of using the LaTeX system is that it distinguishes the document's content from its formatting. Once the content is created, the formatting can be easily replaced. Generating LaTeX code from an image-formatted document requires both the use of computer vision and NLP. This study discovers the boundaries (blocks) of the places where text, tables, and figures are located on an image before making a text classification using the natural language processing methods of these blocks. The next stage of the study determines the reading order to enable meaningful flow. The final stage of the study produces a LaTeX code using the obtained information.

Keywords: Computer vision, text classification, reading order, machine learning

ÖZ

Birkaç on yıl önce insanlar bilgi edinmek için kitap ve dergi gibi basılı kaynakları kullanmaktaydılar. Teknolojinin gelişmesi ile basılı kaynakların yerini dijital dokümanlar almıştır. Bu dokümanlar görüntü biçiminde veya farklı metin formatları şeklinde olabilmektedir. Dijital dokümanları hazırlamak için birçok farklı uygulama bulunmaktadır. Bunlardan bir tanesi LaTeX' tir. LaTeX doküman hazırlama sistemi ve dizgi yazılımıdır. Yüksek kalitede dokümanlar hazırlamak için özellikle bilimsel yayınlar ve matematik alanında kullanılmaktadır. LaTeX ile doküman hazırlanırken içerik bir işaretleme dili kullanılarak hazırlanmaktadır. Bu durum bazı kullanıcılar için bir zorluk oluşturmaktadır. Ancak LaTeX sistemini kullanmanın avantajlarından biri doküman içeriğini biçimlendirmeden ayırmasıdır. Bir kere içerik oluşturulduktan sonra biçimlendirme kolaylıkla değiştirilebilmektedir. Görüntü formatındaki bir dokümandan LaTeX kodunun üretilmesi bilgisayarlı görüş ve doğal dil işleme alanlarının birlikte kullanılmasını gerektirmektedir. Bu çalışmada öncelikle görüntü üzerinde metin, tablo ve şekillerin bulunduğu yerlerin sınırları (bloklar) tespit edilmiştir. Sonrasında bulunan bu blokların doğal dil işleme metotları kullanılarak metin sınıflama yapılmıştır. Bir sonraki aşamada anlam akışının bozulmaması için okuma sırası tespit edilmiştir. Son aşamada elde edilen bilgiler kullanılarak LaTeX kodu üretilmiştir.

Anahtar Kelimeler: Bilgisayarlı görüş, metin sınıflama, okuma sırası, makine öğrenmesi

Submitted : 01.03.2023

Revision Requested : 24.08.2023

Last Revision Received : 31.08.2023

Accepted : 19.09.2023

Published Online : 26.10.2023



This article is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0)

1. INTRODUCTION

A few decades ago, people used such resources as books and magazines to obtain information. With the current developments in technology and the widespread use of the Internet, digital documents are used more often than printed documents. Different applications are also available for preparing digital documents. One such tool is LaTeX, a document preparation system and typesetting software (CTAN Team). LaTeX is used to create high-quality documents, particularly in the fields of science and mathematics. When creating a document using LaTeX, one writes the contents of the document in a plain text file using a simple markup language to indicate the structure of the document (e.g., headings, paragraphs, lists) and to include such things as mathematical equations and citations. This plain text file is then processed by the LaTeX software, which converts it into a typeset document in one's chosen format (e.g., PDF). One of the main advantages of using LaTeX is that it separates the contents of a document from its formatting. This means one can focus on writing the content and let LaTeX handle the document's typesetting details. This makes creating consistent, high-quality documents easier and allows facilitates later changes. The production of LaTeX code from an image of a document is obtainable using computer vision and natural language processing (NLP).

This study focuses on academic publications. Due to their structure, the first page of academic publications contains sections such as the title, author information, abstract, and keywords. The middle pages of a publication contain content groups such as headings, paragraphs, tables, figures, and lists in single- or double-column format. The last page or pages of a publication contain a unique reference list. In order for a digital academic publication to be analyzed, these content groups should initially be found in the digital document as a content block, and then the content should be parsed as text or figures. The contents of the blocks identified as text should be classified according to the type of the page (left justified, centered, or right justified). In addition, in order to not disrupt the flow of a digital publication, the reading order must also be correctly identified. LaTeX code should be created in the final stage, using all the information collected so far.

This study does not include certain features within its scope, such as font type and font size, which should be represented in a digital document. In addition, converting the equation and table of content groups of a digital document into LaTeX code requires a completely different study on the subject. The operations performed in this study will prepare a starting point and an expandable working basis for future studies.

2. Literature Review

LaTeX is a document preparation system. Widely preferred word processing programs such as Microsoft Word and Libre Office Writer can be expressed as follows: What you see is what you get (Klatsky, 2003). An attribute of these programs is that they allow the user to display a document on a computer monitor in exactly the format the document will take when printed. With LaTeX, the content of the document is written in plain text. The desired formatting is made using LaTeX codes, with the result displayed after compilation. In this respect, LaTeX can be compared to the Hypertext Markup Language (HTML) used to develop web pages.

LaTeX was founded in 1978 by Donald E. Knuth, and the development process has been non-stop ever since (CTAN Team). The starting point was the insufficient print quality of the documents. Over the years, LaTeX has been adopted by academic circles in particular for writing books and articles and continues to be developed as an open-source code today. The main reasons why LATEX is preferred are as follows: It can be published professionally, it is a standard, it can be used across platforms, and it is constantly evolving and expanding due to its open-source nature.

In order to convert a document's image file to LaTeX code, the first thing to do is to have blocks such as text or shapes on the image. One of the libraries developed for this process is LayoutParser (Shen et al., 2021). Using deep learning methods, this open-source library facilitates digital document analysis. Another feature that makes LayoutParser stand out is its pre-trained artificial neural network models. These are PubLayNet for academic documents (Zhong et al., 2019), PRImA for journals and academic reports (PRImA), and TableBank for business and academic documents (Li et al., 2019). These pre-trained models have common classes, as well as their own specialized classes. For example, the TableBank pre-trained model finds only tables. LayoutParser also has an optical character recognition (OCR) feature that uses the Tesseract library (Wang et al., 2021).

Another study related to the discovery of blocks in pictures involves Layout (Y. Xu et al., 2020). That study stated that in order to classify the blocks found in the picture, the texts contained in the blocks should be taken into account in addition to their formal characteristics. For example, the bidirectional encoder representations for transformers (BERT) is a pre-trained NLP models that is used to understand texts.

For the purpose of finding blocks, two deep learning models have been used for segmentation (C. Xu et al., 2021). The first deep learning model determined the locations of the blocks using Mask region-based convolutional neural

networks (R-CNN). The second deep learning model uses feature pyramid networks (FPNs) to perceive objects at different scales, and the blocks are classified as a result of the collaboration of these two deep learning models.

Clark & Divvala (2016) claimed that shapes are important in academic publications with a different focus, and their main goal was to identify pictures and the explanatory text underneath them. Instead of using a deep learning model in the development stage, they developed OCR and a rules-based detection method.

Deivalakshmi et al. (2013) proposed an algorithm-based system for detecting text and non-text blocks in a digital document. Horizontal and vertical lines are detected and removed. The blocks are then revealed by grouping text or non-text content using the dilation method. Pictures have different texture properties than text. Texture properties occur with gray-level co-occurrence matrices (GLCMs) in the found blocks. A GLCM is able to work on grayscale images by creating a matrix using just the position numbers of pixels in the vertical, horizontal, and diagonal directions for every pixel in the entire image or in divisions. The next stage classifies the blocks as text or non-text using the k-means clustering algorithm.

Kavasidis et al. (2019) used convolutional neural networks (CNNs) to detect tables and figures in digital documents. Nowadays, a rapid increase in data production is taking place that continues to accelerate. At this point, no person or group of people can monitor this information flow. Instead of trying to get information in digital documents with NLP methods, the solution to this problem is to conclude that the tables and figures in academic publications provide more information than the textual contents of the document.

Deng et al. (2019) used the encoder-decoder model was used to detect tables in documents. They extracted the feature map of the image using CNN in the encoder model. Meanwhile, the decoder model detects table cells using the attention mechanism and the standard long- and short-term memory (LSTM) model.

One of the strongest points of LaTeX involves equations. Finding equations in a picture and expressing them as LaTeX code would be quite convenient. To perform this process, Wang and Liu, (2021) again used the encoder-decoder model, extracting the feature map using CNN as the encoder. A LaTeX code block was produced with an LSTM using a soft attention mechanism as a decoder.

Pang et al.'s (2021) study on converting equations to LaTeX code used a transformer model instead of the encoder-decoder models most researchers use. The main reason for this was that the encoder-decoder model cannot know the entire context and hierarchical positions of the symbols. The study consisted of three stages: extracting the entire context, using a transformer model that reveals the dependencies between location information and symbols, and using a mask-based attention mechanism decoder to produce LaTeX code.

Another process performed in extracting information from digital documents is to determine the fonts of the texts. This feature provides structured information about the document rather than the content. To do this, Wang et al.'s (2015) study used a stacked convolutional autoencoder (SCAE), a special type of CNN.

Ding et al. (2019) conducted on OCR systems, stating that OCR models using CNN and LSTM achieved a high accuracy rate. However, the model was said to occupy a large amount of disk space and to have a high computational cost. The method they proposed was to reduce the LSTM side of the model without making changes to the feature extraction process. Finally, they used a compression method to reduce the size of the model.

Safnuk and Hu's (2018) research was conducted on PDF files with the goal of obtaining LaTeX code by reverse engineering the PDF files produced with LaTeX code. A deep learning model was used to understand the relationships in the metadata of the source PDF file and to produce appropriate codes. The LaTeX code of the model output a digital document that had been successfully reconstructed.

Apart from the literature study, real-world applications have also been examined. Web applications are found that convert word processing document formats (e.g., docx) to LaTeX code. These applications contain text information in XML format in the source file as well as font, font size, and all visual features related to the text. The main difficulty in converting a document's image file to LaTeX code is the availability of these features. Again, by performing OCR on the digital document provided by different applications, all the found characters are placed in their positions on the page with fixed coordinates. These studies were unable to produce a code similar to LaTeX code written by a human.

3. Method

3.1. Dataset

The current study has found no conducted stud to have a suitable dataset. For this reason, 103 academic publications in PDF format were accessed on the web environment. These were then used during the literature search on the subject.

The PDFs collected for training the system under development were first converted to an image file so that each page is a separate image. The information then needs to be extracted from the pictures. This information involves the

location of the text, title, lists, tables, and image blocks, as well as the text contained in the blocks. In order to find the classes of the blocks, the study uses the LayoutParser library (Shen et al., 2021) instead of labeling them individually. This library allows the location and classes of the blocks in each image to be found and recorded in CSV format in accordance with the segmentation system to be developed later using OCR as the last operation on the texts the blocks contain. Table 1 shows the dataset sample records. In order to create the data set, 11,965 blocks were extracted from 1,141 images.

Table 1. *Dataset Sample Records*

Index	Box Index	Box X1	Box X2	Box Y1	Box Y2	Box Text	Box Type	Page Size	Page Number
200	10	951	1574	333	1419	Intelligent Multimedia...	List	2200x1700	3
201	0	225	1440	1792	1848	Fig. 2. Results: a) Input...	Text	2200x1700	4
202	2	219	1450	196	1742	fay\n'nie\n'ni f'n\n \n'n...	Figure	2200x1700	4
203	3	226	843	215	1375	RT ee aa ates caret\n'...	Figure	2200x1700	5
205	9	112	1545	425	1824	Classification Kate for...	Table	2200x1700	5

3.1.1. Data Preprocessing

In order to distinguish between figures and text in the system to be developed in the next stage, the ratio of the area covered by a block to the entire area of the image in the data set and the number of words contained in the block are needed. Firstly, the width and height are calculated from the required columns of the block expressing the coordinates as a rectangle and added to the dataset table as a new column. Then, the page area is calculated using the page width and height from the data set. Finally, the area occupied by the block on the page as a percentage was found and added to the data set. At the development stage, the positions of the blocks and other unnecessary information were discarded, and a data set was formed.

The texts in the blocks contain meaningless characters left over from the OCR process, as well as characters that are not important for the number of words, such as punctuation marks. These need to be cleared in order to reach a correct number of words. The text is then converted to words, and the word count is calculated. Finally, a parameter is obtained using the ratio of the block to the page area and the number of words; this is then added to the data set. Table 2 was formed once the basic data analysis had been performed on the obtained data set.

Table 2. *Percentages for the Number of Words in the Field*

Index	Box Index	Box X1	Box X2	Box Y1	Box Y2	Box Text	Box Type	Page Size	Page Number
200	10	951	1574	333	1419	Intelligent Multimedia...	List	2200x1700	3
201	0	225	1440	1792	1848	Fig. 2. Results: a) Input...	Text	2200x1700	4
202	2	219	1450	196	1742	fay\n'nie\n'ni f'n\n \n'n...	Figure	2200x1700	4
203	3	226	843	215	1375	RT ee aa ates caret\n'...	Figure	2200x1700	5
205	9	112	1545	425	1824	Classification Kate for...	Table	2200x1700	5

Block types are arranged in such a way that their shapes and text (e.g., list, table, text, and title) are combined. In the experiments, the Table Block detection was unsuccessful.

3.1.2. Page and Block Types

Academic publications do not have a complete set of structural standards, but certain structures are common. For example, the first page of an academic publication should contain a main title and an abstract. However, keywords are not found in every publication format. As another example, the last page should contain reference information. Due to these differences, instead of using the same text classification model for each page of the digital document, three different classes have been determined for the pages of a digital document. These are the first pages, the middle pages, and the last (reference) pages. After determining the page type, the models to be developed for the text classes suitable for that page type will be used. For this reason, the selection was made from the data set obtained at the beginning. Afterward, the data set was then recreated by applying the OCR process to the entire page.

Data with block type figures are decoded from the middle pages data set. This stage will take no action regarding that data class. The middle pages' text classes are decoded as text, titles, tables, or lists.

For the first pages, the block types in the dataset are determined as main title, author(s), keywords, title, and text. For the last pages (i.e., references), the block types in the dataset are text, reference, title, and table.

3.2. Training the Model

This study, carries out different model trainings within its scope to perform different tasks. The model trainings divide the datasets into 80% training and 20% testing. The study uses the following machine learning models: decision tree, support vector machine (SVM), k-nearest neighbor (KNN), random forest, and linear classifier. The decision tree model has internal nodes that can be taken as tests on input data patterns and leaf nodes that can be taken as categories. These tests are filtered down through the tree to get the right output-to-input pattern (Navada et al., 2011). SVM is a useful methodology for finding the best possible surface to separate positive samples from negative samples (Ali et al., 2016). KNN is categorized as an unknown document. The KNN classifier ranks the document's neighbors among the training documents and uses the class labels of the k-most similar neighbors (Uğuz, 2011). As the name implies, the random forest classifier consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest in this algorithm gives a class prediction, and the class with the most votes among the trees becomes the model's predicted class (Akpan & Starkey, 2021). Linear classifier is a method for dividing data into classes by finding a linear combination of attributes (Doğan et al., 2019).

3.2.1. Training the Text and Figure Classifications Model

Table 3 provides the algorithms used to perform the figures and text classifications of the block type and the complexity matrix formed as a result of the prediction applied to the dataset test section.

Table 3. Text/Figure Model Evaluation Values

Classification Model	Accuracy
Decision Tree	0.9736
SVM	0.9690
KNN	0.9803
Random Forest	0.9791

Looking at the results from the machine learning models, the most successful model is KNN.

3.2.2. Training the Page Type Classification Model

In order to use the dataset prepared for determining the page type when training the model, the data are converted into a vector showing the representation of words. The accuracy values of the trained models are given in Table 4.

Table 4. Page Type Model Training Accuracy Values

Model	Accuracy
Decision Tree	0.955307
SVM	0.944134
Random Forest	0.910615
Linear	0.966480

Looking at the results, the linear classifier is the most successful model. The complexity matrix for this model is given in Table 5.

Table 5. Linear Classifier Model Training Complexity Matrix

		Estimation		
		First Pages	Middle Pages	Last Pages
Real Results	First Pages	16	3	0
	Middle Pages	1	130	0
	Last Pages	0	2	27

3.2.3. Training the Middle Pages Text Classification Model

With an 87% accuracy, the SVM classifier was the most successful model in the trainings conducted using the text classification methods. When examining the results, table blocks were seen to have the lowest accuracy of about 33% success in being classified.

3.2.4. Training the First Pages Text Classification Model

The linear classifier was the most successful model in the trainings conducted using the text classification methods, with an 82% accuracy rate. When examining the results, the keyword and main title blocks were seen to have the lowest accuracy in being classified, with a success rate of about 75%.

3.2.5. Training the Last Pages (i.e., Reference Pages) Text Classification Model

The SVM classifier was the most successful model in the trainings conducted using the text classification methods, with an 89% accuracy. When examining the results, failure occurred in the table class due to a sufficient sample size not being provided.

3.3. Segmentation

The first stage in producing LaTeX code from a digital document's image file is the presence of content regions (blocks) on the image. The OpenCV library was used for this task. The color scale of the pictures were converted to grayscale so that the boundaries of the blocks could be found more easily. The pictures consist of three channels (i.e., red, green, and blue [RGB]). For each pixel, there are values for these three channels. For the grayscale operation, the OpenCV RGB multiplies each channel by an empirical number, as shown in Equation 1 (*Recommendation ITU-R BT.601-7, 2011*). R is the red channel. G is the green channel. B is the blue channel, and Y is the grayscale channel of the picture.

$$RGBtoGray : Y \leftarrow 0.299 \cdot R + 0.587 \cdot G + 0.114 \cdot B \quad (1)$$

The Otsu (1979) threshold determination method was used to convert the grayscale image to binary toning (black and white). This method first extracts the histogram of the picture. By assuming the threshold value of each class in the histogram, pixel values with smaller thresholds are then calculated as the background, and pixel values with bigger thresholds are calculated as the foreground. The variance value is calculated for the remaining classes as shown in Equation 2. The goal is to minimize the intra-class variance and maximize the inter-class variance. The variable t is a step. $P(i)$ is the probability found from the histogram, q_1 and q_2 are the cumulative sum of the classes. The sum of q_1 and q_2 equals 1, μ_1 and μ_2 are means, and σ_1^2 and σ_2^2 are variances.

$$\begin{aligned}
 q_1(t) &= \sum_{i=1}^t P(i) & q_2(t) &= \sum_{i=t+1}^I P(i) \\
 \mu_1(t) &= \sum_{i=1}^t \frac{iP(i)}{q_1(t)} & \mu_2(t) &= \sum_{i=t+1}^I \frac{iP(i)}{q_2(t)} \\
 \sigma_1^2(t) &= \sum_{i=1}^t [i - \mu_1(t)]^2 \frac{P(i)}{q_1(t)} & \sigma_2^2(t) &= \sum_{i=t+1}^I [i - \mu_2(t)]^2 \frac{P(i)}{q_2(t)}
 \end{aligned} \quad (2)$$

The peculiarity of the Otsu method is that the threshold value is not a fixed value but rather is determined dynamically. The pixels above the threshold value will be white, and the pixels below the threshold value will be black. This method is used to obtain the coordinates of the individual blocks by examining the pixels.

This study has found contours to be unnecessary. One example is the page number found in a picture, or informational messages placed at the top and bottom of a publication downloaded on the web environment stating which academic institution through which the website connection has been established. These contours should not be taken into account. In order to remove them from the contour list, the following situations were observed: If the area occupied by the stroke is less than 10% of the page, if the aspect ratio is less than 20%, if the y-point value of the stroke is less than 3% of the height of the page, or if the y-point value of the stroke is greater than 95% of the height of the page, this stroke is not included in the block list being created. As a separate process, OCR was applied to the image fragments that were created using contour coordinates. The OCR results found the number of words by removing special characters and punctuation marks.

3.4. Reading Order

The order of reading plays an important role in understanding a document. If the document is more than one column, the number of columns should be found, and then the contents (e.g., text, figures, tables) should be identified. The pseudo-code of the reading order algorithm is as follows:

```

GET → blocks(), pageWidth
FOR i TO blocks.count STEP i++
    IF blocks.width(i)-blocks.x(i) > pageWidth /2
        fullColumns.add(blocks.id(i))
    IF blocks.width (i) < pageWidth /2
        leftColumns.add(blocks.id(i))
    IF blocks.x(i) > pageWidth /2
        rightColumns.add(blocks.id(i))
ENDFOR
sortTopToBottom(fullColumns)
sortTopToBottom (leftColumns)
sortTopToBottom (rightColumns)
IF blocks.count = fullColumns.count
    readingOrder = fullColumns
ELSEIF fullColumns.count = 0
    readingOrder = leftColumns + rightColumns
ELSE
    FOR fullColumns.count TO m STEP m—
        FOR rightColumns.count TO n STEP n—
            IF fullColumns.y(m) < rightColumns.y(n)
                readingOrder.add(rightColumns (n))
            ENDFOR
        FOR leftColumns.count TO k STEP k—
            IF fullColumns.y(m) < leftColumns.y(k)
                readingOrder.add(leftColumns (k))
            ENDFOR
        ENDFOR
    readingOrder = reverse(readingOrder)
ENDIF
OUT → readingOrder

```

After the segmentation of the document in the previous processing step, the process of finding the reading order is then carried out using the x and y coordinates and width and height information from which the blocks were formed. In addition, page margin information and inter-column space information, if any, are also found while creating the reading order.

3.5. Text/Figure Classification

For the blocks found as a result of segmentation and placed in reading order, the next stage is to find out which blocks are text and which are shapes. For this process, the values required for normalization are calculated from the data used in the training of the model, then the normalization process is applied to the newly found page block area ratio and word number values. The machine learning model prediction results are then added as a new column to the dataset. In the next stage, the blocks that have been classified as shapes are cropped from the source image according to their coordinates and saved as a new image file. The ID number for that block is given as the name.

3.6. Page Type Classification

OCR is performed on the entire source image to determine the page type. In order to determine the page type, the OCR result made using the vector state of the dataset used in the model's training as a source is converted into a vector, and then predictions are made on the model. The prediction results can be one of three classes: first pages, middle pages, and last pages.

3.7. Classification Of Text According to Page Type

Three different machine learning models have been trained to perform the text classification process according to page type. The predictions are performed using one of these models based on the page type information that was found. The extracted data is then saved to the source folder in CSV format. The page border information (if present), the space between the columns, the width and height of the source image, the information about the columns (if present), and the page type information are saved to the source folder in JSON format. The study has been carried out in this way, with all the operations performed up to this stage having been recorded.

3.8. Generating the LaTeX Code

The presence of special characters that are also used for writing commands in LaTeX code (e.g., % \$ &) in the text will cause compilation errors. In order to prevent errors that may occur, the back space character (i.e., \) is added before these characters should they occur with a function written in the text column in the dataset. The generated LaTeX code is then saved to the source folder in a file with the extension .tex. The text classifications are taken from a list based on page type, and the process then is started.

According to the information from the JSON file for the first pages section, the blocks found as such are identified as single or double columns. The necessary code is then added for the page borders and, if necessary, for the blank information between the columns. The necessary packages are then added. If the source image is classified as a first page, the main page and author information are then added there. LaTeX codes have been added for the middle pages section in accordance with the text classifications contained in the dataset.

3.9. Application Development

The development stages are mentioned in detail. The pseudo-code for the application algorithm is as follows:

4. Findings

Figure 1 shows the .tex file for the compiled PDF output examples; this file's output is formed as a result of a source digital publication given as input to the developed application.

```

GET → image
grayImage = doGrayScale(image)
blackWhiteImage = otsuThreshold(grayImage)
dilationImage = dilation(blackWhiteImage)
blocks() = findContour(dilationImage)
readingOrder = findReadingOrder(blocks())
blocks() = findTextFigure(blocks())
pageType = findPageType(image)
IF pageType = "firstPage"
    blocks() = firstPageTextClassification(blocks())
ELSEIF pageType = "middlePage"
    blocks() = middlePageTextClassification(blocks())
ELSEIF pageType = "lastPage"
    blocks() = lastPageTextClassification(blocks())
ENDIF
OPEN texFile AS tex
    tex.addPreamble(preambleInformation)
    IF pageType = "firstPage"
        tex.preambleAdd("Main Title", "Author")
        tex.addBody("Title", "Text", "Keyword", "Figure")
    ELSEIF pageType = "middlePage"
        tex.addBody("Title", "Text", "List", "Table", "Figure")
    ELSEIF pageType = "lastPage"
        tex.addBody("Title", "Text", "Referance", "Table", "Figure")
CLOSE texFile
OUT → texFile

```

Figure 1a shows the original document and Figure 1b shows the PDF output of the developed application. The application developed in this example was found to be successful. A two-column page structure has been created. Blocks are found that have been made with errors. In addition, no exact match is found due to features that were not taken into account in this application (e.g., font type, paragraph spaces).

4.1. Segmentation

When evaluating the blocks obtained by finding contours in OpenCV, the results are concluded to be a partial success. Successful results were seen to have been obtained in some experiments, while the results from others were unsatisfactory. If the figures in the digital document have an irregular fragmented structure or occur on the first pages of the document, the segmentation performance for the main heading is observed to be low. The reason why the segmentation system does not exhibit repeatable output is that the documents do not have a specific standardized appearance.

4.2. Text/Figure Classification

The classification was theoretically developed by taking into account the ratio of a shape block to the total area of the document and the ratio of the area of the block and the number of words contained in it. This is theoretically true. However, errors are seen to have occurred in the estimates made on real samples. For example, the outputs given unsuccessfully regarding segmentation are seen for text blocks can be classified as shapes.

4.3. Reading Order

Since the reading order algorithm works heuristically, it works as a rules-based algorithm in accordance with the information coming from segmentation. As a result of the experiments carried out, errors are seen to be able to occur due to errors in segmentation.

incorporated into encoder-decoder framework for calculating the fixed-length context vector, i.e., the variable-length representations could be summed using the attention as the weighting coefficients. After adopting an attention mechanism into encoder-decoder, salient regions in the static representation can dynamically rise to the forefront. The attention mechanism plays an indispensable role in image captioning for obtaining a state-of-the-art performance. For example, [42] proposed the "hard" attention for image captioning system to know where to attend and its effectiveness was shown through attention visualization. [43] proposed the concept of "attention correctness" to strength the alignment for image captioning. In [44], an adaptive attention was implemented via a visual search so that the captioning system could also know when to attend, and with a similar motivation, [45] also proposed a global-local attention so that model could selectively pay attention to spatial objects and context information.

D. Neural Network-Based Approaches for OHMER

The generality of the attention based encoder-decoder framework suggests that OHMER may also be one proper application. Recently, [46], [47] used the attention based encoder-decoder model for OHMER and significantly outperformed the best system on CROHME 2014. In [46] the proposed model consisted of a FCN encoder and a GRU decoder equipped with a coverage-based attention model while [47] employed a CRNN as the encoder and the decoder is equipped with a coarse-to-fine attention model. However, both [46] and [47] treated the HMEs input as static images which ignores the handwriting dynamics (namely the temporal order and trajectory). As we can see in Fig. 1, besides the symbol/shape information, the writing order is also preserved in the online sequential data, which is important information and can not be recovered from the static image. Therefore, to capture the dynamic information to reduce handwritten ambiguities, [48] proposed to employ a GRU encoder that directly takes the raw sequential data as input. Validated on CROHME 2014, [48] showed a significant improvement of recognition accuracy over [46], [47].

This study is an extension of the previous work in [48] with the following new contributions. 1) We propose to employ a temporal attention to teach the parser when to rely on the representations extracted by tracker and when to just rely on the built-in language model. 2) To compute the temporal attention, the spatial attention is slightly adjusted. Meanwhile, we newly introduce an attention guider to help improve the learning of spatial attention. 3) We blend a FCN watcher into TAP by considering the strong complementarity between static-image based input and dynamic-trace based input. By processing HMEs from two different modalities, the strengths of [46] and [48] can be fully utilized simultaneously. 4) We use an extra official text dataset containing only \LaTeX notations to train an additional language model for enhancing our parser. 5) More experiments and analyses are included.

III. NETWORK ARCHITECTURE OF TAP

In this section, we elaborate the proposed TAP architecture which parses a mathematical expression structure into a \LaTeX

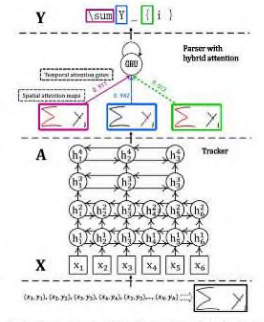


Fig. 2. Overall architecture of Track, Attend, and Parse. X denotes the input sequence in Section III-A. A denotes the annotation sequence in Section III-C. Y denotes the output sequence in Section III-C.

string by tracking a sequence of online handwritten points. As illustrated in Fig. 2, the raw data is a sequence of points containing xy-coordinates which can be visualized as the bottom-right image by drawing the trajectory. A preprocessing is first applied to extract trajectory information from raw sequential data. The tracker is a stack of bidirectional GRU while the parser combines a GRU based language model and a hybrid attention mechanism. As for the hybrid attention mechanism, spatial attention can potentially well learn the alignment between input traces and output string while temporal attention can well know when to rely on the product of spatial attention and when to just rely on the language model. For example, in Fig. 2, the purple, blue and green rectangles denote three symbols with the red color representing the spatial attention probabilities of each handwritten symbol (higher color denotes higher probability) and the probabilities linking to rectangles represent their reliability produced by temporal attention. When predicting the math symbol " \sum ", the spatial attention model aligns well to the stroke of " \sum " (in the purple spatial attention map) which corresponds to the human intuition and the temporal attention probability linking to the purple rectangle is extremely high as the spatial attention map is accurate. Conversely, when predicting the math symbol " \int ", there is no object for spatial attention model to attend to, leading to an inaccurate spatial attention map. Therefore the temporal attention probability linking to the green spatial attention map is small which tells the parser should rely on the built-in language model at this time.

incorporated into encoder-decoder framework for calculating the fixed-length context vector, i.e., the variable-length representations could be summed using the attention as the weighting coefficients. After adopting an attention mechanism into encoder-decoder, salient regions in the static representation can dynamically rise to the forefront. The attention mechanism plays an indispensable role in image captioning for obtaining a state-of-the-art performance. For example, [42] proposed the "hard" attention for image captioning system to know where to attend and its effectiveness was shown through attention visualization. [43] proposed the concept of "attention correctness" to strength the alignment for image captioning. In [44], an adaptive attention was implemented via a visual search so that the captioning system could also know when to attend, and with a similar motivation, [45] also proposed a global-local attention so that model could selectively pay attention to spatial objects and context information.

D. Neural Network-Based Approaches for OHMER

The generality of the attention based encoder-decoder framework suggests that OHMER may also be one proper application. Recently, [46], [47] used the attention based encoder-decoder model for OHMER and significantly outperformed the best system on CROHME 2014. In [46] the proposed model consisted of a FCN encoder and a GRU decoder equipped with a coverage-based attention model while [47] employed a CRNN as the encoder and the decoder is equipped with a coarse-to-fine attention model. However, both [46] and [47] treated the HMEs input as static images which ignores the handwriting dynamics (namely the temporal order and trajectory). As we can see in Fig. 1, besides the symbol/shape information, the writing order is also preserved in the online sequential data, which is important information and can not be recovered from the static image. Therefore, to capture the dynamic information to reduce handwritten ambiguities, [48] proposed to employ a GRU encoder that directly takes the raw sequential data as input. Validated on CROHME 2014, [48] showed a significant improvement of recognition accuracy over [46], [47].

This study is an extension of the previous work in [48] with the following new contributions. 1) We propose to employ a temporal attention to teach the parser when to rely on the representations extracted by tracker and when to just rely on the built-in language model. 2) To compute the temporal attention, the spatial attention is slightly adjusted. Meanwhile, we newly introduce an attention guider to help improve the learning of spatial attention. 3) We blend a FCN watcher into TAP by considering the strong complementarity between static-image based input and dynamic-trace based input. By processing HMEs from two different modalities, the strengths of [46] and [48] can be fully utilized simultaneously. 4) We use an extra official text dataset containing only \LaTeX notations to train an additional language model for enhancing our parser. 5) More experiments and analyses are included.

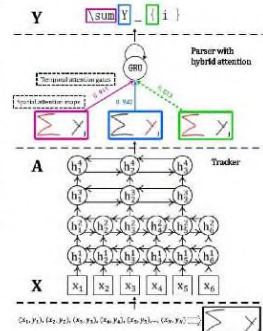


Fig. 1. Fig. 2. Overall architecture of Track, Attend, and Parse. X denotes the input sequence in Section III-A. A denotes the annotation sequence in Section III-C. Y denotes the output sequence in Section III-C.

III. NETWORK ARCHITECTURE OF TAP

In this section, we elaborate the proposed TAP architecture which parses a mathematical expression structure into a \LaTeX

Parser with hybrid attention

string by tracking a sequence of online handwritten points. As illustrated in Fig. 2, the raw data is a sequence of points containing xy-coordinates which can be visualized as the bottom-right image by drawing the trajectory. A preprocessing is first applied to extract trajectory information from raw sequential data. The tracker is a stack of bidirectional GRU while the parser combines a GRU based language model and a hybrid attention mechanism. As for the hybrid attention mechanism, spatial attention can potentially well learn the alignment between input traces and output string while temporal attention can well know when to rely on the product of spatial attention and when to just rely on the language model. For example, in Fig. 2, the purple, blue and green rectangles denote three symbols with the red color representing the spatial attention probabilities of each handwritten symbol (higher color denotes higher probability) and the probabilities linking to rectangles represent their reliability produced by temporal attention. When predicting the math symbol " \sum ", the spatial attention model aligns well to the stroke of " \sum " (in the purple spatial attention map) which corresponds to

(a)

(b)

Figure 1. Sample application of (a) original document and (b) application output.

4.4. Page Type Classification

The machine learning model applied for finding page types gave completely accurate results in the performed experiments. No erroneous estimates were made.

4.5. Finding Text Classes Specific to the Page Type

For a source document whose page type is a first page, two sample results are shown in Figures 2a and 2b. Segmentation problems regarding finding the main title and author blocks negatively affect the classification process. The errors experienced in segmentation are seen to negatively affect the rest of the system.

For a source document whose page type is a middle page, experiments have shown no repeatable success to occur regarding detecting list and table blocks. For a source document whose page type is a last page, two sample results are shown in Figures 3a and 3b. Experiments have shown that tables and reference blocks are confused with text blocks. No repeatable success has occurred in detecting these blocks.

4.6. Generating the LaTeX Code

LaTeX code is heuristically produced. Therefore, the outputs created in the previous stages of the application are used in this step. Because the errors that occurred in the previous steps are cumulatively transferred to this step, no accurate metric has been found for measuring this stage.

5. Discussion and Conclusion

When evaluating the application on different documents, repeated high success was not achievable. One of the reasons for this is that the system has a complex structure consisting of many stages. This situation causes errors that occur at one point to accumulate and negatively affect the rest of the system.

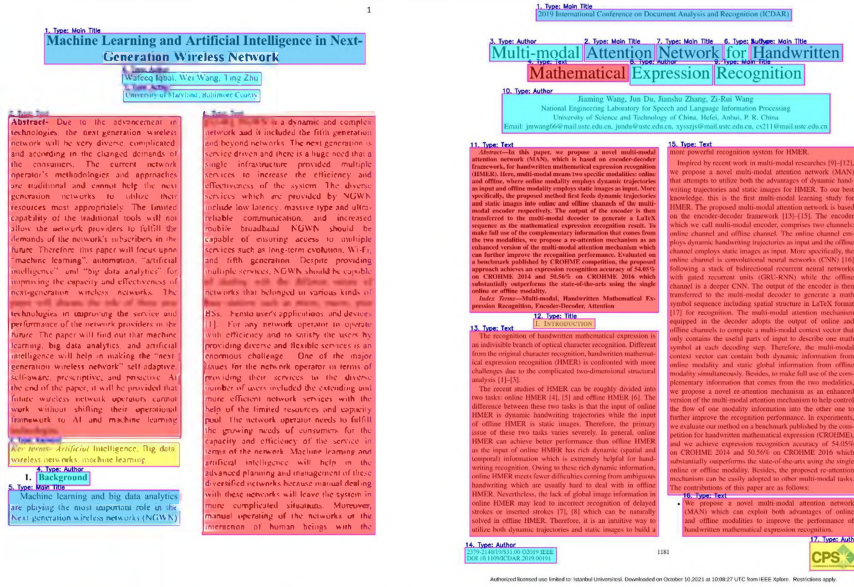


Figure 2. Segmentation results for (a) a successful example and (b) an unsuccessful example.



Figure 3. (a) Segmentation and text classification results for last pages and (b) an example.

Due to its nature, an academic publication consists mainly of text blocks. These text blocks are divided into classes within themselves. Separating these blocks of text from one another is often quite a difficult task. Reference pages are given in Figures 3a and 3b as examples. The presence of different reference systems makes success here difficult. Although NLP methods are used to separate a reference block from a list or a regular paragraph structure, this is a still a monumental task.

During the literature review and the examined commercial applications, focus on one topic is always observed. For example, in a commercial solution that was developed only for converting equations to LaTeX code, the equation block was marked by having the user select it in the browser, after which it is converted to LaTeX code. This is also available in solutions that produce output in JSON format with metadata obtained from PDF files. Apart from these, studies are found to have converted the text obtained with OCR at a very simple level to LaTeX code heuristically. The application developed in this study has been comprehensive and sophisticated based on the mentioned studies.

Comparisons can be made in terms of the segmentation and classification of blocks with the LayoutParser library, which uses the deep learning method in this study to create a dataset. LayoutParser can use different pre-trained models. In accordance with our study, the PublayNet model was preferred. This model has been trained by IBM laboratories using more than 1 million academic publications.

For the comparison process, making a comparison of a document classified as a middle page will be appropriate because the text classes for the documents classified as first and last pages in the developed application were unique. Figure 4b shows a comparison of the system developed in the study with the results of the LayoutParser model used in Figure 4a.



Figure 4. (a) Comparison of the LayoutParser Result and (b) the Developed Application Result.

When examining the system outputs given in Figures 4a and 4b, the LayoutParser made errors in both segmentations and also failed to take some content into account. The work this study has done on these examples was more successful.

When comparing the results from this study with those from the LayoutParser regarding the samples, the success rate is seen to vary. However, the system developed here has entailed a more comprehensive study with text classes specific to reading order and page types. Although a certain success was achieved with the computer vision method applied in this study, the success achieved using the system had varied results among the selected samples.

With regard to the application developed in this study, this study has been shown machine learning models to still be useful with short training times, small model sizes, and fast response times in datasets containing keywords. When additionally considering the long training period and large model sizes of popular NLP models, the NLP models developed for this study can be said to be very practical. As such, the work done in this study is thought to be able to continue being developed in the future or to lay the foundation for future studies.

Peer Review: Externally peer-reviewed.

Author Contributions: Conception/Design of Study- M.K., T.E., M.D.; Data Acquisition- M.K.; Data Analysis/Interpretation- M.K.; Drafting Manuscript- M.K.; Critical Revision of Manuscript- T.E., M.D.; Final Approval and Accountability- M.K., T.E., M.D.

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.

ORCID IDs of the authors / Yazarların ORCID ID'leri

Murat Kazanç	0000-0002-8405-0181
Tolga Ensari	0000-0003-0896-3058
Mustafa Dağtekin	0000-0002-0797-9392

REFERENCES

- Akpan, U. I., & Starkey, A. (2021). Review of classification algorithms with changing inter-class distances. *Machine Learning with Applications*, 4, 100031. <https://doi.org/10.1016/j.mlwa.2021.100031>
- Ali, F., Kwak, K.-S., & Kim, Y.-G. (2016). Opinion mining based on fuzzy domain ontology and Support Vector Machine: A proposal to automate online review classification. *Applied Soft Computing*, 47, 235–250. <https://doi.org/10.1016/j.asoc.2016.06.003>
- Clark, C., & Divvala, S. (2016). PDFFigures 2.0. *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*, 143–152. <https://doi.org/10.1145/2910896.2910904>
- CTAN Team. (n.d.). *What are TEX and its friends?* Retrieved May 8, 2022, from <https://www.ctan.org/tex>
- Deivalakshmi, S., Palanisamy, P., & Vishwanathan, G. (2013). A novel method for text and non-text segmentation in document images. *2013 International Conference on Communication and Signal Processing*, 255–259. <https://doi.org/10.1109/iccsp.2013.6577054>
- Deng, Y., Rosenberg, D., & Mann, G. (2019). Challenges in End-to-End Neural Scientific Table Recognition. *2019 International Conference on Document Analysis and Recognition (ICDAR)*, 894–901. <https://doi.org/10.1109/ICDAR.2019.00148>
- Ding, H., Chen, K., & Huo, Q. (2019). Compressing CNN-DBLSTM models for OCR with teacher-student learning and Tucker decomposition. *Pattern Recognition*, 96, 106957. <https://doi.org/10.1016/j.patcog.2019.07.002>
- Doğan, M. İ., Orman, A., Örcü, M., & Örcü, H. H. (2019). Çok gruplu sınıflandırma problemlerine regresyon analizi ve matematiksel programlama tabanlı yeni bir yaklaşım. *Gazi Üniversitesi Mühendislik-Mimarlık Fakültesi Dergisi*. <https://doi.org/10.17341/gazimmfd.571643>
- Kavasidis, I., Pino, C., Palazzo, S., Rundo, F., Giordano, D., Messina, P., & Spampinato, C. (2019). A Saliency-Based Convolutional Neural Network for Table and Chart Detection in Digitized Documents. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11752 LNCS, 292–302. https://doi.org/10.1007/978-3-030-30645-8_27
- Klatsky, S. (2003). WYSIWYG. *Aesthetic Surgery Journal*, 23(4), 274–275. [https://doi.org/10.1016/S1090-820X\(03\)00150-X](https://doi.org/10.1016/S1090-820X(03)00150-X)
- Li, M., Cui, L., Huang, S., Wei, F., Zhou, M., & Li, Z. (2019). *TableBank: A Benchmark Dataset for Table Detection and Recognition*. <http://arxiv.org/abs/1903.01949>
- Navada, A., Ansari, A. N., Patil, S., & Sonkamble, B. A. (2011). Overview of use of decision tree algorithms in machine learning. *2011 IEEE Control and System Graduate Research Colloquium*, 37–42. <https://doi.org/10.1109/ICSGRC.2011.5991826>
- Otsu, N. (1979). A Threshold Selection Method from Gray-Level Histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1), 62–66. <https://doi.org/10.1109/TSMC.1979.4310076>
- Pang, N., Yang, C., Zhu, X., Li, J., & Yin, X.-C. (2021). Global Context-Based Network with Transformer for Image2latex. *2020 25th International Conference on Pattern Recognition (ICPR)*, 4650–4656. <https://doi.org/10.1109/ICPR48806.2021.9412072>
- PRImA. (n.d.). Retrieved May 22, 2022, from <https://www.primaresearch.org/>
- Recommendation ITU-R BT.601-7. (2011, March). <https://www.itu.int/dmspubrec/itu-r/rec/bt/R-REC-BT.601-7-201103-1!!PDF-E.pdf>
- Safnuk, B., & Hu, G. (2018). Reconstructing LaTeX Source Files from Generated PDFs - a Neural Network Approach. *2018 IEEE 16th International Conference on Industrial Informatics (INDIN)*, 890–895. <https://doi.org/10.1109/INDIN.2018.8472050>
- Shen, Z., Zhang, R., Dell, M., Lee, B. C. G., Carlson, J., & Li, W. (2021). *LayoutParser: A Unified Toolkit for Deep Learning Based Document Image Analysis*. <http://arxiv.org/abs/2103.15348>
- Uğuz, H. (2011). A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm. *Knowledge-Based Systems*, 24(7), 1024–1032. <https://doi.org/10.1016/j.knosys.2011.04.014>
- Wang, Z., & Liu, J. C. (2021). Translating math formula images to LaTeX sequences using deep neural networks with sequence-level training. *International Journal on Document Analysis and Recognition*, 24(1–2), 63–75. <https://doi.org/10.1007/s10032-020-00360-2>
- Wang, Z., Xu, Y., Cui, L., Shang, J., & Wei, F. (2021). *LayoutReader: Pre-training of Text and Layout for Reading Order Detection*. <http://arxiv.org/abs/2108.11591>
- Wang, Z., Yang, J., Jin, H., Shechtman, E., Agarwala, A., Brandt, J., & Huang, T. S. (2015). DeepFont: Identify Your Font from An Image. *Proceedings of the 23rd ACM International Conference on Multimedia*, 451–459. <https://doi.org/10.1145/2733373>
- Xu, C., Shi, C., Bi, H., Liu, C., Yuan, Y., Guo, H., & Chen, Y. (2021). A Page Object Detection Method Based on Mask R-CNN. *IEEE Access*, 9, 143448–143457. <https://doi.org/10.1109/ACCESS.2021.3121152>

- Xu, Y., Li, M., Cui, L., Huang, S., Wei, F., & Zhou, M. (2020). LayoutLM: Pre-training of Text and Layout for Document Image Understanding. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 20, 1192–1200. <https://doi.org/10.1145/3394486.3403172>
- Zhong, X., Tang, J., & Yepes, A. J. (2019). *PubLayNet: largest dataset ever for document layout analysis*. <http://arxiv.org/abs/1908.07836>

How cite this article

Kazanc, M., Ensari, T. & Dagtekin, M. (2023). Converting Image Files to LaTeX Format using computer vision, natural language processing, and machine learning. *Acta Infologica*, 7(2), 253-266. <https://doi.org/10.26650/acin.1258719>

Correlation Analysis of the Relationship between Demographic Variables, Computer Self-Efficacy, and Information-Seeking Behavior of Nigerian University Students

Tunde Toyese Oyedokun¹ , Medinant Dolapo Laaro² , Zainab Olanihun Ambali³ ,
Olabisi Fadeke Adesina³ 

¹(Ag. University Librarian), Thomas Adewumi University, University Library, Kwara State, Nigeria

²(College Librarian), Kwara State College of Arabic and Islamic Legal Studies, College Library, Kwara State, Nigeria

³(Academic Librarian), University of Ilorin, University Library, Kwara State, Nigeria

Corresponding author : Tunde Toyese OYEDOKUN

E-mail : tunde.oyedokun@tau.edu.ng

ABSTRACT

The study evaluated demographic variables, computer self-efficacy, and information-seeking behavior of undergraduate students at the University of Ilorin, Ilorin, Nigeria. The study used a descriptive survey design of the correlational type, and the instrument for collecting data was a questionnaire. Undergraduates of the University of Ilorin, Ilorin, Nigeria constitute the unit of analysis and their population stood at 45,885. Multi-stage sampling techniques that include stratified and purposive sampling were adopted. The sampling size was set at 394, but 366 returned questionnaires were found useful for analysis. The result of findings on the level of computer self-efficacy among undergraduate students at the University of Ilorin indicated that students are highly versed in using computers as a tool, sorting out information from search results, knowing how to access information databases and information repositories, and easily finding the information they need with their computer skills. On students' information-seeking behavior, most users agreed that they consult the library when seeking information that could assist them in their course of study and academic program, and they are willing to pay for relevant information and always check for currency and relevance of the information sources they're using. There is a strong and significant positive relationship between computer self-efficacy and information-seeking behavior. There is no significant relationship between demographic variables and computer self-efficacy. Likewise, there was no significant relationship between demographic variables and information-seeking behavior.

Keywords: Computer self-efficacy, information-seeking behavior, demographic variables, undergraduate students

Submitted : 06.07.2022

Revision Requested : 07.07.2022

Last Revision Received : 31.08.2023

Accepted : 24.08.2023

Published Online : 20.11.2023



This article is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0)

1. INTRODUCTION

Institutions of higher learning are growing more and more reliant on modern technology, particularly with the use of computers as tools for information gathering, searching, and exploration. The unrestricted dissemination of information is improved by digital information, and as a result, computer use has received more attention. An electronic device called a computer is used to digitize, store, document, and retrieve electronic information sources and can be used to network with other computers in order to share information resources and, more crucially, for academic and research purposes. As such, computer use has become increasingly pervasive in today's digital age as a tool for information search and information retrieval, and as such has proven to be an essential educational tool. Students must therefore have computer self-efficacy skills in order to perform effectively in their information-seeking tasks.

Computer self-efficacy is defined by Adekunjo and Unuabor (2018) as the degree to which a person believed that they could use a computer and the related technologies to carry out general or specialized information processing activities. Furthermore, computer self-efficacy is crucial since its operational knowledge would give students a higher chance of educational excellence. Computer self-efficacy, in a more technically restricted sense, refers to people's perceptions, beliefs, and assessments of their capacity to use computers and other electronic devices like laptops and smartphones, as well as internet browsing, to carry out specific tasks and resolve uncertainties. This demonstrates unequivocally how crucial it is for students to be computer-literate. Computer self-efficacy is a foundational component of students' usage and mastery of computers for information retrieval and searching that is built on an already established feeling of self-efficacy (Abdullah & Mustafa, 2019). Any student at a higher level who wants to excel and progress academically in this digital age should be able to explore the digital environment (Chen, 2017).

However, if computer self-efficacy refers to an assessment of one's own computer skills, then this idea may affect one's information-seeking behavior, readiness, and persistence of effort in information searching. Going by forgoing assumption, students' behavioral reactions also, such as their attitude toward computer use, are impacted by their computer self-efficacy. Given the low confidence in their own abilities, some students may feel puzzled or anxious when using computers; On the other hand, those with a high level of computer self-efficacy tend to be more interested in and inclined to use computers for both broader and specific tasks. Whatever the case, one's decision on what behavior to engage in while seeking information, the search effort, persistence in attempting that information-seeking, and the effectiveness of the information search are all driven by one's computer self-efficacy. Information searching, browsing, sourcing, and navigation are all included in the information-seeking process. We seek information to broaden our knowledge about our environment, increase our understanding of subjects of interest in the world, and to pursue professional and personal goals.

Information-seeking is a complex process that involves social, informational, and interactive behavior. It is a deliberate search for information that aims to achieve specific objectives by gathering data from designated information sources. Needs assessment, problem clarification, source selection, query design, search execution, result analysis, question answering, and review are all parts of this process and method for solving problems. Because of this, finding or searching for information to resolve uncertainty is not always an easy task. Information searchers have two options: they can either handle their own processes directly or use an intermediary. In order to solve problems, the information seeker analyzes, extracts, and integrates the required knowledge with what is already available. The process might well be repeated if the issue isn't resolved. However, this is due to the internal limitations of the person; either the drive to continue seeking out more information or the decision to stop.

Information-seeking behavior is an omnibus term that includes a range of activities designed to convey information demands together with the effort focused on information searching, selection, and assessment processes, up to the final use of such information. An individual's persistent tendency and attitude toward sourcing and gathering information for personal use and/or knowledge update can be characterized as information-seeking behavior. In essence, it involves actively looking through available information in the hopes of discovering the answer to a particular question. According to Tubachi (2018), the processes that establish users' information needs, search habits, and subsequent usage of that information can be summed up as information-seeking behavior. In a nutshell, information-seeking behavior is concerned with figuring out how a person searches for and uses information, the channels they use, and the variables that determine how they use it.

Information users engage in information-seeking behavior in response to needs they perceive and, in doing so, place demands on formal or informal information sources or services, with varying degrees of success. When a person perceives a need for information in the context of their environment and realizes that there is a knowledge gap that needs to be filled in order to solve an issue, they engage in information-seeking behavior in an effort to address the perceived need. A multitude of information systems, whether manual or computer-based, may be evaluated during

the search process. In the context of the current study, university students are expected to be able to recognize their information needs, search for information, and choose and assess the information that is available before using it. Due to the need to complete coursework, prepare for lectures, seminars, and workshops, and produce final-year research projects, students' information-seeking behavior is typically deliberate.

Be that as it may, self-efficacy influences individuals' thinking patterns, emotions, and behavior, whereas the information-seeking behavior of students is not an exception to this. This is more reason why information searching goes beyond mastering a set of information search techniques but rather includes emotional state during the search, vicarious experiences of others, and social feedback received from people that determine the success of the performance of a search (Bronstein & Tzivian, 2013). More importantly, information-seeking behavior is motivated by the searcher's sense of self-efficacy. With the introduction of computers, particularly their use in managing and handling information, along with new formats for information sources, information retrieval tools, and information search techniques, students are now expected to be computer self-efficacious.

In today's digital age, the effective use of information and communication technologies (ICTs) is crucial for academic success and professional growth. Nigerian university students, like their counterparts around the world, rely heavily on computers and the Internet to access information, complete assignments, and engage in various academic activities. However, the extent to which students effectively utilize these resources is influenced by their demographic characteristics and their level of computer self-efficacy. While studies have explored the relationship between demographic variables, computer self-efficacy, and information-seeking behavior in various contexts, limited research has focused on Nigerian university students. Nigeria, with its diverse population and rapidly growing ICT sector, presents a unique context for investigating these relationships. Understanding the interplay between demographic variables, computer self-efficacy, and information-seeking behavior among Nigerian university students can provide valuable insights into the factors that influence their use of technology and their information-seeking practices.

Given the dearth of research in this specific context, this study aims to examine the correlation between demographic variables, computer self-efficacy, and information-seeking behavior among Nigerian university students. By investigating these relationships, we seek to contribute to the existing body of knowledge on the factors influencing students' engagement with technology and their information-seeking practices. The findings of this study will not only provide insights into the Nigerian context but also have implications for educational institutions, policymakers, and researchers interested in promoting effective technology use and information literacy among university students.

1.1. Statement of the Research Problem

Many studies examined students' information-seeking behavior from a variety of angles in an effort to comprehend how they find and use information, the sources they use, and the factors that either discourage or encourage information use, such as psychosocial factors, socioeconomic factors, availability and access to information systems, among other things. This is because information-seeking behavior is an area of dynamic interest that fascinates information scientists, communication scientists, sociologists, and psychologists, and as a result, there has been a lot of research done in that subject area (Giade, Khalid & Abdullah, 2018; Suki & Suki, 2016; El-Maamiry, 2016; El-Maamiry, 2016; El-Maamiry, 2016; Bronstein, 2014; Bronstein & Tzivian, 2013).

Many factors affect students' information-seeking behavior, it is desirable to understand the reasons for which information is needed, the environment in which they operate, their skills in identifying the needed information, the channels and sources preferred for acquiring information, and barriers to information use. Additionally, a user's information-seeking behavior is influenced by their level of education, source awareness, experience, access to libraries and other information sources, and the length of time spent seeking information. However, only a few studies have found a link between computer self-efficacy and information-seeking behavior when performing specific tasks on a computer (Adekunjo & Onuabor, 2018; Malliari, 2012). As of when this study was carried out, there is a dearth of studies on correlation analysis of the relationship between demographic variables, computer self-efficacy and information-seeking behavior among university students, specifically in Africa and by extension, in Nigeria. It is in light of this that this study was carried out to fill the empirical gap observed in the literature and general body of knowledge.

1.2. Objective of the Study

The study's broad objective is to establish a relationship between demographic variables, computer self-efficacy, and information-seeking behavior among students at the University of Ilorin. The specific objectives are to:

1. ascertain the degree of computer self-efficacy among undergraduate students at the University of Ilorin in Ilorin, Kwara State, Nigeria;

2. ascertain the information-seeking behavior of undergraduate students of the University of Ilorin, Ilorin, Kwara State, Nigeria;
3. establish the relationship between computer self-efficacy and information-seeking behavior of undergraduate students of the University of Ilorin, Ilorin, Kwara State, Nigeria;
4. establish the relationship between demographic variables (gender, course of study, and study level) and computer self-efficacy of students of the University of Ilorin, Ilorin, Kwara State; and
5. establish the relationship between demographic variables (gender, course of study, and study level) and information-seeking behavior of University of Ilorin students.

2. LITERATURE REVIEW

Information is essential and a fundamental component in the accomplishment of human endeavor. A hazy sense of something being missing led to the discovery of facts that helped clear up confusion, making it a crucial tool. Information is a source of knowledge that people use to try to improve their lives. Information need, according to Doraswamy (2017), is an understanding of an increasingly vague awareness of something missing in one's existing knowledge base. Perceived gaps in one's knowledge area are focused on the need for information. A person's desire for knowledge is influenced by three things: their reasons for looking for information, the use they will make of it once they find it, and how they will use it after they have it. The main motivation for people to seek out information is a need for it. When there is a perception that something is lacking, which prompts the search for information, an information need is created. A person with an information need has a particular gap in his or her understanding of the world, or what we would refer to as a lack of preparation to engage meaningfully in interactions with the environment. In support of this, Khan and Shafique (2011) emphasized that an individual's or group's desire to discover and receive information to meet a need is known as having an information need.

Acknowledging the reason for information search, the totality of human behavior in relation to the sourcing and channel of information and communication is what is referred to as information-seeking behavior. In corroboration of the foregoing, Tubachi (2018) associated information-seeking behavior with a purposeful search for information. A specific type of problem-solving behavior known as information-seeking entails identifying and analyzing the information needed to formulate search strategies, carry out the search, and assess the outcome. Information-seeking behavior refers to the seeker's psychological actions that include looking for, finding, obtaining, and using information. Information seekers actively and purposefully seek out current information from library resources, including online sources. In agreement with that, Padma, Ramasamy, and Sakthi (2013) emphasized that efforts geared toward identifying information needs, obtaining such information, assessing it, and deciding on what information to use to satisfy those needs are collectively referred to as information-seeking behavior. In a nutshell, information-seeking behavior primarily focuses on who needs what information, what kind of information, for what purpose, and how information is accessed, evaluated, and used.

Beliefs about one's own capacity for success affect how people feel, think, act, and motivate themselves. In many aspects, having a strong sense of efficacy improves one's ability to achieve things. High-confidence individuals regard challenging tasks as tasks to be completed rather than as threats to be averted. Such a successful viewpoint encourages intrinsic interest and total immersion in activities. Such individuals establish high standards for themselves and remain steadfastly committed to them. When they fail, they intensify and continue their attempts. They swiftly regain their sense of effectiveness after failures or losses. They attribute failure to insufficient effort or deficient knowledge and skills, which are attainable (Suki & Suki, 2016).

Technology has altered how we communicate, educate ourselves, and act in relation to the efficacy expectation of the conviction that one can carry out behavior necessary to produce desired results (Bandura, 2012). When using technology, self-efficacy is particularly important because, if a person is not confident in their own abilities, they will quickly give up all learned skills when they don't see immediate results. Computer self-efficacy was developed to integrate self-efficacy into performing general and specific tasks on or with a computer. The belief that a person has about their knowledge, skills, and capacity to use a computer to carry out specific tasks is what is referred to as computer self-efficacy. On yet another ground, computer self-efficacy is the assessment of one's capacity to utilize a computer that is not based on past performance but rather on potential future performance. It goes beyond simple component sub-skills, such as basic operating system performance, and instead incorporates judgment of the ability to apply those skills to broader and specific tasks. Additionally, it excludes basic sub-skills, such as inputting formulas in a spreadsheet or formatting drives and diskettes. Instead, it takes into account assessments of the ability to apply those skills to more complex and elaborate tasks, like searching for information, analyzing data, and writing reports. Both generic and task-specific computer self-efficacy are discussed in the literature. In contrast to broad computer

self-efficacy, which frequently encompasses a variety of computer applications, task-specific computer self-efficacy is most closely aligned with Bandura’s original definition of self-efficacy (Hatlevik, Throndsen, Loi & Gudmundsdottir, 2018; Kass, 2014).

Computer self-efficacy is an individual’s belief in their ability to use computers to perform specific tasks successfully. Students’ computer self-efficacy levels significantly influenced their engagement with technology and their overall academic performance. Computer self-efficacy is important in predicting students’ intention to use technology and their adoption of ICTs. Also, an important factor is information literacy skills in navigating and utilizing information effectively. In consonance with forgoing, Brostein (2014) pinpointed that higher levels of computer self-efficacy were positively correlated with proactive information-seeking behavior among university students. Similarly, Kanjiani (2021) demonstrated that computer self-efficacy significantly influenced individuals’ engagement in online information-seeking and their perceived usefulness of information sources.

In summary, the reviewed literature highlights the significance of information need, information-seeking behavior, and computer self-efficacy in the context of university students. It emphasizes the role of computer self-efficacy in influencing students’ engagement with technology and information-seeking behavior. Understanding these concepts can contribute to improving students’ information literacy skills and their ability to effectively access and utilize information resources. The literature emphasizes that information need arises when individuals have a perceived gap in their knowledge or understanding, prompting them to search for information to fill that gap. Information-seeking behavior is described as purposeful and problem-solving behavior, where individuals actively search for, obtain, and use information to satisfy their needs. The literature also highlights the importance of beliefs about one’s own capacity for success, referred to as self-efficacy, in influencing individuals’ behavior and motivation. Specifically, computer self-efficacy is identified as an individual’s belief in their ability to successfully use computers to perform specific tasks. The literature review indicates that higher levels of computer self-efficacy are positively associated with proactive information-seeking behavior.

2.1. Conceptual Framework

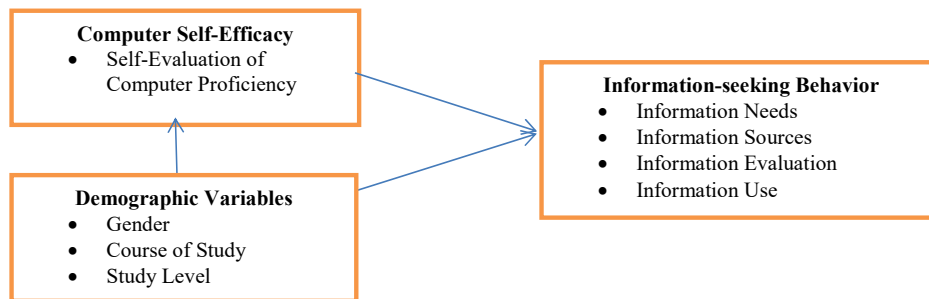


Figure 1. Conceptual model on the relationship between computer self-efficacy, demographic variables and information-seeking behavior.

Figure 1 demonstrates the relationships between the study’s variables and identifies the direction of the enquiry, which can be referred to as the study’s conceptual framework. The conceptual model compared the variables’ relationships. Using a computer to do an information search makes a wider variety of informational resources accessible on a worldwide scale. The ability to look for information online at any time and from any location means that there are no time restrictions on information searches. It provides quick access to current and relevant information. Understanding computer self-efficacy can help to shed light on students’ information-seeking behavior with regard to the usefulness of computers. Students’ confidence in their computer skills may influence their capacity to properly retrieve accurate and relevant information. Perhaps an information seeker’s conviction in their computer’s ability to find information impacts their aptitude and conduct in doing so (Suki & Suki, 2016). Student information-seeking behavior and computer self-efficacy are influenced by demographic factors. Students’ levels of computer skill and information-seeking behaviors may be influenced by characteristics like gender, academic inclination, and course level.

2.2. Empirical Studies

Suki and Suki (2016) researched how perceived self-efficacy affected information retrieval; they found that social feedback had the biggest impact, followed by one’s own self-evaluation and emotional state. Similarly, El-Maamiry

(2016) examined self-efficacy in the information-seeking behavior of students at the State University of Zanzibar, Tanzania. The results showed that mastery experience received the highest ratings, followed by vicarious experience, emotional state, and social feedback, with social feedback receiving the lowest ratings. The results of the analysis revealed considerable differences between students at various levels of study and departments, with seniors having more self-efficacy than juniors and computer science students having higher self-efficacy than students in the natural sciences. Similar to the foregoing, Berkant's (2016) study examines how faculty of education students perceive their computer self-efficacy as well as how they feel about using computers in the classroom. This study uses a correlational survey model and is descriptive in nature, and 414 students from a Turkish university's faculty of education made up the unit of analysis. The findings reveal that students who own their own personal computers have more positive computer attitudes and higher computer self-efficacy beliefs; that class level variables have no effect on students' computer attitudes and self-efficacy beliefs; and that the amount of time spent each day using a computer and prior computer experience are related to both computer attitudes and self-efficacy beliefs. Finally, the results show that male students have higher computer self-efficacy beliefs.

Bronstein (2014) investigated the self-efficacy perceptions of Israeli library and information science students with respect to their information-seeking behavior based on Bandura's four sources of self-efficacy information through an online survey of 205 students that anonymously completed the questionnaire. The research's findings demonstrate that participants reported high levels of self-efficacy and that three of the four sources of self-efficacy data were crucial in helping them establish those beliefs. No gender-based differences were found in self-efficacy perceptions or a number of socio-demographic factors. Age, educational level, and the sources with greater influence were shown to be significantly correlated. Postgraduate students claimed that their mastery experiences, affective states, and social feedback had a greater impact on them.

Okoh and Ijiekhuamhen (2014), research on Federal University of Petroleum Resources students' information-seeking behavior and the researchers employed a standardized questionnaire to gather information. The frequency, bar chart, and percentage methods were used to evaluate the data collected for ease of comprehension. The study's findings show that the majority of the respondents conduct their research using Google, print textbooks, library resources, and social media. Poor search skills, unstable electricity, and a lack of computer literacy were found to be variables impacting respondents' information-seeking behavior. Likewise, Fasola and Olabode (2013) conducted research on students' information-seeking behavior at Ajayi Crowther University in Oyo, Nigeria. Their findings showed that 66% of respondents sought information for academic purposes, 62.8 % of respondents preferred to seek and use information resources at libraries, and that (55.4%) wanted information materials relevant to their course of study. Most respondents (53.2%) said they were satisfied with the library's ability to meet their information needs.

3. Methodology

This is quantitative research, and a descriptive survey design of the correlational type was adopted. The target population of this study consisted of all the undergraduates at the University of Ilorin, and according to the report of the academic planning unit of the university, as of the 2020/2021 academic session, the total number of undergraduates in the University of Ilorin was 45,885. The sampling technique adopted for this study is multi-stage sampling, first with stratified random sampling because of the heterogeneous nature of the target population, where every strata of the population must be duly represented. On this note, the researcher grouped the targeted population alongside undergraduate students' faculties. The second stage of sampling is purposive sampling, when five faculties, namely, the Faculty of Arts, Faculty of Communication and Information Sciences, Faculty of Engineering, Faculty of Management Sciences, and Faculty of Life Sciences, are selected from fifteen faculties at the University of Ilorin, Ilorin, Nigeria. The selected faculties represent the four branches of knowledge, namely, arts and humanities, engineering and technology, social sciences, and pure sciences.

The sample size for the study is 366, which is acceptable according to Israel's (2003) sample size determination table with a precision of $\pm 7\%$ for a population size of 50000. The sample size was allocated proportionately. The researcher ensures that the five selected faculties are represented in the sample and that only those respondents that are available and willing to participate in the study constitute the sample size. The instrument for data collection was a structured and validated questionnaire. The researcher administered 30 copies of the questionnaire to the undergraduates at Kwara State University, Malete, using the split-half odd number method to determine the reliability coefficient of the instrument, and the Cronbach Alpha yielded a reliability co-efficient of 0.723 for computer self-efficacy and 0.734 for information-seeking behavior, which was considered reliable and credible for subsequent administration. The researcher personally administered the questionnaire with the help of two trained research assistants. The data collected for the

study was analyzed using descriptive and inferential statistics such as frequency counts, percentages, mean score, standard deviation, and Product Moment Correlation (PPM).

4. ANALYSIS AND INTERPRETATION

In this section, we present a comprehensive analysis of the correlation between demographic variables, computer self-efficacy, and information-seeking behavior among Nigerian university students. The study investigates the intricate interplay between these factors to shed light on the dynamics that influence students’ utilization of information resources in the digital age. The Tables below encapsulate the descriptive and statistical correlation coefficients obtained from the survey responses of a diverse sample of undergraduate students of the University of Ilorin, Ilorin, Nigeria. Each correlation coefficient represents the strength and direction of the linear relationship between specific demographic variables, levels of computer self-efficacy, and various dimensions of information-seeking behavior.

Table 1. Demographic Characteristic of the Respondents

		Frequency	Percentage
Gender:			
	Male	197	53.8
	Female	169	46.2
Total		366	100
Age:			
	16-20	141	38.5
	21-25	173	47.3
	26-30	29	7.9
	31 and Above	22	6
Total		366	100
Level of Study:			
	100 Level	56	15.3
	200 Level	92	25.1
	300 Level	106	29
	400 Level	82	22.4
	500 Level	30	8.2
Total		366	100
Faculties:			
	Arts	101	27.6
	Engineering and Technology	74	20.2
	Social Science	60	16.4
	Life Science	81	22.1
	Communication and Information Sciences	50	13.7
Total		366	100

Source: Field Survey

Table 1 presents the demographic characteristics of the respondents. In terms of the gender demographic, male participants constitute the majority group with 197 (53.8%), while their female counterparts constitute the least group with 169 (46.2%). On the age of respondents, the majority of participants fall within the age range of 21–25 with 173 (47.3%), followed by the age range of 16–20 with 141 (38.5%), and the age range of 26–30 with 29 (7.9%). While participants within the age range of 31 & above constitute the least group with 22 (6%), The majority of the respondents are in the 300 level of their academic program with 106 (29%) representatives, followed by the 200 level with 92 (25.1%), the 400 level with 82 (22.4%), and the 500 level students constitute the least group with 30 (8.2%). On the faculty of respondents, the majority of the participants are from the Faculty of Art with 101 (27.6%); followed by the Faculty of Life Science with 81 (22.1%); followed by the Faculty of Engineering & Technology with 74 (20.2%); and the Faculty of Social Sciences with 60 (16.4%). Participants from the Faculty of Communication and Information Sciences constitute the least group with 50 (13.7%).

The majority of the participants in the study are male, comprising 53.8% of the respondents. This suggests that the findings and conclusions drawn from the study may reflect the experiences and perspectives of male students more prominently. The distribution of participants across different age ranges indicates that the majority of respondents fall within the age range of 21-25, followed by the age range of 16-20. These findings imply that the study primarily focuses on young adults and may not capture the experiences of older students. The distribution of participants across different academic levels and programs provides insights into the sample composition. The study primarily includes undergraduate students in their intermediate years of study. The representation of different academic levels and programs allows for analysis of potential variations in computer self-efficacy and information-seeking behavior based on these factors.

Research Question 1: What is the level of computer self-efficacy of undergraduate students in the University of Ilorin, Ilorin, Kwara State, Nigeria?

Table 2. Level of Computer Self-Efficacy

S/N	Statements	Strongly Agree	Agree	Disagree	Strongly Disagree	Mean	Std. D
1	I am self-reliant when it comes to searching for information using a computer.	155 (42.1%)	156 (42.6%)	34 (9.3%)	21 (5.7%)	3.22	.84
2	I can easily find the information I need with my computer skills	135 (36.9%)	173 (47.3%)	21 (5.7%)	37 (10.1%)	3.11	.72
3	I'm confident of the quality of information in my search results	164 (44.8%)	188 (51.4%)	5 (1.4%)	9 (2.5%)	3.38	.47
4	A computer enables me to search for information better.	155 (42.3%)	139 (38%)	51 (13.9%)	21 (5.7%)	3.17	.87
5	I have mastered Boolean operators, truncation, and wildcat which assisted me in information searching on the internet.	128 (35%)	181 (49.5%)	49 (13.4%)	8 (2.2%)	3.17	.74
6	I'm sure I can select the right keyword for each and every information search.	133 (36.3%)	186 (50.8%)	40 (10.9%)	7 (1.9%)	3.21	.71
7	I know how to access information databases and information repositories.	101 (27.6%)	204 (55.7%)	56 (15.3%)	5 (1.4%)	3.09	.69
8	Sorting out information from search results can be problematic for me sometimes	90 (24.6%)	201 (54.9%)	65 (17.8%)	10 (2.7%)	3.02	.68
9	I am very versed in using a computer as a tool	78 (21.3%)	204 (55.7%)	74 (20.2%)	10 (2.7%)	2.96	.54
Average Mean Score						3.15	

Source: Field survey

Note: The coefficients for each response category represent the numerical values assigned to those categories. Strongly Agree: Assigned a coefficient of 4, Agree: Assigned a coefficient of 3, Disagree: Assigned a coefficient of 2, and strongly Disagree: Assigned a coefficient of 1. These coefficients represent the levels of agreement or disagreement associated with each response category.

The statistical method used is descriptive statistics, specifically mean scores, to analyze and discuss participants' level of computer self-efficacy. The statistical method used is descriptive analysis. Descriptive statistics involves summarizing and presenting data to provide insights into the central tendency, variability, and distribution of a dataset. In this case, the mean scores of different variables are being discussed to understand the participants' perceptions of their computer self-efficacy.

Table 2 presents the participants' responses on their level of computer self-efficacy. From the nine variables used to measure users' level, only five variables scored a mean above the average mean score (AMS=3.15). I'm confident of the quality of information in my search results has the highest mean score (X=3.38); followed by I am self-reliance when it comes to searching information using a computer with a mean score (X=3.22); followed by I'm sure I can select the right keyword for each and every information search with a third ranking of a mean score (X=3.21). I have mastered Boolean operators, truncation and, wildcat which assisted me in information searches on the internet and Computer enables me to search information better ranked fourth and fifth with a mean score (X=3.17) and (X=3.17) respectively. I am very versed in using a computer as a tool ranked the least in ninth position with a mean score (X=2.96); followed by Sorting out information from search results can be problematic for me sometimes in eighth position with a mean score (X=3.02). I know how to access information databases and information repositories and I can easily find the information I need with my computer skills ranked seventh and sixth respectively with a mean score (X=3.09) and (X=3.11). The implications highlight the varying levels of computer self-efficacy among participants and identify specific areas of strength and areas that may require improvement. These insights can help inform interventions or training programs to enhance participants' computer skills and self-efficacy in information retrieval and processing.

Research Question 2: What is the information-seeking behavior of undergraduate students in the University of Ilorin, Ilorin, Kwara State, Nigeria?

The statistical method used is descriptive statistics, specifically mean scores, to analyze and discuss participants' level of information-seeking behavior. The statistical method used is descriptive analysis, similar to the previous interpretation.

Table 3 presents the participants' responses to their level of information-seeking behavior. From the nine variables used to measure users' level, only five variables scored a mean above the average mean score (AMS=3.27). Participants had the highest seeking behavioral level on willingness to spend a maximum amount of time in seeking information with a mean score (X=3.42); followed by extracting information only from relevant and useful sources, which ranked second

Table 3. Level of Information-seeking Behavior

S/N	Statements	Strongly Agree	Agree	Disagree	Strongly Disagree	Mean	Std. D
1	I sought information that could assist me in my course of study and academic program.	124 (33.9%)	192 (52.5%)	46 (12.6%)	4 (1.1%)	3.20	.68
2	I consult the library when I'm seeking information.	118 (32.2%)	189 (51.6%)	55 (15%)	4 (1.1%)	3.15	.70
3	The Internet is my companion and first port of call when in need of information.	153 (41.8)	167 (45.6)	39 (10.7%)	7 (1.9%)	3.28	.74
4	In acquiring information, I also make use of informal sources (media, friends, and family)	150 (41%)	179 (48.9%)	34 (9.3%)	3 (0.8%)	3.31	.67
5	I'm willing to spend the maximum amount of time seeking information.	168 (45.9)	186 (50.8)	10 (2.7)	2 (0.5)	3.42	.58
6	I'm willing to pay for relevant information.	118 (32.2)	209 (57.1)	34 (9.3)	5 (1.4)	3.20	.67
7	I support the open access initiative because of free access to full-text documents.	132 (36.1)	211 (57.7)	23 (6.3)	-	3.30	.59
8	I always checked for currency and relevance of the information sources I'm using	109 (29.8)	223 (60.9)	31 (8.5)	3 (0.8)	3.20	.65
9	I always extract information only from sources I found useful and relevant.	149 (40.7)	196 (53.6)	19 (5.2)	2 (0.5)	3.34	.61
Average Mean Score						3.27	

Source: Field survey

with a mean score (X=3.34) and making use of informal sources such as family and friends with a mean score (X=3.31), which ranked third. I support open access initiatives because of free access to full-text documents and The Internet is my companion and first port of call when in need of information ranked fourth and fifth with a mean score (X=3.30) and (X=3.28) respectively. I consult the library when I'm seeking information ranked the least at ninth position with a mean score (X=3.15); followed by I always checked for currency and relevance of the information sources I'm using, which ranked eighth with a mean score (X=3.20). I'm willing to pay for relevant information and I sought information that could assist me in my course of study and academic program ranked seventh and sixth with a mean score (X=3.203) and (X=3.197) respectively. The implications reveal the diversity in participants' information-seeking behaviors, including their time investment, source selection, support for open access, reliance on the Internet, and limited use of the library. Understanding these behaviors can inform the design of information resources and services that align with participants' preferences and facilitate effective information-seeking practices.

Research Question 3: Is there a relationship between computer self-efficacy and information-seeking behavior of students of the University of Ilorin, Ilorin, Kwara State, Nigeria?

Table 4. Relationship between Computer Self-Efficacy and Information-seeking Behavior

Variables	1	2	\bar{x}	SD.
1 Computer Self-Efficacy	1	.459**	16.66	3.09
2 Information-seeking Behavior	.459**	1	15.62	3.20

** . Correlation is significant at the 0.01 level (1-tailed)

*. Correlation is significant at the 0.05 level (2-tailed)

The test of hypothesis used is the Pearson correlation coefficient (r) to assess the relationship between Computer Self-Efficacy and Information-seeking Behavior. The correlation coefficient is represented by the value of .459**, which indicates a positive and significant correlation between the two variables. The correlation coefficient ranges from -1 to 1, where values closer to 1 represent a strong positive correlation, values closer to -1 represent a strong negative correlation, and values close to 0 represent no correlation.

Table 4 above, posits there is a strong significant positive relationship between computer self-efficacy and information-seeking behavior among undergraduate students at the University of Ilorin (r =.459, p<.01).

Research Question 4: Is there a relationship between demographic variables (gender, age, level of study, and faculty) and computer self-efficacy of students of the University of Ilorin, Ilorin, Kwara State, Nigeria?

Table 5. Relationship between Computer Self-Efficacy and Information-seeking Behavior

Variables	1	2	3	4	5	\bar{x}	SD.
1 Gender	-					1.46	.50
2 Age	-.078	-				1.81	.82
3 Level of Study	-.025	.302**	-			2.83	1.18
4 Faculty	.170**	-.120*		-		2.74	1.42
5 Computer Self-Efficacy	.044	-.081	-.083	.067	-	16.66	3.09

** . Correlation is significant at the 0.01 level (1-tailed)

* . Correlation is significant at the 0.05 level (2-tailed)

The test of hypothesis used in Table 5 is the Pearson correlation coefficient (r) to examine the relationship between Demographic Variables (Gender, Age, Level of Study, and Faculty) and Computer Self-Efficacy.

Table 5 shows that there is no significant relationship between demographic variables and computer self-efficacy [gender ($r = .044$, $p < .05$); age ($r = -.081$, $p < .05$); level of study ($r = -.83$, $p < .05$) and faculty ($r = .067$, $p < .05$)]. However, there is a strong positive significant relationship between age and level of study ($r = .302$, $p < .01$), gender and faculty ($r = .170$, $p < .05$), and age and faculty ($r = -.120$, $p < .05$).

Research Question 5: Is there a relationship between demographic variables (gender, age, level of study, and faculty) and information-seeking behavior of students of the University of Ilorin, Ilorin, Kwara State, Nigeria?

Table 6. Relationship between Demographic Variables and Information-seeking Behavior

Variables	1	2	3	4	5	\bar{x}	SD.
1 Gender	-					1.46	.50
2 Age	-.078	-				1.81	.82
3 Level of Study	-.025	.302**	-			2.83	1.18
4 Faculty	.170**	-.120*		-		2.74	1.42
5 Information-seeking Behavior	-.016	-.072	-.034	.091	-	15.62	3.20

** . Correlation is significant at the 0.01 level (1-tailed)

* . Correlation is significant at the 0.05 level (2-tailed)

The test of hypothesis used in Table 6 is the Pearson correlation coefficient (r) to examine the relationship between Demographic Variables (Gender, Age, Level of Study, and Faculty) and Information-seeking Behavior.

Table 6 shows that there is no significant relationship between demographic variables and information-seeking behavior [gender ($r = -.016$, $p < .05$); age ($r = -.072$, $p < .05$); level of study ($r = -.034$, $p < .05$) and faculty ($r = .091$, $p < .05$)].

5. DISCUSSION

Out of the nine variables used to measure users' level of computer self-efficacy, only five variables scored a mean above the average mean score (AMS) of 3.15. This indicates that participants' perceptions and confidence levels vary across different aspects of computer self-efficacy. The variable "I'm confident of the quality of information of my search results" received the highest mean score ($X=3.38$). This suggests that participants generally feel confident about the accuracy and reliability of the search results they obtain through computer-based information searches. The variable "I am self-reliant when it comes to searching information using a computer" ranked second with a mean score ($X=3.22$). This implies that participants feel capable and independent in conducting information searches using computer technology. The variable "I'm sure I can select the right keyword for each and every information search" ranked third with a mean score ($X=3.21$). This indicates that participants express confidence in their ability to choose appropriate keywords to retrieve relevant information during their searches. Participants reported having familiarity with Boolean operators, truncation, and wildcards, which assist them in information searches on the internet. These variables ranked fourth and fifth with mean scores of ($X=3.17$) and ($X=3.17$) respectively. This suggests that participants have some level of knowledge and expertise in utilizing these search techniques. The variable "I am very versed in using the computer as a tool" ranked the lowest in ninth position with a mean score ($X=2.96$). This implies that participants may perceive themselves to have a relatively lower level of expertise or proficiency in using a computer. The variable "Sorting out information from search results can be problematic for me sometimes" ranked eighth with a mean score ($X=3.02$). This indicates that participants occasionally face difficulties in organizing and extracting relevant information from search results. Participants expressed moderate confidence in their ability to access information databases and repositories, as reflected in the seventh and sixth rankings with mean scores of ($X=3.09$) and ($X=3.11$) respectively.

This suggests that they possess a certain level of competence in finding the information they need using their computer skills.

On students' information-seeking behavior, from the nine variables used to measure users' level of information-seeking behavior, only five variables scored a mean above the average mean score (AMS) of 3.27. This indicates that participants' behaviors and preferences in information-seeking vary across different aspects. The variable "willingness to spend the maximum amount of time in seeking information" received the highest mean score ($X=3.42$). This suggests that participants demonstrate a high level of commitment and are willing to invest considerable time in their information-seeking activities. The variable "extracting information only from relevant and useful sources" ranked second with a mean score ($X=3.34$). This implies that participants prioritize obtaining information from sources that are deemed relevant and trustworthy, indicating a discerning approach to information selection. Participants reported making use of informal sources such as family and friends for obtaining information. This variable ranked third with a mean score ($X=3.31$), indicating that participants perceive these sources as valuable and accessible in their information-seeking process. Participants expressed support for the open access initiative due to the availability of free access to full-text documents. This variable ranked fourth with a mean score ($X=3.30$), suggesting a positive attitude towards the accessibility and availability of information. The variable "The Internet is my companion and first port of call when in need of information" ranked fifth with a mean score ($X=3.28$). This implies that participants heavily rely on the Internet as their primary source of information, emphasizing its convenience and accessibility. The variable "I consult the library when I'm seeking information" ranked the lowest in ninth position with a mean score ($X=3.15$). This suggests that participants may not frequently utilize traditional library resources in their information-seeking behavior, opting for other sources instead. Participants reported slightly lower attention to checking the currency and relevance of the information sources they use. This variable ranked eighth with a mean score ($X=3.20$), indicating a potential area for improvement in critically evaluating the information's timeliness and relevance. Participants expressed a moderate willingness to pay for relevant information. This variable ranked seventh with a mean score ($X=3.203$), suggesting that while participants recognize the value of information, their inclination to pay for it may not be particularly high. Participants indicated a moderate level of seeking information that could assist them in their course of study and academic program. This variable ranked sixth with a mean score ($X=3.197$), highlighting the importance placed on acquiring information relevant to their educational pursuits.

A strong and significant positive relationship was found between computer self-efficacy and information-seeking behavior. This implies that as the level of computer self-efficacy increases, the level of information-seeking behavior will also increase and vice versa. This result aligned with the study of Okoh and Ijiekhuamhen (2014) which identifies computer proficiency as part of the factors that influence students' information-seeking behavior.

No significant relationship was found between participants' demographic variables (gender, age, level of study, and faculty) and computer self-efficacy; or demographic variables (gender, age, level of study, and faculty) and information-seeking behavior. The findings of the study are partially consonant with the findings of a study carried out by Bronsten (2014) who reported that correlations between self-efficacy percepts and several socio-demographic variables revealed no gender-based differences. Also, Berkant (2016) study reported that class-level variables have no effect on students' computer attitudes and self-efficacy beliefs, and that the amount of time spent each day using a computer and prior computer experience are related to both computer attitudes and self-efficacy beliefs. Meanwhile, the study is in contrast to El-Maamiry's (2016) study which reported considerable differences between students at various study levels and departments, with seniors having more self-efficacy than juniors and computer science students having higher self-efficacy than students in the natural sciences. The reason for this disparity may be the mandatory tablet given to all undergraduate students of the University of Ilorin, Ilorin, Nigeria, regardless of their course of study, combined with the availability of internet connection on the university campus, which means students are connected to the internet and well acquainted with information searching online.

6. CONCLUSION

The study investigated the interrelationship between demographic variables, computer self-efficacy, and information-seeking behavior of undergraduates at the University of Ilorin, Ilorin, Nigeria with the premise that, with the belief in one's knowledge of computer operation and application, undergraduate students would record success with their information search, information retrieval, information evaluation, and information utilization, perhaps with some degree of influence from demographic and academic factors. Computer self-efficacy was found to inform undergraduate students' decisions as to what information-seeking behavior to undertake when in need of information. But demographic variables, on the other hand, were found to be unrelated to what inspires computer self-efficacy of students and even information-seeking behavior of undergraduates at the University of Ilorin, Ilorin, Nigeria. Academically, students'

information-seeking behavior entails deliberate information-seeking since they must finish course assignments, get ready for lectures, seminars, and workshops, and write research projects for their final year. The impact of their computer proficiency on their information-seeking behavior, therefore, is crucial to their academic achievement.

7. RECOMMENDATIONS

The succeeding recommendations are given based on the report of findings below:

1. Despite participants' overall confidence in the quality of search results, there is a need to further develop their skills in critically evaluating the accuracy and reliability of the information they retrieve. Training programs or workshops focused on information evaluation techniques can be provided to improve participants' ability to assess the credibility of search results.
2. Participants demonstrated moderate confidence in accessing information databases and repositories. It is important for the university library to ensure easy access to relevant and reliable information sources, including online databases, academic journals, and other reputable resources. Providing access to such resources can empower participants to make the most of their computer skills and enhance their information-seeking capabilities.
3. Sorting out information from search results is still problematic for some students, hence, information literacy skills needed to be prioritized for students.
4. Creating opportunities for participants to engage in collaborative learning and knowledge sharing can be beneficial. This can involve setting up discussion forums, online communities, or workshops where participants can share their experiences, challenges, and strategies for effective information searching. Peer learning can help participants learn from each other and gain valuable insights and techniques for improving their computer self-efficacy.
5. Participants expressed a preference for extracting information only from relevant and useful sources. To further enhance their information literacy, it is recommended to provide training on critical evaluation skills, enabling students to assess the credibility, reliability, and relevance of different information sources. This can empower them to make informed decisions about the sources they rely on for their academic work.
6. Although participants reported lower utilization of library resources, efforts should be made to increase awareness and promote the value of libraries as valuable information hubs. This can involve collaborating with librarians to develop engaging library orientations, workshops, and targeted outreach programs that highlight the benefits and resources available in the library.
7. Since computer self-efficacy was found to be positively related to information-seeking behavior, it is crucial to provide educational interventions that enhance students' computer proficiency. This can involve offering computer literacy courses, workshops, or online tutorials that cover basic computer skills, internet navigation, and effective use of information search tools. By improving computer proficiency, students can feel more confident and capable in utilizing technology for information-seeking purposes.

Peer Review: Externally peer-reviewed.

Author Contributions: Conception/Design of Study- T.T.O., M.D.L., Z.O.A., O.F.A.; Data Acquisition- T.T.O., M.D.L., Z.O.A., O.F.A.; Data Analysis/Interpretation- T.T.O., M.D.L., Z.O.A., O.F.A.; Drafting Manuscript- T.T.O., M.D.L., Z.O.A., O.F.A.; Critical Revision of Manuscript- T.T.O., M.D.L., Z.O.A., O.F.A.; Final Approval and Accountability- T.T.O., M.D.L., Z.O.A., O.F.A.; Material and Technical Support- T.T.O., M.D.L., Z.O.A., O.F.A.; Supervision- T.T.O., M.D.L., Z.O.A., O.F.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.

ORCID IDs of the authors

Tunde Toyese Oyedokun 0000-0001-5306-038X
Medinant Dolapo Laaro 0000-0003-2717-1523
Zainab Olanihun Ambali 0000-0001-9070-7592
Olabisi Fadeke Adesina 0000-0002-8000-5623

REFERENCES

- Abdullah, Z.D. & Mustafa, K.I. (2019). The underlying factors of computer self-efficacy and the relationship with students' academic achievement. *International Journal of Research in Education and Science*, 5(1), 346-354. <http://files.eric.ed.gov/fulltext/EJ1199492.pdf>
- Adekunjo, O.A. & Unuabor, S.O. (2018). Computer self-efficacy and attitude toward internet use among library and information science postgraduate students in two library schools in Nigeria. *American Journal of Educational Research and Review*, 3(24), 1-13. <https://escipub.org/index.php/AJERR/article/download/344/330>
- Bronstein, J. & Tzivian, L. (2013). Perceived self-efficacy of library and information science professionals regarding their retrieval skills. *Library and Information Science Research*, 36(2), 151-158.
- Bronstein, J. (2014). The role of perceived self-efficacy in the information seeking behavior of library and information science students. *The Journal of Academic Librarianship*, 40 (2), 101-106. <https://iranarze.ir/wp-content/uploads/2016/08/4789-English.pdf>
- Chen, I. (2017). Computer self-efficacy, learning performance, and the mediating role of learning engagement. *Computers in Human Behavior*, 72, 362-370. <https://doi.org/10.1016/j.chb.2017.02.059>
- Doraswamy, N.A. (2017). Information needs and information seeking behavior of Faculty of JNTUH affiliated engineering colleges with reference to special reference to Telangana State. In J. Sharma (Ed.), *Library technologies, service and resources current global trends* (p.318). New Delhi: Excel India Publisher.
- El-Maamiry, A.A. (2016). Self-efficacy in the information seeking behavior of State University of Zanzibar students: A case study. *International Journal of Information Dissemination and Technology*, 6(2), 99-102. <https://www.ijidt.com/index.php/ijidt/article/view/6.2.6/389>
- Fasola, O. S. & Olabode, S. O. (2013). Information seeking behavior of student of Ajayi Crowther University, Oyo, Nigeria. *Brazilian Journal of Information Science*, 7(2), 47-60. <https://brapci.inf.br/index.php/res/download/48708>
- Giade, M.Y., Khalid, Y.I.A. & Abdullah, N. (2018). Determining factors of perceived self-efficacy in information seeking practices through Facebook. *Malaysian Journal of Library and Information Science*, 23(3), 35-47. Retrieved from: <http://dx.doi.org/10.22452/mjlis.vol23no3.3>
- Hatlevik, O.E., Thronsen, I., Loi, M. & Gudmundsdottir, G.B. (2018). Students' ICT self-efficacy and computer and information literacy: Determinants and relationships. *Computers & Education*, 118, 107-119. <https://doi.org/10.1016/j.compedu.2017.11.011>
- Israel, G. D. (2003). *Determining sample size*. Retrieved on February 18th, 2018 from: www.sut.ac.th/im/data/read6.pdf.
- Kanjiani, H. H. (2021). Relationship between information seeking skills and research self-efficacy of postgraduate students. *International Journal of Social Sciences: Current and Future Research Trends*, 10(01), 36-48. https://ijsscfrjournal.isrra.org/index.php/Social_Science_Journal/article/view/857
- Kass, K.D. (2014). Computer self-efficacy: Instructor and student perspective in a university setting. A dissertation submitted to the graduate faculty, Iowa State University, Ames, Iowa. <https://dr.lib.iastate.edu/bitstreams/90b88d7c-f184-40b2-b2d5-eb49fbd794f7/download>
- Khan, S. A. & Shafique, F. (2011). Information need and information seeking behavior: A survey of college faculty at Bahawalpur. *Library Philosophy and Practice (e-journal)*. Retrieved on February 18th, 2018 from: www.webpages.uidaho.edu/mmbolin/khan-shaque.htm
- Malliari, A. (2012). IT self-efficacy and computer competencies of library and information science students. *The Electronic Library*, 30(2), 608-622. <http://dx.doi.org/10.1108/02640471211275675>
- Padma, P., Ramasamy, K. & Sakthi, R. (2013). Information need and information seeking behavior of postgraduate student of school of economics at Madurai Kamara University: A user survey. *International Journal of Education Research Technology*, 4, 33-42.
- Suki, N.M. & Suki, N.M. (2016). Library patrons' emotions after information retrieval: Effects of perceived self-efficacy. *Program Electronic Library and Information Systems*, 50(3), 288-302. <http://dx.doi.org/10.1108/PROG-07-2014-0045>
- Tubachi, P. (2018). *Information seeking behavior: An overview*. https://www.researchgate.net/publication/330521546_INFORMATION_SEEKING_BEHAVIOR_AN_OVERVIEW

How cite this article

Oyedokun, T.T., Dolapo Medinat, L., Ambali, Z.O., Adesina, O.F. (2023). Correlation analysis of the relationship between demographic variables, computer self-efficacy, and information-seeking behavior of Nigerian university students. *Acta Infologica*, 7(2), 267-280. <https://doi.org/10.26650/acin.1141249>

A Deep Learning-Based Classification Study for Diagnosing Corneal Ulcers on Ocular Staining Images

Oküler Boyama Görüntülerinde Kornea Ülserinin Teşhisi İçin Derin Öğrenmeye Dayalı Bir Sınıflandırma Çalışması

Onur Sevli¹ 

¹(Assoc. Prof. Dr.) Burdur Mehmet Akif Ersoy University, Computer Engineering Department, Burdur, Türkiye

Corresponding author : Onur SEVLİ
E-mail : onursevli@mehmetakif.edu.tr

ABSTRACT

Corneal ulcer is a common disease worldwide and is one of the leading causes of corneal blindness. Diagnosis of the disease requires expertise, and the number of experienced ophthalmologists is not sufficient, especially in underdeveloped countries. For this reason, it is necessary to develop technology-based decision support systems in the diagnosis of the disease. However, the number of studies on this subject is not sufficient. In this study, CNN-based classifications were performed using corneal ulcer images obtained by an ocular staining technique, consisting of 712 samples and three classes. In addition to the AlexNet and VGG16 state-of-the-art architectures, which are widely used in the literature, a CNN model proposed for this study was used for classification. In the classifications performed by applying data augmentation, 95.34% accuracy with AlexNet, 98.14% with VGG16, and 100% accuracy with the proposed model was obtained. The findings were compared with similar studies in the literature. It was concluded that the accuracy rates obtained with all of the models used in the study were generally higher than similar studies in the literature, and the accuracy obtained with the proposed CNN model was higher than all of the peers. In addition, the success of the proposed model compared to other models with more complex structures revealed that it is not always necessary to use complex architectures for high accuracy.

Keywords: Corneal ulcer diagnosis, convolutional neural network, classification

ÖZ

Kornea ülseri dünya genelinde yaygın görülen bir hastalık olup kornea körlüğünün önce gelen nedenlerindedir. Hastalığın teşhisi uzmanlık gerektirmekte olup, özellikle az gelişmiş ülkelerde tecrübeli oftalmolog sayısı yeterli sayıda değildir. Bu durum hastalığın teşhisinde etkin ve uzmanlara destek sistemlerin oluşturulmasını gerekli kılmaktadır. Ancak henüz bu konuda yapılmış olan çalışmaların sayısı yeterli düzeyde değildir. Bu çalışmada 712 adet ve 3 türden oluşan, oküler boyama tekniği ile elde edilen kornea ülser görüntüsü kullanılarak CNN tabanlı sınıflandırmalar gerçekleştirilmiştir. Literatürde yaygın kullanılan AlexNet ve VGG16 daha derin state-of-art mimarileri yanında bu çalışma için önerilen bir CNN modeli kullanılmıştır. Veri artırımı uygulanarak gerçekleştirilen sınıflandırmalarda AlexNet ile 95.34%, VGG16 ile 98.14%, ve önerilen model ile 100% doğruluk elde edilmiştir. Elde edilen bulgular literatürdeki benzer çalışmalarda karşılaştırılmıştır. Tüm modeller ile elde edilen doğruluk oranlarının literatürdeki çalışmaların genelinden yüksek olduğu, önerilen CNN modeli ile elde edilen doğruluğun ise emsallerin tamamından yüksek olduğu sonucuna ulaşılmıştır. Ayrıca önerilen modelin daha karmaşık yapıdaki diğer modellere nazaran da yüksek başarı sergilemiş olması, daha minimal mimarilerle de yüksek başarı elde edilebileceğini ortaya koymuştur.

Anahtar Kelimeler: Kornea ülseri teşhisi, evrimsel sinir ağı, sınıflandırma

Submitted : 10.09.2022
Revision Requested : 06.10.2022
Last Revision Received : 04.04.2023
Accepted : 12.06.2023
Published Online : 14.08.2023



This article is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0)

1. INTRODUCTION

The cornea is the hard and transparent layer located in front of the iris, belonging to the dioptric system of the eye. The fibers of the cornea, which has a fibrous structure consisting of collagen, are located in the stroma layer, which forms a large part of its thickness. It is the first layer of the eye that refracts incoming light and accounts for about two-thirds of the total refractive power of the eye. In addition, thanks to its hard structure, it protects other parts of the eye (Maurice, 1957). A corneal ulcer is a condition that occurs as a result of deterioration of the epithelial layer or corneal stroma of the cornea due to inflammatory or infective causes (Chen & Yuan, 2010). Ocular surface diseases, damage caused by corneal surgery or contact lens use, adnexal diseases, and other traumas are among the risk factors for the formation of corneal ulcers (Amescua et al., 2012). Damage to corneal tissues due to viral, bacterial, or fungal sources causes corneal ulcers. Studies reported that viral cases in the formation of corneal ulcers were more common in developed countries, and bacterial and fungal cases were more common in developing countries (Garg & Rao, 1999).

Corneal ulcer is a common eye problem worldwide and is the second leading cause of ocular morbidity (Song et al., 2014). Corneal ulcers can seriously damage eye health, causing cornea scars, perforation, endophthalmitis, and visual trouble. Corneal ulcer is among the leading causes of corneal blindness (Katara et al., 2013). Failure to diagnose the disease in a timely and correct manner and to apply the correct treatment on time may cause irreversible damage to the eye (Diamond et al., 1999; Cohen et al., 1987).

Corneal ulcer is one of the important problems threatening eye health, especially in developing countries, and the annual average corneal ulcer cases in these countries reaches 1.5 million (Basak et al., 2005). Diagnosis of a corneal ulcer is critical and is performed by experienced professionals. However, the number of experienced ophthalmologists around the world, especially in geographical regions with limited resources, is not sufficient and this makes the early diagnosis of the disease difficult. While early diagnosis increases the success of treatment, correct analysis of the morphological structure resulting from the disease is effective in determining the correct treatment procedures. An accurate distinction must be made between different ulcer stages and types to reduce the risk of permanent vision damage or blindness.

The ocular staining technique is used in the diagnosis of corneal ulcers as well as in the diagnosis of various eye diseases. In this technique, topical dyes are widely used to characterize ocular surface diseases and to quantify their severity (Bron et al., 2015). Quantitative analysis of corneal disorders is made more easily by examining colored eye surfaces under a slit lamp microscope. Although the manual diagnosis of a corneal ulcer is reliable, it requires high sensitivity, takes time, and the results obtained may vary in terms of the reviewers. In this case, the right treatment decision may not be made, or the treatment process could be delayed. Delayed or incorrect/incomplete treatment causes progression of the disease and the formation of irreversible defects. For this reason, it becomes necessary to develop intelligent support systems that will help experts make decisions effectively, quickly and with high accuracy.

With the development of technology, artificial intelligence techniques are widely used in the medical field as well as in many other fields. Machine learning, a sub-discipline of artificial intelligence, provides stable predictions about new situations by learning from existing data. Deep learning, which is a machine learning technique, can successfully reveal the complex hierarchy in the nature of data with the help of deep neural networks. The Convolutional Neural Network (CNN) is a deep learning method used especially in computer vision. In the literature, there are studies using different machine learning techniques and CNN for the diagnosis of corneal ulcers. However, the number of these studies using artificial intelligence for corneal ulcer diagnosis is still limited.

Noting the number of corneal ulcer cases in developing countries, Saini et al. (2003) collected a total of 106 corneal ulcer images from patients living in India for their study. The study achieved 90.7% accuracy in the classification study with artificial neural networks (ANNs) for corneal ulcer diagnosis, using the dataset consisting of the images collected. Akram and Debnath (2019), captured images of faces with a digital camera and then segmented the eye region on these images. A study was carried out to detect the presence of corneal ulcers using the fragmentary images. By using data augmentation on a total of 513 images, a binary classification was performed as a corneal or non-corneal ulcer with the proposed CNN model. The average accuracy value obtained for the two classes as a result of 40 epochs is 98.99%. Kim et al. (2019), proposed a CNN-based diagnostic model to determine the degree of corneal ulceration in dogs. They performed classifications with three different degrees normal, superficial, and deep on a total of 1,040 images collected at Korea Konkuk University Veterinary Medical Teaching Hospital. A 92% accuracy was achieved with the ResNet50 model with their classifications using different transfer learning models.

In the literature, the SUSTech-SYSU dataset is widely used in addition to the study-specific datasets on the detection of corneal ulcers. In this dataset, there are 712 eye images obtained using the ocular staining technique. These images belong to three different types of corneal ulcers: flaky corneal ulcers (FCU), point-like corneal ulcers (PCU), and

point-flaky mixed corneal ulcers (PFCU). Different applications were made for segmentation and classification with this dataset, which was used frequently in recent studies.

Segmentation is defined as determining the boundaries of the target region on the image. Corneal ulcer segmentation on ocular staining images is important for the quantitative assessment of ocular surface defects. To realize this critical and challenging task, Wang et al. (2021) performed a segmentation study based on the Adjacent Scale Fusion method. In this study, which was carried out on the SUSTech-SYSU dataset, the Dice Coefficient value of 80.73% was reached. Portela et al. (2021) performed a segmentation study for corneal ulcer detection with a dataset of ocular staining images specific to their study. Using U-NET and DexiNet architectures, they obtained an average of 70.50% Dice Coefficient in the study with a total of 449 FCU type disease images. The PFCU and FCU type disease images are more difficult to distinguish and this results in reduced diagnostic success. Wang et al. (2021) proposed a segmentation network to distinguish FCU and PFCU type images with higher success using the SUSTech-SYSU dataset. They reached a Dice Coefficient of 89.14% with this network called CU-SegNet, which was based on the encoder-decoder structure. Diagnosis of corneal ulcers becomes more difficult due to large differences in shape, blurred borders, and noise interference. Addressing this problem, Wang et al. (2021) performed a segmentation study on the SUSTech-SYSU dataset with a semi-supervised GAN using the Semi-MSST-GAN. A Dice Coefficient of 90.93% was reached with this model, which was then compared with different techniques.

In the literature, besides the segmentation studies on the SUSTech-SYSU dataset, there are classification studies performed with different techniques. Tang et al. (2020), performed a classification on this dataset using a modified VGG network. Eighty-eight . eighty-nine percent accuracy, 92.27% precision, and 71.93 recall values were the results obtained from the classification of images consisting of three different classes; the FCU, PCU, and PFCU. Gross et al. (2021) proposed a specific CNN model for the classification of the same dataset. The highest accuracy value reached with a proposed model was 92.73%. Teeyapan (2021) performed classifications using the SUSTech-SYSU dataset with the transfer learning method. In the study where different architectures were tested, the ResNet50 model provided the highest result with 95.10% accuracy.

Li et al. (2021), suggested a deep learning-based method for early and accurate diagnosis, noting that corneal ulceration is one of the major causes of corneal blindness worldwide. Classifications were performed with the DenseNet121, InceptionV3 and ResNet50 models on a dataset of 6,567 samples, consisting of corneal images with normal cornea, ulcerated cornea and other abnormalities. The highest success was obtained with the DenseNet121 as 96% Cohen's kappa coefficient.

Diagnosis of corneal ulcers can be challenging for specialists. Sajeev and Prem Senthil (2021) proposed a CNN-based method for classifying corneal ulcers of bacterial and viral origin. They classified the dataset consisting of a total of 446 corneal ulcer images belonging to these two classes, with different input sizes and CNN architectures with two or three convolution layers. The highest accuracy obtained was 81.2% with the model with 64x64 input size and three convolution layers.

Xu et al. (2021), stated that corneal ulcer is an emergency that needs to be treated quickly, so a study was completed with a classification study using the deep learning on images with ulcers. Classifications were done on 115,408 microscopic images collected from 10,609 patients, using the VGG16, GoogLeNet and DenseNet models, at the image-level and patch-level. With DenseNet, the most successful model, they achieved an accuracy of 61.04% at image-level and 66.30% at patch-level. The results of the study were compared with the diagnoses of ophthalmologists and it revealed that the method they proposed gave more successful results.

This study performed a CNN-based application for the successful diagnosis of corneal ulcers using ocular staining images. The SUSTech-SYSU dataset, which is widely used in literature, was the preferred method of work. Classification studies were performed on three different types of images using two known state-of-the-art architectures, as well as a less complex proposed CNN model for this study. The main motivation of this study is to achieve a higher success compared to similar studies in literature and to present a more effective solution. In addition, the sub-objective is to demonstrate that a less complex model proposed for this study outperforms the more complex models. The following sections contain the dataset information, the classification methods used, findings and discussion.

2. MATERIAL AND METHOD

2.1. Dataset

The dataset used in the study included eye images obtained by the ocular staining technique, created by Deng et al. (2020) for the detection of corneal ulcers. The images in the dataset were obtained from patients with various types and grades of corneal ulcers at the Zhongshan Ophthalmic Center of Sun Yat-sen University, China. No distinction

was made regarding external conditions such as age and gender of the patients whose eye images were taken. In this data set, which included a total of 712 images, there were data samples belonging to three different classes of corneal ulcers. These classes, in which the data samples belong, are flaky corneal ulcers (FCU), point-flaky mixed corneal ulcers (PFCU), and point-like corneal ulcers (PCU). There were 91, 263, and 358 images in each class, respectively. The graph showing the dataset class distributions is given in Fig. 1.

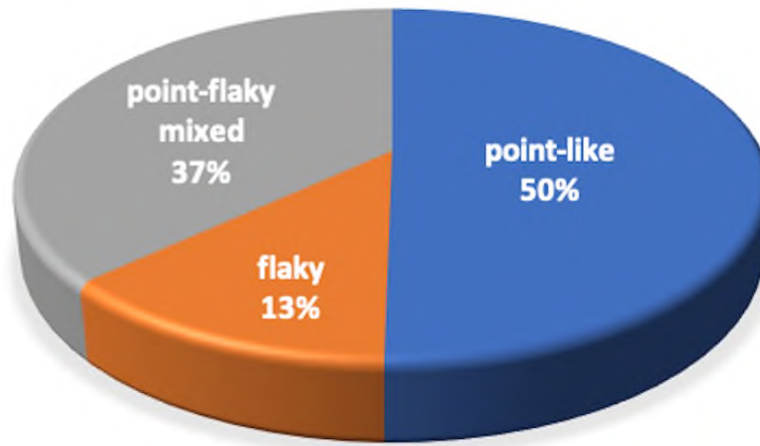


Figure 1. Dataset class distributions

The type of cases for the data set were divided as follows; PCU 50%, PFCU 37%, and FCU 17%. The dataset was not balanced in terms of the number of images in these classes. Image samples of each class in the dataset are given in Fig. 2.

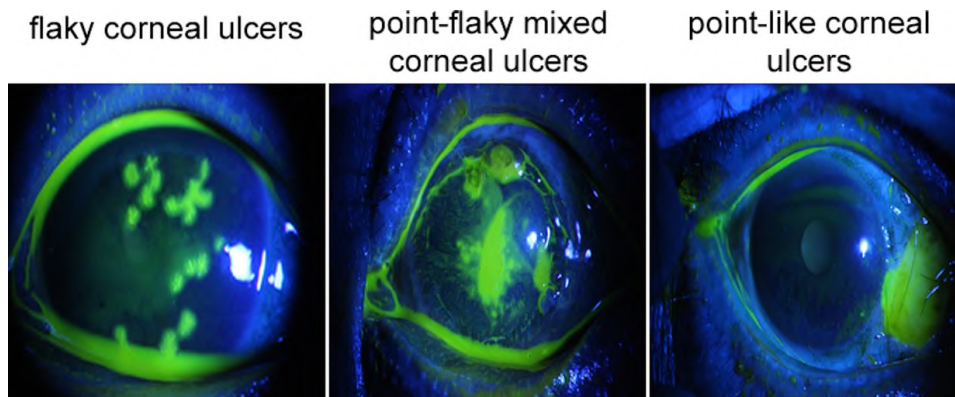


Figure 2. Sample images from the dataset

2.2. Models Used for Classification

Deep learning, which is increasingly used in many fields, is a method of machine learning performed with artificial neural networks consisting of many layers. Deep neural network models, which can contain different numbers of layers and processing units depending on the structure of the problem to be solved (Aksoy, 2021), have a wider learning capacity than classical machine learning techniques. Deep learning models provide high success in revealing complex hierarchical structures and making consistent classifications.

The Convolutional Neural Network (CNN) is a deep learning method that is widely used in solving classification and regression problems in image analysis and has gained popularity with its high success. Input, convolution, pooling, fully connected, and classification layers make up a typical CNN.

The CNN model's inputs are the pixels of the image to be processed. In the convolution layer, feature detectors, also known as filters, are stridden over the input pixels to reveal a subset of features. Convolution is the main operation of a CNN model that enables feature extraction from the image. Dimension reduction is achieved in the pooling stage

using filters applied to the input matrix. The reduction action is carried out by a filter window, also known as a pool, which takes the maximum, minimum, or average of the remaining pixel values in the pool. Activation functions in the weighted layers of the neural network increase nonlinearity.

In the training of machine learning models, the problem of overfitting a certain class may arise due to the structure of the dataset. One of the methods that can be applied to alleviate this problem is the dropout operation. In the dropout process, a certain percentage of neurons in a neural network are randomly disabled during training, increasing the adaptability of the network to different situations. The fully connected layer is involved in the transition to the classification stage. The model in the classification layer tries to predict which class the input sample belongs to.

The number of layers in a CNN model and the number of processing units in each layer are two configurable parameters that change depending on the situation. Training a neural network model with an appropriate amount and variety of data is one of the key steps to obtaining highly accurate results. State-of-the-art CNN models are frequently used in various studies as they successfully classify approximately 14 million images in an ImageNet dataset and provide highly accurate findings when applied to other fields. In this study, two architectures commonly referred to in the literature, AlexNet and VGG16 were used. It was created by improving the architecture.

AlexNet architecture is a CNN model developed by Alex Krizhevsky et al. (2017) that provides high accuracy in classifying the ImageNet dataset. AlexNet architecture consists of 14 layers, eight of which are weighted. There are five convolution layers in the model, three of which are followed by a max-pooling layer. The model has a total of 65 thousand neurons and more than 60 million parameters. AlexNet was among the top five with only a 17% error rate in the ILSVRC-2012 image processing competition and outperformed its successor by 10.9

VGG16 architecture is a CNN model developed by Simonyan and Zisserman (2014). It was developed by enhancing the AlexNet model and using a significant number of 3x3 filters in place of filters with huge core sizes. VGG16 architecture has 13 convolutional layers and five max-pooling layers, and three dense layers in the classification part. It is called VGG16 because it has a total of 16 weighted layers. The VGG16 model, which contains over 138 million parameters, was among the top five models with the highest accuracy of 92.7% in the ImageNet dataset.

It may be necessary to increase the number of layers and components of a CNN model to increase classification accuracy. However, the perception that this increase will always increase the classification success of the model is not correct. The increase in the number of layers and components also increases the number of parameters that need to be calculated in training of the model. The more parameters, the longer the training period of the model. Ideally, the deep learning method is expected to create a model that provides the highest performance with the fewest parameters. In addition to two high complexity state-of-the-art architectures used in this study, a less complex CNN architecture was proposed. The block diagram showing the general structure of the proposed CNN model is given in Fig. 3.

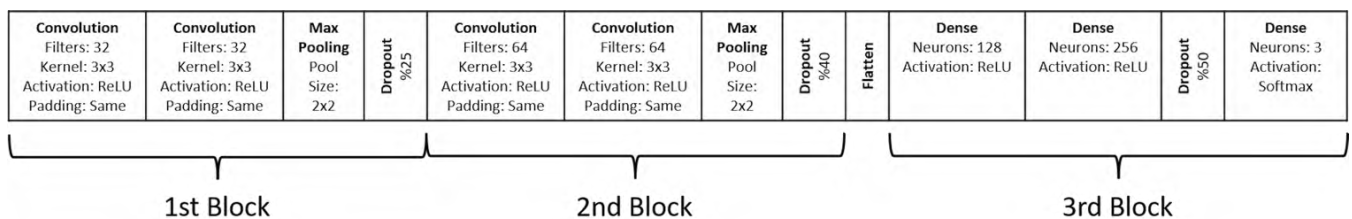


Figure 3. Block diagram of the proposed CNN model

The proposed model is considered in three main blocks. The first block contains two convolutions; a pooling, and a dropout layer. The first two layers, the convolution layers, each contain 32 filters of 3x3 size. In these layers, the ReLU activation function is used, and the same padding is applied. In the third layer, max-pooling is applied with a pool size of 2x2. Then 25% dropout is performed.

In the first two layers of the second block, there are convolution operations with 64 filters of 3x3 size. The ReLU is used as an activation function and the same padding is applied. These layers are followed by a max-pooling layer with a pool size of 2x2. Then a 40% dropout is performed.

After the first two blocks of the model, the extracted features are flattened and transferred to the classifier network, which is the third block. The classification process is performed with a neural network. In the first two layers of the classifier, there are dense layers containing 128 and 256 neurons using the ReLU activation function. Then a 50% dropout is performed. The output layer, a dense layer that contains as many neurons as the number of classes in the dataset, applies the Softmax activation function.

The classification layer used in the proposed CNN model was also used in the classification layer of the other two architectures used in this study. In the proposed model, there are 65,568 parameters excluding the classification layer. The number of parameters of the classification layer is 25,724,035. The total number of trainable parameters of the model, which includes a total of seven weighted layers, including the classification layer, is 25,789,603.

3. EXPERIMENTAL STUDY AND FINDINGS

The 712 corneal ulcer images, consisting of three types colored with the ocular staining technique, were classified with AlexNet, one of the state-of-the-art models widely used in the literature, and the VGG16 model, which was created by improving this architecture, as well as a less complex proposed CNN model used for this study. The state-of-art models used in the classification were fine-tuned. In the preprocessing stage before classification, image rescaling, the normalization of image pixels, and encoding of the labels were performed. Each image in the original dataset is colored and has a size of 2,592x1,728 pixels. Before classification, each image is resized to 224x224 pixels, which is ideal for the CNN architectures used. After rescaling, the pixel values forming the images were normalized with the min-max method. Each label in the dataset consisting of three different classes was numerically encoded. As a result of encoding, the FCU type was labeled as 0, PFCU type as 1, and PCU type as 2.

There were 91 FCU, 263 PFCU, and 358 PCU images in the dataset. In its original form, the dataset had an imbalanced class distribution. When working with imbalanced datasets, the classifier model tends to learn the dominant class and may be weak in learning minority classes. In order to overcome this problem data augmentation was applied using the original images in the dataset by providing the class balance of the dataset. Data augmentation was achieved by applying processes such as rotation, flipping, shifting, reflecting, and scaling on the original images at certain rates. The data augmentation in this study was done by applying 10% rotation, 10% zoom, and 10% shift horizontally and vertically. The processes applied and their ratios were experimental, and were preferred for this study because they gave good results.

The same structural neural network classifier was used after each of the 3 CNN architectures used in this study. This co-classifier has three dense layers. After the first two dense layers, there is a 50% dropout layer. There are 128 neurons in the first dense layer and 256 neurons in the second dense layer, and the ReLU was used as an activation function in both layers. In the dense layer at the output of the classifier, there are three neurons representing the number of classes in the dataset, and the Softmax was used as an activation function.

In the training phase of all three models, common parameters were used and all of them were trained under equal conditions. The values of the parameters were obtained experimentally in a way that would give ideal results for the problem that was to be solved in the study. The hyperparameters and their values used in the training of the models are given in Table 1.

Eighty percent of the data set was used as training and 20% as a test set. To evaluate the performance of each CNN model within an ideal period, the number of epochs was set as 100. As a result of the 100 epoch training, accuracy graphs, and confusion matrices that were obtained from each model were given. The success of each model was reported with different metrics obtained from the confusion matrices.

Table 1. Hyperparameters used in the training phase and their values

Parameter	Value
Batch size	16
Number of epochs	100
Optimizer	Adam
Optimizer parameters	lr=0.00001, beta1 = 0.9, beta_2=0.999, verbose=1, epsilon=None, decay=0.0
Learning rate (LR) reduction	ReduceLROnPlateau
LR reduction metrics	patience=3, verbose=1, factor=0.5, min_lr=0.00001

The confusion matrix provides detailed information about the extent to which the model used can distinguish between the classes in the dataset. A truly positive data sample is called True Positive (TP) if it is positively predicted by the classifier, and False Negative (FN) if it is negatively predicted by the classifier. Similarly, if the data sample with a negative class is predicted negatively by the classifier, it is called True Negative (TN), and if it is incorrectly predicted as a positive, it is called False Positive (FP). The metrics produced to express the performance of the model with these values are given in (1), (2), (3), and (4).

$$Accuracy = (TP + TN) / (TP + TN + FP + FN) \tag{1}$$

$$Precision = TP / (TP + FP) \tag{2}$$

$$Recall = TP / (TP + FN) \tag{3}$$

$$F1 - Score = 2 * (Precision * Recall) / (Precision + Recall) \tag{4}$$

The accuracy metric characterizes the overall success of the classifier. The precision is the hit rate on samples that the model classifies as positive. The recall (also called sensitivity) shows how many of the true positive values are correctly determined. The F1-Score refers to the balance between precision and recall.

The train and validation accuracy graphs and confusion matrix obtained after 100 epoch training and testing processes of the AlexNet model with the specified parameters are given in Fig. 4.

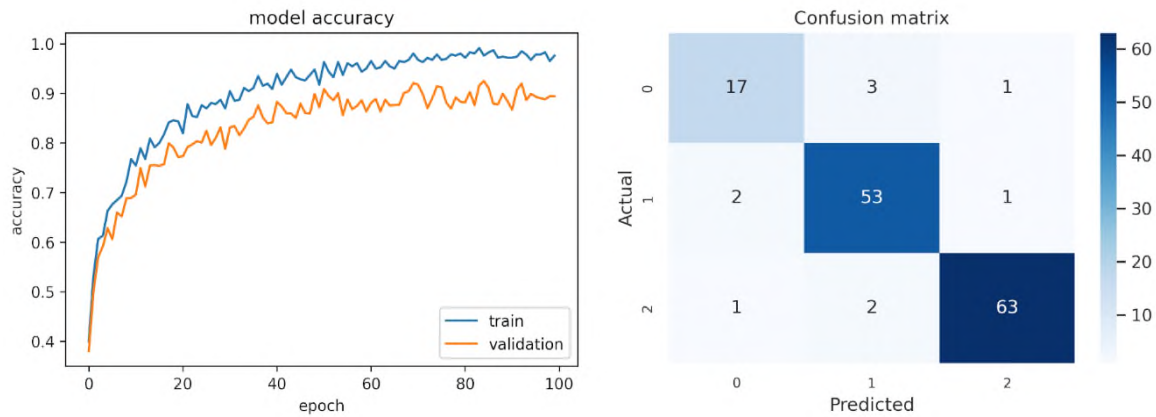


Figure 4. Accuracy graphs and confusion matrix obtained with the AlexNet model

When the accuracy graphs of the model were examined, it showed that the training and validation scores are increasing even though there are oscillations. Although the validation result was slightly lower, it increased in parallel with the training accuracy, indicating that the model did not fall into an overfit condition. Since the optimization of the AlexNet model is more limited compared to the other state-of-the-art model used, its performance is also relatively lower. With this model, the accuracy value reached as a result of 100 epochs was 95.34

When the confusion matrix of the model was examined, it showed that the rate of distinguishing the FCU type labeled as 0 is lower. This model was observed to have more difficulty in distinguishing between the FCU and PFCU classes. The model distinguished the PFCU class at a slightly higher rate, and it was observed that this class was confused with the FCU and to a lesser extent with the PCU class. The class that the model was able to distinguish clearly was the PCU labeled as 2. The model confused this class more with PFCU and less with FCU class. In the general evaluation, it was observed that the success of this model, which was built with AlexNet architecture, was limited compared to the other models. The metrics calculated according to the results obtained from the confusion matrix of the AlexNet model are given in Table 2.

Table 2. Measurements obtained with the AlexNet model

Label	Class	Precision	Recall	F1-Score
0	flaky_corneal_ulcers (FCU)	0.85	0.8095	0.8293
1	point_flaky_mixed_corneal_ulcers (PFCU)	0.9138	0.9464	0.9298
2	point like corneal ulcers (PCU)	0.9692	0.9545	0.9618

When the measurements obtained with the AlexNet model were evaluated, it showed that the PCU is the class that can be distinguished at the highest rate. The precision obtained for this class is 96.92%, recall is 95.45% and F1-Score is 96.18%. The second class with the highest distinction rate was the PFCU. Precision 91.38%, recall 94.64% and F1-Score 92.98% for this class. The class in which the model has the lowest success in distinction is the FCU. The precision obtained for this class is 85%, recall 80.95%, and F1-Score 82.93%.

The accuracy graphs and confusion matrix obtained after 100 epoch training and testing processes of the VGG16 model with the specified parameters are given in Fig. 5.

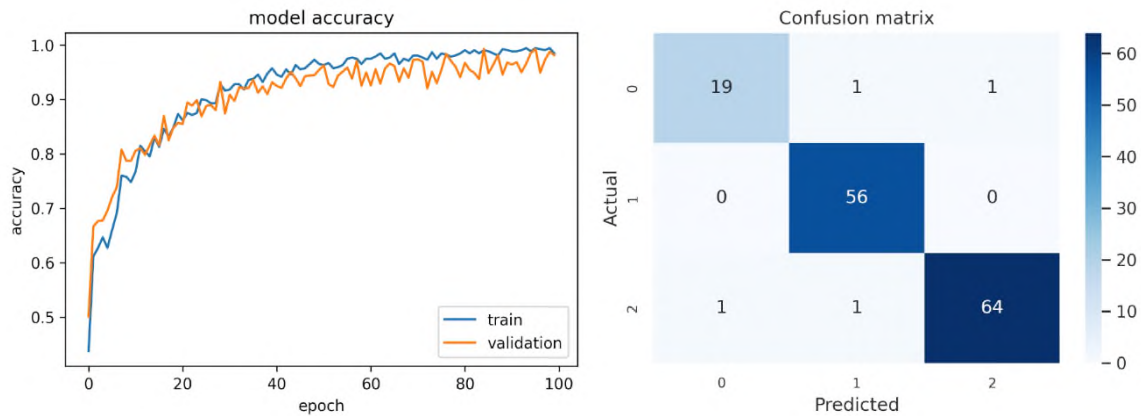


Figure 5. Accuracy graphs and confusion matrix obtained with the VGG16 model

When the accuracy graphs of the VGG16 model were examined, it showed that it is more stable than the AlexNet. The rate of increase in accuracy is relatively higher. Findings were more consistent as expected because the VGG16 architecture is an improved form of AlexNet. The fact that both training and validation accuracies are increased indicates that the model did not have an overfit condition. The accuracy reached with the VGG16 model was 98.14

When the confusion matrix was examined, it showed that the success of the model in distinguishing the FCU class labeled as 0 is relatively low. This class was confused with the PFCU and PCU classes in similar ratios. The model showed the highest success in distinguishing the samples belonging to the one labeled PFCU class. All data samples of the PFCU type were completely distinguished from other types. The PCU class with label 2 is highly distinguishable but confused with the PFCU and FCU classes in equal ratios. The metrics calculated according to the results obtained from the confusion matrix of the VGG16 model are given in Table 3.

Table 3. Measurements obtained with the VGG16 model

Label	Class	Precision	Recall	F1-Score
0	flaky_corneal_ulcers (FCU)	0.95	0.9048	0.9268
1	point_flaky_mixed_corneal_ulcers (PFCU)	0.9655	1.0	0.9825
2	point_like_corneal_ulcers (PCU)	0.9846	0.9697	0.9771

When the measurements obtained with the VGG16 model are examined, the most successful result in terms of the hit rate in the samples classified as positive by the model was obtained for the PCU with 98.46%. This is followed by the PFCU with 96.55% and FCU with 95%. In the correct determination of true positive values, the most successful result was obtained by PFCU at 100%. This is followed by PCU with 96.97% and FCU with 90.48%. In the F1-Score, which shows the balance of these two conditions, the highest success was obtained for PFCU with 98.25%, PCU with 97.71%, and FCU with 92.68%. The most successful results obtained with the VGG16 model were in the PFCU class, and the lowest successful results were in the FCU class.

The CNN model proposed for this study had a simpler architecture compared to the other two models used, AlexNet and VGG16. In this study, it was tested whether a minimal architecture model could compete with more complex models. The accuracy graphs and confusion matrix obtained after 100 epoch training and testing processes using the same hyperparameters as the other models are given in Fig. 6.

When the accuracy graphs of the proposed model were examined, it showed that it increases more rapidly and steadily than the other two models used. According to the accuracy graphs, there was no overfit situation for the proposed model.

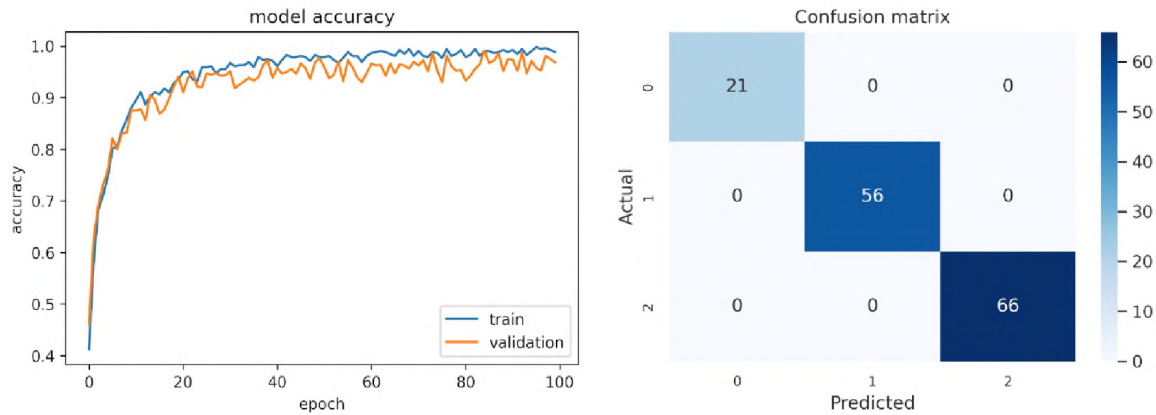


Figure 6. Accuracy graphs and confusion matrix obtained with the proposed model

The oscillations in the graph are less than in the other two models, and the curve tends to flatten in a shorter time. The accuracy value obtained with the proposed model was 100%.

When the confusion matrix of the proposed model was examined, it showed that all classes are clearly distinguished from each other. Any instance of a class was not confused with any other class. Other metrics calculated accordingly are given in Table 4.

Table 4. Measurements obtained with the proposed model

Label	Class	Precision	Recall	F1-Score
0	flaky_corneal_ulcers (FCU)	1.0	1.0	1.0
1	point_flaky_mixed_corneal_ulcers (PFCU)	1.0	1.0	1.0
2	point like corneal ulcers (PCU)	1.0	1.0	1.0

When the measurements obtained were examined, it showed that all classes could be distinguished from each other with 100% success. The fact that the proposed model was less complex than other models and provided higher success showed that the idea about model complexity increasing success is not always true. However, this result is also related to the nature of the dataset used and the preprocessing performed in this study. Therefore, its performance in different problem situations should be tested.

The accuracy values obtained with all models used and the average values of other metrics for all classes are summarized in Table 5.

Table 5. Summary of measurements obtained with all models used

Model	Accuracy (%)	Precision(%)	Recall (%)	F1-Score (%)
AlexNet	95.34	91.10	90.35	90.72
VGG16	98.14	96.67	95.82	96.24
Proposed Model	100	100	100	100

When Table 5 is examined, it shows that the lowest classification success achieved was 95.34%. Although the AlexNet model, which delivered this accuracy is a complicated model, it is less sophisticated than the VGG16 model. The VGG16 is an improved model according to AlexNet and achieved the second highest accuracy rate of 98.14%.

When the examples that the models misclassified were examined, it showed that they belonged to PFCU type cases. Examples of misclassified images are given in Fig. 7.

The PFCU cases combined the characteristics of both PCU and FCU types. PCU typically occurred as small (1 mm or less), sharp-edged and low-depth spots. FCU, on the other hand, produced scratches or a crusty, scaly appearance, usually located in the middle or periphery of the cornea. Due to the coexistence of the features of the FCU and PCU types in the PFCU type, the fact that a feature of any type could suppress the other in general may cause the image class to be determined incorrectly. There are other studies in the literature that support this situation and report that it could be difficult to distinguish the PFCU type from the FCU and PCU types due to the common features (Wang et al., 2021).

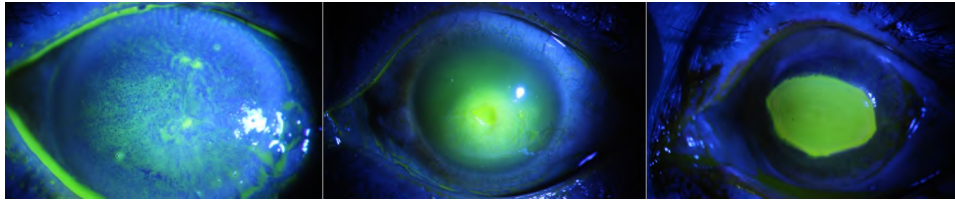


Figure 7. Misclassified sample images

However, both AlexNet and VGG16 models have a deeper and more complex structure, which requires the use and calculation of more parameters in model training. Having more parameters also complicates the optimization, these delays reaching high accuracy values and could cause more oscillations in the result graphs. Although there is a perception that model complexity increases classification accuracy, it was observed in this study that higher accuracy and higher performance could be achieved with a simpler model. The classifier models' goal is to obtain the highest accuracy possible, but a quick response from the model is also anticipated. A model with a simple structure demonstrated quicker reactions with fewer parameters. The proposed model had a simpler structure than the others and delivered the perfect mix between performance and classification success. The proposed model provided 100% accuracy with fewer parameters and reached high accuracy in a shorter time with better performance. This is shown on the model's accuracy graph. While having a high success rate for this problem condition, it should be mentioned that the proposed model has not yet been evaluated in other scenarios. This minimal model may be constrained if the dataset's number of classes and classification complexity rise.

The results obtained in this study with similar studies carried out for the classification of corneal ulcers on eye images colored with the ocular staining technique in the literature are given in Table 6 comparatively.

Table 6. Comparison of this study with similar studies in the literature

Reference	Dataset	Method	Best accuracy (%)
Saini et al. (2003)	study-specific (106 images)	ANN	90.7
Akram and Debnath (2019)	study-specific (513 images)	Proposed CNN	98.99
Kim et al. (2019)	study-specific (1040 images)	ResNet50	92
Tang et al. (2020)	SUSTech-SYSU	Modified VGG	88.89
Xu et al. (2021)	study-specific (115408 images)	DenseNet	66.30
Sajeev and Prem Senthil (2021)	study-specific (446 images)	Proposed CNN	81.2
Gross et al. (2021)	SUSTech-SYSU	Proposed CNN	92.73
Teeyapan (2021)	SUSTech-SYSU	ResNet50	95.10
<i>Literature average</i>			88.24
This study	SUSTech-SYSU	AlexNet	95.34
		VGG16	98.14
		Proposed CNN	100

The number of studies on corneal ulcers and artificial intelligence in the literature is limited. Most of these studies were carried out in the last three years. Besides the SUSTech-SYSU dataset, the originally collected datasets were also used. The methods used were artificial neural networks and deep neural network models in different architectures. The most successful result in the table was obtained as 98.99% accuracy in the study performed by Akram and Debnath (2019). Although the result obtained with their proposed CNN model is close to the result of the VGG16 model used in this study, it is lower than the result obtained with the proposed CNN model in this study. In addition, the dataset used in that study includes 513 samples, which is less than the number of samples used in this study. Therefore, the generalizability of its success is lower.

Another study with high accuracy was carried out by Teeyapan (2021). In that study, the same dataset was used and an accuracy of 95.10% was obtained in the classification performed with the ResNet50 model. This 95.10% accuracy is lower than any classification in this study. Although both studies use the same data set, this study showed that the

data processing and classification techniques used outperformed that study. Another study using the ResNet50 model was carried out by Kim et al. (2019) on a data set approximately 1.5 times larger, but the accuracy rate remained at 92%. This rate was lower than all the results in this study.

Gross et al. (2021) used the same data set and performed classifications with a proposed CNN model. The accuracy rate of 92.73% achieved was lower than all the models used in this study, and it is approximately 7% lower when compared to the proposed model of this study. Another study using a proposed model was carried out by Sajeev and Prem Senthil (2021). The dataset size used in that study was about two-thirds of this study. The 81.2% accuracy they obtained is at least 14% lower than all the models used in this study, and the limited data set they used indicates lower generalizability.

Tang et al. (2020) performed classification on the same dataset with a variation of the VGG16 architecture used in this study. The accuracy of 88.89% obtained was approximately 10% lower than the accuracy obtained with the VGG16 model used in this study. In addition, it was up to 12% lower than the other two models used in this study. The study conducted by Xu et al. (2021) was carried out using the DenseNet model on a much larger data set compared to other studies in the literature. The accuracy rate remained at 66.30% due to the increase in the number of samples, the increase in the workload, and the performance limitation of the model used. This score they obtained is 30% lower than the general average of this study.

The average accuracy of other studies in literature is approximately 88.24%. In this study, even the lowest-performing AlexNet model has a 7% higher value than this rate. The 100% accuracy provided by the proposed model is 12% higher than the average and also higher than all other studies.

The proposed model of this study showed a higher performance with a minimal structure compared to the complex models both in this study and in the literature.

However, the proposed model yielded effective results in corneal ulcer classification on the ocular staining images, but its performance may vary for different problems. Although it is expected to show high success for similar problems, it should be tested in different problem situations. In addition, although the datasets used in this subject in the literature do not generally contain a very high number of data samples, the fact that the number of original images used in this study is not very high could be considered a limitation. In this study, data augmentation was applied to alleviate the problem.

4. CONCLUSION

A corneal ulcer is a common eye problem, and an accurate diagnosis of the disease reduces the risk of permanent eye damage. However, the diagnosis of the disease requires special expertise. Especially in undeveloped countries, the scarcity of experienced ophthalmologists increases the need for artificial intelligence-based decision support systems for accurate diagnosis. In this study, a deep learning-based approach is presented for the classification of corneal ulcers with high success through ocular staining images. Classifications were performed on the SUSTech-SYSU dataset, consisting of 712 samples of three different types of corneal ulcers, with two different state-of-the-art models and a proposed CNN model. An accuracy of 95.34% was obtained with the AlexNet model, 98.14% with the VGG16, and 100% with the proposed model. When the findings were compared with similar studies in the literature, it was found that the average of the three models used was higher than the other studies, and the proposed model gave better results than all of the existing studies. This study contributes to the literature containing a limited number of studies on this subject. It also revealed that high accuracy can be achieved with models with less complexity for certain problems. In future studies, the proposed model will be tested for performance with similar medical image analyses and for solving different problems.

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.

ORCID ID of the author / Yazarın ORCID ID'si

Onur Sevli 0000-0002-8933-8395

REFERENCES

- Akram, A., & Debnath, R. (2019). An Efficient Automated Corneal Ulcer Detection Method using Convolutional Neural Network. 2019 22nd International Conference on Computer and Information Technology (ICIT), 1–6.
- Aksoy, B. (2021). Estimation of Energy Produced in Hydroelectric Power Plant Industrial Automation Using Deep Learning and Hybrid Machine Learning Techniques. *Electric Power Components and Systems*, 49(3), 213–232. <https://doi.org/10.1080/15325008.2021.1937401>
- Amescua, G., Miller, D., & Alfonso, E. C. (2012). What is causing the corneal ulcer? Management strategies for unresponsive corneal ulceration. *Eye*, 26(2), 228–236. <https://doi.org/10.1038/eye.2011.316>
- Basak, S. K., Basak, S., Mohanta, A., & Bhowmick, A. (2005). Epidemiological and microbiological diagnosis of suppurative keratitis in Gangetic West Bengal, eastern India. *Indian Journal of Ophthalmology*, 53(1), 17–22.
- Bron, A. J., Argüeso, P., Irkeç, M., & Bright, F. V. (2015). Clinical staining of the ocular surface: Mechanisms and interpretations. *Progress in Retinal and Eye Research*, 44, 36–61. <https://doi.org/10.1016/j.preteyeres.2014.10.001>
- Chen, J., & Yuan, J. (2010). Strengthen the study of the ocular surface reconstruction. *Chinese Journal of Ophthalmology*, 46(1), 3–5.
- Cohen, E. J., Laibson, P. R., Arentsen, J. J., & Clemons, C. S. (1987). Corneal ulcers associated with cosmetic extended wear soft contact lenses. *Ophthalmology*, 94(2), 109–114.
- Deng, L., Lyu, J., Huang, H., Deng, Y., Yuan, J., & Tang, X. (2020). The SUSTech-SYSU dataset for automatically segmenting and classifying corneal ulcers. *Scientific Data*, 7(1), 23. <https://doi.org/10.1038/s41597-020-0360-7>
- Diamond, J., Leeming, J., Coombs, G., Pearman, J., Sharma, A., Illingworth, C., Crawford, G., & Easty, D. (1999). Corneal biopsy with tissue micro homogenisation for isolation of organisms in bacterial keratitis. *Eye*, 13(4), 545–549.
- Garg, P., & Rao, G. N. (1999). Corneal ulcer: Diagnosis and management. *Community Eye Health*, 12(30), 21–23. PubMed.
- Gross, J., Breitenbach, J., Baumgartl, H., & Buettner, R. (2021). High-performance detection of corneal ulceration using image classification with convolutional neural networks. E 54th Hawaii International Conference on System Sciences, 3416–3425.
- Katara, R. S., Patel, N. D., & Sinha, M. (2013). A clinical microbiological study of corneal ulcer patients at western Gujarat, India. *Acta Medica Iranica*, 399–403.
- Kim, J. Y., Lee, H. E., Choi, Y. H., Lee, S. J., & Jeon, J. S. (2019). CNN-based diagnosis models for canine ulcerative keratitis. *Scientific Reports*, 9(1), 1–7.
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6), 84–90.
- Li, Z., Jiang, J., Chen, K., Chen, Q., Zheng, Q., Liu, X., Weng, H., Wu, S., & Chen, W. (2021). Preventing corneal blindness caused by keratitis using artificial intelligence. *Nature Communications*, 12(1), 1–12.
- Maurice, D. M. (1957). The structure and transparency of the cornea. *The Journal of Physiology*, 136(2), 263.
- Portela, H. M. B., MS Veras, R. de, Vogado, L. H. S., Leite, D., Sousa, J. A. de, Paiva, A. C. de, & Tavares, J. M. R. (2021). A Coarse to Fine Corneal Ulcer Segmentation Approach Using U-net and DexiNed in Chain. Iberoamerican Congress on Pattern Recognition, 13–23.
- Saini, J. S., Jain, A. K., Kumar, S., Vikal, S., Pankaj, S., & Singh, S. (2003). Neural network approach to classify infective keratitis. *Current Eye Research*, 27(2), 111–116.
- Sajeev, S., & Prem Senthil, M. (2021). Classifying infective keratitis using a deep learning approach. 2021 Australasian Computer Science Week Multiconference, 1–4.
- Simonyan, K., & Zisserman, A. (2014). Very Deep ConvNets for Large-Scale Image Recognition. CoRR.
- Song, X., Xie, L., Tan, X., Wang, Z., Yang, Y., Yuan, Y., Deng, Y., Fu, S., Xu, J., Sun, X., & others. (2014). A multi-center, cross-sectional study on the burden of infectious keratitis in China. *PLoS One*, 9(12), e113843.
- Tang, N., Liu, H., Yue, K., Li, W., & Yue, X. (2020). Automatic classification for corneal ulcer using a modified VGG network. 2020 International Conference on Artificial Intelligence and Computer Engineering (ICAICE), 120–123.
- Teeyapan, K. (2021). Deep learning-based approach for corneal ulcer screening. The 12th International Conference on Computational Systems-Biology and Bioinformatics, 27–36.
- Wang, T., Wang, M., Zhu, W., Wang, L., Chen, Z., Peng, Y., Shi, F., Zhou, Y., Yao, C., & Chen, X. (2021). Semi-MsST-GAN: A Semi-Supervised Segmentation Method for Corneal Ulcer Segmentation in Slit-Lamp Images. *Frontiers in Neuroscience*, 15.
- Wang, T., Zhu, W., Wang, M., Chen, Z., & Chen, X. (2021). Cu-segnet: Corneal ulcer segmentation network. 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), 1518–1521.
- Wang, Z., Lyu, J., Luo, W., & Tang, X. (2021). Adjacent Scale Fusion and Corneal Position Embedding for Corneal Ulcer Segmentation. *International Workshop on Ophthalmic Medical Image Analysis*, 1–10.
- Xu, Y., Kong, M., Xie, W., Duan, R., Fang, Z., Lin, Y., Zhu, Q., Tang, S., Wu, F., & Yao, Y.-F. (2021). Deep sequential feature learning in clinical image classification of infectious keratitis. *Engineering*, 7(7), 1002–1010.

How cite this article

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>

Performance Evaluation of Magnitude-Based Fuzzy Analytic Hierarchy Process (MFAHP) Method

Magnitüde Bağlı Bulanık Analitik Hiyerarşi Süreci (MBAHS) Yöntemi Performans Değerlendirmesi

Barış Tekin Tezel¹  Ayşe Övgü Kınay¹ 

¹(Assist. Prof.), Dokuz Eylül University, Faculty of Science, Department of Computer Science

Corresponding author : Barış Tekin TEZEL
E-mail : baris.tezel@deu.edu.tr

ABSTRACT

In Analytic Hierarchy Process (AHP), which is a very common method in Multi-Criteria Decision Making (MCDM) problems, the use of fuzzy set theory, which allows human judgment to be expressed more realistically, has gained popularity in recent years. However, this situation causes more computational complexity due to the way fuzzy numbers are expressed and the operators used. In this study, the results of real various problems in a hierarchical structure with the Magnitude Based Fuzzy Analytic Hierarchy Process (MFAHP) were compared with the results of the Modified Fuzzy Logarithmic Least Squares method (MFLLSM) and Buckley's Geometric Means method (GM), which are two known methods to obtain accurate weight values. The results show that there is no statistically significant difference between MFAHP and the results of these two methods. In the performance comparison, although it is known that it produces incorrect results, unfortunately, the results of Chang's Extent Analysis method on fuzzy AHP (FEA) are also included because it is a widely used method. As another important finding of this study, it can be said that MFAHP is faster than both methods when the running times are compared. Finally, software for the calculations of these methods mentioned in the study has been developed and link shared.

Keywords: Fuzzy multi-criteria decision making, fuzzy analytic hierarchy process, performance evaluation

ÖZ

Çok Kriterli Karar Verme (ÇKKV) problemlerinde oldukça yaygın bir yöntem olan Analitik Hiyerarşi Sürecinde (AHS), insan yargısının daha gerçekçi bir şekilde ifade edilmesini sağlayan bulanık küme teorisinin kullanımı son yıllarda önem kazanmıştır. Ancak bu durum, bulanık sayıların ifade edilme şekli ve kullanılan operatörler nedeniyle daha fazla hesaplama karmaşıklığına neden olmaktadır. Bu çalışmada, hiyerarşik yapıdaki çeşitli gerçek hayat problemlerinin Magnitüde Bağlı Bulanık Analitik Hiyerarşi Süreci (MBAHS) ile elde edilen sonuçların doğruluğu, doğru ağırlık değerleri elde etmekte kullanılan iki yöntem olan Modifiye Bulanık Logaritmik En Küçük Kareler yöntemi (MFLLSM) ve Buckley'nin Geometrik Ortalamalar yöntemi (GM) sonuçları ile karşılaştırılmıştır. Sonuçlar, MBAHS ile bu iki yöntemin sonuçları arasında istatistiksel olarak anlamlı bir fark olmadığını göstermektedir. Performans karşılaştırmasında hatalı sonuçlar ürettiği bilirse de ne yazık ki yaygın olarak kullanılan bir yöntem olduğu için bulanık AHS'de Chang'in Extent Analizi (CEA) yöntemi sonuçları da yer almaktadır. Bu çalışmanın bir diğer önemli bulgusu olarak çalışma süreleri karşılaştırıldığında da MBAHS'nin her iki yöntemden daha hızlı olduğu söylenebilir. Son olarak çalışmada adı geçen bu yöntemlerin hesaplamalarının yapılabileceği bir yazılım geliştirilmiş ve bağlantısı paylaşılmıştır.

Anahtar Kelimeler: Bulanık çok kriterli karar verme, bulanık analitik hiyerarşi süreci, performans değerlendirme.

Submitted : 26.12.2022
Revision Requested : 05.01.2023
Last Revision Received : 07.06.2023
Accepted : 10.07.2023
Published Online : 24.11.2023



This article is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0)

1. INTRODUCTION

In its most general definition, decision-making problems are described as the problem of one or more decision-makers choosing one or more of the alternatives by considering various criteria and possibly the sub-criteria of these criteria. However, as the number of decision-makers, criteria, and alternative concepts, which are the fundamental concepts of this definition, increases, the problem's solution becomes quite complex. For this reason, Multi-Criteria Decision Making (MCDM) methods are used to help decision-makers choose alternatives, especially in complex decision-making processes, by considering the criteria analytically.

The Analytic Hierarchy Process (AHP) method proposed by Saaty (1977; 1980) is widely preferred among the different MCDM methods. Prioritization, resource allocation, business process reengineering, quality management, and planning are some of the domains where the AHP method is applied (Vaidya & Kumar, 2006; Korkmaz, Gökçen & Çetinyokuş, 2008; Amiri, 2010; Dweiri, Kumar, Khan & Jain, 2016).

The reasons for the widespread use of this method are the hierarchical structure of the problems, the ease of calculation, and the calculation of both criterion weights and alternative priorities. AHP provides a hierarchical structure for a problem, starting with the goal and continuing through criteria, sub-criteria, and alternatives. This method enables decision-makers to systematically evaluate relations and see more clearly what to compare. Thus, the global priorities of the alternatives are determined by making pairwise comparisons of the necessary elements. Pairwise comparisons, on the other hand, are frequently ambiguous. Because the number of elements to be compared increases, it becomes difficult to compare each pair with exact numbers (Xu & Liao, 2013). As it is known, linguistic expressions in Saaty's relative importance scale are represented in ascending order of importance, from 1 for "equally important" to 9 for "extremely more important". However, studies on fuzzy AHP have started to increase rapidly with the use of fuzzy numbers with the thought that they can better reflect these linguistic expressions instead of the integers defined for the linguistic expressions in the scale. As a result, fuzzified techniques have gained popularity in recent decades, and various fuzzy AHP (FAHP) methods have been presented (Liu, Eckert & Earl, 2020). However, integrating the concept of fuzziness into the AHP complicates the computational process. The use of fuzzy sets in AHP, on the other hand, makes the computing process more difficult. For this reason, FAHP methods, which give accurate results similar to the classical AHP method and can be applied easily at the same time, should become widespread.

The main motivation of this study is to evaluate the results of the Magnitude Based Fuzzy Analytic Hierarchy Method (MFAHP), a new FAHP method proposed by (Kinay & Tezel, 2022), on real examples. In this evaluation, the results of the Geometric Mean Method (GM) (Buckley, 1985) and Modified Fuzzy Logarithmic Least Squares Method (MFLLSM) (Wang, Elhag & Hua, 2006) were considered. Because using these two methods, accurate results are obtained. However, both methods, especially MFLLSM, are difficult to apply, especially in social sciences, and the computational load is high. In addition, Fuzzy Extent Analysis (FEA) (Chang, 1996) results are also included in the comparisons. This method is widely used due to its easy application, but unfortunately, it produces wrong results and the necessity of not using it has been mentioned in various studies (Liu et al., 2020; Ahmed & Kiliç, 2019). Although the computational load is significantly lower than other methods, studies are still being carried out to cope with the rapidly increasing computational load depending on the problem structure and representation. In summary, a parallel computing method has been developed (Ballı & Bahadır, 2013), which allows the system to run faster and increases efficiency and performance for FEA operations that require a large number of computation-intensive operations. In this study, according to the results obtained from real numerical examples, the fact that the working time of the MFAHP method is not significantly different from the FEA can be considered an important advantage of the MFAHP method over the MFLLSM and GM methods, where accurate results are obtained. Therefore, the research hypothesis of our study is that the MFAHP method will provide accurate results while requiring less computation time compared to the GM and MFLLSM methods on real examples.

The main research contributions (RC) of this study can be summarized as follows:

- **RC1:** It is statistically shown that the results obtained using the MFAHP method are as accurate as the results obtained using the GM and MFLLSM methods.
- **RC2:** It is shown that the MFAHP method is faster in terms of computation time.
- **RC3:** Software has been developed and shared for the solution of all methods used in the study.

The rest of this paper is structured as follows. Related studies are presented in Section 2. The information used in the content of the study and the application of the MFAHP method on an example and the results are described in Section 3. The performance comparisons of the MFAHP with the MFLLSM, the GM, and the FEA are presented in Section 4. And finally, conclusions will be highlighted in Section 5.

2. RELATED WORKS

In (Mardani, Jusoh & Zavadskas, 2015), the applications and methods of fuzzy Multi-Criteria Decision Making (FMCDM) methods have been reviewed and it has been said that AHP and FAHP methods are the most preferred methods in decision-making problems. In 1983, the first method of FAHP was proposed (Van Laarhoven & Pedrycz, 1983). The Fuzzy Logarithmic Least Squares Method (fuzzy LLSM) is used to derive weights in the type of triangular fuzzy numbers from pairwise comparison matrices containing triangular fuzzy numbers. Buckley (1985) used the trapezoidal numbers and the geometric mean approach to achieve the fuzzy pairwise comparison. Kwiesielewicz (1996) presented a generalized pseudo-inverse approach that used spectral decomposition to solve the fuzzy LLSM. In (Boender, 1989), a change in the normalization process is proposed to prevent deviations in weight values resulting from the normalization process used in the fuzzy LLSM method. Ruoning and Xiaoyan (1996) developed a fuzzy LLSM depending on the notion of distance in a fuzzy evaluation scale. In (Chang, 1996), the extent analysis method on fuzzy AHP (FEA) was proposed by Chang by obtaining synthetic extent values of pairwise comparisons. Büyüközkan et al. (2004) provide a survey of FAHP algorithms with their main features, advantages, and disadvantages. The fuzzy LLSM approach presented in (Van Laarhoven & Pedrycz, 1983) was developed by (Wang, Luo & Hua, 2008) and it was named modified fuzzy LLSM (MFLLSM). This method can be expressed as a constrained nonlinear optimization model proposed such that normalized triangular fuzzy weights can be obtained.

It is also mentioned in (Wang, Luo & Hua, 2008) that real weights cannot be obtained with the FEA method, and this may lead to wrong decisions. However, Kubler et al. (2016) emphasize that owing to its simplicity of use, it is still a popular method in many domains. According to Ahmed & Kiliç (2019), FEA produces the least accurate results.

The main purpose of this study is to evaluate the performance of MFAHP results on real examples in a hierarchical structure and to show that this method gives accurate results and has a low computational load. In order to derive weight values or in other words the priority vectors, in the MFAHP method, the magnitude value of each fuzzy number was considered as suggested in (Abbasbandy & Hajjari, 2009). Studies show that in fuzzy AHP methods, comparison judgments are expressed as triangular fuzzy numbers in pairwise comparison matrices at a rate of 91% (Lie et al., 2020). Therefore, for the magnitude calculation used in the MFAHP method, the method suggested by (Abbasbandy & Hajjari, 2009) was preferred, which gives sufficient results in the comparison of fuzzy triangular numbers.

In this study, all examples used in calculations were obtained from articles published in indexed journals. Therefore, we did not check again the consistencies of the fuzzy pairwise comparison matrices which are used as the preference relations in the examples mentioned in Section 4.

3. BASIC CONCEPTS AND METHODS

3.1. Fuzzy Membership Function

The concept of fuzzy sets, first proposed in (Zadeh, 1965), is used to solve problems with ambiguous descriptions. Fuzzy sets can be thought of as a general representation of crisp sets and these are the sets of objects defined by a membership function. The membership function determines the degree of belonging of the elements to the related set. The degree of belonging of the elements to the set can take all membership degrees from "does not belong to the set" to "belongs to the set". That is, the degree of belonging of the elements to the related set is defined in $[0,1]$.

The following is the definition of the triangular fuzzy membership function as used in the examples in Section 4 of this study.

Definition 1. $A = (l, m, u)$ on $U = (-\infty, \infty)$ is expressed as a triangular fuzzy number, and its membership function $\mu_A: U \rightarrow [0,1]$ is given as:

$$\mu_A(x) = \begin{cases} \frac{(x-l)}{(m-l)}, & l < x < m \\ 1, & x = m \\ \frac{(u-x)}{(u-m)}, & m < x < u \\ 0, & otherwise \end{cases} \quad (1)$$

3.2. Extent Analysis on Fuzzy AHP

The extent analysis on fuzzy AHP (FEA) was proposed by Chang (1996) and it is summarized as follows.

The following is a fuzzy pairwise comparison matrix with judgments expressed as triangular fuzzy membership functions:

$$A = (a_{ij})_{n \times n} = \begin{bmatrix} (1,1,1) & (l_{12}, m_{12}, u_{12}) & \dots & (l_{1n}, m_{1n}, u_{1n}) \\ (l_{21}, m_{21}, u_{21}) & (1,1,1) & \dots & (l_{2n}, m_{2n}, u_{2n}) \\ \vdots & \vdots & \ddots & \vdots \\ (l_{n1}, m_{n1}, u_{n1}) & (l_{n2}, m_{n2}, u_{n2}) & \dots & (1,1,1) \end{bmatrix} \quad (2)$$

where $a_{ij}=(l_{ij},m_{ij},u_{ij})$, and $a_{ji}=a_{ij}^{-1}=(1/u_{ij},1/m_{ij},1/l_{ij})$ for $i,j = 1, \dots, n, i \neq j$. First, the sum values for the rows of the fuzzy pairwise comparison matrices are obtained as follows:

$$RS_i = \sum_{j=1}^n a_{ij} = \left(\sum_{j=1}^n l_{ij}, \sum_{j=1}^n m_{ij}, \sum_{j=1}^n u_{ij} \right), i = 1, \dots, n. \quad (3)$$

Then in the second step, the row sums are normalized as in Equation (4).

$$S_i = \sum_{j=1}^n a_{ij} \otimes \left[\sum_{k=1}^n \sum_{j=1}^n a_{kj} \right]^{-1} = \frac{RS_i}{\sum_{j=1}^n RS_j} = \left(\frac{\sum_{j=1}^n l_{ij}}{\sum_{k=1}^n \sum_{j=1}^n l_{kj}}, \frac{\sum_{j=1}^n m_{ij}}{\sum_{k=1}^n \sum_{j=1}^n m_{kj}}, \frac{\sum_{j=1}^n u_{ij}}{\sum_{k=1}^n \sum_{j=1}^n l_{kj}} \right), i = 1, \dots, n. \quad (4)$$

In the third step, Equation (5) is used to calculate each possibility value:

$$V(S_i \geq S_j) = \begin{cases} 1 & , \quad m_i \geq m_j \\ \frac{(u_i - l_j)}{(u_i - m_i) + (m_j - l_j)} & , \quad l_j \leq u_i, i, j = 1, \dots, n; j \neq i \\ 0 & , \quad otherwise \end{cases} \quad (5)$$

where $S_i=(l_i,m_i,u_i)$ and $S_j=(l_j,m_j,u_j)$ and $V(S_i \geq S_j)$ value is shown in Fig.1.

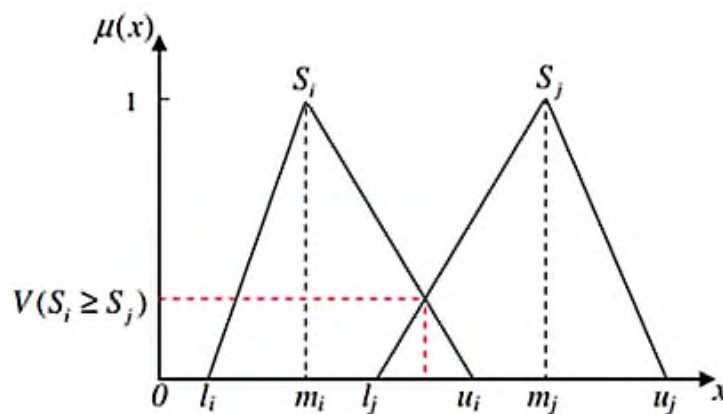


Figure 1. . Graphical representation of $V(S_i \geq S_j)$.

In the fourth step, the degree of possibility of S_i over all other (n-1) fuzzy numbers is calculated as:

$$V(S_i \geq S_j | j = 1, \dots, n; j \neq i) = \min_{j \in \{1, \dots, n\}, j \neq i} V(S_i \geq S_j, i = 1, \dots, n. \quad (6)$$

Finally, the weight values are obtained as follows.

$$w_i = \frac{V(S_i \geq S_j \mid j = 1, \dots, n; j \neq i)}{\sum_{k=1}^n V(S_k \geq S_j \mid j = 1, \dots, n; j \neq k)}, i = 1, \dots, n. \quad (7)$$

where the weight values are crisp values.

3.3. Geometric Mean Method

The Geometric mean method was proposed by Buckley in 1985. In the first step, for each fuzzy pairwise comparison matrix expressed as in Equation (2), the geometric mean of each criterion is calculated as follows:

$$z_i = \left(\prod_{j=1}^n a_{ij} \right)^{1/n}, i = 1, \dots, n. \quad (8)$$

Then, weight values r_i of each criterion or each alternative are obtained by Equation (9) as follows:

$$r_i = z_i \otimes [z_1 \oplus z_2 \oplus \dots \oplus z_n]^{-1} \quad (9)$$

In the third step, obtained weight values are converted into crisp values by using the Center of Area defuzzification method as in Equation (10):

$$S_i = \frac{l_i + m_i + u_i}{3}, i = 1, \dots, n. \quad (10)$$

In the final step, these weight values are normalized by Equation (11) to obtain the normal weight values.

$$w_i = \frac{S_i}{\sum_{i=1}^n S_i}, i = 1, \dots, n. \quad (11)$$

where the weight values are crisp values.

3.4. Modified Fuzzy Logarithmic Least Squares Method

The MFLLSM is improved by Wang et al. (2006) to determine the local fuzzy weights of the fuzzy pairwise comparison matrix in Equation (2). Each decision problem that was used in calculations in Section 4 has only one decision-maker. Therefore, the way the method is defined in (Wang et al., 2008) is as follows.

$$\min J = \sum_{i=1}^n \sum_{j=1, j \neq i}^n \left((\ln w_i^L - \ln w_j^U - \ln l_{ij})^2 + (\ln w_i^M - \ln w_j^M - \ln m_{ij})^2 + (\ln w_i^U - \ln w_j^L - \ln u_{ij})^2 \right) \quad (12)$$

$$s.t. \begin{cases} w_i^L + \sum_{j=1, j \neq i}^n w_j^U \geq 1, \\ w_i^U + \sum_{j=1, j \neq i}^n w_j^L \leq 1, \\ \sum_{i=1}^n w_i^M = 1 \quad i = \dots, n \\ \sum_{i=1}^n (w_i^L + w_i^U) = 2, \\ w_i^U \geq w_i^M \geq w_i^L > 0. \end{cases} \quad (13)$$

After this stage, the global fuzzy weights can be obtained by solving the following two linear programming models and an equation for each decision alternative A_i ($i=1, \dots, n$) as follows:

$$w_{A_i}^L = \min_{W \in \Omega_W} \sum_{j=1}^m w_{ij}^L w_j, i = 1, \dots, n, \quad (14)$$

$$w_{A_i}^U = \min_{W \in \Omega_W} \sum_{j=1}^m w_{ij}^U w_j, i = 1, \dots, n, \quad (15)$$

$$w_{A_i}^M = \sum_{j=1}^m w_{ij}^M w_j^M, i = 1, \dots, n, \tag{16}$$

where $\Omega_W = W = (w_1, \dots, w_m)^T \mid w_j^L \leq w_j \leq w_j^U, \sum_{j=1}^m w_j = 1, j=1, \dots, m$ is the set of weights, (w_j^L, w_j^M, w_j^U) is the normalized fuzzy weight of criterion j ($j=1, \dots, m$), and $(w_{ij}^L, w_{ij}^M, w_{ij}^U)$ is the normalized fuzzy weight of alternative A_i concerning the criterion j ($i=1, \dots, n; j=1, \dots, m$).

3.5. Magnitude-Based Fuzzy Analytic Hierarchy Process

In many disciplines, ranking fuzzy numbers is an important part of decision-making, and numerous scholars have created various ranking approaches (Abbasbandy & Hajjari, 2009; Wang & Kerre, 2001a; Wang & Kerre, 2001b; Chutia & Chutia, 2017; Bortolan & Degani, 1985).

In FAHP methods, preferred values in fuzzy pairwise comparison matrices are usually normal triangular fuzzy numbers. This means that the height of the triangular membership function is equal to 1. For this reason, the magnitude of a fuzzy number proposed by (Abbasbandy & Hajjari, 2009), which forms the basis of the MFAHP method proposed by (Kinay & Tezel, 2022), has been preferred because it is easy and effective for FAHP calculations. The MFAHP method is defined in (Kinay & Tezel, 2022) as follows.

In the first step, row sum values for each fuzzy pairwise comparison matrix are obtained using Equation (3).

The second step is to apply the normalization process as stated in (Wang et al., 2008; Wang & Elhag, 2006) as follows.

$$S_i = \frac{RS_i}{\sum_{j=1}^n RS_j} = \left(\frac{\sum_{j=1}^n l_{ij}}{\sum_{j=1}^n l_{ij} + \sum_{k=1, k \neq i}^n \sum_{j=1}^n u_{kj}}, \frac{\sum_{j=1}^n m_{ij}}{\sum_{k=1}^n + \sum_{j=1}^n, m_{kj}}, \frac{\sum_{j=1}^n u_{ij}}{\sum_{j=1}^n u_{ij} + \sum_{k=1, k \neq i}^n \sum_{j=1}^n l_{kj}} \right), i = 1, \dots, n. \tag{17}$$

In the third step, the magnitude values of each S_i value are calculated as given in Equation (18).

$$Mag(S_i) = \frac{l_i + 10m_i + u_i}{12}, i = 1, \dots, n. \tag{18}$$

In the last step, magnitude values for each S_i value are normalized by Equation (19).

$$w_i = \frac{Mag(s_i)}{\sum_{j=1}^n Mag(s_j)}, i = 1, \dots, n. \tag{19}$$

where the weight values are crisp values.

3.6. An example application of the MFAHP method

In this subsection, the computations of the MFAHP method were demonstrated using the example of the shipping registry selection problem given in (Celik, Er & Ozok, 2009). This problem has three main criteria (C1, C2, and C3). The main criteria have four, three, and three sub-criteria (C11-C14, C21-C23, and C31-C33), respectively, and it is intended to determine the most preferred one among the four alternatives (A1-A4). Detailed information about this problem, its hierarchical structure, and fuzzy pairwise comparison matrices can be found in (Celik, Er & Ozok, 2009).

The MFAHP method will be illustrated step-by-step using the fuzzy pairwise comparison matrix values which are generated for the three criteria in Table 1.

Table 1. Fuzzy pairwise comparison matrix for three criteria

Criteria	C1	C2	C3
C1	(1, 1, 1)	(5/2, 3, 7/2)	(3/2, 2, 5/2)
C2	(2/7, 1/3, 2/5)	(1, 1, 1)	(2/3, 1, 3/2)
C3	(2/5, 1/2, 2/3)	(2/3, 1, 3/2)	(1, 1, 1)

As the first step, RS_i values are calculated for each criterion using Equation (3), and the results were obtained as follows;

$$RS_1=(5.0000,6.0000,7.0000)$$

$$RS_2=(1.9524,2.3333,2.9000)$$

$$RS_3=(2.0667,2.5000,3.1667)$$

Table 2. Results obtained with MFAHP.

Local weights of alternatives with respect to C1					
	C11	C12	C13	C14	Weights
Weights	0.2007	0.2012	0.2567	0.3413	
A1	0.1235	0.1478	0.2012	0.1042	0.1417
A2	0.3779	0.2839	0.2012	0.3484	0.3035
A3	0.3550	0.4117	0.2012	0.3228	0.3159
A4	0.1436	0.1566	0.3964	0.2245	0.2387
Local weights of alternatives with respect to C2					
	C21	C22	C23		Weights
Weights	0.2315	0.5527	0.2158		
A1	0.1725	0.3352	0.3762		0.3064
A2	0.2294	0.2332	0.2966		0.2460
A3	0.3411	0.2751	0.1636		0.2663
A4	0.2570	0.1565	0.1636		0.1813
Local weights of alternatives with respect to C3					
	C31	C32	C33		Weights
Weights	0.2516	0.2516	0.4969		
A1	0.3324	0.2960	0.1403		0.2278
A2	0.2499	0.3788	0.4529		0.3832
A3	0.1673	0.2056	0.1786		0.1826
A4	0.2503	0.1195	0.2282		0.2064
Global weights of alternatives					
	C1	C2	C3		Weights
Weights	0.5510	0.2167	0.2323		
A1	0.1417	0.3064	0.2278		0.1974
A2	0.3035	0.2460	0.3832		0.3096
A3	0.3159	0.2663	0.1826		0.2742
A4	0.2387	0.1813	0.2064		0.2188

In the second step, the normalization operation in Equation (17) is applied for all three rows of the matrix (the rows represent the main criteria in this matrix) as follows.

$$S_1 = RS_1 \otimes [RS_1 \oplus RS_2 \oplus RS_3]^{-1} = \left(\frac{5.0000}{(5.0000 + 6.0667)}, \frac{6.0000}{(10.8333)}, \frac{7.0000}{(7.0000 + 4.0190)} \right) = (0.4518, 0.5539, 0.6356) \quad (20)$$

$$S_2 = RS_2 \otimes [RS_1 \oplus RS_2 \oplus RS_3]^{-1} = \left(\frac{1.9524}{(1.9524 + 10.1667)}, \frac{2.3333}{(10.8333)}, \frac{2.9000}{(2.9000 + 7.0667)} \right) = (0.1611, 0.2154, 0.2910) \quad (21)$$

$$S_3 = RS_3 \otimes [RS_1 \oplus RS_2 \oplus RS_3]^{-1} = \left(\frac{2.0667}{(2.0667 + 9.9000)}, \frac{2.5000}{(10.8333)}, \frac{3.1667}{(3.1667 + 6.9524)} \right) = (0.1727, 0.2308, 0.3129) \quad (22)$$

In the third step, the magnitude values of the normalized row totals are calculated using Equation (18) as follows.

$$Mag(S_1) = \frac{0.4518 + 10 * 0.5539 + 0.6356}{12} = 0.5522 \quad (23)$$

$$Mag(S_2) = \frac{0.1611 + 10 * 0.2154 + 0.2910}{12} = 0.2171 \quad (24)$$

$$Mag(S_3) = \frac{0.1727 + 10 * 0.2308 + 0.3129}{12} = 0.2328 \quad (25)$$

In the last step, normalized weight values are obtained with Equation (19) as follows.

$$W=(0.5510,0.2167,0.2323)^T$$

Global weights are found for each alternative by multiplying the local weights according to the hierarchical structure, starting from the sub-criteria. As a result, local and global weight values of MFAHP were obtained as in Table 2.

4. PERFORMANCE ANALYSIS

Table 3. The weights obtained by FEA, MFLLSM, MFAHP, and GM.

	FEA		MFLLSM		MFAHP		GM	
	Weights	Rank	Weights	Rank	Weights	Rank	Weights	Rank
(Celik, Er & Ozok, 2009)	0.0429	4	0.1935	4	0.1974	4	0.1941	4
	0.3583	2	0.3133	1	0.3096	1	0.3125	1
	0.3878	1	0.2710	2	0.2742	2	0.2682	2
(Kahraman, Cebeci & Ruan, 2004)	0.2110	3	0.2232	3	0.2188	3	0.2253	3
	0.0418	3	0.3088	2	0.3155	2	0.3091	2
	0.6179	1	0.2845	3	0.2821	3	0.2829	3
(Arikan & Dağdeviren, 2013)	0.3404	2	0.4081	1	0.4024	1	0.4080	1
	0.1672	1	0.1613	2	0.1606	2	0.1613	2
	0.1337	4	0.1345	5	0.1340	5	0.1327	5
(Arif et al., 2021)	0.1295	5	0.1349	4	0.1355	4	0.1342	4
	0.1235	6	0.1292	7	0.1287	6	0.1286	7
	0.1640	2	0.1642	1	0.1645	1	0.1630	1
(Büyükoğuzkan, Çifçi & Gülerüz, 2011)	0.1628	3	0.1495	3	0.1487	3	0.1505	3
	0.1193	7	0.1311	6	0.1280	7	0.1297	6
	0.2560	1	0.2664	1	0.2489	2	0.2693	1
(Dong, Li & Zhang, 2015)	0.2560	1	0.2517	2	0.2497	1	0.2455	2
	0.1970	3	0.1471	3	0.1670	3	0.1473	3
	0.1806	4	0.1376	4	0.1489	4	0.1416	4
(Dong, Li & Zhang, 2015)	0.0894	5	0.1063	5	0.0994	5	0.1062	5
	0.0210	6	0.0934	6	0.0862	6	0.0902	6
	0.1339	3	0.1882	4	0.1803	4	0.1841	4
(Aydoğan, Demirtas & Dağdeviren, 2015)	0.3147	2	0.3050	1	0.3015	1	0.3027	1
	0.4307	1	0.3046	2	0.2944	2	0.3008	2
	0.1207	4	0.2105	3	0.2238	3	0.2124	3
(Isaai et al., 2011)	0.4099	1	0.3609	1	0.3626	1	0.3600	1
	0.2672	3	0.3100	3	0.3031	3	0.3070	3
	0.3229	2	0.3310	2	0.3342	2	0.3329	2
(Aydoğan, Demirtas & Dağdeviren, 2015)	0.4234	1	0.3553	1	0.3569	1	0.3530	2
	0.2513	3	0.2953	3	0.2876	3	0.2930	3
	0.3253	2	0.3512	2	0.3555	2	0.3540	1
(Isaai et al., 2011)	0.1617	2	0.3163	2	0.3153	2	0.3173	2
	0.6959	1	0.4020	1	0.3996	1	0.3996	1
	0.1425	3	0.2827	3	0.2852	3	0.2831	3
(Jaganathan, Erinjeri & Ker, 2007)	0.4686	1	0.3648	1	0.3658	1	0.3662	1
	0.2796	2	0.3095	3	0.3154	3	0.3073	3
	0.2518	3	0.3284	2	0.3187	2	0.3265	2
(Prašćević & Prašćević, 2016)	0	4	0.0738	4	0.0719	4	0.0735	4
	0.3467	1	0.3392	2	0.3725	1	0.3549	1
	0.3328	2	0.3647	1	0.3169	2	0.3474	2
(Schra, Brar & Kaur, 2013)	0.3205	3	0.2257	3	0.2387	3	0.2242	3
	0.6372	1	0.4206	1	0.4227	1	0.4221	1
	0.2926	2	0.3281	2	0.3297	2	0.3295	2
(Yuen & Henry, 2008)	0.0702	3	0.2516	3	0.2476	3	0.2484	3
	0.2338	2	0.3191	2	0.2920	2	0.2972	2
	0.1952	3	0.2842	3	0.2689	3	0.2842	3
(Tyagi et al., 2017)	0.5711	1	0.3967	1	0.4391	1	0.4186	1
	0.2214	3	0.3282	2	0.3249	2	0.3298	2
	0.4157	1	0.3600	1	0.3607	1	0.3558	1
(Aydoğan, Delice & Papajorgji, 2013)	0.3630	2	0.3127	3	0.3143	3	0.3144	3
	0	3	0.1703	3	0.1848	3	0.1711	3
	0	3	0.1971	2	0.2311	2	0.1955	2
(Aydoğan, Delice & Papajorgji, 2013)	0.8475	1	0.4746	1	0.4288	1	0.4726	1
	0.1525	2	0.1615	4	0.1553	4	0.1608	4
	0.4270	1	0.3113	1	0.3169	1	0.3089	1
(Aydoğan, Delice & Papajorgji, 2013)	0.0898	4	0.2125	3	0.2068	4	0.2115	3
	0.1121	3	0.2088	4	0.2100	3	0.2086	4
	0.3711	2	0.2707	2	0.2663	2	0.2710	2

The global weights of the examples used to compare these four methods are shown in Table 3. When all the examples' global weights are considered, MFLLSM, MFAHP, and GM provide remarkably similar results and ranks, with a few

exceptions. FEA, on the other hand, has the same rankings in just five of fifteen examples. In the example with seven alternatives (Arikan & Dağdeviren, 2013), the top five rows in ranked alternatives are the same for MFAHP, MFLLSM, and GM results. However, for the same example, only the third-ranking value of the FEA method is the same as the other methods.

At this stage, it is important to analyze these similarities and differences observed in Table 3. Therefore, the analysis of variances method (ANOVA) was used to determine the similarity of the results obtained by FEA, MFLLSM, MFAHP, and GM. It determines statistical differences between the mean differences of absolute error values of global weights between the methods. It is important to note that while ANOVA can show that at least two of the groups are significantly different from each other, it cannot show which groups are different from each other. Therefore, a post hoc test was performed to analyze the results further. The equality of variances was first checked using Levene's test, and the normality assumption was checked using the Q-Q plot. The results show that the assumptions of variance homogeneity and normality were violated in all instances. Therefore, Kruskal-Wallis Test for independent samples, given in Table 4, was used, indicating that the methods do not produce significantly similar results based on absolute error values.

But it does not indicate which subgroups are causing this difference. As a result, Mann-Whitney post hoc tests were used to further investigate Kruskal-Wallis results, as given in Table 5. Also, graphical representations of mean differences are presented in Fig.2. Mann-Whitney tests and Fig.2 shows that FEA produces significantly different results than others. When MFLLSM, MFAHP, and GM are compared, it is seen that they have remarkably similar performance.

Table 4. Kruskal-Wallis test result for independent samples.

Total N	342
Test Statistic	187.401 ^a
Degree of Freedom	5
Asymptotic Sig.(2-sided test)	0.000

^a The test statistic is adjusted for ties.

Table 5. Mann-Whitney tests for pairwise comparisons of absolute value of weight differences.

	Test Statistic	Std. Test Statistic	Sig.	Adj. Sig. ^a
MFLLSM-GM vs MFAHP-GM	-42.430	-2.291	0.022	0.329
MFLLSM-GM vs MFLLSM-MFAHP	51.877	2.801	0.005	0.076
MFLLSM-GM vs FEA-MFAHP	171.772	9.275	0.000	0.000*
MFLLSM-GM vs FEA-MFLLSM	175.009	9.450	0.000	0.000*
MFLLSM-GM vs FEA-GM	176.018	9.504	0.000	0.000*
MFAHP-GM vs MFLLSM-MFAHP	9.447	0.510	0.610	1.000
MFAHP-GM vs FEA-MFAHP	129.342	6.984	0.000	0.000*
MFAHP-GM vs FEA-MFLLSM	132.579	7.159	0.000	0.000*
MFAHP-GM vs FEA-GM	133.588	7.213	0.000	0.000*
MFLLSM-MFAHP vs FEA-MFAHP	119.895	6.474	0.000	0.000*
MFLLSM-MFAHP vs FEA-MFLLSM	123.132	6.649	0.000	0.000*
MFLLSM-MFAHP vs FEA-GM	124.140	6.703	0.000	0.000*
FEA-MFAHP vs FEA-MFLLSM	3.237	0.175	0.861	1.000
FEA-MFAHP vs FEA-GM	-4.246	-0.229	0.819	1.000
FEA-MFLLSM vs FEA-GM	-1.009	-0.054	0.957	1.000

Each row tests the null hypothesis that the Samp-1 and Samp-2 dist.s are the same. Asymptotic significances (2-sided tests) are displayed. The significance level is .05.

^a Significance values have been adj. by the Bonferroni correction for multiple tests.

* Absolute value of weight differences is significant at the 0.05 level.

As a result, FEA is not a suitable method to obtain priorities from the fuzzy pairwise comparison matrix in that it can assign an unreasonable zero value as weights of some essential sub-criteria and criteria. These assignments lead to inaccurate results. Because the weights obtained by FEA do not indicate the relative importance of alternatives or criteria, this method is not recommended for use.

On the other hand, we claimed that MFAHP produces results close to the results of MFLLSM, and GM while reducing

the cost of computation to the level of FEA at least. MFLLSM involves complicated calculations, which cannot be performed easily without using professional optimization software packages, to obtain the local fuzzy weights by solving a constrained nonlinear optimization model for each fuzzy comparison matrix although it makes a correct decision and handles all these problems. However, MFAHP, GM, and FEA have a much lower processing load. This situation easily can be seen in Table 6 and in Fig.3.

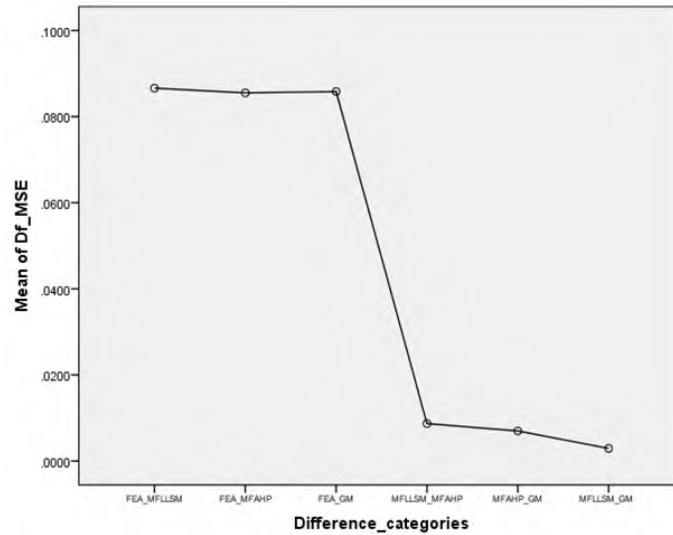


Figure 2. Graphical representation of mean differences of weights.

Table 6. CPU time of the implementations(in seconds)(Sorted in ascending order according to the values of the MFAHP)

Ref No	FEA	MFLLSM	MFAHP	GM
(Isaai et al., 2011)	1.82E-05	1.580142	1.05E-05	1.500E-04
(Büyükoğuzkan & Cifci, 2011)	1.70E-05	3.219868	1.16E-05	1.421E-04
(Prašević & Prašević, 2016)	2.40E-05	1.792041	1.37E-05	2.049E-04
(Dong, Li & Zhang, 2015)	2.11E-05	1.797778	1.39E-05	1.918E-04
(Dong, Li & Zhang, 2015)	2.16E-05	1.793572	1.42E-05	2.103E-04
(Sehra, Brar & Kaur, 2013)	2.24E-05	1.812290	1.43E-05	2.353E-04
(Jaganathan, Erinjeri & Ker, 2007)	3.42E-05	2.298624	1.94E-05	3.886E-04
(Aydoğan, Delice & Papajorgji, 2013)	2.77E-05	2.300450	1.97E-05	3.294E-04
(Arikan & Dagdeviren, 2013)	5.28E-05	6.117062	4.05E-05	5.963E-04
(Celik, Er & Ozok, 2009)	6.23E-05	7.209692	4.24E-05	6.851E-04
(Aydoğan, Demirtas & Dagdeviren, 2015)	7.78E-05	9.279611	5.55E-05	8.203E-04
(Tyagi et al., 2017)	11.06E-05	11.689248	7.62E-05	13.782E-04
(Kahraman, Cebeci & Ruan, 2004)	10.09E-05	13.257331	7.80E-05	14.571E-04
(Yuen & Henry, 2008)	12.28E-05	12.222307	9.49E-05	12.977E-04
(Arif et al., 2021)	26.79E-05	21.035452	16.61E-05	33.436E-04

In addition, the Kruskal-Wallis test for independent samples, presented in Table 7, and Mann-Whitney post hoc tests in Table 8 confirm that there is no difference between MFAHP and FEA in terms of running time among these four approaches, whereas GM is significantly different from MFAHP and FEA, too.

Table 7. Kruskal-Wallis test result for independent samples for running times.

Total N	60
Test Statistic	49.318 ^a
Degree Of Freedom	3
Asymptotic Sig.(2-sided test)	0.000

^a The test statistic is adjusted for ties.

When MFAHP is preferred, the decision process seems to overcome the problems of FEA, and a similar decision is made without the high computational cost as in MFLLSM and GM. All approaches were programmed in C# language.

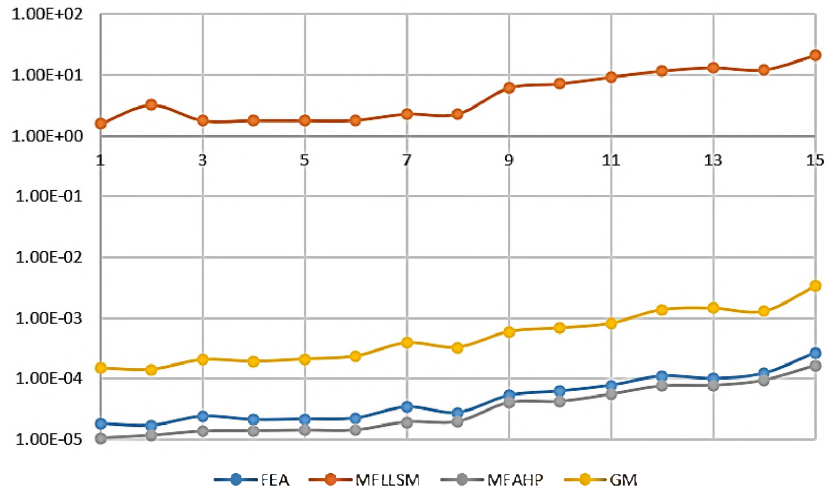


Figure 3. Graphical representation of CPU time of the implementations.

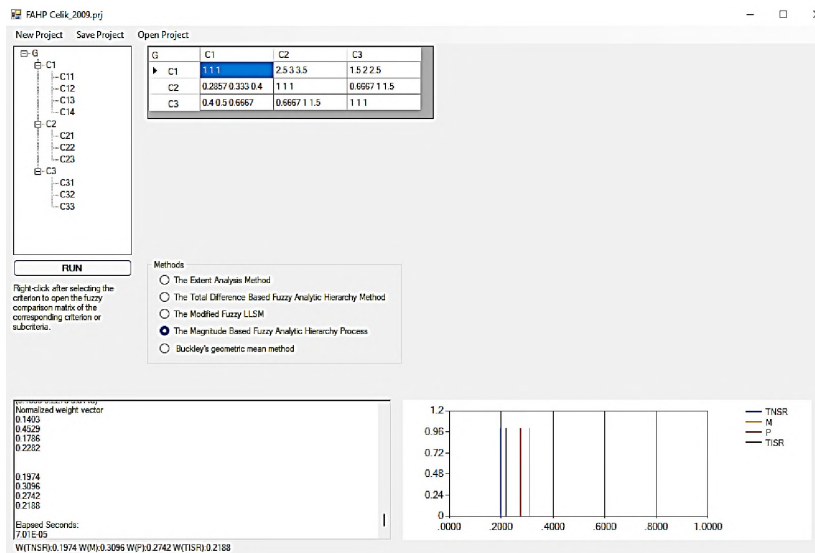


Figure 4. A screenshot of the application.

Table 8. Pairwise comparisons of running times.

	Test Statistic	Std. Test Statistic	Sig.	Adj Sig. ^a
MFAHP-FEA	5.267	0.826	0.409	1.000
MFAHP-GM	-24.333	-3.816	0.000	0.001*
MFAHP-MFLLSM	39.867	6.252	0.000	0.000*
FEA-GM	-19.067	-2.990	0.003	0.017*
FEA-MFLLSM	-34.600	-5.426	0.000	0.000*
GM-MFLLSM	15.533	2.436	0.015	0.089

Each row tests the null hypothesis that the Samp-1 and Samp-2 distributions are the same. Asymptotic significances (2-sided tests) are displayed. The significance level is .05.

^a Significance values have been adj. by the Bonferroni correction for multiple tests.

* Absolute value of weight differences is significant at the 0.05 level.

The CPU of the computer, where all the experiments were performed, is an AMD ThreadRipper 1950x clocked @3.7 GHz and the main memory consists of 64 GB of DDR4 RAM.

An application that has implementations of all above-mentioned algorithms, and the data of all examples can be downloaded from <https://github.com/baristezel/FAHP>. A screenshot of the application is shown in Fig.4. We hope that it may aid in the understanding of the MFAHP method and its contributions.

Also, the global weight values obtained by these methods for the example in (Celik, Er & Ozok, 2009), as shown in Fig.5. In this figure, it can be seen that MFAHP results are very close to the midpoints of MFLLSM results and GM results.

5. DISCUSSION AND CONCLUSION

In this study, instead of the FEA method, which is frequently used in the solution of complex FMCDM problems and leads to incorrect results, the MFAHP method, which allows for obtaining appropriate and efficient results, has been evaluated. FEA, MFLLSM, MFAHP, and GM are compared by using numerical examples. As a result, it has been shown that MFAHP can solve problems arising from FEA and has similar center value results that are more easily calculated compared to MFLLSM and GM results. Even though the GM appears to produce results that are more similar to MFLLSM in some instances, pairwise comparison tests revealed no significant differences between MFAHP, GM, and MFLLSM results. On the other hand, when the running time values are examined, it is seen that the MFAHP gives results in a shorter time than MFLLSM and GM, while there is no statistically significant difference between MFAHP and FEA. While it has been observed that there is no statistically significant difference between MFAHP, GM, and MFLLSM according to global weight differences between methods, the MFAHP approach is both faster and much simpler than others.

It can be argued that the main contribution of this study is to demonstrate, through statistical significance, that the MFAHP method is capable of producing results that are as accurate as those obtained by both the GM method and the more challenging-to-understand and -implement MFLLSM. It was observed that there was no comparison of the methods with MFLLSM in the literature. In this sense, the comparison of the efficiency of the MFAHP method with the MFLLSM, which solves the problem as a constrained nonlinear optimization model and gives accurate results, is a prominent feature of this study. Moreover, statistical analysis demonstrates that the MFAHP method achieves the result values in a shorter time than both the GM and the MFLLSM. In other words, MFAHP method results can be obtained as accurately as MFLLSM and GM results and as fast as FEA. Finally, to aid other researchers in conducting research or applications in this field, we have developed software that calculates all the methods discussed in this study.

Since triangular fuzzy numbers are generally used in comparison matrices in FMCDM problems, the limitation of this study is that the examples used to compare the results of the methods in our study contain only such fuzzy numbers. However, it will be an important contribution to examine the effectiveness of the MFAHP method for other types of fuzzy numbers, especially in future studies.

Overall, the MFAHP algorithm is comparable to the MFLLSM and GM methods in terms of weight calculation accuracy for triangular fuzzy numbers, while also demonstrating superior performance in computational efficiency.

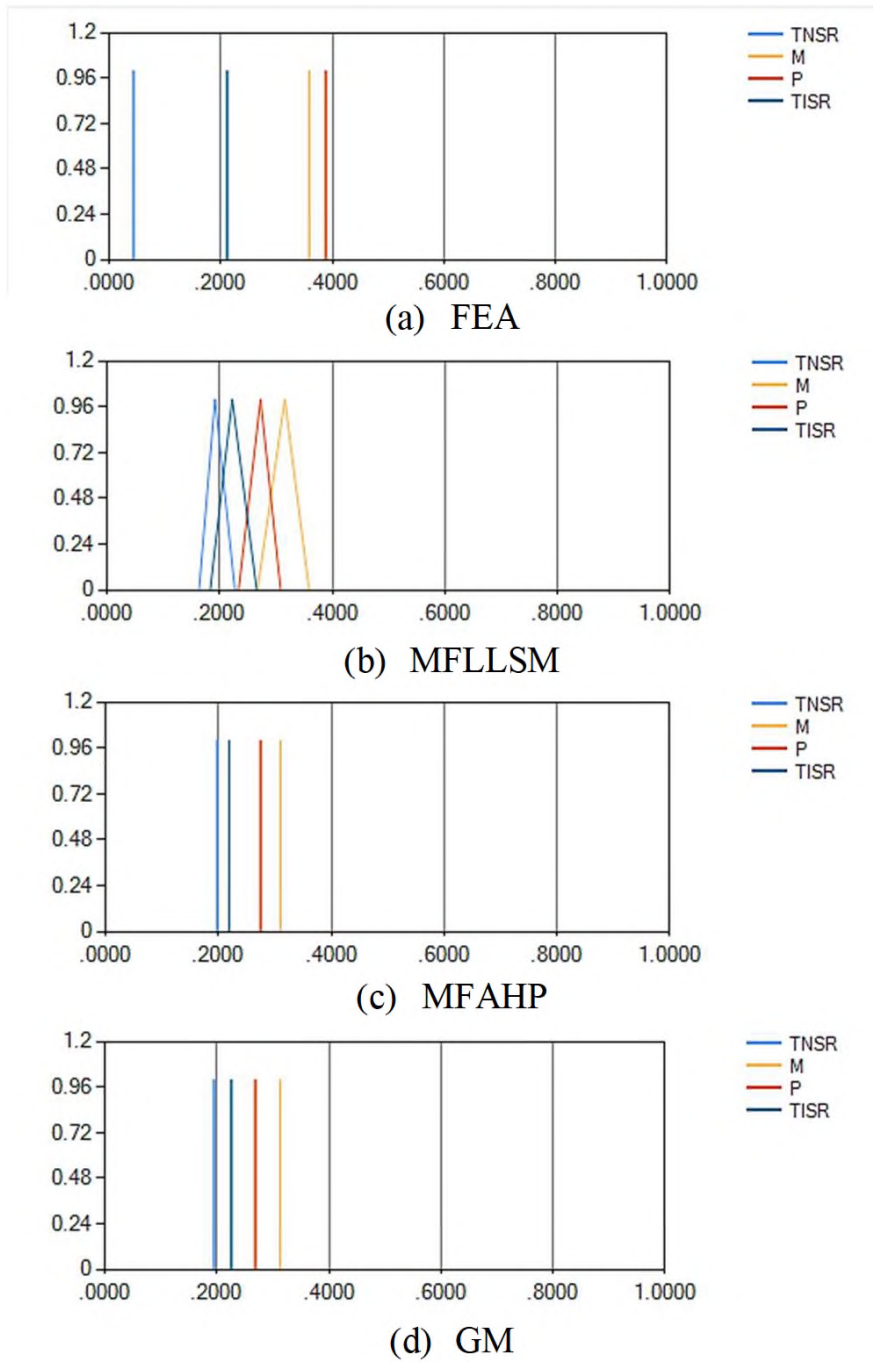


Figure 5. Graphical representation of the global weights for the example in (Celik, Er & Ozok, 2009).

Peer Review: Externally peer-reviewed.

Author Contributions: Conception/Design of Study- B.T.T., A.Ö.K.; Data Acquisition- B.T.T., A.Ö.K.; Data Analysis/Interpretation- B.T.T., A.Ö.K.; Drafting Manuscript- A.Ö.K.; Critical Revision of Manuscript- B.T.T.; Final Approval and Accountability- B.T.T., A.Ö.K.; Material and Technical Support- B.T.T.

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.

ORCID IDs of the authors / Yazarların ORCID ID'leri

Barış Tekin Tezel 0000-0003-4873-7848

Ayşe Övgü Kınay 0000-0001-9908-8652

REFERENCES

- Abbasbandy, S., & Hajjari, T. (2009). A new approach for ranking of trapezoidal fuzzy numbers. *Computers & Math. with Applications*, 57(3), 413-419.
- Ahmed, F., & Kilic, K. (2019). Fuzzy Analytic Hierarchy Process: A performance analysis of various algorithms. *Fuzzy Sets and Systems*, 362, 110-128.
- Amiri, M. P. (2010). Project selection for oil-fields development by using the AHP and fuzzy TOPSIS methods. *Expert systems with applications*, 37(9), 6218-6224.
- Arif, J. M., Ab Razak, M. F., Mat, S. R. T., Awang, S., Ismail, N. S. N., & Firdaus, A. (2021). Android mobile malware detection using fuzzy AHP. *Journal of Information Security and Applications*, 61, 102929.
- Arikan, R., Dağdeviren, M., & Kurt, M. (2013). A fuzzy multi-attribute decision making model for strategic risk assessment. *International Journal of Computational Intelligence Systems*, 6(3), 487-502.
- Aydogan, E. K., Delice, E. K., & Papajorgji, P. (2013). An effective approach for evaluating usability of Web sites. In *Enterprise Business Modeling, Optimization Techniques, and Flexible Information Systems* (pp. 97-107). IGI Global.
- Aydogan, E. K., Demirtas, O., & Dagdeviren, M. (2015). A New Integrated Fuzzy Multi-Criteria Decision Model for Performance Evaluation. *Business and Management Studies*, 1(1), 38-55.
- Ballı, S., & Karasulu, B. (2013). Bulanık karar verme sistemlerinde paralel hesaplama. *Pamukkale Üniversitesi Mühendislik Bil. Dergisi*, 19(2), 61-67.
- Boender, C. G. E., De Graan, J. G., & Lootsma, F. (1989). Multi-criteria decision analysis with fuzzy pairwise comparisons. *Fuzzy Sets and Systems*, 29(2), 133-143.
- Bortolan, G., & Degani, R. (1985). A review of some methods for ranking fuzzy subsets. *Fuzzy sets and systems*, 15(1), 1-19.
- Buckley, J. J. (1985). Fuzzy hierarchical analysis. *Fuzzy sets and systems*, 17(3), 233-247.
- Büyükköçkan, G., Kahraman, C., & Ruan, D. (2004). A fuzzy multi-criteria decision approach for software development strategy selection. *International journal of general systems*, 33(2-3), 259-280.
- Büyükköçkan, G., Çifçi, G., & Güleriyüz, S. (2011). Strategic analysis of healthcare service quality using fuzzy AHP methodology. *Expert systems with applications*, 38(8), 9407-9424.
- Celik, M., Er, I. D., & Ozok, A. F. (2009). Application of fuzzy extended AHP methodology on shipping registry selection: The case of Turkish maritime industry. *Expert Systems with Applications*, 36(1), 190-198.
- Chutia, R., & Chutia, B. (2017). A new method of ranking parametric form of fuzzy numbers using value and ambiguity. *Applied Soft Comp.*, 52, 1154-1168.
- Dong, M., Li, S., & Zhang, H. (2015). Approaches to group decision making with incomplete information based on power geometric operators and triangular fuzzy AHP. *Expert Systems with Applications*, 42(21), 7846-7857.
- Dweiri, F., Kumar, S., Khan, S. A., & Jain, V. (2016). Designing an integrated AHP based decision support system for supplier selection in automotive industry. *Expert Systems with Applications*, 62, 273-283.
- Isaai, M. T., Kanani, A., Tootoonchi, M., & Afzali, H. R. (2011). Intelligent timetable evaluation using fuzzy AHP. *Expert systems with App.*, 38(4), 3718-3723.
- Jaganathan, S., Erinjeri, J. J., & Ker, J. I. (2007). Fuzzy analytic hierarchy process based group decision support system to select and evaluate new manufacturing technologies. *The International Journal of Advanced Manufacturing Technology*, 32(11), 1253-1262.
- Kahraman, C., Cebeci, U., & Ruan, D. (2004). Multi-attribute comparison of catering service companies using fuzzy AHP: The case of Turkey. *International journal of production economics*, 87(2), 171-184.
- Kınay, A. O., & Tezel, B. T. (2022). Modification of the fuzzy analytic hierarchy process via different ranking methods. *International Journal of Intelligent Systems*, 37(1), 336-364.
- Korkmaz, I., Gökçen, H., & Çetinyokuş, T. (2008). An analytic hierarchy process and two-sided matching based decision support system for military personnel assignment. *Information Sciences*, 178(14), 2915-2927.
- Kubler, S., Robert, J., Derigent, W., Voisin, A., & Le Traon, Y. (2016). A state-of-the-art survey & testbed of fuzzy AHP (FAHP) applications. *Expert Systems with Applications*, 65, 398-422.

- Kwiesielewicz, M. (1996). The logarithmic least squares and the generalized pseudoinverse in estimating ratios. *European Journal of Operational Research*, 93(3), 611-619.
- Liu, Y., Eckert, C. M., & Earl, C. (2020). A review of fuzzy AHP methods for decision-making with subjective judgements. *Expert Systems with Applications*, 161, 113738.
- Saaty, T. L. (1977). A scaling method for priorities in hierarchical structures. *Journal of mathematical psychology*, 15(3), 234-281.
- Saaty, T.L. (1980) *The Analytic Hierarchy Process: Planning, Priority Setting, Resources Allocation*. McGraw-Hill, New York.
- Vaidya, O. S., & Kumar, S. (2006). Analytic hierarchy process: An overview of applications. *European Journal of operational research*, 169(1), 1-29.
- Xu, Z., & Liao, H. (2013). Intuitionistic fuzzy analytic hierarchy process. *IEEE transactions on fuzzy systems*, 22(4), 749-761.
- Wang, Y. M., Elhag, T. M., & Hua, Z. (2006). A modified fuzzy logarithmic least squares method for fuzzy analytic hierarchy process. *Fuzzy Sets and systems*, 157(23), 3055-3071.
- Chang, D. Y. (1996). Applications of the extent analysis method on fuzzy AHP. *European Journal of Operational Research*, 95(3), 649-655.
- Mardani, A., Jusoh, A., & Zavadskas, E. K. (2015). Fuzzy multiple criteria decision-making techniques and applications—Two decades review from 1994 to 2014. *Expert Systems with Applications*, 42(8), 4126-4148.
- Van Laarhoven, P. J., & Pedrycz, W. (1983). A fuzzy extension of Saaty's priority theory. *Fuzzy sets and Systems*, 11(1-3), 229-241.
- Ruoning, X., & Xiaoyan, Z. (1996). Fuzzy logarithmic least squares ranking method in analytic hierarchy process. *Fuzzy Sets and Systems*, 77(2), 175-190.
- Wang, Y. M., Luo, Y., & Hua, Z. (2008). On the extent analysis method for fuzzy AHP and its applications. *European Journal of Operational Research*, 186(2), 735-747.
- Zadeh, L. A. (1965). Fuzzy sets. *Information and control*, 8(3), 338-353.
- Wang, X., & Kerre, E. E. (2001). Reasonable properties for the ordering of fuzzy quantities (I). *Fuzzy sets and systems*, 118(3), 375-385.
- Wang, X., & Kerre, E. E. (2001). Reasonable properties for the ordering of fuzzy quantities (II). *Fuzzy sets and systems*, 118(3), 387-405.
- Wang, Y. M., & Elhag, T. M. (2006). On the normalization of interval and fuzzy weights. *Fuzzy sets and systems*, 157(18), 2456-2471.
- Prašćević, N., & Praščević, Ž. (2016). Application of fuzzy AHP method based on eigenvalues for decision making in construction industry. *Tehnički vjesnik/Technical Gazette*, 23(1), 57-64.
- Sehra, S. K., Brar, D., Singh, Y., & Kaur, D. (2013). Multi criteria decision making approach for selecting effort estimation model. arXiv preprint arXiv:1310.5220.
- Tyagi, S., Agrawal, S., Yang, K., & Ying, H. (2017). An extended Fuzzy-AHP approach to rank the influences of socialization-externalization-combination-internalization modes on the development phase. *Applied Soft Computing*, 52, 505-518.
- Yuen, K. K., & Lau, H. C. (2008). Software vendor selection using fuzzy analytic hierarchy process with ISO/IEC 9126. *IAENG International journal of computer science*, 35(3).

How cite this article

Tekin Tezel, B., Kinay, A.O. (2023). Performance evaluation of Magnitude-Based Fuzzy Analytic Hierarchy Process (MFAHP) method. *Acta Infologica*, 7(2), 293-307. <https://doi.org/10.26650/acin.1224496>

Vision-Based Amateur Drone Detection: Performance Analysis of New Approaches in Deep Learning

Görüntü Tabanlı Amatör Drone Tespiti: Derin Öğrenmede Yeni Yaklaşımların Performans Analizi

Ahmet Aydın¹ , Tarık Talan² , Cemal Aktürk² 

¹(Lect.), Gaziantep Islam Science and Technology University, Vocational School of Technical Sciences, Gaziantep, Türkiye

²(Assist. Prof. Dr.), Gaziantep İslam Bilim ve Teknoloji Üniversitesi, Faculty of Engineering and Natural Sciences, Computer Engineering Department, Gaziantep, Türkiye

Corresponding author : Cemal AKTÜRK

E-mail : cemal.akturk@gibtu.edu.tr

ABSTRACT

Interest in unmanned aerial vehicles (UAVs) has increased significantly. UAVs capable of autonomous operations have expanded their application areas as they can be easily deployed in various fields. The expansion of UAVs' areas of operation also brings safety issues. Although legally prohibited places for UAV flights are defined, measures should be taken to detect violations. This study tested recently proposed methods that are used to detect objects from images on UV images, and their performances were discussed. We tested the models on a new dataset named GDrone that we created by collecting various images of drones. Two tested models, YOLOv6 and YOLOv7, have never been tested with a drone dataset. According to the experimental tests, the most successful model was YOLOv7 architecture, and its mAP (mean Average Precision) was 95.8% on GDrone dataset.

Keywords: Unmanned aerial vehicles, amateur drone detection, convolutional neural networks, UAW dataset

ÖZ

İnsansız hava araçlarına (İHA) olan ilgi önemli ölçüde artmıştır. Otonom çalışabilen İHA'lar, çeşitli alanlara kolaylıkla konuşlandırılabilmeleri nedeniyle uygulama alanlarını genişletmiştir. İHA'ların faaliyet alanlarının genişlemesi, aynı zamanda güvenlik sorunlarını da beraberinde getirmektedir. İHA uçuşları için yasaklanmış olan yerler yasal olarak tanımlanmış olsa da ihlallerin tespitine yönelik tedbirlerin alınması gerekmektedir. Bu çalışmada, ultraviyole görüntüler üzerinde nesnelerin tespit edilmesi için kullanılan ve son zamanlarda önerilen yöntemler test edilmiş ve performansları tartışılmıştır. Modelleri, çeşitli drone görüntülerini toplayarak oluşturduğumuz GDrone isimli yeni bir veri seti üzerinde test ettik. Test edilen YOLOv6 ve YOLOv7 modelleri daha önce bir drone veri seti ile test edilmemiştir. Deneysel testlere göre en başarılı model YOLOv7 mimarisidir ve GDrone veri kümesindeki mAP (ortalama hassasiyet) değeri %95,8 olarak belirlendi.

Anahtar Kelimeler: İnsansız hava araçları, amatör drone tespiti, evrişimli sinir ağları, İHA veri seti

Submitted : 29.03.2023

Revision Requested : 17.08.2023

Last Revision Received : 14.09.2023

Accepted : 22.09.2023

Published Online : 21.11.2023



This article is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0)

1. INTRODUCTION

Unmanned aerial vehicles (UAVs) are small, hard-to-track vehicles that pose a security risk and expose experts and authorities working in this field to various security risks (Vattapparamban et al. 2016). UAVs, which were first used in the entertainment industry or for filming purposes, are used in traffic monitoring, photography, communication (Khan, Park, and Gonzalez 2017; Tang et al. 2017)), disaster relief (Al-Hourani, Kandeepan, and Jamalipour 2015) and autonomous driving. Thanks to their capabilities and programmability, they are also used in vegetation monitoring (Onishi and Ise 2021). On the contrary, the UAV can also be used for maliciously damaging purposes. These purposes include military espionage activities intelligence gathering, physical attacks on public facilities, infrastructure networks, or people in crowded environments. In addition, terrorist elements can use drones to transport weapons, ammunition, explosives, and even radioactive material. This shows that even the physical presence of drones can have dire consequences. The use of UAVs for espionage purposes can also pose a serious threat to the security of states, as they provide images from areas that may be considered sensitive points of a country (military and police units) (Al-Emadi et al. 2019). Because these security threats are not negligible or acceptable risks, detecting the presence and identity of a drone in the air is an unavoidable problem that must be addressed in terms of public safety. For this reason, the importance of UAV detection systems is increasing day by day. The development of malicious drone detection systems, anti-drone systems, addresses the problems of detecting drones, determining the type of drone, and locating the drone. Anti-drone systems can also ensure that the drone conducting a malicious flight is neutralized by force of arms or by disrupting its signal (Carter and Schwartz 2010).

Drones leave three different traces when hovering. These traces are: the sounds they make, the radio frequencies, and the physical images they emit to find their flight path or to receive commands. Sounds (Liu et al. 2017), images (Pham and Nguyen 2020) or radio frequencies (Al-Sa'd et al. 2019) can be used for drone detection. Sound-based drone detection is an altitude-dependent process that may not be detected if the drone is flying at a high altitude. However, ambient noise (birds, wind noise, etc.) may suppress the sound of the drone and prevent it from being detected. Radio frequency detection systems, on the other hand, are a method that uses signals between the drone's control unit and the drone. However, since autonomous drones do not have a separate control unit, such detection is impossible. In addition, many communication devices use radio frequencies. This can make the detection of a drone using radio frequencies very complicated. Drone detection systems on images have convincing elements. Another difficulty in detecting drones on images is distinguishing them from other objects that may be in the air, such as kites or birds. Because they may be various backgrounds that make the drones detection more difficult.

In this study, new methods for UAV detection on captured images were tested and their performances were discussed. Since images contain various backgrounds, in this way, the success of existing and new models in detection a drone under challenging conditions was measured. To sum up, the contributions of this study to the literature are listed below:

- A new dataset containing 600 drone images was proposed.
- The performance of the latest architectures such as YOLOv7, YOLOv6 was measured for the first time on a drone dataset and compared with other methods.
- An object detector with the fewest parameters, the lowest computational cost, the highest efficiency and performance was proposed.

2. LITERATURE REVIEW

Object detection is one of the most important tasks that occupies a wide space in the field of computer vision. Despite advances in deep learning, localizing an object is a very challenging process. While existing solutions use machine learning methods based on manual feature extraction, feature maps are autonomously extracted using convolutional neural networks in conjunction with deep learning studies (Aydin, Salur, and Aydin 2021). With the advent of modern object detectors, studies in the field of object detection have come into focus. These object detectors are divided into two classes: single-stage object detectors and two-stage object detectors. Object detectors such as R-CNN and Fast R-CNN (Girshick 2015), fall into the two-stage class, while object detectors such as YOLO (Redmon et al. 2016) and SSD (Liu et al. 2016), are single-stage object detectors. Two-stage object detectors make a site proposal in the first step and perform object inspection in those regions in the second step. However, object detectors such as YOLO and SSD operate less expensively, faster, because the region recommendation step is eliminated. Real-time drone detection systems are needed because drones are small and difficult to detect and must be detected quickly to prevent violations. Using single-stage object detectors for these real-time systems is more advantageous because of their speed.

Because drones are valuable tools, they can serve practical purposes in many professions, but this does not prevent the malicious use of drones. To prevent these malicious purposes, drone detection is of great importance. There are various

studies on drone detection using combination of sound, audio and video, radio frequency, audio, and thermal imaging. However, these studies require some pre-processing. Studies on image data have focused on real-time systems. In this way, detection operations can be performed without preprocessing. For this reason, this section focuses on literature studies on image data.

With their modification of the YOLOv4 (Wang, Bochkovskiy, and Liao 2023) algorithm, (Liu et al. 2021) have developed a highly successful detection detector suitable for real-time use. For the data collected in their study, their model achieved the highest mean average precision (mAP) of 93.6%. The researchers claimed that the model runs at 43 FPS (frames per second). In another study, (Zheng et al. 2021) trained with different architectures for the dataset they created. By using the Grid R-CNN architecture, the researchers achieved a success rate of 90.1%. Similarly, (Sahin and Ozer 2021) performed a detection model on a dataset with 10 classes. In the model, they used the YOLOv3 architecture. In the study by (Behera and Raj 2020) the drone data set was analyzed using the YOLOv3 model. As the best result over 150 epochs, the researchers achieved 74% success in their studies. Another study was conducted by (Lee, La, and Kim 2018), achieved an 89% success rate in drone detection using the Haar Feature-based Cascade Classifier, which they adapted to their object detection model. In the another study, (Nalamati et al. 2019) used both Single Shot Detector (SSD) models and various CNN-based architectures such as ResNet-101 and Inception with Faster-RCNN to detect drones in long-range surveillance video. Due to sparse data, the best success rate of researchers using transfer learning was 0.49 mAP in experiments with Faster-RCNN. Müller (2017) investigated the suitability of image differentiation and background subtraction techniques for extracting and examining candidate regions in video data obtained from static and moving cameras (Müller 2017).

3. METHOD

3.1. Data Collection

The proposed GDrone dataset contains 600 images with different backgrounds obtained from various sources. Since these images are from the areas where drones are used in daily life, it is expected to reflect realistic results. The images in the dataset are sampled from various videos. Each image has a size of 416x416. All images in the dataset were carefully manually labeled by a research group specifically set up for this task. The drone dataset has a more difficult structure than other datasets, as it contains images of different daylight levels taken in different environments and drone images located at different distances. Emphasis is placed on the widely influential DJI phantom 4. Some sample images can be seen in Fig 1. The dataset created in the study can be accessed by other researchers using the link <https://github.com/ahmetmericaydin/GDrone-Dataset>

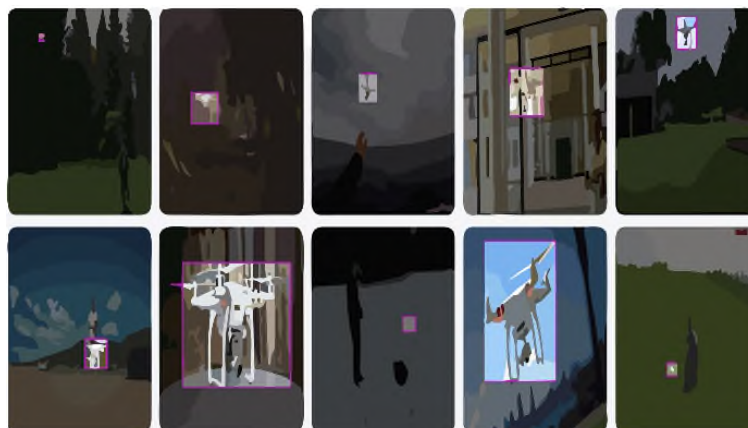


Figure 1. Some sample of images from the GDrone dataset

The GDrone dataset contains drone images at different distances and under different lighting conditions. In addition, the GDrone dataset contains different backgrounds such as sea and sky. The proposed dataset contains different types of drone images. Since the structures or sizes of drones may be different, this is one of the challenges in drone detection. One of the main purposes of creating a dataset is to evaluate different datasets. Considering these aspects, the GDrone dataset can be classified as a suitable dataset for evaluating the performance of drone detection methods.

In the study, 70% of the data was used for training, 20% for validation, and 10% for testing. The algorithms were run on a remote server with a Tesla T4 graphics card and 16 GB RAM.

3.2. Data Augmentation

Data augmentation is one of the common methods to improve accuracy variance in training data. Applying translation and rotation operations to the data set ensures that the existing data set is tripled. Various methods can be applied in data augmentation and the size of the dataset can be further increased. However, since the goal of the study was to achieve effective and efficient results with a short training time, the data augmentation method was used less frequently. It can be observed that the data augmentation applied to the GDrone dataset leads to an increase in the mean average precession (mAP) value. By applying the augmentation procedure to the training dataset, a total of 1200 images were obtained in the training dataset. No augmentation is applied to the test data.

3.3. YOLO

The YOLO (You Only Look Once) architecture, developed with the rapid developments in deep learning in recent years, is one of the effective methods used and developed for object detection. YOLO, developed for real-time object detection, continues to be further developed in different variants. There are versions developed by many scientists due to the various changes in the architecture and improvements in the number of parameters, high performance and performance. The fast operation of the YOLO algorithm is due to the fact that it can estimate the class and weight of the objects on the image by scanning the image once. The YOLO algorithm creates bounding boxes to identify objects on the image. While doing this, the midpoint where the object intersects in the image is used. In this way, bounding boxes are obtained.

Height, width, class and frame centers are created to define a bounding frame (box). Each bounding box consists of certain parameters. Each box also creates its own prediction score.

$$y = (pc, bx, by, bh, bw, c) \tag{1}$$

The bx given in Equation 1 represents the center of the by -frame. C indicates the desired classes. bh and bw indicate the height and width of the frame. As specified in Equation 1, the object belonging to the class is included in the frame, and the object is determined according to the class to which it belongs. If the number of overlapping boxes is more than one, the correct boundaries are drawn based on the maximum number of overlaps IoU (Intersection over Union) and inserted into the boxes. If the prediction score is one, the boundaries of the object were predicted correctly. However, depending on the overflow rate at the borders, the percentage overflow rate is subtracted as an error margin and the success rate is determined accordingly.

3.4. Overview of YOLOv7

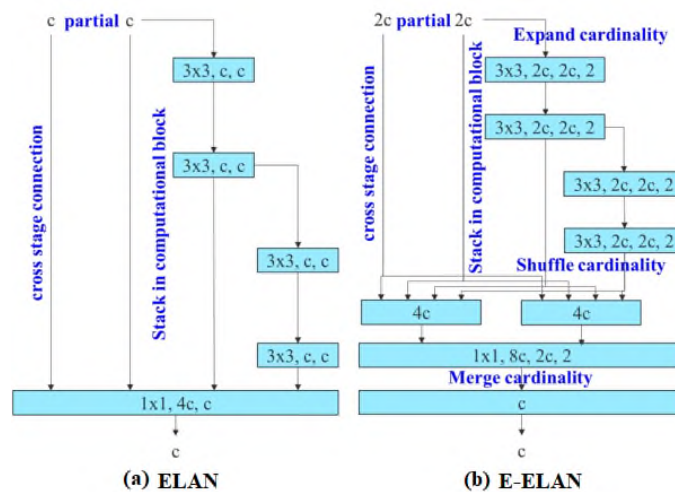


Figure 2. ELAN and E-ELAN architectures ((Wang et al. 2022))

The latest object detector, YOLOv7 (Wang et al. 2023), outperforms other known object detection models in both

speed and accuracy with the dataset COCO. The YOLOv7 architecture is based on the extended (E-ELAN) architecture based on ELAN (Wang, Liao, and Yeh 2022). Figure 2 shows the architectures ELAN and E-ELAN.

Figure 2 shows the improvements made to the architecture of ELAN. The researchers in (Wang et al. 2022) extended the architecture of ELAN and applied it to the YOLOv7 model. In this way, they have improved the performance. In contrast, the E-ELAN architecture proposed in YOLOv7 architecture uses the cardinality of extend, shuffle and combine to improve the learning capability of the network. ELAN, on the other hand, provides performance by simply changing the architecture in the computational block. It uses RepConv (Ding et al. 2021) for merging and scaling. The YOLOv7 architecture is an architecture in which different architectures are developed and adapted. Comparisons of the YOLOv7 architecture with other object detectors are presented in (Wang et al. 2022).

4. METRICS

To measure the performance of the proposed model mAP is used. Since all images in the dataset contain drones, we used mAP, which is commonly used to evaluate whether any searched object is detected. Object detection models also make classification; classification metrics are also included. mAP is a widely used performance metric in the field of computer vision and information retrieval, particularly for tasks such as object detection and image retrieval. mAP is calculated by Equation (3).

$$Precision(P) = \frac{TP}{(FP + TP)} \quad (2)$$

$$mAP = \frac{1}{n} \sum_{k=1}^{k=n} AP_k \quad (3)$$

where TP stands for true positive, TN for true negative, FP for false positive, FN for false negative, n is the number of classes and AP_k is the average precision of class k.

5. EXPERIMENTS RESULTS

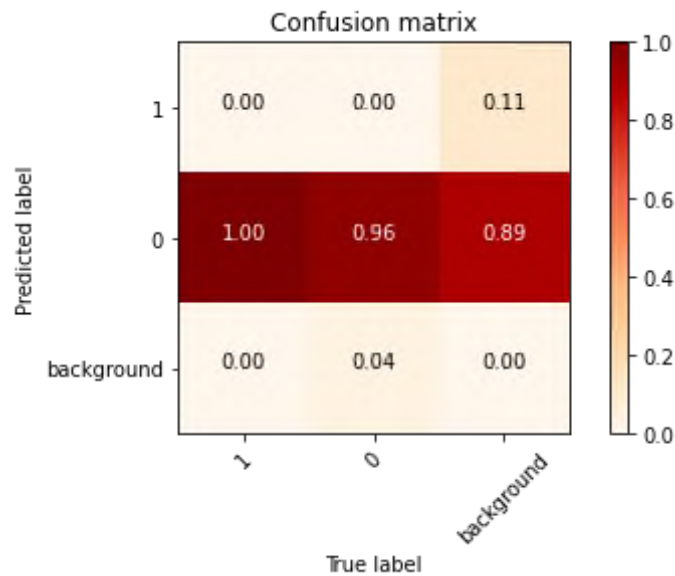


Figure 3. Dataset showing the performance of the YOLOv7 model on the GDrone dataset

This section shows how successful the YOLOv7 architecture is in using image data from the GDrone dataset. The YOLOv7 model has 415 layers. In the experimental tests, images of size 640x640 were applied to the model as input. As parameters, the model was run with 8 batch sizes and 100 cycles. The experimental results showed that the YOLOv7 model achieved a mAP accuracy of 95.8% for the GDrone dataset. The actual complexity values shown in Figure 3 illustrate the performance on a known test data set.

When Figure 3 is examined, it can be observed that approximately 96% of the estimated data is correct, but there is an almost 4% margin of error due to the dataset having various backgrounds and consisting of different drones. The findings show that the dataset is an effective dataset in measuring the effect of a model. In Figure 4, other performance measures obtained from the study of the model are given.

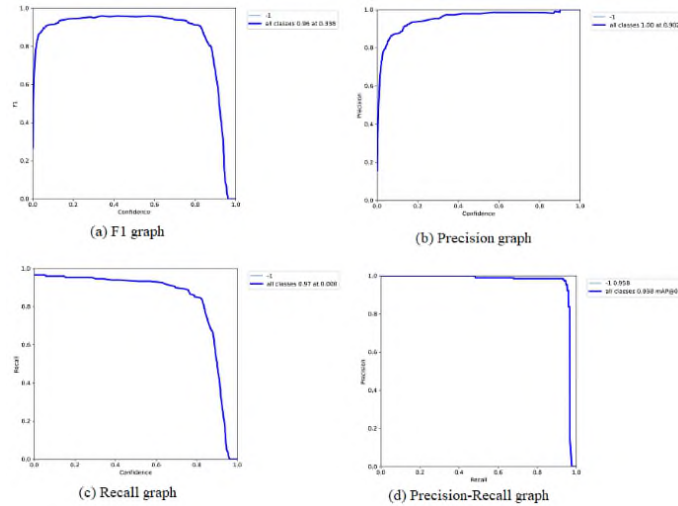


Figure 4. a) Represents the model’s F1 graph b) Represents the model’s precision graph c) Represents the model’s recall graph d) Represents the model’s precision-recall graph

Figure 4 (a) shows the F1 success plot of the YOLOv7 model. The F1 success rate of the YOLOv7 model is 96%. The F1 plot shown represents the harmonic mean of the precision value of the model and the recall value, and is a value used when comparing models. The recall value is the value that indicates how many data were positively predicted based on how many of the data were defined as positive. The precision value in Figure 5 (b) produced results greater than 90%. The precision value indicates how many of the positively predicted values are actually positive. Figure 5 (c) shows a recall value of 97%. The recall value is the value that indicates how many of the data presented as positive were actually predicted to be positive. In Figure 5 (d), the YOLOv7 model provides a result with an accuracy of 95.8% mAP. This result shows the average sensitivity. The YOLOv7 model has shown that it can be applied to an embedded real-time system and perform with 95.8% mAP and 52 FPS.

In different model studies with the same parameter values on the same data set, the success rate (mAP) for the YOLOS (Fang et al. 2021) model is 90.1%, for the YOLOv6 model 91.8%, and for the YOLOv5 model 89.1%. Table I shows the performance of the models in comparison.

Table 1. Performance Plot of the Latest Object Detectors Implemented On Gdrone

Model	Image Size	GFlops	Parameter (M)	mAP(%)
YOLOv5	416x416	7.3	16.8	94.1
YOLOS	416x416	-	6.5	90.1
YOLOv6	416x416	18.62	17.19	91.8
YOLOv7	640x640	103.2	36.48	95,8

Table 1 shows that the YOLOv5 model achieves the least success, while the YOLOS model achieves more successful results than the YOLOv5 model, even though it has the fewest parameters. Although the YOLOv6 model is more successful than the YOLOv5 and YOLOS models, it appears to be less powerful compared to the real-time performance of the other models. However, it was found and shown that the YOLOv7 model performs better than the other models tested in terms of both performance and success. Table II compares the success of the different models in similar studies.

When the results in Table 2 are examined, the successes of the architectures tested on different datasets are shared. In (Zheng et al. 2021), the YOLOv3 model gave the worst result when tested on the Det-Fly dataset. Immediately after, in

Table 2. Performance of the Models on Different Datasets

Model	Image Size	GFlops	Parameter (M)	mAP(%)
YOLOv5	416x416	7.3	16.8	94.1
YOLOS	416x416	-	6.5	90.1
YOLOv6	416x416	18.62	17.19	91.8
YOLOv7	640x640	103.2	36.48	95,8

source (Behera and Raj 2020), the success of the YOLOv3 model is reported to be 74%. Source (Liu et al. 2021) states that the Pruned YOLOv4 model works effectively and the author obtained results of over 94% when experimenting with his own dataset. The YOLOv5 model used in the study numbered (Zhao et al. 2021) can be considered the most effective study in the literature to date. However, considering different improvements and studies with different datasets, it is observed that the YOLOv7 architecture achieves the best result with a mAP success rate of 95.8%. Since the YOLOv7 architecture has not yet been used on a similar dataset in the literature, it is not possible to compare YOLOv7. However, the dataset we have is open source and is made available to researchers for study and comparison. Figure 5 shows the test results of YOLOv7.

**Figure 5.** Some examples of test results of the YOLOv7 model

As can be seen from Figure 5, factors such as different backgrounds and the distance of the drone from the camera were considered in terms of suitability for real-world conditions. Since drones are expected to be a threat in any region, the diversity of the dataset is one of the most important elements for the study. The fact that the proposed model works with 95.8% success and 52 FPS shows that it can work in a real-time system.

6. CONCLUSION

Artificial intelligence technology, which is widely used today, is used as a driving force for revolutionary changes in many fields such as production, education, health and defense industry (Talan 2021). With this technology, it is not only limited to increasing computerized computing capacity, but also used to develop human-like thought and behavior processes. Analyzing people's emotions (Korkmaz, Aktürk, and Talan 2023), detecting perceived objects, facial recognition and acting like a human are among these processes. Especially applications such as object recognition, face recognition, biometric recognition (Aktürk, Aydemir, and Rashid 2023) and classification are frequently used in the health and security sectors. In this study, performance analysis of image processing algorithms was carried out by focusing on the detection of UAVs. Because it is possible to say that the increasing prevalence of UAVs in parallel with the development of technology brings with it some security threats. In other words,

It is an unavoidable fact that the development of UAVs and such vehicles carries risks and may pose serious risks to states such as various irregularities, illegal activities and ultimately public safety. Examples of serious security problems are the transport of prohibited substances, the transfer of explosives to target areas, or espionage activities. Therefore, UAVs or drones must be detected effectively. In this study, the performance of new object detectors is evaluated and a

high-sensitivity model is proposed for real-time applications. Various models were created with the data obtained from different sources and the success of these models were compared. The YOLOv7 model has an improvement rate of 1.7% compared to the YOLOv5 model. The proposed method appears to be an effective and accurate tool. In this study, a drone dataset was added as a contribution to the iteration. In addition, the YOLOv7 and YOLOv6 models were tested on a drone dataset for the first time, the performances of these methods were compared and the results were presented. It has been shown that the YOLOv7 model can be used to defend against incoming threats, especially in organizations with high security needs and in regions exposed to threats.

Peer Review: Externally peer-reviewed.

Author Contributions: Conception/Design of Study- A.A., T.T., C.A.; Data Acquisition- A.A., T.T., C.A.; Data Analysis/Interpretation- A.A., T.T., C.A.; Drafting Manuscript- A.A., T.T., C.A.; Critical Revision of Manuscript- A.A., T.T., C.A.; Final Approval and Accountability- A.A., T.T., C.A.; Material and Technical Support- A.A., T.T., C.A.; Supervision- A.A., T.T., C.A.

Conflict of Interest: : The authors have no conflict of interest to declare.

Grant Support: This study was supported by the BAP (The Scientific Research Projects) under Grant No: 2021-FM-02.

REFERENCES

- Aktürk, Cemal, Emrah Aydemir, and Yasr Mahdi Hama Rashid. 2023. "Classification of Eye Images by Personal Details with Transfer Learning Algorithms." *Acta Informatica Pragensia* 12(1):32–53.
- Al-Emadi, Sara, Abdulla Al-Ali, Amr Mohammad, and Abdulaziz Al-Ali. 2019. "Audio Based Drone Detection and Identification Using Deep Learning." Pp. 459–64 in 2019 *15th International Wireless Communications & Mobile Computing Conference (IWCMC)*. IEEE.
- Al-Hourani, Akram, Sithamparanathan Kandeepan, and Abbas Jamalipour. 2015. "Stochastic Geometry Study on Device-to-Device Communication as a Disaster Relief Solution." *IEEE Transactions on Vehicular Technology* 65(5):3005–17.
- Al-Sa'd, Mohammad F., Abdulla Al-Ali, Amr Mohamed, Tamer Khattab, and Aiman Erbad. 2019. "RF-Based Drone Detection and Identification Using Deep Learning Approaches: An Initiative towards a Large Open Source Drone Database." *Future Generation Computer Systems* 100:86–97.
- Aydin, Ahmet, Mehmet Umut Salur, and İlhan Aydin. 2021. "Fine-Tuning Convolutional Neural Network Based Railway Damage Detection." Pp. 216–21 in *IEEE EUROCON 2021-19th International Conference on Smart Technologies*. IEEE.
- Behera, Dinesh Kumar, and Arockia Bazil Raj. 2020. "Drone Detection and Classification Using Deep Learning." Pp. 1012–16 in 2020 *4th International Conference on Intelligent Computing and Control Systems (ICICCS)*. IEEE.
- Carter, Ashton B., and David N. Schwartz. 2010. *Ballistic Missile Defense*. Brookings Institution Press.
- Ding, Xiaohan, Xiangyu Zhang, Ningning Ma, Jungong Han, Guiguang Ding, and Jian Sun. 2021. "Repvgg: Making Vgg-Style Convnets Great Again." Pp. 13733–42 in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.
- Fang, Yuxin, Bencheng Liao, Xinggang Wang, Jiemin Fang, Jiyang Qi, Rui Wu, Jianwei Niu, and Wenyu Liu. 2021. "You Only Look at One Sequence: Rethinking Transformer in Vision through Object Detection." *Advances in Neural Information Processing Systems* 34:26183–97.
- Girshick, Ross. 2015. "Fast R-Cnn." Pp. 1440–48 in *Proceedings of the IEEE international conference on computer vision*.
- Khan, Gul Zameen, Eun-Chan Park, and Ruben Gonzalez. 2017. "M 3-Cast: A Novel Multicast Scheme in Multi-Channel and Multi-Rate Wifi Direct Networks for Public Safety." *IEEE Access* 5:17852–68.
- Korkmaz, Adem, Cemal Aktürk, and Tarik Talan. 2023. "Analyzing the User's Sentiments of ChatGPT Using Twitter Data." *Iraqi Journal For Computer Science and Mathematics* 4(2):202–14.
- Lee, Dongkyu, Woong Gyu La, and Hwangnam Kim. 2018. "Drone Detection and Identification System Using Artificial Intelligence." Pp. 1131–33 in 2018 *International Conference on Information and Communication Technology Convergence (ICTC)*. IEEE.
- Liu, Hansen, Kuangang Fan, Qinghua Ouyang, and Na Li. 2021. "Real-Time Small Drones Detection Based on Pruned Yolov4." *Sensors* 21(10):3374.
- Liu, Hao, Zhiqiang Wei, Yitong Chen, Jie Pan, Le Lin, and Yunfang Ren. 2017. "Drone Detection Based on an Audio-Assisted Camera Array." Pp. 402–6 in 2017 *IEEE Third International Conference on Multimedia Big Data (BigMM)*. IEEE.
- Liu, Wei, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. 2016. "Ssd: Single Shot Multibox Detector." Pp. 21–37 in *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*. Springer.
- Müller, Thomas. 2017. "Robust Drone Detection for Day/Night Counter-UAV with Static VIS and SWIR Cameras." Pp. 302–13 in *Ground/Air Multisensor Interoperability, Integration, and Networking for Persistent ISR VIII*. Vol. 10190. SPIE.
- Nalamati, Mrunalini, Ankit Kapoor, Muhammed Saqib, Nabin Sharma, and Michael Blumenstein. 2019. "Drone Detection in Long-Range Surveillance Videos." Pp. 1–6 in 2019 *16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE.
- Onishi, Masanori, and Takeshi Ise. 2021. "Explainable Identification and Mapping of Trees Using UAV RGB Image and Deep Learning." *Scientific Reports* 11(1):903.

- Pham, Giao N., and Phong H. Nguyen. 2020. "Drone Detection Experiment Based on Image Processing and Machine Learning." Pp. 779–88 in *Proceedings of the IEEE conference on computer vision and pattern recognition*.
- Redmon, Joseph, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. "You Only Look Once: Unified, Real-Time Object Detection." Pp. 779–88 in *Proceedings of the IEEE conference on computer vision and pattern recognition*.
- Sahin, Oyku, and Sedat Ozer. 2021. "Yolodrone: Improved Yolo Architecture for Object Detection in Drone Images." Pp. 361–65 in *2021 44th International Conference on Telecommunications and Signal Processing (TSP)*. IEEE.
- Talan, Tarik. 2021. "Artificial Intelligence in Education: A Bibliometric Study." *International Journal of Research in Education and Science* 7(3):822–37.
- Tang, Fengxiao, Zubair Md Fadlullah, Nei Kato, Fumie Ono, and Ryu Miura. 2017. "AC-POCA: Anticoordination Game Based Partially Overlapping Channels Assignment in Combined UAV and D2D-Based Networks." *IEEE Transactions on Vehicular Technology* 67(2):1672–83.
- Vattapparamban, Edwin, Ismail Güvenç, Ali I. Yurekli, Kemal Akkaya, and Selçuk Uluğaç. 2016. "Drones for Smart Cities: Issues in Cybersecurity, Privacy, and Public Safety." Pp. 216–21 in *2016 international wireless communications and mobile computing conference (IWCMC)*. IEEE.
- Wang, Chien-Yao, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. 2023. "YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors." Pp. 7464–75 in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Wang, Chien-Yao, Hong-Yuan Mark Liao, and I. Hau Yeh. 2022. "Designing Network Design Strategies through Gradient Path Analysis." *ArXiv Preprint ArXiv:2211.04800*.
- Zhao, Jianqing, Xiaohu Zhang, Jiawei Yan, Xiaolei Qiu, Xia Yao, Yongchao Tian, Yan Zhu, and Weixing Cao. 2021. "A Wheat Spike Detection Method in UAV Images Based on Improved YOLOv5." *Remote Sensing* 13(16):3095.
- Zheng, Ye, Zhang Chen, Dailin Lv, Zhixing Li, Zhenzhong Lan, and Shiyu Zhao. 2021. "Air-to-Air Visual Detection of Micro-UAVs: An Experimental Evaluation of Deep Learning." *IEEE Robotics and Automation Letters* 6(2):1020–27.

How cite this article

Aydin, A., Talan, T., Akturk, C. (2023). Vision-based amateur drone detection: performance analysis of new approaches in deep learning. *Acta Infologica*, 7(2), 308-316. <https://doi.org/10.26650/acin.1273088>

Digital Transformation and Innovation in Health for Future Health Services: Turkey Global Innovation Index Time Series Analysis Between 2018 and 2022

Geleceğin Sağlık Hizmetleri için Sağlıkta Dijital Dönüşüm ve Inovasyon: 2018-2022 Yılları Arası Türkiye Inovasyon İndeksi Zaman Serileri Analizi

Ayça Asena Özdemir¹ , Zehra Alakoç Burma² 

¹(Lect. Dr.), Mersin University, Faculty of Medicine, Department of Medical Education, Mersin, Türkiye
²(Assoc. Prof. Dr.), Mersin University, Yenişehir Campus, Mersin Vocational School, Mersin, Türkiye

Corresponding author : Zehra ALAKOÇ BURMA
E-mail : zalakocburma@gmail.com

ABSTRACT

Health significantly impacts an individual's overall well-being, ability to sustain life, and life experience. The field of healthcare is currently undergoing a digital transformation, with innovations and advancements. Digitising patient records, e-prescriptions, telehealth, and medical imaging technologies provide better healthcare services and more effective tools. The aim is to improve the early detection of diseases, optimize treatment processes, personalize treatment plans, and make healthcare more accessible. Innovations such as robotic surgery, biotechnology, artificial intelligence, and gene editing make surgical interventions more precise and effective, while improving diagnosis and treatment processes. The process of drug development is accelerating, with promising approaches emerging for the treatment of genetic diseases, and the development of health technologies and service models. The widespread adoption of innovative solutions and transformations in the healthcare sector is aimed at positively impacting the outcomes of future health services. The article examines digital transformation and innovation in healthcare, analyzing time series data for Turkey's global innovation indicators between 2018 and 2022 and forecasting values for 2023. The study investigates our rank in the world, innovation potential, positive and negative indicators, and correlations between indicators for each measure. The results have significant importance in understanding and evaluating our country's innovation performance and its contribution to digital transformation in healthcare, as well as the future innovation strategies. The aim of the study is to provide guidance for researching ways to transform our healthcare systems and provide more effective healthcare services.

Keywords: Digital transformation in healthcare, healthcare innovation, time series analysis

Submitted : 31.08.2023
Revision Requested : 25.09.2023
Last Revision Received : 26.09.2023
Accepted : 06.10.2023
Published Online : 25.10.2023



This article is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0)

ÖZ

Sağlık, bireyin genel refahını, yaşamını sürdürme yeteneğini ve yaşam deneyimini büyük ölçüde etkiler. Günümüzde sağlık alanında da dijital dönüşüm ve inovasyonlar yaşanmaktadır. Hasta kayıtlarının dijitalleştirilmesi, e-reçete, tele-sağlık, tıbbi görüntüleme teknolojileri ile daha iyi hizmet sunma, daha etkili araçlar sağlama, hastalıkların erken teşhis edilmesi, tedavi süreçlerinin optimize edilmesi, tedavi planlarının kişiselleştirilmesi ve sağlık hizmetlerinin daha erişilebilir hale gelmesi amaçlanmaktadır. Robotik cerrahi, biyoteknoloji, yapay zekâ ve gen düzenleme gibi inovasyonlarla cerrahi müdahaleler daha hassas ve etkili hale gelmekte, tanı ve tedavi süreçleri iyileştirilmekte, ilaç geliştirme süreçleri hızlanmakta, genetik hastalık tedavilerinde umut verici yaklaşımların ortaya çıkması, sağlık teknolojileri ve hizmet modellerinin geliştirilmesi hedeflenmektedir. Sağlık alanındaki dönüşüm ve inovatif çözümlerin yaygınlaşması; gelecekteki sağlık hizmetlerinin sonuçlarını olumlu yönde etkileyecektir. Makalede, sağlıkta dijital dönüşüm ve inovasyon incelenmiş, Türkiye'nin 2018-2022 yılları arasındaki küresel inovasyon göstergelerinin zaman serileri analizi yapılmış ve 2023 yılı değerleri tahmin edilmiştir. Her gösterge için dünyadaki sıramız, inovasyon potansiyelimiz, iyi ve kötü olduğumuz göstergeler ve göstergelerin birbiri ile ilişkileri incelenmiştir. Çıkan sonuçlar; ülkemizin inovasyon performansını ve bunun sağlıkta dijital dönüşüme olan katkısı ve gelecekteki inovasyon stratejileri için sağlayabileceği etkilerini anlama ve değerlendirme açısından büyük önem taşımaktadır. Çalışmanın amacı; ülkemizin sağlık sistemlerini dönüştürme ve daha etkili bir sağlık hizmeti sunma yollarının araştırılmasına rehber olacaktır.

Anahtar Kelimeler: Sağlıkta dijital dönüşüm, sağlıkta inovasyon, zaman serileri analizi

1. INTRODUCTION

Today, the healthcare sector is undergoing a profound transformation brought about by the rapid development of digital technologies and innovative methodologies. The convergence of health and technology, commonly referred to as health technology, has already begun and will continue to reshape the delivery, accessibility, and experience of healthcare. This transformation in healthcare involves not only the expansion of traditional medical practices but also the creation of novel solutions to complex health-related problems. At the core of this transformation are two key concepts: Digital Transformation and Innovation in Healthcare. "Digital Transformation in Healthcare" refers to the use of information and communication technologies for healthcare processes, while "Innovation in Healthcare" refers to the application of innovative approaches to healthcare services. Digital Transformation and Health Economics: A Bibliometric Analysis on Digital Health, Tunçsiper (2023) reached various results. When the author analysed the studies on digital health in three different axes as the countries where the authors are located, authors and keywords in more detail, he found that the most frequently used words and phrases in the studies of the authors are keywords such as "innovation", "digital transformation", "digital health" and "artificial intelligence".

Many of the transformative technologies in healthcare have their roots in advances in information and communication technologies. It is worth noting that the considerable success of healthcare services owes much to these information and communication technologies. As the world increasingly uses technology to reshape healthcare systems, fostering a dynamic interaction between digital transformation and innovation, nations are turning to these areas to improve patient care and healthcare outcomes.

In the past century, enhancements in healthcare have resulted in a twofold increase in life expectancy in both high-income and developing economies (Roser, 2019; Ma, 2019; Shetty, 2019). The increase in life expectancy has contributed to the enlargement of the global labor force, spurred economic growth, and enhanced the quality of life for numerous individuals (WIPO, 2015a; Sampat, 2019). As societies experience economic growth, prosperity enables improved health and an elevated quality of life, thereby extending access to functional healthcare systems for a greater number of individuals in low- and middle-income economies (Kenny, 2011; WIPO, 2015a).

Historically, the arenas for health innovation, along with the innovation pathways themselves, have predominantly been centered in high-income economies, primarily within Europe and North America (Tannoury et al., 2017). Some of the leading countries in pharmaceutical patents include Switzerland, the United Kingdom, and the United States. The Netherlands and the United States are the leading countries for medical technology patents, while Switzerland and the United Kingdom are the leading countries for biotechnology patents. The spread of medical innovation to developing economies depends on their ability to foster innovation, and thereby economic growth. In terms of global sectoral investment, healthcare investments are second only to information technology (IT). Pharmaceutical, biotechnology, and medical device companies rank among the primary global corporate contributors to research and development (R&D), collectively dedicating more than US\$100 billion each year. This amount constitutes nearly 20% of the annual R&D outlays by the leading 2,500 R&D firms globally across various sectors (Hernández et al., 2018, R&D Magazine, 2018.—Top investors such as Roche (Switzerland), Johnson and Johnson (U.S.) and Merck US (U.S.) invested on average around US\$10 billion in R&D last year). This positions medical technologies within the top five rapidly expanding

technology domains since 2016, with the remaining four IT-related fields (WIPO, 2018. — see Patent applications and grants worldwide WIPO [World Intellectual Property Organization] (2018). World Intellectual Property Indicators 2018. Geneva: World Intellectual Property Organization).

However, medical technologies are increasingly innovating in adjacent health-related sectors such as IT and software applications, including innovations such as mechanical heart valves, artificial organs, digital health technologies, and 3D devices. The majority of medical innovations are concentrated in developed countries, where these innovations further strengthen their already robust economies. We desire to shift medical innovation to developing countries, including Turkey. This study attempts to assess Turkey's efforts in digital transformation and innovation in healthcare, using global innovation indicators. The paper analyzes Turkey's progress in innovation, its achievements in healthcare, and the challenges it faces. In addition, the study provides insights into Turkey's positioning in global innovation rankings and the impact of digital transformation on these rankings.

This article provides a comprehensive assessment of Turkey's performance in global innovation indices, with a particular focus on the implications for innovation in healthcare. The research findings have significant implications for Turkey's digital transformation and innovation initiatives in terms of healthcare infrastructure, investments, and goals. As a result, Turkey's digital transformation and innovation in healthcare efforts have the potential to improve the quality of national healthcare services while enhancing the country's standing in global innovation rankings. The primary objective of this study is to shed light on Turkey's achievements or shortcomings in this area and to answer the following questions for shaping future innovation strategies: What is Turkey's comparative position in each indicator and sub-indicator of the Global Innovation Index? What are the areas of notable progress, areas of regression, and areas of minimal change over the five years? What are the potential societal and economic impacts of innovation in healthcare? What are the barriers that must be overcome to fully realize this potential?

While digital transformation and innovation in healthcare hold great promise, they bring about certain challenges. It is paramount that healthcare professionals, including physicians, nurses, lab technicians, and operators of imaging equipment such as MRIs, gain the skills, understanding, and aptitude to use these evolving technologies. Comprehensive education, and training on issues such as digitization, data security, patient confidentiality, and adapting to new technologies is essential for all healthcare professionals in all settings to enable digital transformation. The collaboration between healthcare professionals, technologists, and managers is emerging as a key determinant for the success of this transformation.

The process of digital transformation and innovation in healthcare is of great significance as it strives to improve the quality of healthcare services, enhance the patient experience, and achieve favorable health outcomes. Advances in these areas will enable a future healthcare paradigm that is not only more effective, higher quality, and economically viable, but also sustainable, accessible, and personalized. These advances, in turn, generate economic growth and contribute to an improved quality of life for citizens. It is noteworthy that many countries around the world devote a significant proportion of their GDP to healthcare, a proportion that will continue to grow as populations age, placing a significant burden on both the economy and the labor, social, and psychological fabric of these countries. Immediate investment in innovation in healthcare delivery is therefore essential to alleviate these burdens.

2. HEALTHCARE AND KEY GOALS

The World Health Organization (WHO) established a definition of health in its Constitution, ratified in 1948. Accordingly, "Health is not merely the absence of disease or infirmity, but a state of complete physical, mental and social well-being." Health refers to a state of complete physical, mental, and social well-being. This extensive definition encompasses the condition of the individual, reflecting not only the absence of disease or physical infirmity but also the presence of mental and social well-being. Healthcare embodies a holistic state in which physiological, psychological, and social elements converge to influence an individual's daily functioning and ability to lead a fulfilling life. Therefore, healthcare is a fundamental concept that profoundly shapes an individual's overall well-being and life experiences.

The WHO categorizes healthcare into four primary dimensions: diagnosis, treatment, outcome, and health. Today, remarkable changes and innovations are seen in these four health domains due to the emergence of cutting-edge technologies, including electronic health records (EHRs), TeleHealth, mobile health applications, data analytics, cloud technology, Internet, artificial intelligence (AI), machine learning, 3D printing, robotic surgery, wearable technologies, e-prescribing systems, and e-drug tracking, as shown in Table 1.

Table 1. Current innovations in the four steps to a healthy planet (The Global Innovation Index 2019)

Diagnosis	Treatment	Outcome	Wellness
<ul style="list-style-type: none"> • Two-way data transmission from the patient • Wearable tech for monitoring • AI for diagnosis, reducing skill needed • Telehealth, reducing need for proximity 	<ul style="list-style-type: none"> • Focused factories (Industrialization) • Digital therapeutics • AI for treatment selection • Data on social determinants of health • Drug discovery in silico • Faster global trials • Cell and gene therapies • Oncology advances • Precision medicine 	<ul style="list-style-type: none"> • Real-world evidence • Value-based care 	<ul style="list-style-type: none"> • Prevention incentives • Interventions for social determinants of health

3. DIGITAL TRANSFORMATION AND INNOVATION IN HEALTHCARE

The foundation of digital transformation and innovation in healthcare lies in information and communication technologies.

3.1. Digital Transformation in Healthcare

Digital transformation in healthcare refers to the modernization of healthcare systems and the increased integration of digital technologies to improve the efficiency, effectiveness, accessibility, and patient-centeredness of healthcare services. This transformation involves technology-enabled changes that have various implications for healthcare systems around the world, which include the adoption of **Electronic Health Records (EHRs)**, the establishment of TeleHealth platforms, the integration of artificial intelligence (AI)-enabled diagnostic tools and the proliferation of wearable health devices.

The range of applications is vast. Electronic Health Records (EHRs) facilitate the digital storage and sharing of patients' medical data, streamline patients' access to their health history, facilitate information exchange among physicians, and improve the management of treatment processes. **TeleHealth**, which extends remote healthcare services, especially to people who live in remote areas or face accessibility challenges, facilitates consultations and treatment through video or audio communications that connect patients with healthcare professionals. **Mobile Health Applications** empower individuals to monitor and manage their health, with features such as health tracking, healthy lifestyle promotion, and medication reminders. **Data Analytics** supports early disease detection and the development of more effective treatments by providing insights into disease trends, epidemiological data, and treatment outcomes through comprehensive analysis of big data. **Artificial Intelligence and Machine Learning** are helpful in correct and quick medical decision-making, particularly in areas such as medical image analysis, disease diagnosis, and treatment recommendations. **Robotic Surgery** systems reduce surgical risks and enable surgeons to perform more precise and controlled procedures. **E-Prescribing and E-Medication Tracking** enable digital prescribing by physicians and allow patients to obtain medications from pharmacies. This process may facilitate medication reminders and side-effect monitoring. Finally, **Blockchain Technology** strengthens data privacy and security, ensuring the secure storage and exchange of patients' health data.

3.2. Innovation in Healthcare

Healthcare innovation encompasses the complex process of conceiving, cultivating, and implementing novel concepts, procedures, products, or services in the healthcare sector. The convergence of healthcare and innovation seeks to improve patient care, facilitate early disease detection, advance precision medicine, refine treatment modalities, optimize healthcare delivery, expand access to healthcare services, and improve the overall efficiency of healthcare systems, which has the potential to be a major catalyst for transforming healthcare systems. Innovation spans multiple dimensions, including novel treatment modalities, cutting-edge medical technologies, innovative medical procedures, advances in pharmaceuticals, drone-based drug delivery services, innovative care delivery models, streamlined operational workflows, and the seamless integration of cutting-edge technologies. Innovation in healthcare manifests itself in three primary areas: medicine, technology, management, and service delivery.

The range of applications is vast. **Medical Technology Advancements** aim to improve the effectiveness and precision of patient care through the development of novel medical devices, diagnostic techniques, and treatment methods. **Digital Health Applications** introduce innovations in areas such as mobile health applications, patient monitoring, promotion of healthy lifestyles, and management of patient-drug interactions. **Genetic Research and Personalized Medicine**

aim to predict disease susceptibility and tailor treatment strategies based on an individual's unique genetic makeup. **TeleHealth and Remote Care** provide remote patient monitoring, counseling, and treatment options to improve patient access to healthcare services. **Data Analytics** uses the scrutiny of vast data sets to identify disease trends, conduct epidemiological studies, and evaluate treatment outcomes. **Nanotechnology and Drug Development** uses technologies such as nanoparticles and microfluidic devices to drive innovation in the development of more effective and precisely targeted drugs. **Robotic Surgery and Artificial Intelligence-Assisted Diagnostics** are beneficial to improve surgical precision and speed up and improve the accuracy of diagnostic procedures. **Healthcare Management and Efficiency** are driving innovation in new management paradigms and business processes to improve the efficiency of hospital management and healthcare delivery. Finally, the integration of innovation into healthcare practice holds the promise of early disease detection, tailored interventions, and expanded access to healthcare services.

3.3. Distinguishing Digital Transformation and Innovation in Healthcare

Digital transformation and innovation in healthcare are distinct but interrelated concepts. Despite their similar goals, digital health transformation and innovation in healthcare represent discrete approaches and focal points in improving healthcare services. These two concepts can be distinguished along three key aspects. First, their focus differs. Digital transformation primarily seeks to integrate technology into healthcare processes, while innovation primarily seeks to create novel and more effective solutions. Second, their methodologies differ. Digital transformation focuses on digitizing existing healthcare practices and creating data-driven workflows, while innovation requires inventive thinking and disruptive approaches. Finally, their areas of application differ. Digital transformation involves technology-enabled solutions such as electronic health records, TeleHealth applications, and medical devices, while innovation covers a broader spectrum, ranging from medical treatments to healthcare administration.

Digital transformation in healthcare refers to the transition from traditional methods to digital platforms, which involves the use of digital technologies to promote data-centric processes aimed at making healthcare services more efficient, accessible, and coordinated. This process includes actions such as digitizing health-related data, automating hospital operations, and implementing electronic health records. For example, the conversion of paper-based patient records to electronic health records (EHRs) enables better management and secure sharing of healthcare services, such as the secure storage of medical data on digital platforms.

Conversely, innovation in healthcare seeks to improve the effectiveness and efficiency of healthcare services by integrating innovative ideas, methods, or technologies into the healthcare landscape. Technological advances in informatics contribute to innovation in healthcare. Artificial intelligence, robotics, remote diagnostics, genomics, big data analytics, mobile healthcare, stem cell research, regenerative medicine, biomarkers, and nanotechnology applications are all paving the way for advances in healthcare. Innovation aims to refine healthcare, optimize disease management, and improve the patient experience. For example, a pharmaceutical company's development of a cancer treatment that offers an alternative to conventional therapies is an example of innovation in healthcare. The creation of the mRNA vaccine for COVID-19 during the pandemic is an example of innovation in healthcare.

Innovation in healthcare can be achieved in several areas, including medicine, technology, management, and service delivery. Figures 1, 2, and 3 illustrate the potential for innovation in each of these three areas. These three figures were highlighted in the GII-2019 report as "Promising Fields for Medical Innovation and Technologies."

The future of medical innovation and its role in improving health outcomes, including reducing mortality and increasing productivity and quality of life, will depend on the policies and institutions established by national and global stakeholders to support research and innovation efforts. Technological progress is widely recognized as a key source of long-term economic growth. The Nobel laureate William Nordhaus has quantified the economic value of the increase in life expectancy over the past century, claiming that it rivals the economic growth of all other sectors combined. To illustrate, treatments for depression and advances in hip replacements have significantly reduced morbidity and improved overall quality of life. Certain emerging medical technologies, for instance, contraceptive pills, have profoundly transformed the workforce and societal structures (Bailey, 2006). Presently, a substantial body of scholarly work scrutinizes the feasibility and sustainability of the "technological imperative" in the field of medicine (Cutler et al., 2001). This has the potential to open new avenues for disease prevention, diagnosis, and treatment in healthcare.

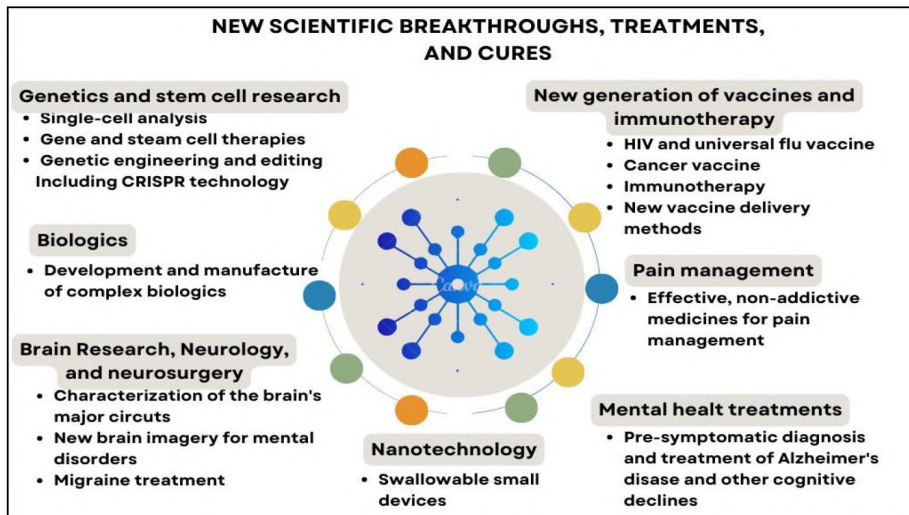


Figure 1. New scientific breakthroughs, treatments, and cures (The Global Innovation Index 2019)

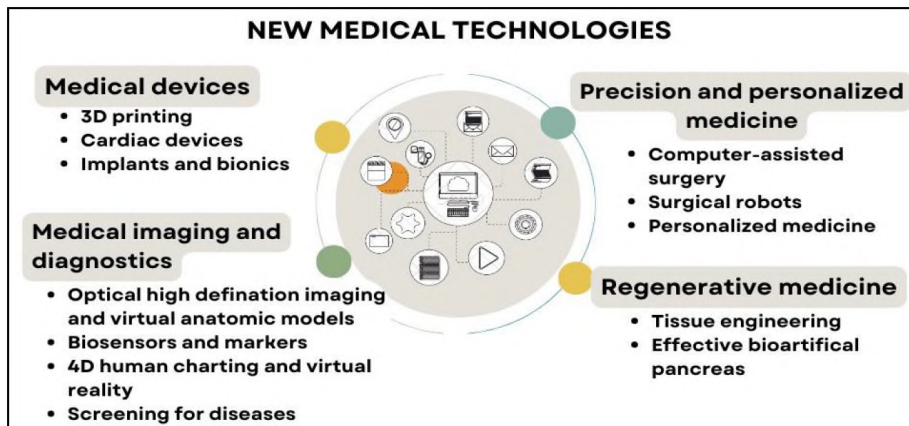


Figure 2. New medical technologies (The Global Innovation Index 2019)

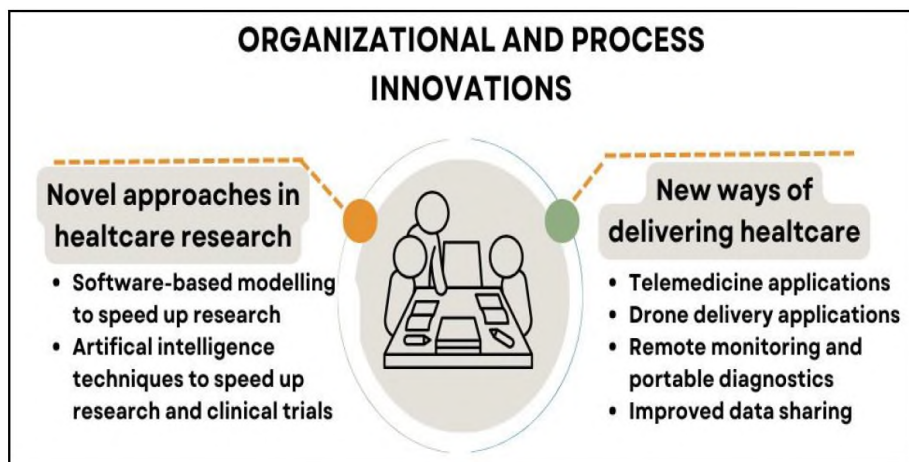


Figure 3. Organizational and process innovations (The Global Innovation Index 2019)

4. GLOBAL INNOVATION INDEX TURKEY AND TIME SERIES ANALYSIS

4.1. Materials: The Global Innovation Index (GII)

This study relies on data from the years 2018 to 2022, inclusive of the "Global Innovation Index (GII)" reports released in 2022. These reports are a collaborative effort involving Cornell University, The Business School for the World (INSEAD), the Confederation of Indian Industry (CII), and the World Intellectual Property Organization (WIPO), a specialized agency of the United Nations. It's noteworthy that this report has been consistently published since 2007. The details of the 2022 reports are at this website: . You can find previous reports on the About the GII | Global Innovation Index page.

The principal objective of the Global Innovation Index (GII) report is to evaluate the preparedness of nations in terms of promoting innovation and to furnish valuable insights for governments, corporations, and individuals, aiding them in optimizing the benefits of innovation. The report assigns scores to diverse sub-indicators, covering domains such as organizational innovation, human capital, and business development within each country. Subsequently, it ranks nations based on these scores. Furthermore, the report encompasses a range of policies and recent findings that are relevant to creating an environment conducive to innovation, with the aim of inspiring countries to embrace innovative practices.

The Global Innovation Index (GII) is structured around two fundamental components: the Innovation Input Sub-index (IISI) and the Innovation Output Sub-index (IOSI). The IISI comprises five essential dimensions: institutions, human capital & research, infrastructure, market sophistication, and business sophistication. These five foundational pillars encompass various aspects of a nation's economic framework that facilitate and nurture activities driven by innovation. The IOSI is divided into two distinct segments: knowledge technology output and creative output. Innovation outputs represent the concrete results arising from innovative activities occurring within the economy. Within each pillar, there exist three sub-pillars, each comprising individual indicators. The collective indicators for each year culminate in a comprehensive assessment. Indicator scores are assessed on a scale ranging from 1 to 7, where 7 denotes an exceptional performance, and 1 signifies a lower standing. Countries are then ranked based on these scores, with the highest-ranking country securing the 1st position. A lower rank index value corresponds to a more favorable indicator, while higher values indicate areas for improvement. Table 2 offers an exhaustive breakdown of the main and subsidiary topics encompassed within the Global Innovation Index (GII). Each entry in this table is assigned a numerical value, which is used in the time series analysis.

4.2. Methods: Time Series

Time series analysis constitutes a discipline within the field of statistics and, occasionally, within the realm of econometrics. However, its methodologies find application across virtually every scientific domain. A time series refers to a sequence of measurements that are recorded over successive intervals of time (Akdi, 2003). Essentially, a time series comprises numeric data points representing the values of variables that are observed sequentially from one time period to the next. While continuous realization of these observed values is not mandatory, it is essential for monitoring the progression of the series at consistent time intervals (Granger and Newbold, 1977). Time series data are analyzed and used for forecasting through various statistical models. In our study, we applied appropriate models, including ARIMA (AutoRegressive Integrated Moving Average), Holt's Exponential Smoothing, and Brown's Single Exponential Smoothing, tailored to the specific structure of our data. ARIMA models are used for analyzing time series data and forecasting future values. These models capture elements such as level, trend, and seasonal variation of the time series data for forecasting future values. When properly fitted to the data, the ARIMA model provides reliable forecasts based on historical data. Holt's exponential smoothing method analyzes time series data and extracts level and trend components, with the primary objective of identifying and encapsulating these specific characteristics of the time series. Brown's Single Exponential Smoothing method forecasts time series data only concerning the level component, focusing on detecting variations in the level component of the time series. It does not incorporate seasonal or trend components and is typically used to analyze less complex time series data (Gardner Jr., E. S., 1985).

This study involved the construction of a time series model based on Turkey's rankings among 126 countries in 2018, 126 countries in 2019, 131 countries in 2020, 131 countries in 2021, and 132 countries in 2022, as determined by "The Global Innovation Index" spanning the years 2018 to 2022. To perform forecasting on this time series data, the ARIMA, Holt, and Brown statistical methods were employed.

The time series analysis focused on modeling Turkey's "The Global Innovation Index" rank values from 2018 to 2022. It encompassed rank estimations for each year within the 2018-2022 period, along with forecasting values for the year 2023, presented with confidence intervals. In addition, the P values of the indicators within the forecasting

Table 2. Global Innovation Index

Global Innovation Index	4.3.1. Applies tariff rate, weighted mean, %
Innovation Output Sub-Index	4.3.2. Non-agricultural mkt access weighted tariff, %
Innovation Input Sub-Index	4.3.3. Intensity of local competition
Innovation Efficiency Ratio	5. Business sophistication
1. Institutional	5.1. Knowledge workers
1.1. Political Environment	5.1.1. Knowledge-intensive employment, %
1.1.1. Political stability*	5.1.2. Firms offering format training, % firms
1.1.2. Government effectiveness*	5.1.3. GERD performed by business, % GDP
1.1.3. Press Freedom*	5.1.4. GERD financed by business, %
1.2. Regulatory environment	5.1.5. GMAT test takers/mnpop. 20-34
1.2.1. Regulatory quality*	5.2. Innovation linkages
1.2.2. Rule of law*	5.2.1. University/Industry research collaboration ¹
1.2.3. Cost of redundancy dismissal, salary weeks	5.2.2. State of cluster development ¹
1.3. Business environment	5.2.3. GERD financed by abroad, %
1.3.1. Ease of starting a business*	5.2.4. JV-strategic alliance deals/tr PPP\$ GDP
1.3.2. Ease of resolving insolvency*	5.2.5. Patent families filed in 3+ offices/bn PPP\$ GDP
1.3.3. Ease of paying taxes*	5.3. Knowledge absorption
2. Human Capital & research	5.3.1. Royalty & license fees payments, % total trade
2.1. Education	5.3.2. High-tech imports less re-imports, %
2.1.1. Expenditure on education, %oGDP	5.3.3. Comm.computer&info.services imp., %total trade
2.1.2. Gov't expenditure/pupils,secondary,%GDP/cap	5.3.4. FDI net inflows, % GDP
2.1.3. School life expectancy, years	6. Knowledge&technology outputs
2.1.4. PISA scales in reading, maths, & Science	6.1. Knowledge creation
2.1.5. Pupil-teacher ratio, secondary	6.1.1. Domestic resident patent app/tr PPP\$ GDP
2.2. Tertiary education	6.1.2. PCT resident patent app./tr PPP
2.2.1. Tertiary enrolment, %ogross	6.1.3. Domestic res utility model app./tr PPP\$ GDP
2.2.2. Graduates in science&engineering, %	6.1.4. Scientific & technical articles/bn PPP\$ GDP
2.2.3. Tertiary inbound mobility, %	6.1.5. Citable documents H index
2.3. Research&development(R&D)	6.2. Knowledge impact
2.3.1. Researchers, headcounts/mn pop	6.2.1. Growth rate of PPP\$ GDP/worker, %
2.3.2. Gross expenditure on R&D, %oGDP	6.2.2. Newbusinesses/th pop. 15-64
2.3.3. QS university ranking, average scope top 3*	6.2.3. Computer software spending, %GDP
3. Infrastructure	6.2.4. ISO 9001 quality certificates/bn PPP\$ GDP
3.1. Information&communication technologies (ICTs)	6.2.5. High-&medium-high-tech manufactures, %
3.1.1. ICT access*	6.3. Knowledge diffusion
3.1.2. ICTuse*	6.3.1. Royalty & license-fees receipts, % total trade
3.1.3. Government's Online service*	6.3.2. High-tech exports less re-exports, %
3.1.4. E-participation*	6.3.3. Comm. computer& info. Services exp., % total trade
3.2. General infrastructure	6.3.4. FDI net outflows, % GDP
3.2.1. Electricity output, kWh/cap	7. Creative outputs
3.2.2. Logistics performance*	7.1. Intangible assets
3.2.3. Gross Capital formation, %oGDP	7.1.1. Domestic res trademark app/bn PPP\$ GDP
3.3. Ecological sustainability	7.1.2. Madrid trademark app. Holders/bn PPP\$ GDP
3.3.1. GDP/unit of energy use, 2005 PPP\$/kg oil eq	7.1.3. ICTs & business model creation ¹
3.3.2. Environmental performance*	7.1.4. ICTs & organizational model creation ¹
3.3.3. ISO 14001 environmental certificates/bn PPP\$ GDP	7.2. Creative goods & services
4. Market sophistication	7.2.1. Cultural & Creative services exports, % total trade
4.1. Credit	7.2.2. National feature films/mn pop. 15-69
4.1.1. Ease of getting credit*	7.2.3. Global ent. & media output/th pop. 15-69
4.1.2. Domestic credit to private sector, %oGDP	7.2.4. Printing & publishing manufactures, %
4.1.3. Microfinance gross loans, %oGDP	7.2.5. Creative goods exports, % total trade
4.2. Investment	7.3. Online creativity
4.2.1. Ease of protecting investors*	7.3.1. Generic top-level domains (TLDD)/th pop. 15-69
4.2.2. Market capitalization, % oGDP	7.3.2. Country-code TLDs/th pop.15-69
4.2.3. Total value of stocks traded, % oGDP	7.3.3. Wikipedia edits/pop. 15-69
4.2.4. Venture Capital deals/tr PPP\$ GDP	7.3.4. Video uploads on YouTube/pop. 15-69
4.3. Trade & competition	* an index ¹ a survey question

The sign (*) in the relevant indicators indicates the index values obtained by various organizations. The sign (!) indicates the values obtained from the survey questions. Some indicator values were obtained from the data of national organizations.

models were categorized into four groups: best (1), no change (2), potential change (3), and worst (4), based on the trend observed in the graphs. Statistical analysis was conducted using the SPSS 21.0 software package.

5. RESULTS

Given the large number of indicators, the results of this study are concisely summarized and presented in tables and graphs. These results are systematically organized and interpreted, taking into account the items corresponding to each sub-index.

Table 3. Global Innovation Index, Innovation Output and Input Sub-Indexes - Rankings and Forecasts for 2023

	2018	2019	2020	2021	2022	2023 Forecasting (95% C.I)	p
GII	50 (53)	49 (50)	51 (46)	41 (43)	37 (40)	36 (23-49)	0.746
IOSI	43 (44)	56 (44)	52 (44)	45 (44)	49 (44)	44 (22-65)	<0.001
IISI	62 (62)	56 (58)	52 (54)	45 (49)	49 (44)	42 (29-55)	0.396

Observed (Predicted)

Table 3 summarizes the 5-year rankings of the Global Innovation Index (GII), Innovation Output, and Input sub-indexes, along with projections for 2023. Turkey's rank in the Global Innovation Index (GII) was 50 out of 126 countries in 2018, maintained a relatively similar position at 49 in 2019, and moved slightly to 51 out of 131 countries in 2020. However, after 2021, the GII scores showed a remarkable momentum, exceeding the expected forecast among 131 countries, and this momentum continued in 2022. The forecast for 2023, which indicates an improvement in Turkey's GII ranking to 36th, seems to indicate a positive trend. However, a statistical analysis of the model shows that there is no statistically significant momentum in Turkey's ranking over the years (Figure 4A) ($p>0.05$).

Regarding the ranking value of the Innovation Output Sub-Index (IOSI), Turkey's ranking was 43 in 2018 and experienced a significant decline in 2019. Since 2020, there has been an upward trend in the ranking. In 2021, Turkey's ranking was close to that of 2018, with a slight decline in 2022. The projection for 2023 shows an increase in rankings, indicating an improving trend for Turkey, with a rapid improvement (Figure 4B) ($p<0.05$).

Regarding the ranking value of the Innovation Input Sub-Index (IISI), Turkey's position was 62 in 2018, with subsequent increases in subsequent years, but a decline in 2022. For the projected year 2023, the expected rank value is 42, which is an improvement compared to the previous years. Nevertheless, the p-value of the model does not indicate any significant positive or negative progress (Figure 4C) ($p>0.10$).

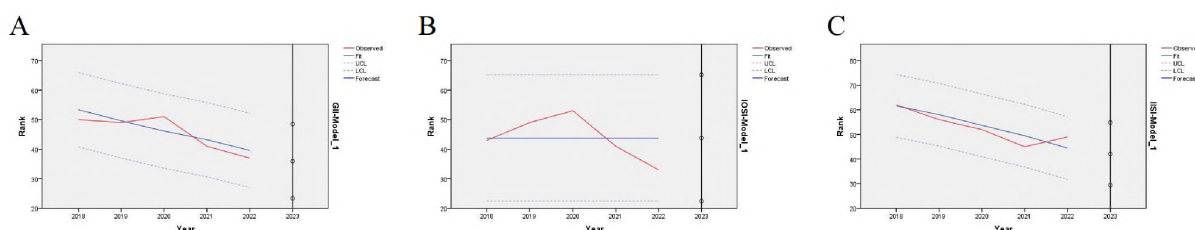


Figure 4. Global Innovation Index (A); Innovation Output Sub-Index (B); and, Innovation Input Sub-Index (C) Time Series Graphics and 2023 Forecasting

The study includes separate assessments of Turkey's rankings in the fundamental indicators of the GII, IISI, and IOSI. Due to the large number of sub-indicators, as opposed to individual assessments for the subsequent index scores of the five core sub-indicators within IISI and the two core sub-indicators within IOSI, they are examined collectively as integral sub-indicators, signifying either Turkey's negative progress, positive progress, or stable positioning.

Table 4 illustrates the 5-year rankings of institutions and projections for 2023 in terms of the basic input indicator. Thus, it shows that Turkey's ranking in 2022 showed a significant decline compared to the previous years. However, the forecast for 2023 predicts an improvement of 7 positions to reach 94th place. The statistical model for institutions suggests that Turkey has made positive progress, although precautions are needed regarding the policies implemented in 2022 ($p<0.05$). Within the core institutions indicator, sub-indices 1.1 Political Environment and 1.2 Regulatory Environment show stationary momentum. However, the forecast predicts a slight improvement in the 1.3 Business Environment indicator, moving from 92nd to 91st place. In particular, among the institutional indicators, a decline is expected in 2023 for both 1.1.1 Political Stability and 1.2.3 Cost of Redundancy Dismissal, and Salary Weeks. However, this decline, particularly in political stability, is significant given Turkey's already high ranking in this area.

Table 5 shows the 5-year rankings of human capital and research, along with projections for 2023, as a basic input indicator, it shows that Turkey's ranking in 2021 showed a significant improvement compared to the previous years.

Table 4. Institutions - Rankings and Forecasts for 2023

	2018	2019	2020	2021	2022	2023 Forecasting (95% C.I)	p
1.	96 (94)	85 (94)	94 (94)	93 (94)	101 (94)	94 (78-110)	<0.001
1.1.	102 (81)	69 (81)	77 (81)	75 (81)	81 (81)	81 (46-116)	<0.001
1.1.1.	124 (94)	79 (94)	92 (94)	89 (94)	87 (94)	94 (46-142)	<0.001
1.1.2.	68 (67)	67 (68)	71 (69)	70 (71)	73 (72)	73 (69-78)	0.731
1.2.	97 (97)	102 (102)	108 (107)	109 (114)	109 (110)	109 (102-116)	0.003
1.2.1.	60 (na)	67 (64)	74 (71)	72 (78)	74 (76)	78 (64-91)	0.207
1.2.2.	71 (72)	76 (74)	82 (77)	78 (81)	83 (83)	85 (75-96)	0.540
1.2.3.	111 (111)	115 (115)	117 (119)	118 (119)	119 (119)	120 (117-123)	<0.001
1.3.	97 (91)	82 (91)	91 (91)	91 (91)	92 (91)	91 (76-106)	<0.001
1.3.1.	66 (61)	63 (68)	62 (74)	62 (78)	99 (81)	91 (41-142)	0.643
1.3.2.	112 (112)	96 (99)	104 (87)	104 (76)	47 (67)	53 (-17-123)	0.738

Observed (Predicted)

Table 5. Human capital and research - Rankings and Forecasts for 2023

	2018	2019	2020	2021	2022	2023 Forecasting (95% C.I)	p
2.	49 (41)	46 (41)	42 (41)	26 (41)	41 (41)	41 (16-65)	0.001
2.1.	82 (47)	73 (47)	7 (47)	6 (47)	66 (47)	47 (-57-150)	0.048
2.1.1.	69	70		65		No estimation	-
2.1.2.	82 (86)	90 (86)			87 (86)	86 (69-104)	0.001
2.1.3.	14 (14)	14 (13)	12 (13)	11 (12)	11 (11)	10 (7-12)	0.360
2.1.4.	49 (50)	49 (47)	41 (45)	41 (42)	41 (39)	37 (28-47)	0.481
2.1.5.	79 (80)	81 (80)	84 (80)	80 (80)	77 (80)	80 (73-87)	<0.001
2.2.	49 (53)	43 (53)	91 (53)	24 (53)	56 (53)	53 (-16-121)	0.009
2.2.1.	3 (3)	3 (3)		2 (2)	2 (2)	2 (0-3)	0.400
2.2.2.	58 (54)	65 (63)	73 (72)	75 (81)	97 (89)	99 (79-119)	0.722
2.2.3.	78 (79)	82 (79)	80 (79)	80 (80)	78 (80)	79 (74-85)	0.448
2.3.	36 (38)	39 (38)	40 (38)	38 (38)	36 (38)	38 (33-43)	<0.001
2.3.1.	46 (46)	44 (45)	46 (44)	43 (44)	42 (43)	42 (37-46)	0.779
2.3.2.	38 (38)	37 (38)	39 (38)	36 (38)	39 (38)	38 (34-41)	<0.001
2.3.3.	27 (29)	31 (29)	33 (29)	29 (29)	29 (29)	29 (23-36)	0.852
2.3.4.	41 (42)	44 (43)	45 (44)	45 (45)	46 (46)	47 (44-51)	0.508

Observed (Predicted)

However, there was a decline in 2022, with the forecast for 2023 suggesting no substantial progress ($p < 0.05$). Within sub-index 2.1 Education, there was a significant increase in rankings in 2020 and 2021 (the pandemic period), but a rapid decline in 2022. The forecast for 2023 shows an improvement over the previous year. The sub-index 2.2 Tertiary education saw significant progress in 2021, but this was not sustained in 2022. The projection for 2023 foresees a slight increase in the tertiary education index. The 2.3 Research and Development (R&D) sub-index has shown modest fluctuations over the years, with the last 5-year rankings close to the expected ranking of 38. A slight decline is expected in 2023 compared to the previous year. Forecasts for the 2023 rankings in 2.1.2 Government expenditure/pupil (secondary), GDP% per capita, and 2.3.2 Gross expenditure on R&D, GDP% show significant improvements compared to 2022. Regrettably, no meaningful model could be constructed for 2.1.1 Expenditure on education, GDP%, which precluded any forecast.

Table 6. Infrastructure - Rankings and Forecasts for 2023

	2018	2019	2020	2021	2022	2023 Forecasting (95% C.I)	p
3.	52 (49)	41 (49)	54 (49)	48 (49)	48 (49)	49 (35-62)	<0.001
3.1.	65 (63)	49 (57)	49 (50)	47 (44)	38 (39)	33 (16-51)	0.487
3.1.1.	67 (69)	69 (68)	66 (67)	66 (66)	64 (65)	64 (60-68)	0.998
3.1.2.	67 (68)	68 (66)	61 (65)	64 (63)	60 (61)	59 (50-68)	0.689
3.1.3.	64 (56)	27 (49)	27 (35)	22 (25)	22 (15)	8 (-39-55)	0.440
3.1.4.	59 (56)	37 (48)	37 (36)	23 (28)	23 (18)	11 (-14-36)	0.436
3.2.	33 (42)	38 (42)	57 (42)	42 (42)	41 (42)	42 (17-67)	<0.001
3.2.1.	56 (56)	54 (56)	54 (56)	57 (56)	57 (56)	56 (51-60)	<0.001
3.2.2.	33 (43)	46 (43)	46 (43)	46 (43)	44 (43)	43 (27-59)	<0.001
3.2.3.	21 (26)	20 (26)	47 (26)	26 (26)	16 (26)	26 (-8-60)	0.009
3.3.	54 (53)	52 (54)	55 (56)	54 (57)	61 (59)	60 (51-69)	0.768
3.3.1.	16 (17)	19 (17)	16 (17)	19 (17)	17 (17)	17 (13-22)	<0.001
3.3.2.	87 (94)	88 (94)	84 (94)	84 (94)	125 (94)	94 (45-143)	<0.001
3.3.3.	70 (65)	67 (65)	57 (65)	66 (65)	66 (65)	65 (52-79)	<0.001

Observed (Predicted)

Table 6 presents an overview of the 5-year infrastructure ranking and offers projections for 2023 as a basic input indicator. Accordingly, Turkey's ranking was 41st in 2019, but the following years showed a decline. The projection for 2023 predicts a slight decline from the previous year. Looking at the sub-indices, a statistically significant decrease similar to the basic input indicator is expected for 3.2 General Infrastructure.

Within the sub-index 3.2.3 Gross capital formation, GDP%, there was a significant decline in 2020, followed by a gradual recovery in the following years. The projection for 2023 foresees a decrease of 10 units compared to the previous year. On the other hand, for the sub-index 3.3.1 GDP/unit of energy use, 2005 PPP\$/kg oil equivalent, Turkey's ranking has shown consistency over the past 5 years, and this ranking is expected to be maintained in 2023. For the sub-indices 3.2.1 Electricity output, kWh/capacity, 3.2.2 Logistics performance, and 3.3.3 ISO 14001 environmental certificates/billion PPP\$ GDP, the forecasts for 2023 show a slight decrease compared to the previous year, while a rapid improvement is expected for the sub-index 3.3.2 Environmental performance.

Table 7. Market Sophistication - Rankings and Forecasts for 2023

	2018	2019	2020	2021	2022	2023 Forecasting (95% C.I.)	p
4.	55 (44)	52 (44)	28 (44)	49 (44)	37 (44)	44 (13-76)	0.001
4.1.	95 (92)	66 (81)	66 (69)	68 (57)	39 (47)	36 (-3-74)	0.738
4.1.1.	70 (40)	29 (40)	34 (40)	34 (40)	35 (40)	40 (-6-87)	0.006
4.1.2.	45 (46)	44 (46)	46 (46)	51 (46)	44 (46)	46 (38-54)	<0.001
4.1.3.	77 (77)	78 (77)	76 (77)	77 (77)		77 (74-80)	<0.001
4.2.	77 (75)	87 (75)	44 (75)	105 (75)	61 (75)	75 (10-140)	0.002
4.2.1.	20 (21)	24 (21)	21 (21)	21 (21)		21 (15-27)	0.958
4.2.2.	61 (60)	56 (58)	54 (56)	55 (54)	53 (53)	51 (44-58)	0.443
4.2.3.	78 (79)	78 (79)		85 (79)	76 (79)	79 (67-92)	<0.001
4.3.	9 (10)	15 (10)	7 (10)	10 (10)	11 (10)	10 (2-19)	0.001
4.3.1.	60 (65)	67 (65)	62 (65)	63 (65)	71 (65)	65 (52-77)	<0.001
4.3.2.	8 (8)	6 (7)	6 (5)	4 (4)	3 (3)	2 (0-4)	0.337
4.3.3.	13 (13)	13 (13)	13 (12)	13 (12)	11 (12)	11 (9-14)	0.691

Observed (Predicted)

Table 7 illustrates the 5-year rankings for market sophistication and presents forecasts for 2023 as a basic input indicator. Accordingly, Turkey's ranking was 28 in 2020, followed by a rapid decline in 2021. Turkey's ranking shows a notable increase in 2022, followed by a predicted decline in 2023. Looking at the 4.2 Investment sub-index, a decline is expected in 2023, similar to the trends observed in the primary inputs index. In the 4.3 Trade & Competition sub-index, on the other hand, Turkey has maintained a commendable ranking over the past 5 years, with further improvements predicted for 2023. However, for the subindices 4.1.1 Ease of getting credit, 4.1.2 Domestic credit to the private sector, GDP%, and 4.2.3 Total value of stocks traded, %GDP, Turkey's ranking is expected to decline in 2023 compared to the previous year, while no significant improvement or regression is expected for the subindex 4.1.3 Microfinance gross loans, GDP%. Subindex 4.3.1 Applies tariff rate, weighted average, % showed a sharp decline in 2022; however, projections for 2023 suggest an increase in the ranking.

Table 8. Business Sophistication - Rankings and Forecasts for 2023

	2018	2019	2020	2021	2022	2023 Forecasting (95% C.I.)	p
5.	72 (75)	71 (67)	57 (61)	46 (52)	47 (43)	37 (18-55)	0.333
5.1.	71 (74)	72 (67)	59 (63)	49 (55)	52 (47)	43 (23-62)	0.395
5.1.1.	72 (73)	71 (71)	73 (69)	69 (67)	63 (65)	63 (54-72)	0.997
5.1.2.	52 (51)	53 (51)	48 (51)	50 (51)	54 (51)	51 (45-58)	<0.001
5.1.3.	36 (37)	37 (36)	36 (35)	33 (34)	33 (33)	32 (28-36)	0.539
5.1.4.	19 (22)	27 (22)	28 (22)	18 (22)	17 (22)	22 (7-36)	0.001
5.1.5.	70 (72)	72 (71)	71 (70)	69 (69)	68 (69)	68 (64-72)	0.999
5.2.	102 (104)	97 (96)	91 (89)	79 (83)	75 (74)	67 (59-76)	0.278
5.2.1.	63 (68)	88 (68)	70 (68)	62 (68)	68 (68)	68 (38-98)	1.000
5.2.2.	56 (60)	76 (60)	64 (60)	48 (60)	58 (60)	60 (31-89)	<0.001
5.2.3.	90 (70)	68 (70)	59 (70)	71 (70)	60 (70)	70 (35-104)	<0.001
5.2.4.	92 (90)	95 (98)	106 (103)	115 (111)	116 (119)	125 (113-137)	0.318
5.2.5.	42 (41)	43 (41)	50 (41)	33 (41)	38 (41)	41 (24-59)	<0.001
5.3.	57 (60)	57 (54)	48 (50)	36 (45)	44 (38)	35 (13-57)	0.450
5.3.1.	71 (80)	74 (73)	76 (66)	56 (58)	44 (51)	44 (15-73)	0.999
5.3.2.	21 (21)	33 (33)	55 (45)	62 (77)	51 (69)	40 (5-75)	0.027
5.3.3.	121 (133)	124 (120)	124 (111)	84 (104)	92 (90)	81 (32-130)	0.529
5.3.4.	88 (93)	89 (93)	97 (93)	100 (93)	91 (93)	93 (78-108)	<0.001
5.3.5.	25 (26)	19 (21)	19 (16)	9 (12)	25 (26)	2 (-7-11)	0.360

Observed (Predicted)

Table 8 shows the 5-year rankings for business sophistication and provides projections for 2023. In terms of the basic

input indicator, significant improvements were observed over the past 5 years, and the projections for 2023 suggest that this positive trend will continue, although without statistical significance ($p > 0.05$). Subindices 5.1.4 GERD financed by business, %, 5.2.2 Status of cluster development, 5.2.3 GERD financed from abroad, %, 5.2.5 Patent families filed in 3+ offices/billion PPP\$ GDP, and 5.3.4 Net FDI inflows, % GDP are projected to decline in 2023 compared to the previous year. On the other hand, sub-index 5.1.2 Firms offering formal training, % firms has shown marginal fluctuations over the past 5 years and is expected to show a slight improvement in 2023 compared to the previous year. Sub-index 5.3.2 High-tech imports fewer re-imports, %, shows a decline from 2018 to 2021 but starts to recover in 2022. This is also reflected in the projections for 2023, which anticipate an improvement in rankings.

Table 9. Knowledge & Technology Outputs - Rankings and Forecasts for 2023

	2018	2019	2020	2021	2022	2023 Forecasting (95% C.I.)	p
6.	52 (53)	59 (53)	57 (53)	50 (53)	47 (53)	53 (39-67)	<0.001
6.1.	41 (41)	38 (40)	40 (39)	37 (38)	37 (37)	36 (31-41)	0.577
6.1.1.	30 (30)	27 (29)	30 (27)	24 (26)	24 (24)	23 (15-30)	0.616
6.1.2.	32 (31)	32 (31)	28 (31)	31 (31)	31 (31)	31 (26-35)	<0.001
6.1.3.	16 (18)	17 (18)	20 (18)	20 (18)	17 (18)	18 (13-23)	<0.001
6.1.4.	59 (57)	60 (57)	54 (57)	52 (57)	60 (57)	57 (47-67)	<0.001
6.1.5.	35 (35)	35 (35)	35 (35)	35 (35)	35 (35)	35 (35-35)	-
6.2.	53 (56)	57 (51)	42 (48)	38 (41)	39 (36)	32 (13-51)	0.484
6.2.1.	33 (47)	46 (38)	37 (31)	12 (23)	10 (14)	6 (-34-45)	0.699
6.2.2.	66 (67)	66 (66)	65 (65)	65 (64)	63 (64)	63 (61-65)	0.999
6.2.3.	20 (20)	20 (20)	20 (20)	20 (20)	20 (20)	20 (20-20)	-
6.2.4.	73 (72)	80 (72)	67 (72)	70 (72)	70 (72)	72 (58-86)	<0.001
6.2.5.	41 (45)	44 (44)	42 (45)	55 (44)		46 (27-65)	0.157
6.3.	90 (107)	112 (97)	96 (90)	73 (82)	67 (73)	64 (16-112)	0.802
6.3.1.		96 (98)	90 (87)	76 (78)		56 (43-69)	0.302
6.3.2.	63 (63)	63 (63)	64 (63)	61 (63)	63 (63)	63 (60-66)	1.000
6.3.3.	122 (131)	122 (120)	124 (112)	94 (106)	93 (95)	86 (51-121)	0.564
6.3.4.	63 (63)	73 (71)	81 (80)	94 (89)	93 (99)	105 (90-120)	0.446

Observed (Predicted)

Table 9 presents an overview of the 5-year Knowledge & Technology outputs ranking and offers projections for 2023. As a basic input indicator, there was an increasing trend after 2019, but projections for 2023 show a decline compared to 2022. Sub-indices including 6.1 Knowledge Creation, 6.2 Knowledge Impact, and 6.3 Knowledge Diffusion are projected to increase in the 2023 forecasts compared to previous years, although these trends lack statistical significance. On the other hand, the sub-indices 6.1.3 Domestic res utility model app./tr PPP\$ GDP and 6.2.4 ISO 9001 quality certificates/bn PPP\$ GDP are expected to experience slight decreases compared to 2022. However, the sub-index 6.1.2 PCT resident patent app./tr PPP is expected to have the same rank. In particular, the subindex 6.1.4 Scientific & technical articles/bn PPP\$ GDP achieved its best ranking in the last 5 years, reaching 52nd place in 2021. Forecasts for 2023 suggest a slight improvement over 2022.

Table 10 shows the 5-year rankings and 2023 forecasts for Creative output. As a basic input indicator, a rapid increase was observed in 2022 compared to previous years, and it is predicted to rise to 9th place in the 2023 forecast. Conversely, the 7.1 Intangible Assets and 7.3 Online Creativity sub-indices show significant declines in 2020, followed by improvements. Nevertheless, the forecasts for 2023 show a decline compared to the previous year. The sub-indices include 7.1.1 Domestic Trademark Apps/billion PPP\$ GDP, 7.1.3 ICTs & Business Model Creation and 7.3.3 Wikipedia Edits/pop. 15-69 are expected to decline in 2023. Conversely, the 7.2.5 Creative goods exports, % of total trade sub-index is expected to remain constant in 2023 compared to 2022. In addition, sub-indices such as 7.2.1 Cultural and creative services exports, % total trade, 7.2.4 Printing and publishing manufactures, %, 7.3.2 Country code TLDs/th pop.15-69, and 7.3.4 Video uploads to YouTube/pop. 15-69 are projected to increase slightly in 2023 compared to 2022, while the 7.1.4 ICTs & organizational model creation sub-index is projected to increase compared to 2021.

Finally, Table 12 shows the rankings of the grouped innovation indicators in an evaluation that takes into account the direction of the curves formed based on the p-values of the predicted values in the models created for 2023. This evaluation aims to clarify Turkey's position in terms of the Global Innovation Index and its sub-indices and to provide clearer insights into the areas in which the country needs to improve. The 2023 criteria indicators can be improved with additional support along with the implementation of short-term action plans.

6. CONCLUSIONS AND EVALUATIONS

This study provides an in-depth analysis of Turkey's performance on global innovation indicators from 2018 to 2022, in the context of digital transformation and innovation in healthcare. This analysis reveals the intricate dynamics between

Table 10. Creative Outputs - Rankings and Forecasts for 2023

	2018	2019	2020	2021	2022	2023 Forecasting (95% C.I)	p
7.	39 (na)	40 (33)	50 (34)	35 (44)	15 (29)	9 (-35-53)	0.452
7.1.	11 (17)	20 (17)	31 (17)	18 (17)	4 (17)	17 (-11-45)	0.021
7.1.1.	14 (11)	13 (11)	17 (11)	6 (11)	6 (11)	11 (-3-25)	0.007
7.1.2.	1 (3)	1 (3)	6 (3)	5 (3)	1 (3)	3 (-4-10)	0.066
7.1.3.	53 (53)	72 (53)	44 (53)	45 (53)	51 (53)	53 (22-84)	<0.001
7.1.4.	75 (93)	98 (93)	100 (93)	100 (93)		93 (54-132)	0.001
7.2.	60 (59)	60 (61)	60 (64)	61 (66)	72 (68)	71 (57-85)	0.722
7.2.1.	75 (74)	46 (74)	92 (74)	82 (74)	77 (74)	74 (27-122)	0.001
7.2.2.	58 (60)	59 (57)	62 (55)	62 (54)	44 (53)	49 (23-75)	0.604
7.2.3.	43 (44)	46 (45)	48 (46)	47 (48)	48 (49)	50 (45-54)	0.561
7.2.4.	62 (62)	71 (71)	73 (80)	75 (75)	70 (77)	65 (51-79)	0.001
7.2.5.	18 (19)	21 (19)	19 (19)	19 (19)	19 (19)	19 (16-22)	<0.001
7.3.	56 (56)	55 (56)	69 (56)	50 (56)	48 (56)	56 (33-78)	<0.001
7.3.1.	36 (36)	36 (36)	36 (36)	36 (36)	37 (37)	37 (36-38)	0.691
7.3.2.	66 (66)	68 (68)	69 (70)	68 (70)	67 (67)	66 (63-69)	0.002
7.3.3.	85 (80)	85 (80)	101 (80)	61 (80)	68 (80)	80 (36-124)	<0.001
7.3.4.	36 (36)	23 (23)	19 (10)	18 (15)	17 (17)	16 (3-29)	<0.001

Observed (Predicted)

Table 11. Status of the Global Innovation Index and its sub-indexes according to P values

Status	P value	Curve direction	Description
1	<0.05		Appropriate policies are being implemented, and it is recommended to maintain the current system.
2	<0.05		The present policies do not have a positive or negative effect. The 2023 criteria indicators can be improved with additional support along with the implementation of short-term action plans.
3	>0.05		The existing policies have been ineffective, and there has been no progress, neither positive nor negative. It is necessary to develop new policies and comprehensive action plans and to implement them.
4	<0.05		Wrong policies have led to a negative trend, hence requiring urgent action plans.

Turkey’s digital transformation initiatives in healthcare, its innovation efforts, and the resulting innovation capability outcomes. By exploring Turkey’s future within the global innovation ecosystem, valuable insights into the interaction between digital transformation and innovation in healthcare have emerged, contributing to a better understanding of Turkey’s innovation trajectory. Moreover, this research provides important guidance for future strategic directions aimed at leveraging digital technologies and innovation to improve healthcare services and outcomes for Turkey’s population. Turkey’s position within the global innovation landscape also serves as an indicator of the synergy between digitalization and innovation in healthcare, and their collective influence on Turkey’s health innovation standing, and its performance on the global innovation stage.

The study has identified Turkey’s strengthening, weakening, and no progress areas across the indicators and sub-indicators of the Global Innovation Index Due to the large number of indicators in the index, the results are not interpreted individually for each indicator and are categorized into four different situations for analysis, as shown in Table 12. A careful examination and interpretation of Table 12 is essential for understanding the results of the study. In terms of the index and sub-index values in the first column, appropriate policies should be implemented and the existing system should be maintained. In terms of the index and sub-index values in the second column, there were no significant positive or negative impacts of current policies. The benchmark indicators for 2023 can be improved through additional support and the implementation of short-term action plans. Regarding the index and subindex values in column 3, we see that our current policies are ineffective and there is neither positive nor negative progress. As shown in the table, most of the indices and sub-indices are grouped under this column. These indicators require new policies and comprehensive action plans. This column requires thorough analysis and evaluation, and urgent action should be taken to improve innovation scores. Finally, in the fourth column, where adverse trends have emerged due to

Table 12. Status of Turkey according to P values of Global Innovation Index and sub-indexes (for 1,2,3,4)

	1	2	3	4
Basic	IOSI		GII IISI	
1. Institutions	1 1.3	1.1 1.2	1.1.2 1.2.1 1.2.2 1.3.1 1.3.2	1.1.1 1.2.3
2. Human Capital&Research	2.1 2.1.2 2.2 2.3.2	2	2.1.3 2.1.4 2.2.1 2.2.2 2.2.3 2.3.1 2.3.3 2.3.4	2.1.5 2.3
3. Infrastructure	3.2.1 3.2.2 3.3.2 3.3.3	3.3.1	3.1 3.1.1 3.1.2 3.1.3 3.1.4 3.3	3 3.2 3.2.3
4. Market Sophistication	4.3 4.3.1	4.1.3	4.1 4.2.1 4.2.2 4.3.2 4.3.3	4 4.1.1 4.1.2 4.2 4.2.3
5. Business Sophistication	5.1.2 5.3.2	-	5 5.1 5.1.1 5.1.3 5.1.5 5.2 5.2.1 5.2.4 5.3 5.3.1 5.3.3 5.3.5	5.1.4 5.2.2 5.2.3 5.2.5 5.3.4
6. Knowledge & Technology Out.	6.1.4	6.1.2 6.1.5 6.2.3	6.1 6.1.1 6.1.5 6.2 6.2.1 6.2.2 6.2.3 6.2.5 6.3 6.3.1 6.3.2 6.3.3 6.3.4	6 6.1.3 6.2.4
7. Creative Outputs	7.1.4 7.2.1 7.2.4 7.3.2 7.3.4	7.2.5	7 7.1.2 7.2 7.2.2 7.2.3 7.2.3 7.3.1	7.1 7.1.1 7.1.3 7.3 7.3.3

misguided policies, immediate action plans are imperative to rectify the situation for the specified index and sub-index values.

In Japan, one of the leading countries in innovation, the role of education and training, along with related social innovations, is one of the four key components within the national innovation system (Freeman, 1987:4). Tunçsiper, Ç. & Bakar, A. (2023) stated that in order for a country to have a healthy society, its economy should be strong, economic growth should be ensured and made sustainable and health services should be provided to meet the needs of the society. Similarly, Turkey needs an innovation miracle to strengthen not only the health sector but all other industries, to stimulate economic growth, to move from the league of developing countries to the league of developed countries, to have a competitive edge, and to carve out a distinctive global identity. To achieve this, significant emphasis must be placed on R&D and education, particularly in the area of digital transformation in education. Turkey still has a long way to go in terms of innovation.

Enterprises and policymakers should emphasize the process of converting research outcomes into commercially feasible applications, which might necessitate the initiation of partnerships between the public and private sectors, fostering an entrepreneurial environment within public research institutions, encouraging the establishment of academic spin-off ventures, and establishing business incubators and centers of excellence (Gelijns et al., 1994; Thune et al., 2016). To provide higher quality healthcare to a larger population at a low cost, it is imperative to accelerate the transformation of healthcare, in a more personalized, digitally integrated, and collaborative manner. IT-driven innovation should play

a key role in expanding essential healthcare services and narrowing the existing gap between developed and developing countries.

Turkey has not reached to the point of satisfaction with respect to innovation. For this to materialize, the various health system actors will have to create and use better channels and to transmit relevant information and feedback. (Barberá-Tomás et al., 2012.) To act as a bridge between research and the application of innovation in a real-life context, medical professionals with experience in research, training in the use of new hardware and software, and training in advanced research technologies—such as 3D modeling—are needed (CSIRO, 2017).

The actors involved in shaping medical innovation need to be reconsidered. Academic healthcare organizations, such as university hospitals, have traditionally been boundary-spanning organizations between care and science. 80. (Lander, 2016; Miller, 2016) The critical role of hospitals and doctors in future demand-led health innovation is undeniable.81(Gulbrandsen et al., 2016; Smits et al., 2008.)

For Turkey, the results of its innovation journey from 2018 to 2022 are of paramount importance. These findings highlight the extent to which Turkey has embraced innovation at the intersection of technology and healthcare. Our findings underscore the impact of Turkey's global innovation performance on digital transformation and innovation in healthcare. Our goal is to improve the healthcare services for individuals, improve expected outcomes, and serve as a guide for policymakers, academics, researchers, computer scientists, and healthcare professionals regarding future strategic paths and action plans. The study, which is expected to serve as a beacon for Turkey, is poised to catalyze transformations and innovations in healthcare.

Peer Review: Externally peer-reviewed.

Author Contributions: Conception/Design of Study- A.A.Ö., Z.A.B.; Data Acquisition- A.A.Ö., Z.A.B.; Data Analysis/Interpretation- A.A.Ö., Z.A.B.; Drafting Manuscript- A.A.Ö., Z.A.B.; Critical Revision of Manuscript- A.A.Ö., Z.A.B.; Final Approval and Accountability- A.A.Ö., Z.A.B.; Material and Technical Support- A.A.Ö., Z.A.B.; Supervision- A.A.Ö., Z.A.B.

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.

ORCID IDs of the authors / Yazarların ORCID ID'leri

Ayça Asena Özdemir 0000-0002-0108-1880
Zehra Alakoç Burma 0000-0002-0376-516X

REFERENCES

- Akdi, Y. (2003). *Time series analysis: unit roots and cointegration*, Knives Bookstore, Ankara.
- Bailey, M. J. (2006). More power to the pill: *The impact of contraceptive freedom on women's life cycle labor supply*. *The quarterly journal of economics*, 121(1), 289-320.
- Barberá-Tomás, D., & Consoli, D. (2012). Whatever works: Uncertainty and technological hybrids in medical innovation. *Technological Forecasting and Social Change*, 79(5), 932-948.
- CSIRO Futures. (2017). Medical technologies and pharmaceuticals: a roadmap for unlocking future growth opportunities for Australia.
- Dunstone, R. L. Commonwealth Scientific and Industrial Research Organization Division of Plant Industry, Canberra, Australia. *JOJOBA*, 157.
- Cutler, D. M., & McClellan, M. (2001). Is technological change in medicine worth it?. *Health affairs*, 20(5), 11-29.
- Freeman, C. (1987). *Technology policy and economic performance: lessons from Japan*. London: Pinter Publishers. s.4.
- Gardner Jr, E. S. (1985). Exponential smoothing: The state of the art. *Journal of forecasting*, 4(1), 1-28.
- Gelijns, A., & Rosenberg, N. (1994). The dynamics of technological change in medicine. *Health affairs*, 13(3), 28-46.
- Granger, C. W. J., & Newbold, P. (2014). *Forecasting economic time series*. Academic press.
- Gulbrandsen, M., Hopkins, M., Thune, T., & Valentin, F. (2016). Hospitals and innovation: Introduction to the special section. *Research Policy*, 8(45), 1493-1498.
- Hernández, H., Grassano, N., Tübke, A., Potters, L., Gkotsis, P., & Vezzani, A. (2018). The 2018 EU Industrial R&D Investment Scoreboard; EUR 29450 EN. *Publications Office of the European Union: Luxembourg*.
- Kenny, C. (2011). *Getting Better: why global development is succeeding—and how we can improve the world even more*. Basic Books.
- Lander, B. (2016). Boundary-spanning in academic healthcare organizations. *Research Policy*, 45(8), 1524-1533.
- Huateng, M. (2019). Application of artificial intelligence and big data in China's healthcare services. *Global innovation index creating healthy lives-the future of medical innovation*. *World Intellectual Property Organization*, 103-110.
- Miller, F. A., & French, M. (2016). Organizing the entrepreneurial hospital: Hybridizing the logics of healthcare and innovation. *Research Policy*,

45(8), 1534-1544.

- Nordhaus, W. D. (2002). The health of nations: the contribution of improved health to living standards. (Working Paper No. 8818). National Bureau of Economic Research.
- R&D Magazine. (2018). 2018 Global R&D Funding Forecast, Winter 2018. Retrieved from <https://www.rdmag.com/>
- Roser, M. (2019). Our world in data: life expectancy. Retrieved from: <https://ourworldindata.org/life-expectancy>
- Sampat, B. (2019). The economics of health innovation: looking back and looking forward. *Global innovation index creating healthy lives-the future of medical innovation. World Intellectual Property Organization*, 81-6.
- Shetty, D. (2019). Innovations in Health-Care Affordability and Delivery—An Indian Perspective. *Global Innovation Index. Global innovation index creating healthy lives-the future of medical innovation. World Intellectual Property Organization*, 163-166.
- Smits, R. E., & Boon, W. P. (2008). The role of users in innovation in the pharmaceutical industry. *Drug discovery today*, 13(7-8), 353-359.
- Tannoury, M., & Attieh, Z. (2017). The influence of emerging markets on the pharmaceutical industry. *Current therapeutic research*, 86, 19-22.
- Tunçsiper, Ç., & Bakar, A. (2023). Sağlık ekonomisi çerçevesinde sağlık harcamaları: Türkiye örneği. *Biga İktisadi ve İdari Bilimler Fakültesi Dergisi*, 4(1), 20-28.
- Tunçsiper, Ç. (2023). Dijital Dönüşüm ve Sağlık Ekonomisi: Dijital Sağlık Üzerine Bibliyometrik Bir Analiz. *Gümüşhane Üniversitesi Sağlık Bilimleri Dergisi*, 12(1), 21-31.
- The Global Innovation Index 2019*. Retrieved February 10, 2023, from <https://www.wipo.int/publications/en/details.jsp?id=4434>
- Thune, T., & Mina, A. (2016). Hospitals as innovators in the health-care system: A literature review and research agenda. *Research policy*, 45(8), 1545-1557.
- WIPO (World Intellectual Property Organization). (2015). *A Look Inside the Economic Growth Engine [Chapter 1]*. *World Intellectual Property Report 2015: Breakthrough Innovation and Economic Growth (pp. 21-46)*. Geneva: World Intellectual Property Organization
- WIPO (World Intellectual Property Organization). (2018). *Patent applications and grants worldwide, World Intellectual Property Indicators 2018*. Geneva: World Intellectual Property Organization).

How cite this article / Atıf Biçimi

Ozdemir, A.A., Alakoc Burma, Z. (2023). Digital transformation and innovation in health for future health services: Türkiye global innovation index time series analysis between 2018 and 2022. *Acta Infologica*, 7(2), 317-332. <https://doi.org/10.26650/acin.1352261>

Comparison of Outlier Detection Methods in Linear Regression: A Multiple-Criteria Decision-Making Approach

Doğrusal Regresyonda Uç Değer Tespit Yöntemlerinin Karşılaştırılması: Çok Kriterli Karar Verme Yaklaşımı

Mehmet Hakan Satman¹ 

¹(Prof.Dr.), Istanbul University, Faculty of Economics, Department of Econometrics, Beyazıt, Istanbul, Türkiye

Corresponding author : Mehmet Hakan SATMAN
E-mail : mhsatman@istanbul.edu.tr

ABSTRACT

This paper focuses on the application of a suite of simulation studies to assess well-known and contemporary outlier detection methods in linear regression. These simulations vary across different parameters, including the number of observations, parameters, levels, and direction of contamination. The recorded final parameter estimates are used to rank the methods using Multiple-criteria decision-making (MCDM) tools. The study reveals that method success varies based on simulation settings. MCDM analysis results indicate a limited set of applicable methods when the contamination structure and level are unknown. Additionally, the most successful methods demand increased computation time, while some alternatives exhibit applicability within shorter durations with median rankings. These findings offer valuable insights for researchers employing regression analysis in scenarios where the underlying model is known, and the possibility of potential outliers exists.

Keywords: outlier detection, robust regression, linear regression, decision analysis

ÖZ

Bu makale, doğrusal regresyonda bilinen ve çağdaş aykırı değer tespit yöntemlerini değerlendirmek için bir dizi simülasyon çalışmasının uygulanmasına odaklanmaktadır. Bu simülasyonlar, gözlem sayılarının, parametre sayılarının ve kirlenmenin yönü ve oranı dahil olmak üzere farklı parametreler için gerçekleştirilmiştir. Kaydedilen nihai parametre tahminleri ve Çok Kriterli Karar Verme (ÇKKV) araçları kullanılarak tahmincilerin sıralanması sağlanmıştır. Çalışma, tahmincilerin başarısının simülasyon ayarlarına bağlı olarak değiştiğini ortaya koymaktadır. ÇKKV analizi sonuçları, kirlenme yönünün ve oranının bilinmediği durumlarda uygulanabilecek tahmincilerin sınırlı sayıda olduğunu göstermektedir. Ayrıca, en başarılı yöntemler artan hesaplama zamanı gerektirirken, bazı alternatifler orta sıralamalarla kısa süreler içinde uygulanabilirlik göstermektedir. Bu bulgular, altta yatan modelin bilindiği ve potansiyel aykırı değerlerin olabileceği senaryolarda regresyon analizi kullanan araştırmacılar için değerli öngörüler sunmaktadır.

Anahtar Kelimeler: uçdeğer teşhisi, dayanıklı regresyon, doğrusal regresyon, karar analizi

Submitted : 14.07.2023
Revision Requested : 09.11.2023
Last Revision Received : 09.11.2023
Accepted : 10.11.2023
Published Online : 14.12.2023



This article is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0)

1. INTRODUCTION

Suppose the linear regression model is

$$y = X\beta + \varepsilon$$

where y is the n -vector of the response variable, X is the design matrix, β is the unknown vector of regression parameters, ε is the i.i.d. error-term with zero mean, p is the number of parameters, and n is the number of observations. The Ordinary Least Squares (OLS) estimator

$$\hat{\beta} = (X'X)^{-1}X'y$$

is an unbiased and efficient estimator of β , that is,

$$E(\hat{\beta} - \beta) = \mathbf{0}$$

and variance of $\hat{\beta}$ is minimum among the other unbiased estimators when some conditions are held. This implies

$$MSE(\hat{\beta}) = [Bias(\hat{\beta})]^2 + Var(\hat{\beta})$$

is minimum where *MSE* is *Mean Square Error*.

When data includes unusual observations (a.k.a. outliers), properties of OLS may drastically change depending on the level of contamination. In the case of single outlier, one-leave-out techniques and regression diagnostics are successfully applied (Belsley et al., 1980; Hadi and Chatterjee, 2015). When the level of contamination is high and known, m -leave-out techniques can be used instead but these class of methods may not be applicable as the number of all possible subsets tend to be quite large where $m < n$ is the number of outliers. In addition to this, the number of outliers is generally unknown.

Outlier detection and robust regression methods seek a solution for the outlier problem in linear regression in different but similar ways. An outlier detection procedure simply performs computation iterations to reveal the outliers. Contrary, a robust regression estimator tries to estimate an outlier-free $\hat{\beta}$ without inherently labelling any observations as clean or contaminated. When an outlier detection algorithm reports a set of contaminated observations then a robust estimate of β can be obtained by removing the contaminated observations from the data. A robust regression estimate of β can also be used to delete observations using a predefined threshold.

In this paper, 17 outlier detection and robust regression methods are simulated using a suite of Monte Carlo studies. In the simulations, Mean Square Error of estimated parameters are evaluated. The methods are ranked in the context of a multi-criteria decision-making analysis (MCDM). It is shown that the algorithms fail in many situations depending on the number of observations, number of parameters, level of contamination, and the direction of outliers. The MCDM analysis shows that only a small subset of techniques are applicable when the properties of outliers are unknown.

In Section 1 the problem and the context of the paper is introduced. In Section 2 we introduce the methods and estimators simulated in this study. The MCDM methods used in the decision analysis are also introduced. In Section 3 simulation and MCDM analysis results are reported. Finally in Section 4, we discuss the results and conclude.

2. MATERIALS AND METHODS

2.1. Outlier Detection Methods

2.1.1. *hs93*, *bacon*, and *bch2006*

hs93, *bacon*, and *bch2006* are multi-stage outlier detection methods and they are introduced in the same place as they follow similar patterns by construct. *hs93* is a multi-stage method and starts with an initial subset with size of $p + 1$ in its first stage (Hadi and Simonoff, 1993). The observations with lowest DFFITS regression diagnostics are used to construct the initial basic subset. In the second step, the initial basic subset is used to construct a basic subset by enlarging the former by adding new observations. In the last stage the subset obtained from the former stages is enlarged until a test statistic exceeds a threshold. The threshold is selected as α -quantiles of Student's T Distribution with degrees of freedom $s - p$ where s is the number of observations held by the latest subset.

bacon (Blocked Adaptive Computationally efficient Outlier Nominators) is a multi-stage outlier detection method (Billor et al., 2000). In the first stage, an initial basic subset is created which is considered as free of outliers. In this

stage, a sample of $p + 1$ observations is created and enlarged until the basic subset includes up to m observations. This $p + 1$ sized sample is constructed through a multivariate outlier detection algorithm which is only applied on the design matrix. The method is iterated until a specific t-statistic reaches a predefined cut-off value. The method requires the parameter m to be set.

bch2006 is a multi-stage outlier detection method and it shares similar patterns to that used in the bacon procedure (Billor et al., 2006). The method initially calculates the Mahalanobis distances for all rows of the design matrix excluding the intercept using the coordinate-wise median instead of the sample mean for the location estimate. Best h observations are selected to build a vector of squared Mahalanobis distances. The generated basic-subset is then fed into an iteratively weighted least squares procedure, and this step is iterated until a maximum number of iterations is reached.

2.1.2. cm97 and ccf

cm97 starts with construction of weights using the diagonal elements of the hat matrix using the formula

$$w_i = \frac{1}{\max(H_{ii}, \bar{p})}$$

where $H = X(X'X)^{-1}X'$ and $\bar{p} = p/n$ (Chatterjee and Mächler, 1997). A weighted least squares regression is applied using the weights w_i . The weights are updated using the formula

$$w_i = \frac{(1 - H_{ii})^2}{\max(|r_i|, m)}$$

until the estimated regression coefficients are stabilized where r_i is the i th residual and m is the sample median of absolute residuals.

ccf is a fast regression method that is robust to outliers and shares a similar logic with the cm97 method. The method starts with a weighted least squares estimation with i th weight is set to $w_i = n/2$ for all observations by default (Barratt et al., 2020). The weights are updated using the formula

$$w_i = \Gamma \text{sign}(e_i^2 - \alpha)$$

where $\alpha = p \times \sum_i^n e_i^2$, and e_i is the i th residual. The authors suggest selecting the Γ parameter as 0.1. The iterations of weight updating are repeated until a predefined maximum number of iterations is reached.

2.1.3. imon2005

imon2005 implements a robust version of the well-known regression diagnostics DFFITS, namely GDFFITs (Rahmatullah Imon, 2005). The method starts with constructing an outlier-free h -subset through a robust fit estimator. The authors suggest using lms but any other robust fitting algorithm can be used instead. It is also suggested that the observations with GDFFITs statistic that exceed $3\sqrt{\frac{p}{h}}$ are labelled as outliers.

2.1.4. ks89

ks89 method starts with calculating Studentized residuals and considers the first p observations regarding the corresponding smallest values (Kianifard and Swallow, 1989). The initial subset is enlarged using the recursive residuals. The recursive residuals are calculated using the formula

$$w_k = (y_k - X'_k \hat{\beta}) / \sqrt{1 + X'_k (X^{*'} X^*)^{-1} X_k}$$

where w_k is the k th recursive residual, X^* is the subset of the design matrix with elements corresponding to first $k - 1$ smallest recursive residuals. The iterations are repeated until $k = n$. The observations that have standardized recursive residuals greater than a specific threshold are labelled as outliers. The threshold can be selected as α -quantiles of a Student's T distribution with degrees of freedom $n - p - 1$.

2.1.5. *lad and quantilereg*

The *lad* (Least Absolute Deviations) estimator minimizes the sum of absolute residuals and has a unique solution obtained by a goal programming context (Narula et al., 1999). Supposing e_i^- and e_i^+ denote the i th residual, $e_i^- > 0$ if the i th residual is negative, $e_i^+ > 0$ if the i th residual is positive, otherwise it fits the regression equation. The linear objective function

$$\min z = \sum (e^- + e^+)$$

is minimized subject to the constraints

$$X\beta + e^- - e^+ = y$$

where $e_i^- \geq 0$, $e_i^+ \geq 0$, $\beta_j \in \mathcal{R}$, $i = 1, 2, \dots, n$, and $j = 1, \dots, p$. Similarly, *quantilereg* (Quantile Regression) estimates a predefined conditional quantile of the response variable y (Yu et al., 2003). *quantilereg* regression parameters minimizes the linear objective function

$$\min z = \sum [(1 - \tau)e^- + \tau e^+]$$

under the same constraints of *lad* where $0 \leq \tau \leq 1$. When τ is set to 0.25, 0.50 or 0.75, well-known conditional quartiles are estimated. Note that any other percentile value can be selected, instead. When τ is set to 0.50, the conditional median of the response variable is estimated given a set of exploratory variables.

2.1.6. *lms, lts, and lta*

lms (Least Median of Squares) estimator

$$\min \text{median } e^2$$

minimizes the sample median of squared residuals (Rousseeuw, 1984), whereas, *lts* (Least Trimmed Squares) estimator

$$\min \sum_{i=1}^h e_i^2$$

minimizes the sum of first h ordered squared residuals where h is at least half of the data (Rousseeuw and Van Driessen, 2006). Similarly *lta* estimator

$$\min \sum_{i=1}^h |e_i|$$

minimizes the sum of the first h ordered absolute residuals (Hawkins and Olive, 1999). Since the objective function of these estimators are not in closed-form, the estimation process requires comprehensive iterations. Rousseeuw (1984) proposed a random sampling based algorithm for *lms*. Rousseeuw and Van Driessen (2006) devised a fast algorithm for *lts* in which a couple of samples of size p are randomly drawn and enlarged to size h using *concentration steps* (c-steps).

2.1.7. *py95*

py95 is a method in which the eigen structure of

$$M = \frac{1}{ps^2} EDHDE$$

matrix is investigated where $s^2 = \sum e^2 / (n - p)$, H is hat matrix, D is $n \times n$ diagonal matrix with elements $1/(1 - H_{ii})$, E is $n \times n$ diagonal matrix with elements e_i , e_i is the i th residual (Peña and Yohai, 1995). Differently, *py95* reports

suspicious observations rather than absolute outliers. Suppose that the v is one of the eigenvectors of M . Let $a_i = v_i/v_{i-1}$ for $i = n, n - 1, \dots, c_1$, $b_j = v_j/v_{j+1}$ for $j = 1, 2, \dots, c_2$, $c_1 = c_2 = \lfloor n/4 \rfloor$, and oc is vector of ordered coordinates. If none of $a_i > k$ for $i \in oc$ and $b_j > k$ for $j \in oc$, then there is no any suspicious observations where k can be selected as 2.5. Otherwise, the method returns the set of suspicious outliers.

2.1.8. *satman2013 and satman2015*

satman2013 is a two-stage method for detecting outliers in linear regression (Satman, 2013). In the first stage of the method, a subset of outlier-free observations is created using a robust covariance matrix estimation inspired by the *Comediance* statistic (Huo et al., 2012). This covariance matrix is calculated in reasonably small times when it is compared to the MVE and MCD (Van Aelst and Rousseeuw, 2009; Rousseeuw and Driessen, 1999) but lacks a couple of nice statistical properties such as rotation invariance. The method continues with a weighted least squares estimation using the weights obtained by the former stage. Finally, the method iterates *c-steps* defined in (Rousseeuw and Van Driessen, 2006) using the clean subset of observations obtained.

Similarly, *satman2015* (Satman, 2015) is also a two-stage method but it differs in constructing the basic subset. Instead using the *Comediance* measure, the method constructs an initial subset using the design matrix by applying a multi-dimensional sorting algorithm, e.g. non-dominated sorting algorithm defined in (Deb, 2015). A $p + 1$ subset of initial subset of observations are selected from the most-middle of the data. This selection method is not invariant to affine transformations.

2.1.9. *smr98 and asm2000*

smr98 algorithm starts with an OLS estimation (Sebert et al., 1998). A single-linkage clustering is then applied on the standardized pairs of \hat{y} and $\hat{\epsilon}$. The cluster tree is cut using the Mojana criterion

$$\bar{h} + 1.25\sigma_h$$

where h is the vector of heights of dendrogram branches. Clusters with the majority of observations are labeled as clean. The standardized pairs of $(\hat{y}, \hat{\epsilon})$ play a role of dimension reduction, so the algorithm works perfectly when the number of regressors is small, e.g. $p = 2$. The performance of the algorithm drastically reduces in higher dimensions. *asm2000* solves this problem by applying a robust fit at the very early steps of the *smr98* algorithm. The clustering stage is based on the robust estimates of \hat{y} and $\hat{\epsilon}$ (Adnan et al., 2000).

2.2. SIMULATION STUDY

In the simulation study, regression data is created using the following data generating process: The number of observations and the number of regression parameters selected as $n = 100, 500, 1000$ and $p = 5, 10, 25$, respectively. Each single design matrix has 1s in the first column, that is, the models include an intercept term. Exploratory variables and the error term are drawn from independent Normal distributions with zero mean and unit variance. Regression parameters are set to $[5, 5, \dots, 5]$. Regression data is then contaminated either in x - and y - directions with the ratios of $c = 0.10, 0.20, 0.30$. Variables are contaminated using the formula

$$V_i = \max(V) + r_i$$

where r_i is a random value drawn from a Uniform(0, 5) distribution, V is either the response variable or columns of the design matrix excluding the intercept, and $\max(V)$ is the maximum value of V including the V_i . x - outlier observations are contaminated in all dimensions.

Figure 1(a) represents a random data contaminated in x - direction. As the used contamination formula indicates, outlier values are at least distant as the maximum value of the majority of observations. Similarly, Figure 1(b) represents a random data with outliers in y - direction. Note that the configuration of $p = 2$ is never used in simulations but the same logic is applied in greater dimensions of spaces.

If the method \mathcal{M} is a robust regression estimator, then the reported $\hat{\beta}_i$ is used to calculate the MSE values. If the method \mathcal{M} is an outlier detection method and the reported outlier set is \mathcal{S} then OLS parameters are estimated using the complement set of \mathcal{S} . By this setting, all of the methods are considered as regression estimators and low MSE values are the signals and indicators of the low masking and swamping effects of method \mathcal{M} . Since the data generating process differs in the number of parameters, *mean of mean square errors* (mmse) are calculated and presented for each single setting in the simulation results.

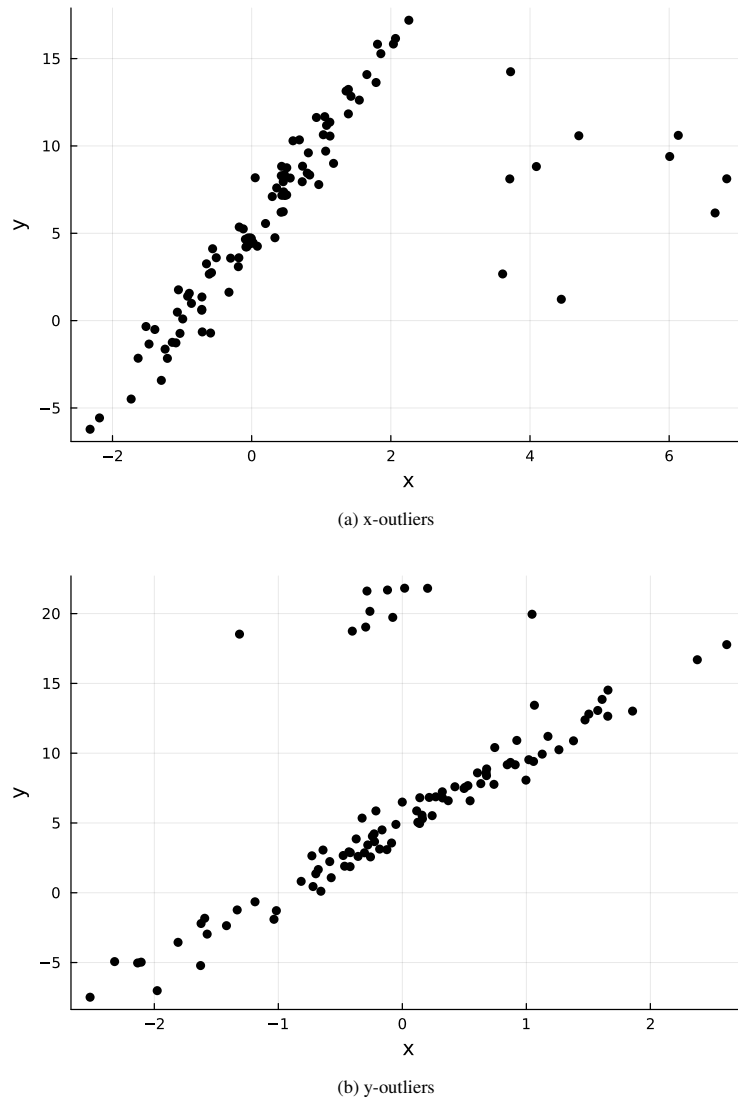


Figure 1: Simulation data with $n = 100$, $p = 2$, and $c = 0.10$

2.3. Multiple-criteria Decision-Making Tools

Suppose a multiple-criteria decision problem is presented as in the Table 1 where C_1, \dots, C_m are criteria, w_1, \dots, w_m are weights of criteria, $g_1(A_k), \dots, g_m(A_k)$ are functions that takes a cost or a gain value for the alternative A_k , and A_1, \dots, A_n are alternatives. Since a sorting operation requires the \leq operator is defined for elements of vector A , Table 1 is not called to be *sortable* because of the operator \leq is not defined in \mathcal{R}^m (or at least it doesn't have an exact and unique definition), A_1, A_2, \dots, A_n are not sortable.

Multiple-criteria decision-making tools are methods which are defined for sorting, a.k.a. ranking, alternatives A_1, \dots, A_n by using different kinds of comparing operators. The TOPSIS method (Hwang and Yoon, 1981) scores the alternatives using the Euclidean distance of weighted normalized A_i vectors to best-ideal and worst-ideal vectors. VIKOR (Opricovic, 1998; Opricovic and Tzeng, 2002) scores the alternatives using the formula

$$v \frac{s_i - \min s}{\max s - \min s} + (1 - v) \frac{r_i - \min r}{\max r - \min r}$$

where s_i and r_i are the sum and maximum of the i th row of weighted normalized decision matrix, respectively, and v can be selected as 0.5. ARAS (Zavadskas and Turskis, 2010) creates an extended decision matrix by adding an additional row that contains ideal values of all alternatives. A vector of *Utility degrees* is then formed to score alternatives. WASPAS (Zavadskas et al., 2012) utilities scores by using the product of normalized decision matrix and row sums. COPRAS (Zavadskas et al., 1994) scores the alternatives using the formula

Criteria	C_1	C_2	\dots	C_m
Weights	w_1	w_2	\dots	w_m
Functions	f_1	f_2	\dots	f_m
A_1	$g_1(A_1)$	$g_2(A_1)$	\dots	$g_m(A_1)$
A_2	$g_1(A_2)$	$g_2(A_2)$	\dots	$g_m(A_2)$
\vdots	\vdots	\vdots	\vdots	\vdots
A_n	$g_1(A_n)$	$g_2(A_n)$	\dots	$g_m(A_n)$

Table 1: A generic multiple-criteria decision problem

$$S_i = \frac{Q_i}{\max Q}$$

where $i = 1, 2, \dots, n$, $Q_i = s_i^+ + \sum_{i=1}^n \frac{s_i^-}{s_i^+ Z}$, $Z = \sum_{i=1}^n 1/s_i^-$, s_i^+ and s_i^- are sums of rows of the normalized decision matrix regarding to the direction of optimization, e.g. either maximization or minimization, respectively.

Selection of criteria weights depends on the researcher and it is generally subjective. CRITIC (Diakoulaki et al., 1995) is an automatic method for selecting the importance level of criteria, a.k.a. weights. CRITIC weights are calculated using

$$w_j = s_j / \sum s$$

where s_j is the score of the j th criterion defined as

$$s_j = N_j \sum \mathcal{F}_j$$

and N_j is standard deviation of j th column of the normalized decision matrix, \mathcal{F}_j is j th column of the matrix \mathcal{F} , $\mathcal{F} = \mathbf{1} - \hat{\Sigma}$, $\hat{\Sigma}$ is the sample correlation matrix of the normalized decision matrix.

2.4. The Software

Simulation study and the multiple-criteria decision-making tools are applied with Julia (Bezanson et al., 2017). Julia is a fast, dynamic and compiled programming language that is mostly used in scientific computing. Selection of the programming language is mostly pragmatic as the simulation study requires 54000 iterations for each single estimator and the required functionality is packed compactly in a single environment. The Julia package *LinRegOutliers* is used in simulations (Satman et al., 2021a). This package implements all of the estimators used in this study purely in Julia. The multiple-criteria decision-making analysis is applied using the Julia package *JMcDM* (Satman et al., 2021b).

The methods of the *LinRegOutliers* package are implemented in a unified way and they are called in a scheme of

`method(X, y)`

where X is the design matrix, y is the response vector, and `method` is either `lms`, `lts`, `hs93`, etc. The *JMcDM* package is implemented in a similar way and a single MCDM method is called like

`method(decisionMat, weights, directions)`

where `decisionMat` is the decision matrix, `weights` is the vector of weights of criteria, and `directions` is the vector of directions of optimizations which can be either `minimum` or `maximum`. `method` is the function name and it can take values `topsis`, `waspas`, `copras`, etc. Use of the methods are explained in a great detail in papers Satman et al. (2021a) and Satman et al. (2021b), respectively.

3. RESULTS

Tables 2 - 4 summarize the simulation results. In these tables, average MSE values of estimates ($\hat{\beta}$) are reported for different contamination ratios ($c = 0.10, 0.20, 0.30$), outlier direction (either in x-space or y-space), and number of parameters ($p = 5, 10, 25$).

Table 2 summarizes the simulation results for $n = 100$. When the contamination in x-space is low ($c = 0.10$) and $p = 5, 10$; bacon, asm2000, lts, and lta have relatively smaller mmse values. hs93 comes into scenes in higher dimensions ($p = 25$). This situation is also current for higher contamination rates ($c = 0.20, 0.30$). When $p = 25$ and contamination rates are higher, hs93, bacon, and lta have better performance.

In the case of y-outliers and $n = 100$, most of the methods are fine except bacon, imon2005, and ccf for small contamination rates ($c = 0.10$). When the contamination rate is increased to 0.20, smr98, py95, ks89, lms, satman2015 tend to have larger mmse values. When the contamination rate is maximum, the winners are lts, asm2000, and hs93 with distant mmse values compared to the remaining ones.

Table 3 summarizes the results for $n = 500$. When the dimensionality and the contamination is low ($p = 5, c = 0.10$); bacon, asm2000, imon2005, satman2013, lts, lms, and lta have better performance in the presence of x-outliers. The list remains the same when $p = 10$. In higher dimensions ($p = 25$) satman2013 is replaced by hs93 by their corresponding mmse values. A small subset of the list survives in higher dimensions and higher contamination rates. bacon, hs93, and lta are successors for $c = 0.20$. When $c = 0.30$, only bacon and hs93 have relatively smaller mmse values.

In the case of y-outliers and $n = 100, p = 5$, and $c = 0.10$, the methods have similar performance by means of mmse. This situation remains the same for higher dimensions and contamination rate. In the worst case of $p = 25$ and $c = 0.30$ hs93, bch2006, satman2013, lts, lad, quantilereg, and cm97 have relatively smaller mmse values and can be considered as applicable.

Table 4 summarizes the results for $n = 1000$. bacon, asm2000, imon2005, smr98, satman2013, lts, lta, and lms have better performance for $p = 5$ and $c = 0.10$ in the presence of x-outliers. In the case of high contamination rates only asm2000, lts, and lta have distant mmse values to the remaining elements of the list. In the worst case of $p = 25$ and $c = 0.30$, hs93 and bacon are well ahead regarding their low mmse values.

In the case of y-outliers and $n = 1000, p = 5$, and $c = 0.10$, all of the methods are applicable. When $c = 0.20$ and $p = 10$ imon2005 and ccf exit the list. In the worst case of $p = 25$ and $c = 0.30$ most of the methods are applicable except ks89, py95, lta, lms, imon2005, ccf, and satman2015.

Success of methods differ regarding the number of observations, the number of parameters, the contamination rate, and the direction of contamination. However, these factors are generally unknown by the researcher, that is, the multivariate data is not visible to plots even a dimension reduction tool is applied to data¹. As a consequence, the researcher is almost blind to direction of outliers and the contamination ratio.

Table 5 represents the scores calculated by TOPSIS, VIKOR, ARAS, WASPAS, and COPRAS methods to the decision matrix of simulation results. In the decision matrix, rows (the alternatives) are the methods. The criteria are formed by the simulation settings. The i th row and the j th column of the decision matrix represents the mmse of the method M_i for regression setting f_j . In Table 5 it is shown that the o1s has the lowest rank by all of the methods since the simulation data is always contaminated. The other methods have higher scores as expected. asm2000 is in the top three for all MCDM methods. hs93 is in the top three for 4 out of 5 methods whereas lta takes a place for 3 out of 5 methods. lta, hs93, asm2000, lts, bacon take a place for at least one MCDM method.

The success of the methods is compared in terms of computation time as well as mmse values. Table 6 represents the average absolute times and relative times elapsed by the methods².

Table 6 shows that the statistical properties and the consumed times of methods are related as the most successful methods hs93 and lta consume more time than the others. ccf is also consistent as it has lower ranks by the MCDM and lower computation times. satman2013 is an interesting method as it takes 8th or 9th row in the rankings with its relatively small computation times. cm97 has similar speed properties with lower rankings. The cheapest-success method is bacon as it takes higher rankings with median computation times.

¹ Classical covariance matrix based methods are not robust to outliers and, for instance, *Principal Component Analysis* requires a robust covariance matrix to be estimated in the presence of outliers. Performance of covariance estimators (by means of 1. Detecting the true outliers, 2. Rejecting the false outliers, 3. Unbiased estimate of the location vector, 4. Efficient estimation of variance, etc.) is another issue and this subject is out of scope of this paper

² The absolute calculation times are average of all computation times in all simulation settings. These elapsed times are measured using a MacBook Pro with 8GB of memory and 2Ghz 8-core CPU. Since the elapsed times differ in many hardware configurations, relative average times are reported. The time consumed by o1s is set to 1x. Other methods' average times are divided by absolute elapsed time of o1s. By this representation, the elapsed times are directly comparable.

4. DISCUSSIONS AND CONCLUSIONS

Robust regression methods take the model and a dataset as input and return the estimate of regression parameters whereas outlier detection methods take the same input and return a set of indices of outliers. When the reported outlier set is omitted from the data, the estimated OLS parameters are considered robust. In this study the well-known and modern outlier detection methods and robust regression methods are simulated for different number of observations, the number of parameters, levels of contamination, and direction of contamination. MSE of estimated parameters are recorded during the simulations. Simulation results show that the success of a method differs regarding the simulation setting. However, the researchers are generally not aware of the underlying data generating process and the contamination structure and selection of the proper method is a decision problem.

Multiple-criteria decision-making (MCDM) tools are generally used by ranking the alternatives using a given set of criteria and importance levels of these criteria. TOPSIS, Vikor, ARAS, WASPAS, and COPRAS are some well-known MCDM tools applied in the decision making literature. In this paper, these MCDM tools are used to rank outlier detection methods by their average MSE (mmse) of parameter estimates. The criteria are formed by each single setting of the data generating process. Since the importance level is unknown or subjective, the CRITIC method is used to determine a set of weights.

The results of the MCDM analysis show that the *ols* estimator has the lowest rank as expected just because the simulation study is performed on the contaminated data. All of the MCDM tools scored *asm2000* in the top three whereas *hs93* is ranked in top three for four out of five listings. *lta* takes the place of three out of five MCDM methods in the top 3. *lta*, *hs93*, *asm2000*, *lts*, *bacon* are top-ranked for at least one MCDM method. If the researcher has no idea of the underlying data generating process, results of these methods can be considered.

The computation times consumed by the methods are also reported. It is shown that the more successful methods take more computation time. *satman2013* is an interesting method as it takes 8th or 9th row in the rankings with its relatively small computation times. *cm97* has similar speed properties with lower rankings. The cheapest-success method is *bacon* as it takes higher rankings with median computation times. If the consumed time is an issue, *bacon* can be used with reasonably small MSE of estimates in many settings.

If the direction and level of contamination is known, results of the simulations are directly comparable. When $n = 1000$, $p = 25$ and the contamination is at the maximum level, *hs93* is the most performant method by means of lower MSE. If the direction of contamination is known and the presence of y -outliers is the case, *hs93*, *bch2006* are the absolute winners with a small time difference.

The simulation results is a confirmation of the previous simulation studies reported in [Billor and Kiral \(2008\)](#) and [Wisnowski et al. \(2001\)](#) in some sights. Instead of reporting the masking and swamping ratios, this study is original as it reports MSE of estimated parameters. The former studies utilize a comprehensive study with a wider range of contamination levels and extra contamination directions and structures. Our study differs as it tests the methods in larger data sets including the ones with $n = 1000$, $p = 25$ and covers a wider and novel set of methods to compare.

Combining the building blocks of successful methods for a faster and more robust outlier detection procedure and developing new methods would be the subject of future works.

Algorithm	$c = 0.10, d = x$			$c = 0.10, d = y$		
	$p = 5$	$p = 10$	$p = 25$	$p = 5$	$p = 10$	$p = 25$
asm2000	0.012	0.012	0.016	0.022	0.030	0.048
bacon	0.012	0.015	0.090	0.012	0.015	3.205
bch2006	16.400	30.081	43.537	0.041	0.040	0.055
ccf	17.682	23.629	32.232	0.135	1.402	7.934
cm97	17.134	24.072	35.438	0.021	0.021	0.028
hs93	0.216	0.470	0.073	0.012	0.012	0.016
imon2005	2.089	23.562	32.354	0.857	1.621	6.242
ks89	17.596	24.594	36.397	0.040	0.074	0.334
lad	17.944	24.881	37.132	0.025	0.026	0.032
lms	0.048	0.046	0.078	0.048	0.048	0.077
lta	0.079	0.083	0.109	0.076	0.084	0.112
lts	0.066	0.060	0.148	0.068	0.061	0.052
ols	18.256	23.713	32.329	2.505	3.776	7.696
py95	13.307	24.140	38.775	0.044	0.068	0.271
quantilereg	18.003	24.949	36.807	0.025	0.025	0.031
satman2013	0.350	13.780	38.723	0.061	0.051	0.046
satman2015	20.823	29.247	48.788	0.059	0.050	0.089
smr98	5.139	21.288	36.532	0.014	0.026	0.139

Algorithm	$c = 0.20, d = x$			$c = 0.20, d = y$		
	$p = 5$	$p = 10$	$p = 25$	$p = 5$	$p = 10$	$p = 25$
asm2000	0.013	0.014	37.395	0.090	0.084	0.202
bacon	0.055	0.017	0.086	0.135	0.491	41.722
bch2006	22.286	30.565	46.163	0.038	0.045	0.105
ccf	19.788	24.718	34.325	2.418	10.052	17.402
cm97	19.884	25.439	37.679	0.054	0.047	0.076
hs93	0.870	0.443	0.019	0.014	0.015	0.030
imon2005	15.372	24.660	34.233	5.648	8.908	17.886
ks89	20.084	26.077	39.141	0.508	1.064	6.031
lad	20.150	26.220	39.578	0.051	0.048	0.075
lms	0.059	0.084	57.019	0.062	0.084	7.460
lta	0.075	0.094	0.229	0.076	0.092	0.225
lts	0.059	0.053	49.205	0.058	0.056	0.053
ols	19.719	24.674	34.308	8.494	10.297	18.027
py95	20.043	26.880	43.595	1.100	1.291	4.736
quantilereg	20.157	26.284	40.096	0.050	0.050	0.075
satman2013	18.126	27.329	42.747	0.049	0.048	0.159
satman2015	22.551	31.344	58.984	0.150	0.728	20.408
smr98	16.419	26.095	40.112	0.043	0.331	3.717

Algorithm	$c = 0.30, d = x$			$c = 0.30, d = y$		
	$p = 5$	$p = 10$	$p = 25$	$p = 5$	$p = 10$	$p = 25$
asm2000	0.016	0.017	51.756	0.307	0.367	0.394
bacon	9.324	7.070	7.990	1.146	4.071	102.800
bch2006	22.895	32.088	48.924	0.038	0.041	5.497
ccf	20.510	25.398	36.658	12.891	21.031	30.651
cm97	20.558	26.297	41.145	0.190	0.268	10.264
hs93	1.205	0.045	0.023	0.021	0.020	1.563
imon2005	19.499	25.498	36.469	14.851	20.330	30.489
ks89	20.942	26.957	41.661	3.172	6.749	24.486
lad	21.029	27.221	43.547	0.126	0.153	7.418
lms	0.096	4.699	82.574	0.089	0.785	39.316
lta	0.071	0.114	44.034	0.072	0.115	28.762
lts	0.049	1.423	71.372	0.049	0.047	0.120
ols	20.249	25.424	36.646	17.996	21.026	29.899
py95	20.998	27.834	54.573	8.208	9.528	24.205
quantilereg	20.993	27.183	43.366	0.128	0.134	8.077
satman2013	21.549	29.471	48.733	0.045	0.055	10.647
satman2015	23.830	35.256	89.485	63.767	63.865	88.561
smr98	20.285	27.583	45.360	1.422	4.714	26.167

Table 2: Average MSE for $n = 100$

Algorithm	$c = 0.10, d = x$			$c = 0.10, d = y$		
	$p = 5$	$p = 10$	$p = 25$	$p = 5$	$p = 10$	$p = 25$
asm2000	0.002	0.002	0.002	0.004	0.004	0.010
bacon	0.002	0.002	0.002	0.002	0.002	0.002
bch2006	7.936	22.874	29.937	0.007	0.009	0.010
ccf	17.371	21.816	24.900	0.090	0.813	3.769
cm97	17.423	21.799	25.203	0.008	0.006	0.004
hs93	0.130	0.327	0.252	0.002	0.002	0.002
imon2005	0.004	0.004	9.546	0.803	0.881	1.254
ks89	17.270	21.878	25.619	0.013	0.018	0.032
lad	17.510	21.991	25.626	0.008	0.006	0.005
lms	0.021	0.027	0.579	0.021	0.027	0.125
lta	0.029	0.042	0.070	0.029	0.043	0.072
lts	0.018	0.016	0.012	0.018	0.017	0.013
ols	17.827	21.803	24.907	2.629	3.022	3.821
py95	14.599	21.698	25.768	0.016	0.021	0.034
quantilereg	17.455	22.000	25.650	0.008	0.006	0.005
satman2013	0.016	3.474	27.198	0.016	0.015	0.011
satman2015	18.312	23.765	28.833	0.016	0.014	0.011
smr98	0.161	8.767	25.019	0.003	0.003	0.003

Algorithm	$c = 0.20, d = x$			$c = 0.20, d = y$		
	$p = 5$	$p = 10$	$p = 25$	$p = 5$	$p = 10$	$p = 25$
asm2000	0.002	0.003	14.801	0.027	0.044	0.028
bacon	0.003	0.003	0.003	0.003	0.011	0.018
bch2006	19.446	24.112	30.776	0.006	0.008	0.009
ccf	19.037	22.466	25.357	2.324	10.111	12.729
cm97	19.029	22.552	25.664	0.032	0.019	0.011
hs93	1.517	0.295	0.106	0.002	0.003	0.003
imon2005	0.524	3.177	25.279	6.378	7.638	9.953
ks89	19.055	22.757	26.167	0.390	0.500	0.803
lad	19.105	22.717	26.133	0.026	0.016	0.011
lms	0.072	1.764	24.534	0.067	0.321	6.204
lta	0.031	0.055	0.148	0.030	0.054	0.146
lts	0.015	0.013	18.859	0.014	0.014	0.012
ols	19.050	22.445	25.345	9.879	11.094	12.625
py95	19.067	22.907	26.260	0.558	0.716	0.777
quantilereg	19.072	22.716	26.229	0.026	0.016	0.011
satman2013	17.174	23.631	28.475	0.013	0.012	0.010
satman2015	20.096	24.512	29.793	0.014	0.012	0.010
smr98	7.287	21.515	25.999	0.003	0.003	0.003

Algorithm	$c = 0.30, d = x$			$c = 0.30, d = y$		
	$p = 5$	$p = 10$	$p = 25$	$p = 5$	$p = 10$	$p = 25$
asm2000	0.003	0.327	26.057	0.188	0.098	0.163
bacon	10.414	7.387	1.362	1.601	0.675	0.585
bch2006	20.182	24.834	31.424	0.006	0.005	0.008
ccf	19.508	22.750	25.615	15.693	23.772	26.795
cm97	19.524	22.869	26.008	0.122	0.070	0.041
hs93	3.272	0.593	0.003	0.003	0.003	0.003
imon2005	5.178	16.416	25.854	18.310	20.939	25.607
ks89	19.591	23.045	26.509	2.876	3.510	5.491
lad	19.606	23.067	26.615	0.075	0.044	0.027
lms	1.799	14.530	27.409	0.539	4.349	16.216
lta	0.033	0.077	8.549	0.034	0.078	7.022
lts	0.011	0.682	34.699	0.011	0.011	0.010
ols	19.496	22.740	25.621	22.363	24.151	26.671
py95	19.591	23.210	26.576	7.666	7.348	6.340
quantilereg	19.619	23.059	26.656	0.075	0.044	0.027
satman2013	19.955	24.459	29.300	0.011	0.010	0.009
satman2015	20.577	25.150	31.301	153.840	111.669	63.315
smr98	14.966	22.961	26.526	0.004	0.300	0.927

Table 3: Average MSE for $n = 500$

Algorithm	$c = 0.10, d = x$			$c = 0.10, d = y$		
	$p = 5$	$p = 10$	$p = 25$	$p = 5$	$p = 10$	$p = 25$
asm2000	0.0010	0.0010	0.0010	0.0010	0.0010	0.0010
bacon	0.0010	0.0010	0.0010	0.0010	0.0010	0.0010
bch2006	5.7756	22.0429	26.7652	0.0036	0.0039	0.0045
ccf	17.5024	21.6381	24.2338	0.0804	0.7822	3.4871
cm97	17.5442	21.6073	24.3706	0.0066	0.0043	0.0029
hs93	1.7994	4.0458	0.9715	0.0010	0.0010	0.0010
imon2005	0.0010	0.0010	0.0010	0.8566	0.8655	1.0107
ks89	17.3462	21.6626	24.6064	0.0088	0.0117	0.0185
lad	17.5348	21.6808	24.5792	0.0060	0.0040	0.0028
lms	0.0230	0.0423	3.5922	0.0224	0.0347	0.3284
lta	0.0214	0.0385	0.0674	0.0220	0.0377	0.0664
lts	0.0096	0.0091	0.0334	0.0094	0.0093	0.0078
ols	17.8770	21.6403	24.2406	2.7914	3.1037	3.6352
py95	16.1162	21.4521	24.6933	0.0190	0.0161	0.0191
quantilereg	17.5272	21.6874	24.5982	0.0060	0.0041	0.0028
satman2013	0.0086	1.2815	25.9011	0.0088	0.0083	0.0064
satman2015	16.4922	22.8783	26.8538	0.0090	0.0077	0.0062
smr98	0.0024	2.7959	23.6374	0.0014	0.0010	0.0010

Algorithm	$c = 0.20, d = x$			$c = 0.20, d = y$		
	$p = 5$	$p = 10$	$p = 25$	$p = 5$	$p = 10$	$p = 25$
asm2000	0.0010	0.0010	12.8159	0.0322	0.0010	0.0295
bacon	0.0010	0.0010	0.0010	0.0010	0.0010	0.0010
bch2006	19.1146	23.0278	27.3760	0.0030	0.0035	0.0040
ccf	18.9962	22.2326	24.5423	2.5844	11.0908	13.1101
cm97	18.9782	22.2848	24.7040	0.0294	0.0156	0.0081
hs93	11.3026	2.2483	0.0011	0.0010	0.0010	0.0013
imon2005	0.0688	0.0441	16.8873	7.0404	8.0550	9.6137
ks89	19.0040	22.3559	24.9464	0.3832	0.4586	0.6018
lad	19.0112	22.3573	24.9449	0.0234	0.0132	0.0073
lms	0.4672	6.1905	24.1882	0.1580	0.9109	7.0917
lta	0.0236	0.0499	0.1384	0.0252	0.0499	0.1381
lts	0.0078	0.0074	15.2878	0.0080	0.0078	0.0070
ols	19.0194	22.2258	24.5409	10.9012	11.8859	12.8084
py95	19.0350	22.5864	25.0380	1.2078	0.6776	0.5984
quantilereg	19.0048	22.3663	24.9542	0.0232	0.0133	0.0073
satman2013	17.3134	23.0274	26.8841	0.0074	0.0063	0.0054
satman2015	19.7148	23.5378	27.3970	0.0072	0.0070	0.0057
smr98	2.5292	19.7529	24.8598	0.0018	0.0019	0.0018

Algorithm	$c = 0.30, d = x$			$c = 0.30, d = y$		
	$p = 5$	$p = 10$	$p = 25$	$p = 5$	$p = 10$	$p = 25$
asm2000	0.001	0.498	24.525	0.094	0.159	0.048
bacon	10.133	7.521	1.123	2.403	0.949	0.345
bch2006	19.733	23.402	27.918	0.003	0.003	0.003
ccf	19.436	22.464	24.730	17.917	26.366	28.234
cm97	19.425	22.520	24.894	0.116	0.062	0.030
hs93	11.553	1.125	0.001	0.045	0.002	0.002
imon2005	1.780	3.854	24.839	20.245	22.903	26.694
ks89	19.472	22.623	25.143	3.101	3.522	4.466
lad	19.480	22.639	25.228	0.070	0.038	0.019
lms	5.995	18.192	25.641	1.536	6.981	17.702
lta	0.028	0.073	7.585	0.028	0.072	6.289
lts	0.006	0.751	30.136	0.006	0.006	0.006
ols	19.425	22.463	24.716	24.620	26.749	28.140
py95	19.484	22.770	25.270	11.084	7.890	5.468
quantilereg	19.447	22.621	25.217	0.069	0.038	0.019
satman2013	19.719	23.656	27.324	0.006	0.005	0.005
satman2015	20.138	23.930	28.055	183.544	153.479	61.311
smr98	10.018	22.276	25.141	0.002	0.002	0.195

Table 4: Average MSE for $n = 1000$

TOPSIS		VIKOR		ARAS		WASPAS		COPRAS	
Algorithm	Score	Algorithm	Score	Algorithm	Score	Algorithm	Score	Algorithm	Score
lta	0.922	lta	0.998	asm2000	0.528	asm2000	0.344	hs93	1.000
hs93	0.922	asm2000	0.974	bacon	0.454	bacon	0.305	lta	0.835
asm2000	0.885	lts	0.960	hs93	0.451	hs93	0.297	asm2000	0.694
lts	0.855	bacon	0.953	lts	0.183	lts	0.134	lts	0.529
lms	0.782	hs93	0.884	smr98	0.159	smr98	0.123	bacon	0.234
bacon	0.756	lms	0.867	lta	0.122	bch2006	0.078	lms	0.169
smr98	0.745	smr98	0.770	bch2006	0.112	lta	0.075	smr98	0.136
satman2013	0.720	imon2005	0.763	imon2005	0.081	satman2013	0.058	satman2013	0.126
cm97	0.682	satman2013	0.741	satman2013	0.081	cm97	0.051	cm97	0.097
lad	0.680	cm97	0.544	quantilereg	0.063	quantilereg	0.051	bch2006	0.097
quantilereg	0.680	bch2006	0.542	lad	0.063	lad	0.051	lad	0.096
bch2006	0.679	lad	0.540	cm97	0.062	imon2005	0.049	quantilereg	0.096
ks89	0.669	quantilereg	0.540	lms	0.050	lms	0.033	ks89	0.084
py95	0.666	py95	0.484	satman2015	0.030	satman2015	0.025	py95	0.079
imon2005	0.584	ks89	0.477	ks89	0.010	ks89	0.010	imon2005	0.047
satman2015	0.505	satman2015	0.415	py95	0.009	py95	0.009	satman2015	0.039
ccf	0.486	ccf	0.074	ccf	0.002	ccf	0.002	ccf	0.036
ols	0.337	ols	0.010	ols	0.001	ols	0.000	ols	0.024

Table 5: Ranking and scores

Algorithm	Absolute time	Relative time
ols	0.00013	1.000
ccf	0.00155	12.116
cm97	0.00774	60.596
satman2013	0.01599	125.211
quantilereg	0.06122	479.287
lad	0.06172	483.183
satman2015	0.06690	523.747
smr98	0.15469	1211.039
ks89	0.24363	1907.391
bacon	0.25505	1996.776
lms	0.42403	3319.777
lts	0.48258	3778.131
imon2005	0.53484	4187.311
asm2000	0.54336	4254.016
py95	0.68249	5343.253
bch2006	1.55553	12178.298
lta	4.65866	36472.915
hs93	6.57872	51505.193

Table 6: Absolute and relative elapsed times by algorithms. Relative average times are calculated due to the ols by setting its time to 1x.

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.

ORCID ID of the author / Yazarın ORCID ID'si

Mehmet Hakan Satman 0000-0002-9402-1982

REFERENCES

- R. Adnan, H. Setan, and M. N. Mohamad. Identifying multiple outliers in linear regression: Robust fit and clustering approach. In *The 10th FIG International Symposium on Deformation Measurements, SESSION X : THEORY OF DEFORMATION ANALYSIS II*, pages 380–389, Orange, California, USA, 2000.
- S. Barratt, G. Angeris, and S. Boyd. Minimizing a sum of clipped convex functions. *Optimization Letters*, 14:2443–2459, 2020.
- D. A. Belsley, E. Kuh, and R. E. Welsch. *Regression diagnostics: Identifying influential data and sources of collinearity*. 1980. ISBN 0-471-05856-4.
- J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah. Julia: A fresh approach to numerical computing. *SIAM review*, 59(1):65–98, 2017. doi:10.1137/141000671.
- N. Billor and G. Kiral. A comparison of multiple outlier detection methods for regression data. *Communications in Statistics—Simulation and Computation*, 37(3):521–545, 2008.
- N. Billor, A. S. Hadi, and P. F. Velleman. Bacon: blocked adaptive computationally efficient outlier nominators. *Computational statistics & data analysis*, 34(3):279–298, 2000.
- N. Billor, S. Chatterjee, and A. S. Hadi. A re-weighted least squares method for robust regression estimation. *American journal of mathematical and management sciences*, 26(3-4):229–252, 2006.
- S. Chatterjee and M. Mächler. Robust regression: A weighted least squares approach. *Communications in Statistics-Theory and Methods*, 26(6):1381–1394, 1997.
- K. Deb. Multi-objective evolutionary algorithms. *Springer handbook of computational intelligence*, pages 995–1015, 2015.
- D. Diakoulaki, G. Mavrotas, and L. Papayannakis. Determining objective weights in multiple criteria problems: The critic method. *Computers & Operations Research*, 22(7):763–770, 1995. doi:10.1016/0305-0548(94)00059-h.
- A. S. Hadi and S. Chatterjee. *Regression analysis by example*. John Wiley & Sons, 2015.
- A. S. Hadi and J. S. Simonoff. Procedures for the identification of multiple outliers in linear models. *Journal of the American statistical association*, 88(424):1264–1272, 1993.
- D. M. Hawkins and D. Olive. Applications and algorithms for least trimmed sum of absolute deviations regression. *Computational Statistics & Data Analysis*, 32(2):119–134, 1999.
- L. Huo, T.-H. Kim, and Y. Kim. Robust estimation of covariance and its application to portfolio optimization. *Finance Research Letters*, 9(3):121–134, 2012.
- C.-L. Hwang and K. Yoon. *Methods for Multiple Attribute Decision Making*. Springer Berlin Heidelberg, 1981.
- F. Kianifard and W. H. Swallow. Using recursive residuals, calculated on adaptively-ordered observations, to identify outliers in linear regression. *Biometrics*, pages 571–585, 1989.
- S. C. Narula, P. H. Saldiva, C. D. Andre, S. N. Elian, A. F. Ferreira, and V. Capelozzi. The minimum sum of absolute errors regression: a robust alternative to the least squares regression. *Statistics in medicine*, 18(11):1401–1417, 1999.
- S. Opricovic. Multicriteria optimization of civil engineering systems, 1998.
- S. Opricovic and G.-H. Tzeng. Multicriteria planning of post-earthquake sustainable reconstruction. *Computer-Aided Civil and Infrastructure Engineering*, 17(3):211–220, may 2002. doi:10.1111/1467-8667.00269.
- D. Peña and V. J. Yohai. The detection of influential subsets in linear regression by using an influence matrix. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):145–156, 1995.
- A. Rahmatullah Imon. Identifying multiple influential observations in linear regression. *Journal of Applied statistics*, 32(9):929–946, 2005.
- P. J. Rousseeuw. Least median of squares regression. *Journal of the American statistical association*, 79(388):871–880, 1984.
- P. J. Rousseeuw and K. V. Driessen. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41(3):212–223, 1999.
- P. J. Rousseeuw and K. Van Driessen. Computing lts regression for large data sets. *Data mining and knowledge discovery*, 12:29–45, 2006.
- M. H. Satman. A new algorithm for detecting outliers in linear regression. *International Journal of statistics and Probability*, 2(3):101, 2013.
- M. H. Satman. Fast online detection of outliers using least-trimmed squares regression with non-dominated sorting based initial subsets. *International Journal of Advanced Statistics and Probability*, 3(1):53, 2015.
- M. H. Satman, S. Adiga, G. Angeris, and E. Akadal. Linregoutliers: A julia package for detecting outliers in linear regression. *Journal of Open Source Software*, 6(57):2892, 2021a. doi:10.21105/joss.02892.
- M. H. Satman, B. F. Yıldırım, and E. Kuruca. Jmcdm: A julia package for multiple-criteria decision-making tools. *Journal of Open Source Software*, 6(65):3430, 2021b. doi:10.21105/joss.03430.

- D. M. Sebert, D. C. Montgomery, and D. A. Rollier. A clustering algorithm for identifying multiple outliers in linear regression. *Computational statistics & data analysis*, 27(4):461–484, 1998.
- S. Van Aelst and P. Rousseeuw. Minimum volume ellipsoid. *Wiley Interdisciplinary Reviews: Computational Statistics*, 1(1):71–82, 2009.
- J. W. Wisnowski, D. C. Montgomery, and J. R. Simpson. A comparative analysis of multiple outlier detection procedures in the linear regression model. *Computational statistics & data analysis*, 36(3):351–382, 2001.
- K. Yu, Z. Lu, and J. Stander. Quantile regression: applications and current research areas. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 52(3):331–350, 2003.
- E. K. Zavadskas and Z. Turskis. A new additive ratio assessment (aras) method in multicriteria decision-making, 2010.
- E. K. Zavadskas, A. Kaklauskas, and V. Sarka. The new method of multicriteria complex proportional assessment of projects, 1994.
- E. K. Zavadskas, Z. Turskis, and J. Antucheviciene. Optimization of weighted aggregated sum product assessment. *Electronics and Electrical Engineering*, 122(6), jun 2012. doi:10.5755/j01.eee.122.6.1810.

How cite this article

Satman, M.H. (2023). Comparison of outlier detection methods in linear regression: A multiple-criteria decision-making approach. *Acta Infologica*, 7(2), 333-347. <https://doi.org/10.26650/acin.1327370>

Query by Image Examination: Classification of Digital Image-Based Forensics Using Deep Learning Methods

Görüntü İncelemesine Göre Sorgulama: Dijital Görüntü Tabanlı Adli Görüntülerin Derin Öğrenme Yöntemleri Kullanılarak Sınıflandırılması

İlker Kara¹ 

¹(Assist. Prof. Dr.), Cankiri Karatekin University, Eldivan Vocational School of Health Services, Cankiri, Türkiye

Corresponding author : İlker KARA

E-mail : karaiKab@gmail.com

ABSTRACT

The continuous increase in the use of information systems and online services has also spurred the forensic examination of digital and image data, which serves as the primary platform for information transfer. In particular, according to the latest reports, the examination of the images obtained from all kinds of recording devices that have the quality of evidence as a result of the forensic case and that can provide the clarification of the incident and the detection of the criminal elements are becoming a critical problem due to the huge amount of data. Our contribution in this study is two-folded. First, we present a new approach that classifies digital images into eight different crime categories using six different models. Second, we have created a new dataset for the classification of crimes and opened it to the public. Throughout the study, we have used our new dataset which has a total of 15,065 image samples from 8 different crime categories including Bet, ChildAbuse, Credit Card and Banking, Drugs, Frightening, Knives, Pornographic and Weapons. In this study, six different models were used to classify crime images. The CNN model was developed by us and five other models used for transfer learning. Pre-trained network model parameters VGG16, VGG19, Xception Model, InceptionResNetV2 and NASNetLarge were used for crime image classification tasks. In addition, the performance of these models is compared using test accuracy and time metrics. Resultly, we achieved prediction accuracy of up to 89.74% using the NASNetLarge model.

Keywords: image processing, deep learning method, image classification, data mining, forensic investigation

ÖZ

Bilgi sistemlerinin ve çevrimiçi hizmetlerin kullanımındaki sonsuz artış, bilgi aktarımı için temel platformlardan biri olan dijital ve görüntü içeren verilerin adli incelemelerini de tetiklemiştir. Adli görüntü inceleme temel olarak bilimsel yöntemlerin ve adli inceleme yazılımlar kullanılarak ilgili görüntüler hakkında delil oluşturulmasını sağlayan bilimsel bir disiplindir. Özellikle, son raporlara göre, adli vaka sonucunda delil niteliği taşıyan ve olayın aydınlanmasını sağlayabilecek her türlü kayıt cihazından elde edilmiş görüntülerin incelenmesi ve suç unsuru olanlarının tespiti artan veri miktarı nedeniyle giderek büyük bir problem haline gelmektedir. Bu çalışmada katkımız iki katkı sunmaktadır. İlk olarak dijital görüntülerin altı farklı model kullanarak sekiz farklı suç kategorisi olarak sınıflandıran yeni bir yaklaşım sunuyor. İkincisi, suçların sınıflandırılması için yeni bir veri kümesinin oluşturularak paylaşımına sunuyor. Çalışma boyunca, Bet, ChildAbuse, kredi kartı ve bankacılık, uyuşturucu, korkutucu, bıçak, pornografik ve silah dâhil olmak üzere 8 farklı suç Kategorisine ait toplam 15.065 görüntü örneğini kapsayan yeni veri setimizi kullanıldı. Suç görüntülerini sınıflandırmak için bu çalışmada 6 farklı model kullanılmıştır. CNN modeli kendimiz ve öğrenmeyi ince ayarlara aktarmak için kullanılan diğer beş model tarafından yaratılmıştır. Görüntü sınıflandırma görevleri için VGG16, VGG19, Xception modeli, InceptionResNetV2 ve NASNetLarge önceden eğitilmiş ağ modeli parametreleri kullanıldı. Ayrıca, bu modellerin performansı test doğruluğu ve zaman ölçümleri kullanılarak karşılaştırılır. Sonuçlar, NASNetLarge modeli kullanarak %89.74'e kadar tahmin doğruluğu elde edilmiştir.

Anahtar Kelimeler: Görüntü işleme, derin öğrenme yöntemi, görüntü sınıflandırması, veri madenciliği, adli inceleme

Submitted : 13.04.2023
Revision Requested : 20.10.2023
Last Revision Received : 14.11.2023
Accepted : 30.11.2023
Published Online : 11.12.2023



This article is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0)

1. INTRODUCTION

Data obtained from images represent an effective and natural communication medium in the current age of technology. Unlike the text data, information obtained from the images are helpful in illuminating the forensic cases as the image itself contains evidence for the criminal elements. As such, forensic images are defined as images obtained from any recording device which can provide evidence and therefore enlighten the criminal event. The source of the forensic image, that can be accepted as evidence by the court, can be any kind of recording devices including professional or compact cameras, mobile phones and security cameras (Choodum et al., 2015). The increase in computers and other related computing systems (e.g. mobile devices, IoTs) that we use constantly in our daily lives on the other hand has led to the formation of a large amounts of images (Hafiz et al., 2020). Although the accumulation of image data is helpful in forensic studies, manual analysis of large volumes of digital images is a significantly tedious task. A forensic investigation can take an average of six months with the analysis of more than 300,000 digital images, among which only an average of 100 images are reported to be related to the crime in question (Ferreira et al., 2020). Consequently, using digital imagery as an aid for decision making and as a support for scientific arguments in the forensic investigations is hindered to a great extent. One way to address this issue is to distinguish and classify objects that may be important in the image data, instead of taking into consideration of all the objects available in a given image. The several approaches proposed to determine the authenticity and classification of images and their origins can be classified into two branches, active image forensics and passive image-based forensics. (Birajdar et al., 2013) (See Figure 1).

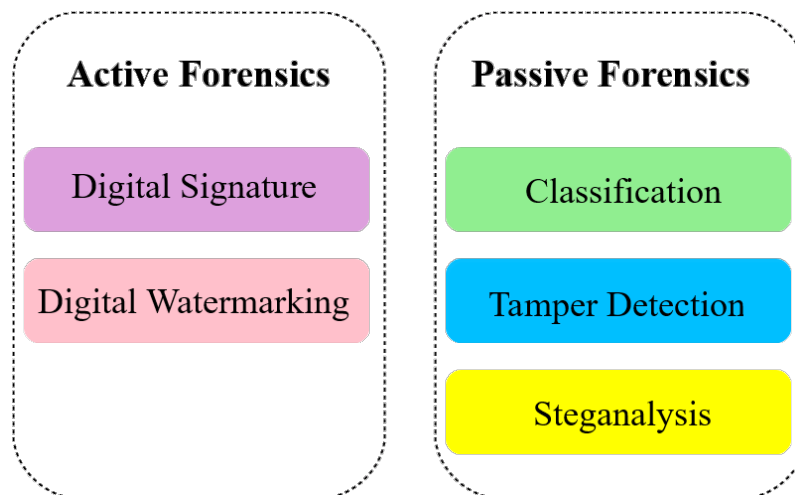


Figure 1. Commonly employed digital image forensics methods in the literature.

Active image forensics require additional priory knowledge of the source of the image. This information requires that the device producing the image contain a digital signature (Birajdar et al., 2013) or a digital watermark (Chandra et al., 2010). Passive image forensics on the other hand is more practical in terms of technology and attempts to determine whether an image is authentic or not based solely on the characteristics of the image without any additional embedded information (Wang et al., 2009). Passive image forensics comes into play once an image has been created and stored in criminal information systems. Depending on the nature of the crime under investigation in digital image forensics, images can be categorized using the passive image forensics approach (Mahalakshmi et al., 2012), image manipulation detection (Thakur et al., 2020), and image source detection (Peng et al., 2013).

Piva, proposed a model for distinguishing and classifying forensic image data and detecting whether the data is manipulated to deceive forensic analysis methods (Piva, 2013). Forensic softwares are needed for the detection and classification of forensic images (Pearson, 2006). Although there are a number of tools designed to classify forensic images, such as Belkasoft (Belkasoft, 2021), X-Ways Forensics (X-Ways Forensics, 2021), most of the time, experts in forensic image analysis experience great difficulties in the examinations made with such forensic software tools due to the irregular and inconsistent data (Cao et al., 2009).

From this perspective, there are several important advantages that serve as a motivating factor for this study in the detection and classification of forensic images. As stated, the differentiation and categorization of forensic image data will offer significant convenience in forensic investigations (Pearson, 2006). Classifying examined digital images based on crime types will also facilitate the detection and identification of evidence. With all this in consideration, we propose

a machine learning-based approach to classify digital images. We employ six different machine learning methods. We propose a robust scheme for classifying eight different crime categories for the samples examined using this approach. To this end, this study mainly presents the three contributions listed below without any commercial concern:

- The method presented in this study concentrates on the categorization of criminal elements in digital images, with a focus on the capture phase.
- We proposed a new dataset of 15,065 items belonging to eight different categories. We also make the proposed dataset publicly accessible for non-commercial purposes (Accessed via link, <http://ilkerkara.karatekin.edu.tr/RequestDataset.html>).
- We analyzed a new dataset consisting of digital images of different subjects. To be more specific, we have applied six different machine learning methods to classify 15,065 unique 224x224 pixel target images belonging to eight different crime categories, including Illegal Betting, Child Abuse, Credit Card and Banking, Drugs, Violent, Knives and firearms, Illegal Pornography.

This article is organized as follows: In chapter two, we reviewed several related studies. Chapter three presents the details of the proposed dataset and chapter four presents the applied methods and the experimental results comparing the classification of six unique machine learning algorithms. Chapter five is discussion and chapter six concludes the study and explains possible future directions.

2. RELATED WORK

Due to the increasing amount of forensic evidence in forensic image examinations, the effectiveness of traditional methods is hindered on a high scale. Automated approaches based on machine learning and deep learning models and automatic classification of forensic images are designed to address these problems.

The widely used traditional machine learning models in the classification of forensic evidence include Bayesian algorithms - BayesNet (Grillo et al., 2009; Marturana et al., 2011; Marturana and Tacconi, 2013) and Naïve Bayes (McClelland and Marturana, 2014), Decision Trees (Marturana et al. 2011; Marturana and Tacconi, 2013; Garfinkel et al. 2010), and K-Nearest Neighbor (Gomez, 2012). One of the deep learning models, Convolutional Neural Networks (CNNs) approach is used in forensic image classification. For this purpose, CNN models were created for weapon classification consisting of forensic images (Olmos et al. 2018; Verma and Dhillon, 2017). In the proposed study by Dey et al. a topological signature-based learning scheme was used for the classification of images with an accuracy of 83.2% (Dey et al., 2017). Lin et al used a CNN-based learning method to determine the authenticity of digital images within the scope of forensic analysis (Lin et al., 2018). Similarly, forensic evidence images used the pre-trained Faster R-CNN model approach (Ren et al., 2015).

The study by Karakuş (2018) proposed a model that provides fast and accurate analysis of image data. The proposed model consists of VGG16 network structure and network layers designed for image classification. In the study, a dataset comprising images with a resolution of 300x300 pixels was utilized. Of these images, 2085 were generated using the Kaggle platform, while 915 were obtained from various other sources. The dataset consisted of a total of 3000 image data, with 1500 images depicting guns and 1500 images depicting knives. While 2000 of the images obtained were used for training purposes, 1000 of them were used for verification purposes. An accuracy rate of 97.8% was obtained in the model. (Kara et al., 2018).

Saber et al. (2020), conducted a study on digital image forgery detection and forensic informatics. This study tried to resolve the question of how to ensure the accuracy of images that can serve as evidence in an investigation process. In this study, the advantages and usage areas of existing forensic image technology, comparative studies, the benefits and harms of forgery detection systems including deep learning and convolutional neural networks were examined. These investigations were elucidated within the sections titled "Digital Image Forgery Detection Methods," "Forensic Approaches," and "Comparative Study," bolstered by prior research. It was emphasized that the process was laborious due to the manipulations performed on the image. As a result of the research, it was found that different image processing techniques such as preprocessing, feature extraction, feature selection and classification are also very useful for the precise detection of forgery. Passive methods prove to be highly effective for forgery detection when compared to active approaches. Among these passive methods, copy-move and image fusion are extensively employed by numerous researchers, owing to their benefits of reduced complexity and enhanced accuracy (Saber et al., 2020).

Ferreira et al. (2020), reported that only 148 images containing illegal content (sexual abuse) were found in a database containing more than 300,000 images and 1100 videos. The study focuses on the use of deep learning techniques to identify image manipulation. As a result of the study, it gave satisfactory results in terms of the increase in the time

spent per image and the increase in the margin of error of the analysis due to the manual examination of the images by forensic experts (Ferreira et al., 2020).

Forensics experts develop analysis tools to help them quickly recognize and classify digital images to focus on possible criminal elements in the evidence examined in investigations. In 2012, a commercial analysis tool called ADF Digital Evidence Investigator trained on tensorflow, an artificial intelligence library, to classify digital images especially for crimes of Child Sexual Abuse Material (CSAM) (Adfsolutio, 2021). Since digital image classification is time consuming and ADF tools are often used to quickly qualify exhibits at the crime scene or in the lab, they used a filter (ignoring icons, thumbnails, and other pixel art) and focused only on CSAM crimes classification. Recent developments of new techniques for classification have shown very promising results even in large datasets such as ImageNet (DDS09) (Adfsolutio, 2021).

Alharbi, aims to increase the classification performance of small-sized forensic images in his study (Sharma et al., 2021). For this purpose, the CIFAR-10 dataset containing 60,000, 32x32 color images was used. Principal component analysis (PCA), KNN (K-Nearest Neighbor), and CNN models were used to classify forensic image contents in the study. As a result of the study, the best result was obtained with CNN with a success rate of 74.10%. Although this rate is promising, the margin of error is still too high.

Del Mar-Raave developed a machine learning prototype capable of recognizing weapons in forensic image content (Del Mar-Raave et al., 2021). Given the multitude of weapon types, the dataset used in the study, comprising 608 forensic images, was refined by exclusively selecting realistic photographs of pistols or firearms. In our study, a similar approach was employed by reducing the dataset through the classification of the type of crime committed. Four ImageNet-trained models (InceptionV3, Xception Model, ResNet, and VGG16) were utilized to assess forensic images for weapon identification. Del Mar-Raave et al. achieved the most successful result with the Xception model, attaining an accuracy of 90% in their tests. Despite the promising results, the study's limitation lies in the relatively small dataset used.

It can be concluded that studies on the classification of forensic evidence focus on the creation and optimization of machine learning and deep learning models. Forensic tool development studies for crime categories are rarely used in the classification of forensic evidence. In this study, we have used a similar approach by Del mar, utilizing six distinct machine learning methods to classify 15,065 unique 224x224 pixel target images associated with eight different crime categories.

3. MATERIAL AND METHODS

In this section, we have introduced our approach crime categories, dataset and classification models.

3.1. Dataset

The concept of crime defines the behaviors and actions that are prohibited by the law, defined as crimes by law, and are punished if committed. Combatting crime and delinquency can be assessed in two main facets: the prevention of crime before it occurs and research, detection, and analysis after a crime has taken place. Forensic experts play a crucial role in the latter, elucidating crimes by scrutinizing suspected individuals and providing insights for criminal court decisions. This process depends on the type of the crime that has been committed. However, factors such as the number and the size of the materials examined or the number of qualified specialist personnel are also important. The increasing use of visuals in many applications due to the advances in the technology manifests itself in the human factor in forensic image examinations, revealing the need for expert personnel. To alleviate this problem, tools such as AccessData FTK, EnCase, Belkasoft (McDown et al., 2016) are mainly employed in forensic analysis. Forensic image reviews, inclusive of an analysis of the tools' advantages and disadvantages, are conducted by expert professionals. In principle, although a classification can be made according to file extensions in the examined digital material, they lack the capacity to make decisions regarding content. From this perspective, forensic image analysis offers several important advantages that motivate this study. Automatic classification can be achieved by employing deep learning models that leverage determinative features selected based on crime types discernible in forensic image content. Within this framework, the concept of classifying forensic image content according to specific characteristics during the application phase proves beneficial in terms of alleviating the workload of experts.

The current study focusses on "Illegal Betting, ChildAbuse, Credit Card and Banking, Drugs, Violent, Knives, Firearms, Illegal Pornographic" categories. Illegal Betting, in other words, "Illegal gambling", is any kind of betting action taking place using technology in sports competitions, and is considered to be illegal if a license/permission is not given by the authorities. Criminals create trap virtual environments that harm the economy and affect individuals

socio-psychologically with illegal betting websites that are not subject to taxation. In the European Police Organization (EUROPOL) 2020 Report, it is estimated that the annual cost of illegal sports betting to the world is approximately 1.69 trillion euros (Europol, 2020). The economic loss is seen as a global problem as the illegal betting industry threatens the economy of all countries. Forensic Informatics Specialists usually carry content analysis by focusing on visuals such as coupons, betting odds, promotional advertisements containing sports activities in visual examinations related to this crime.

ChildAbuse, which is an important legal, medical and social problem, is considered as a serious crime in terms of psycho-social and legal aspects due its short and long-term consequences (Kara, 2017). ChildAbuse crime has many dimensions such as physical, emotional, economic or sexual abuse. In the forensic image analysis of child sexual abuse crime, in addition to forensic informatics experts, it is necessary to decide on factors such as the child's age and biological development in the image (Sanap et al., 2015). In addition, accessing content containing child sexual abuse crime has difficulties. Considering this aspect, the study focused on the physical dimension of ChildAbuse crime.

Credit Card and Banking frauds are defined as the crimes of using someone else's bank or credit cards or producing, selling, transferring cards on behalf of someone else illegally. In addition to the availability of developing virtual cards used in online shopping platforms, the number of people who want to take advantage of these cards is substantial on a global scale. In the forensic image analysis carried out on Credit Card and Banking crime, forensic experts generally focus on transactions patterns, transaction documents and banking or credit card images (Wu et al., 2009).

Drugs are habitual or addictive due to their chemical structure; Drugs and stimulants that cause physical, mental, social and judicial problems are seen as a social problem. Criminals are able to supply drugs of many types to almost all social layers of the society by actively using the newly available technological platforms. Forensic experts in forensic analysis of drugs, generally focus on the drug types (cannabis, amphetamines, Ecstasy, heroin, cocaine/crack, stimulants, Ecstasy, sedatives, hallucinogens, opioids, inhalants, and other substances (Isnard et al., 2001).

Violence includes physical or mental suffering, inhumane acts incompatible with human dignity. Intimidation, insults, threats or sexual harassments can be classified as violence. Forensic informatics experts generally focus on the content of images that have been subjected to violence, physically damaged, and whose bodily integrity has been disrupted in forensic image analysis related to violent crime (Del Mar-Raave et al., 2021).

Injury or death as a result of violent attacks are mostly committed with the use of weapons. The concept of a weapon can be defined very broadly to include knives, swords, handguns, rifles, shotguns, machine guns, anti-aircraft missiles, anti-tank missile/rocket launcher, or chemical weapon. The most frequently used weapons in crimes are firearms and knives. Firearms and knives damage the integrity of the body which can result in injury or even death intentionally or unintentionally. For this reason, the study focused on image analysis containing knives and firearms (de Castro et al., 2010).

Illegal pornography, or "Obscenity," refers to publications, images, or other forms of media that serve the purpose of arousing sexual impulses and are contrary to moral values. The characterization of a product as illegal pornographic in investigations is based on the following criteria: i) it dehumanizes sexuality and renders it brutal, and ii) it diminishes the human being to a psychological-impulse-reaction formation, transforming individuals into explicit objects of sexual lust (Seigfried et al., 2008). However, child pornography is evaluated by opening a different heading in terms of psycho-social and legal aspects (Del Mar-Raave et al., 2021). Forensic Informatics Experts focus on the existence of the above-mentioned qualities in the examinations made on pornographic materials and decide whether the content should be regarded as illegal.

3.2. Gathering Digital Data

The importance of a well curated and correct dataset is vital for data-driven studies. As an attempt to build such a suitable dataset, we cooperated with an information security company located in Turkey. The mentioned company possesses a dedicated team and system for gathering a substantial number of forensic digital image samples. They have generously shared authentic forensic digital images with us. There are 15.065 images in the dataset which belong to 8 different crime categories including Illegal betting, Child Abuse, Credit Card and Banking, Drugs, Violent, Knives and firearms, Illegal Pornographic. Images were collected from real forensic cases classified according to the category of the crime. In the context of related crimes, potentially criminal images were identified on a suspicious computer subjected to forensic case analysis. The suspects in these images were made anonymized in the shared dataset, and the images were presented without violating copyright and privacy principles. The images in the dataset were labeled using CVAT (Computer Vision Annotation Tool) within the company and subsequently manually reviewed by the authors to identify and rectify any inconsistent or noisy data (Figure 2). This process is done to increase the success rate for deep learning models. Moreover, the collected data were analyzed and classified by the authors one by one. While labelling

the images, care was taken that the images were taken from real crime scenes. The dataset used in this study is available via the link given in the introduction section. Therefore, another contribution of this study is the newly created dataset for classification of crimes by using images.

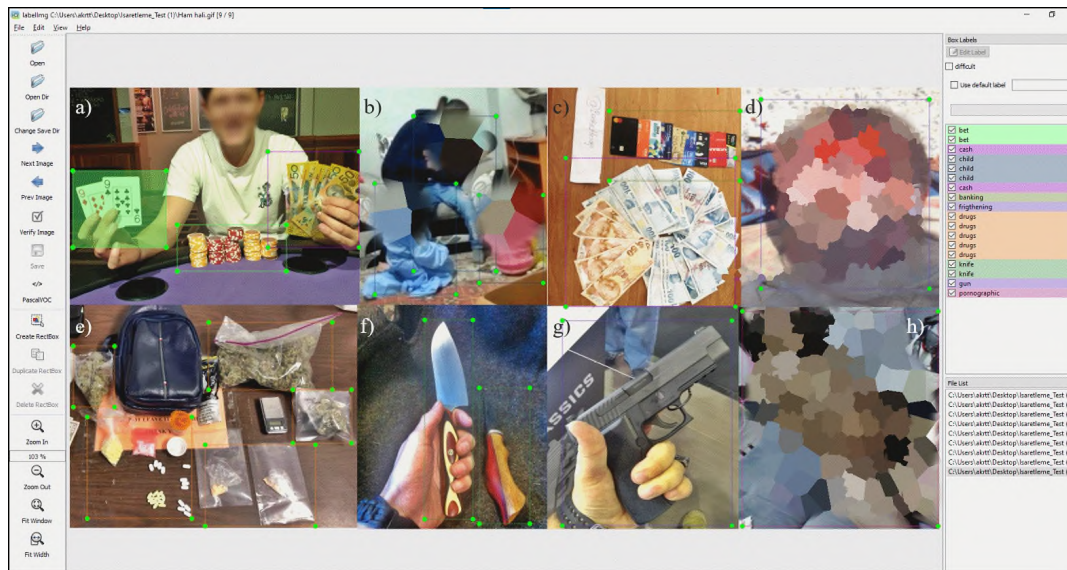


Figure 2. The images in the dataset are labeled with all categories using the CVAT tool. a) Illegal Betting, b) Child Abuse, c) Credit Card and Banking, d) Violent, e) Drugs, f) Knives, g) Firearms, h) Illegal Pornographic.

3.3. Classification Models

In order to classify crime images 6 different models used in this study. The CNN model is created by ourselves and the five other models used for transfer learning to fine-tuning. VGG16, VGG19, Xception Model, InceptionResNetV2 and NASNetLarge pre-trained network model parameters are used for image classification tasks. Further, the performance of these models are compared using test accuracy and time metrics.

VGG16 is a convolutional neural network model proposed by K. Simonyan and A. Zisserman (Zhang et al., 2015). VGG16 achieved a 92.7% test accuracy rate in ImageNet. It is an improved version of AlexNet with changing the kernel-size. VGG19 (Simonyan et al., 2015) is also a network trained on ImageNet dataset but the difference between VGG19 and VGG16 is that the former has 19 deep layers whereas the latter employs 16 deep layers. The Xception Model was presented by Francois Chollet in 2017 (Chollet, 2017). Xception Model is an improved version of inception architecture by changing the standard Inception modules into depth wise Separable Convolutions. InceptionResNetV2 (Szegedy et al., 2017) is another convolutional neural network that was trained using the ImageNet dataset. The network contains 164 deep layers and has learned a rich features of a large dataset of images with the input image size of 299x299 for this model. NASNetLarge (Zoph et al., 2018) was also trained by using the ImageNet dataset with the input image size of 331x331.

The NASNetLarge model has the highest accuracy rate. When evaluated according to time and accuracy criteria, the model with the closest accuracy rate to this model is the Inception ResNetV2 model. While the accuracy is 1.4% lower for the Inception model, the NASNetLarge model is five times more expensive in terms of time. Therefore, for datasets with many images where time is more important, the InceptionResNetV2 model can be used instead of NASNetLarge.

In this study, we fine-tuned these five pre-trained models to train and test our crime image dataset, selecting them based on their highest accuracy rate in "Top-5 Accuracy" (Zoph et al., 2018). Moreover, we have analyzed the effect of deeper networks on image classification tasks.

3.3.1. CNN Algorithm

CNN (Convolutional Neural Network), one of the deep learning algorithms, is a type of artificial neural network. This algorithm, which gives successful results in many areas such as image processing, voice recognition, natural language processing, is especially effective in analyzing and processing visual data.

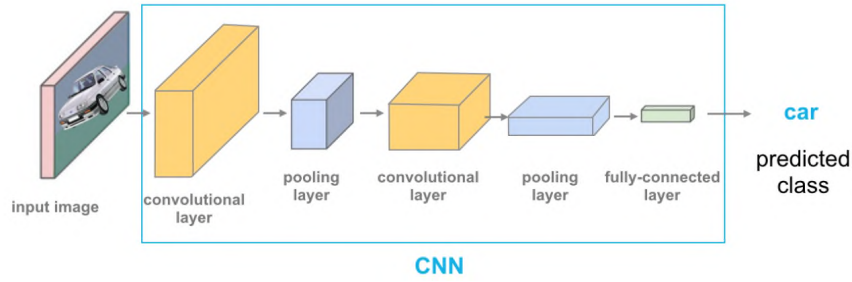
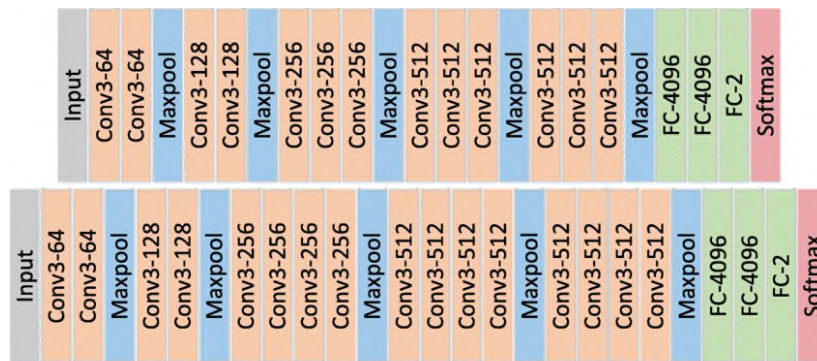


Figure 3. Structure of CNN Algorithm (Paluszek et al., 2020).

3.3.2. VGG16 and VGG19 Models

These models use a multilayer neural network architecture and convolutional neural networks (CNN) for feature extraction. The VGG16 model was developed by the Visual Geometry Group (VGG) at Oxford University. In VGG16, small filters (3x3) are used in the convolution layers. VGG16 consists of 13 convolution layers and 3 fully connected layers. There are 5 max pooling layers with 2x2 dimensions. The last layer is softmax. With the softmax layer, the incoming input data is classified. ReLu is used as the activation function. VGG19 consists of 16 convolution layers and 3 fully connected layers. VGG19, like VGG16, consists of 5 pooling layers and softmax as the last layer. While VGG16 contains 138 million parameters, VGG19 contains approximately 144 million parameters.



Network structures of VGG16 (top) and VGG19 (bottom)

Figure 4. Structure of VGG16 and VGG19 Network Models (Tammina, 2019).

3.3.3. InceptionResNetV2 Model

This model, developed by Google, has a combined architecture between Inception v4 and ResNet. Among the optimisations and innovations made in Deep Networks, the most different one is the ResNet structure where 'residual' connections are made. It has been observed that Inception v4 provides better accuracy but uses fewer parameters. The InceptionResNetV2 model was trained on the ImageNet dataset consisting of more than 5 million images.

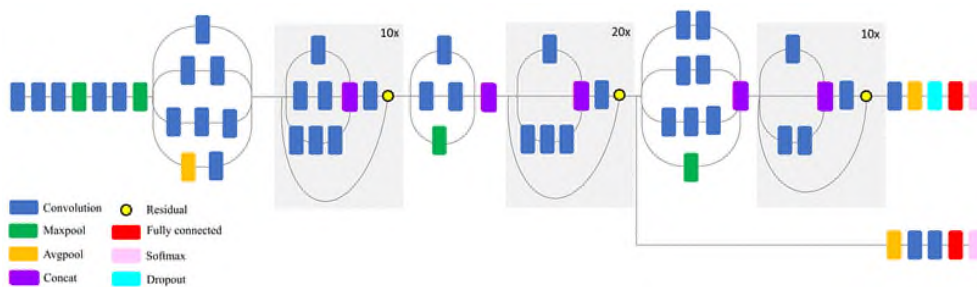


Figure 5. Structure of InceptionResNetV2 Model (Peng et al., 2022).

3.3.4. NASNetLarge Model

This model, developed by Google, is a scalable CNN architecture consisting of basic building blocks (cells) that are fine-tuned through reinforcement learning. It was built using automated machine learning (AutoML). The NASNetLarge model, trained on the ImageNet dataset, exhibits higher accuracy rates compared to other models.

3.3.5. Xception Model

This model, developed by Google, is a hypothesis based on the Inception module, which provides fully decomposable cross-channel and spatial correlations within CNN feature maps. This model is designed to solve the depth problem in convolutional neural network architecture. The Xception model, trained on the ImageNet dataset, demonstrates superior accuracy rates compared to other models.

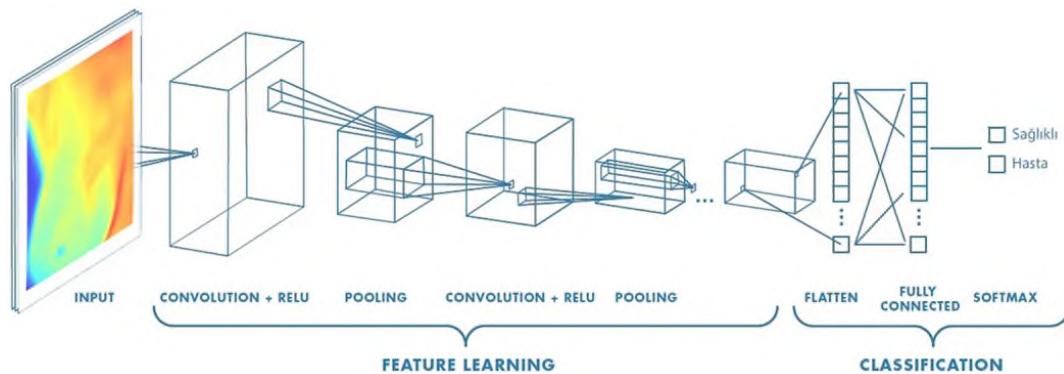


Figure 6. Structure of the Xception Model (Sharma et al., 2018)

4. APPROACH

In this section, the workflow of classification models, phases and conclusion of experiments are presented.

4.1. Workflow

A total of 6 different models, CNN, VGG16, VGG19, Xception Model, InceptionResNetV2 and NASNetLarge, were used on the dataset consisting of Illegal Betting, Child Abuse, Credit Card and Banking, Drugs, Violence, Knives and firearms, Illegal Pornographic images.

The dataset of 15,065 images collected from digital environments through applications such as AccessData FTK, EnCase, Belkasoft was resized to 224x224 resolution. After resizing the images, labelling was performed on the dataset in 8 different categories using the CVAT (Computer Vision Annotation Tool) program. The labelled data were divided into 80% training and 20% testing.

The preprocessed dataset was given as input to the network and 3 dense layers were added to increase the learning capacity of the model. The activation function "softmax", which is used in the last layer of multiclass classifications, and the loss function "categorical_crossentropy", which measures the difference between the actual class labels and the classes predicted by the model, were used as activation functions.

In this research, for all computational works "Python 3.8" is used as the programming language and "Keras library" with "TensorFlow" backend used for Deep Learning algorithms. Additionally, we used "Spyder 4.2.5" on the "Anaconda 2.0.4" platform, meanwhile we used "Lenovo ThinkPad P1 Gen3" with "Intel Core i7-10750H" CPU, "Nvidia Quadro T2000 Max-Q 4GB" and memory of "64 GB DDR4-3200" hardware components. Figure 3 shows the whole steps of the classification process.

4.2. Table of Various Settings

In this study, six different deep learning models are compared to obtain the highest accuracy rate in classifying crime images. CNN, VGG16, VGG19, Xception Model, InceptionResNetV2 and NASNetLarge deep learning models are well known and proven in the literature. Apart from these models, a different CNN model was also used in the project.

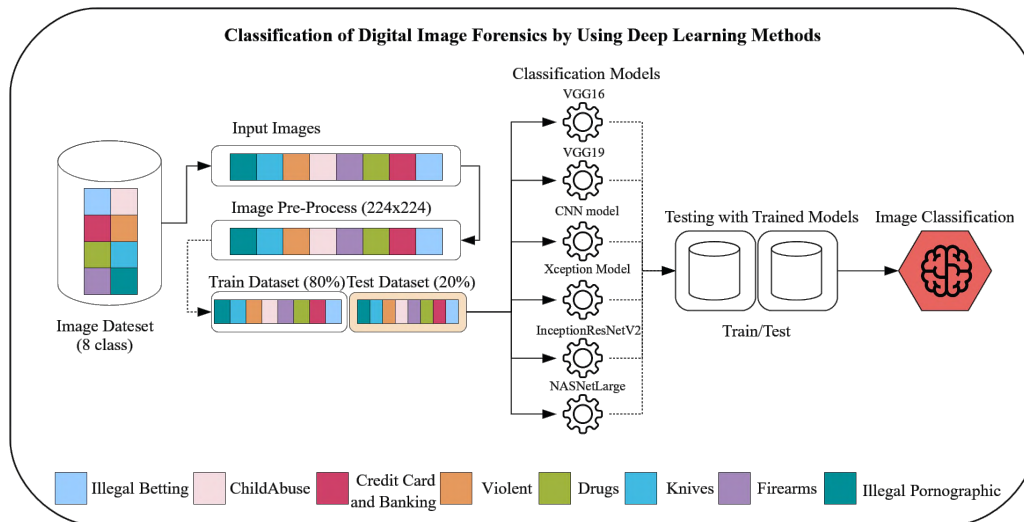


Figure 7. Classification of image dataset.

In the CNN model, it was desired to observe how it would work and result in this study. In this model, which is not expected to give a high accuracy rate, 3 Convolutional layers were applied to emphasize certain features on the pattern and 2 Intensive layers were applied to produce results using features from previous layers. Among the other models used, it gave the lowest result with a running time of 29 minutes and a test accuracy rate of 66.18%. NASNetLarge has the highest result with 91.74% test accuracy rate with pre-trained parameters and 45 minutes running time. When ranking the other models utilized in the project based on the test accuracy rate, VGG19 secured the first position with 81.07%, followed by VGG16 at 83.52%. The Xception Model claimed the third spot with an 88.78% test accuracy rate, while the InceptionResNetV2 model ranked fourth, achieving a test accuracy rate of 90.37%. The accuracy rates and run times of the models tested using 3,013 images and the accuracy rates of the models trained using 12,052 images from the dataset are presented in Table 1.

Table 1. Comparison of deep learning models.

Model	Training Accuracy	Test Accuracy	Time
VGG16	98.73%	83.52%	00:05:15
VGG19	98.12%	81.07%	00:05:16
CNN Model	98.07%	66.18%	00:29:07
Xception Model	97.90%	88.78%	00:22:38
InceptionResNetV2	97.24%	90.37%	00:08:35
NASNetLarge	98.75%	91.74%	00:44:57

5. DISCUSSION

As the number of evidences increase, studies that classify Forensic Images using deep learning methods will benefit experts in forensic investigations in resolving the crimes. The method proposed in this study not only provides an avenue for innovation but also possesses the flexibility to be updated. This adaptability stems from its capacity to extend its application to various crime types, allowing for the integration of different deep learning models tailored for classification.

In addition, through the method proposed, the need for the human factor in the classification of Forensic Images can be reduced, and thus may keep the human errors at a minimum level. It is aimed that the proposed method will help forensic experts by automatically classifying the forensic image analysis process that may contain criminal elements,

just like a smart assistant. In this way, it will contribute to shortening of the decision-making process of the forensic expert by quickly examining the evidence.

Classification of forensic images using the deep learning method also includes complications. Objects defined according to the applicable crime categories may not always contain an element of crime. For example, it would not be healthy to talk about a murder or injury crime in every knife image. Again, the decision on this issue should be up to an expert. The proposed method is not in the position of a decision maker, but in the task of an intelligent assistant that helps the decision maker and accelerates their work. Also, samples such as images obtained from low-resolution security cameras are a major disadvantage for object recognition algorithms. The models trained with low resolution images may not achieve the same rate of success compared to high resolution images. Moreover, no matter how high the resolution is, there is always a margin of error in the algorithms that classify Forensic Images using the deep learning method.

In this study, we use a new CNN learning-based strategy for classification of criminal elements in digital images within the scope of forensic investigations. Existing experiments demonstrate that digital images do not establish a solid foundation for classification. The rationale behind this argument stems from the capability of digital images in forensic investigations to pinpoint where to focus within images, automatically identifying the most distinctive regions. However, our dataset is limited to 8 crime type classes, and we believe our experimental setup require larger datasets to support this argument more strongly. Another noteworthy consideration is the potential of deep learning methods that can be harnessed for various applications. In conclusion, we believe that the use of modern CNN architecture methods can contribute to the classification of digital images based on feature extraction and capturing of the criminal elements.

6. CONCLUSION

In a forensic case, the abundance of data and documents to be extracted from digital materials naturally impacts the analysis process. Furthermore, the human factor in the analysis process is directly proportional to attention and knowledge, which in turn affects the quality of the analysis conducted. In this context, it can be observed that when data is not examined in depth and the volume of data is massive, these processes become practically impossible.

In this study, we conducted a study to classify and categorize digital images, which are sources of information in forensic investigations, according to eight different crime categories using six different models. In this sense, we propose an approach to classify and categorize digital images in forensic investigations. Experiments based on CNN learning indicate that utilizing criminal elements in forensic investigations is a viable method for classification, particularly based on the capture phase.

In support of this study, we prepared and published a publicly available dataset of our new dataset, which includes a total of 15,065 image samples from eight different crime categories used in the study. We also investigated the effect of various CNN learning-based methods and found that the NASNetLarge model gave the best results.

In conclusion, the proposed approach with an accuracy of 91.74% shows promising results, offering a reasonable detection time of 00:44:57 seconds. We believe that the digital forensics will gain more popularity in the near future due to the increase in digital image cases.

In future work, we plan to explore approaches that have better classification capabilities based on CNN learning and improve accuracy, in which different crimes include detection of different crime categories and possible manipulations.

Another interesting idea we are planning to test is an automated digital image analysis approach, which could allow us to automatically generate image properties of the examined digital forensic images.

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.

ORCID ID of the author / Yazarm ORCID ID'si

İlker Kara 0000-0003-3700-4825

REFERENCES

- adfsolutio, ns<https://www.adfsolutions.com/>, 2021.
- Belkasoft, <https://belkasoft.com/>, 2021.
- Birajdar, G.K., & Mankar, V.H. (2013). Digital image forgery detection using passive techniques: *A survey. Digital investigation*, 10(3), 226-245.
- Cao, H., & Kot, A.C. 2009. Accurate detection of demosaicing regularity for digital image forensics. *IEEE Transactions on Information Forensics and Security*, 4(4), 899-910.
- Chandra, M., Pandey, S., Chaudhary, R. (2010). Digital watermarking technique for protecting digital images. *In 2010 3rd International Conference on Computer Science and Information Technology* 7, 226-233. IEEE.
- Choodum, A., Boonsamran, P., NicDaeid, N., Wongniramaikul, W. (2015). On-site semi-quantitative analysis for ammonium nitrate detection using digital image colourimetry. *Science & Justice*, 55(6), 437-445.
- Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1251-1258).
- Coulbaly, S., Kamsu-Foguem, B., Kamissoko, D., Traore, D. (2019). Deep neural networks with transfer learning in millet crop images. *Computers in Industry*, 108, 115-120.
- Europol, (2020). How Are Organised Crime Groups Involved in Sports Corruption?. <https://www.europol.europa.eu/newsroom/news/how-are-organised-crime-groups-involved-in-sports-corruption>.
- de Castro Polastro, M., da Silva Eleuterio, P.M. (2010). Nudetective: A forensic tool to help combat child pornography through automatic nudity detection. *In 2010 Workshops on Database and Expert Systems Applications*, 349-353. IEEE.
- Del Mar-Raave, J. R., Bahşi, H., Mršić, L., Hausknecht, K. 2021. A machine learning-based forensic tool for image classification-A design science approach. *Forensic Science International: Digital Investigation*, 38, 301265.
- Dey, T., Mandal, S., Varcho, W. (2017). Improved image classification using topological persistence. *In Proceedings of the conference on Vision, Modeling and Visualization*, 161-168).
- Ferreira, W.D., Ferreira, C.B., da Cruz Júnior, G., Soares, F. (2020). A review of digital image forensics. *Computers & Electrical Engineering*, 85, 106685.
- Hafiz, R., Haque, M. R., Rakshit, A., Uddin, M. S. (2020). Image-based soft drink type classification and dietary assessment system using deep convolutional neural network with transfer learning. *Journal of King Saud University-Computer and Information Sciences*. 34(5), 1775-1784.
- Garfinkel, S.L., Parker-Wood, A., Huynh, D., Migletz, J. (2010). An automated solution to the multiuser carved data ascription problem. *IEEE Transactions on Information Forensics and Security*, 5(4), 868-882.
- Grillo, A., Lentini, A., Me, G., Ottoni, M. (2009). Fast user classifying to establish forensic analysis priorities. *In 2009 Fifth International Conference on IT Security Incident Management and IT Forensics*, 69-77. IEEE.
- Gomez, L. S. M. (2012). Triage in-Lab: case backlog reduction with forensic digital profiling. *In Proceedings of the Argentine Conference on Informatics and Argentine Symposium on Computing and Law*, 217-225.
- Isnard, A., Council, T. C. (2001). Can surveillance cameras be successful in preventing crime and controlling anti-social behaviours. In *Character, Impact and Prevention of Crime in Regional Australia Conference*.
- Kara, I. (2017). A Review About Child Abuse Crimes Committed Through Internet In Turkey. *Int J Forensic Sci Pathol*, 5(3), 337-340.
- Karakuş, S., Kaya, Ö. Ü., Ertam, Ö. Ü. F., Talu, M. F. (2018). *Derin Öğrenme Yöntemlerinin Kullanılarak Dijital Deliller Üzerinde Adli Bilişim İncelemesi*.
- Keras, <https://keras.io/api/applications/>, 2021.
- Kuhle, L.F., Oezdemir, U., Beier, K.M. (2021). Child Sexual Abuse and the Use of Child Sexual Abuse Images. In *Pedophilia, Hebephilia and Sexual Offending against Children* (pp. 15-25). Springer, Cham.
- Lin, X., Li, J.H., Wang, S.L., Cheng, F., Huang, X.S. (2018). Recent advances in passive digital image security forensics: A brief review. *Engineering*, 4(1), 29-39.
- Mahalakshmi, S.D., Vijayalakshmi, K., Priyadharsini, S. (2012). Digital image forgery detection and estimation by exploring basic image manipulations. *Digital Investigation*, 8(3-4), 215-225.
- Marturana, F., Tacconi, S. (2013). A Machine Learning-based Triage methodology for automated categorization of digital media. *Digital Investigation*, 10(2), 193-204.
- Marturana, F., Me, G., Berte, R., Tacconi, S. (2011). A quantitative approach to triaging in mobile forensics. *In 2011 IEEE 10th International Conference on Trust, Security and Privacy in Computing and Communications* (pp. 582-588). IEEE.
- McClelland, D., Marturana, F. (2014). A Digital Forensics Triage methodology based on feature manipulation techniques. *In 2014 IEEE International Conference on Communications Workshops (ICC)* (pp. 676-681). IEEE.
- McDown, R.J., Varol, C., Carvajal, L., Chen, L. (2016). In-Depth Analysis of Computer Memory Acquisition Software for Forensic Purposes. *Journal Of Forensic Sciences*, 61, S110-S116.
- M Kirchner & R. Böhme (2007). Tamper hiding: Defeating image forensics In *International Workshop on Information Hiding*, Springer, Berlin, Heidelberg. (2007), pp.326-341.
- Olmos, R., Tabik, S., Herrera, F. (2018). Automatic handgun detection alarm in videos using deep learning. *Neurocomputing*, 275, 66-72.
- Paluszek, M., & Thomas, S. (2020). Practical Matlab deep learning. A Project-Based Approach, Michael Paluszek and Stephanie Thomas.

- Pearson, H. (2006). Forensic software traces tweaks to images. *Nature*, 439(7076), 520-522.
- Peng, F., Liu, J., Long, M. (2013). Identification of natural images and computer generated graphics based on hybrid features. In *Emerging Digital Forensics Applications for Crime Detection, Prevention, and Security* (pp. 18-34). IGI Global.
- Peng, C., Liu, Y., Yuan, X., & Chen, Q. (2022). Research of image recognition method based on enhanced inception-ResNet-V2. *Multimedia Tools and Applications*, 81(24), 34345-34365.
- Piva, A. (2013). An overview on image forensics. *International Scholarly Research Notices*, 2013.
- Rahimzadeh, M., Parvin, S., Safi, E., Mohammadi, M.R. (2021). Wise-SrNet: A Novel Architecture for Enhancing Image Classification by Learning Spatial Resolution of Feature Maps. arXiv preprint arXiv:2104.12294.
- Ren, S., He, K., Girshick, R., Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 91-99.
- 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.
- Saber, A. H., Khan, M. A., & Mejbil, B. G. (2020). A survey on image forgery detection using different forensic approaches. *Advances in Science, Technology and Engineering Systems Journal*, 5(3), 361-370.
- Sharma, M., & Vig, L. (2018). Automatic classification of low-resolution chromosomal images. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops* (pp. 0-0).
- Sanap, V.K., & Mane, V. (2015). Comparative study and simulation of digital forensic tools. *Int J Comput Appl*, 975, 8887.
- Seigfried, K.C., Lovely, R.W., Rogers, M.K. (2008). Self-Reported Online Child Pornography Behavior: A Psychological Analysis. *International Journal of Cyber Criminology*, 2(1).
- Sharma, A., Singh, A., Choudhury, T., Sarkar, T. (2021). Image Classification using ImageNet Classifiers in Environments with Limited Data.
- Simonyan, K., Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.
- Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A.A. (2017). Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence*.
- Tammina, S. (2019). Transfer learning using vgg-16 with deep convolutional neural network for classifying images. *International Journal of Scientific and Research Publications (IJSRP)*, 9(10), 143-150.
- Thakur, R., Rohilla, R. (2020). Recent advances in digital image manipulation detection techniques: A brief review. *Forensic Science International*, 312, 110311.
- Verma, G.K., Dhillon, A. (2017). A handheld gun detection using faster r-cnn deep learning. In *Proceedings of the 7th International Conference on Computer and Communication Technology* (pp. 84-88).
- x-ways, <https://www.x-ways.net/>, 2021.
- Wu, L.T., Parrott, A.C., Ringwalt, C L., Yang, C., Blazer, D.G. (2009). The variety of ecstasy/MDMA users: results from the National Epidemiologic Survey on alcohol and related conditions. *The American Journal on Addictions*, 18(6), 452-461.
- Zhang, X., Zou, J., He, K., Sun, J. (2015). Accelerating very deep convolutional networks for classification and detection. *IEEE transactions on pattern analysis and machine intelligence*, 38(10), 1943-1955.
- Zoph, B., Vasudevan, V., Shlens, J., Le, Q.V. (2018). Learning transferable architectures for scalable image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 8697-8710).
- Wang, W., Dong, J., Tan, T. (2009). A survey of passive image tampering detection. In *International Workshop on Digital Watermarking* (pp. 308-322). Springer, Berlin, Heidelberg.
- Dateset, <https://ilkerkara.karatekin.edu.tr/RequestDataset.html> dataset, 2021.

How cite this article

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

The Efficiency of Regularization Method on Model Success in Issue Type Prediction Problem

Sorun Türü Tahmini Probleminde Düzenleştirme Yönteminin Model Başarısı Üzerindeki Etkisi

Ali Alsaç¹ , Mehmet Mutlu Yenisey² , Murat Can Ganiz³ , Mustafa Dağtekin⁴ , Taner Ulusinan⁵ 

¹(M.Sc.), Istanbul University-Cerrahpasa, Institute of Graduate Studies, Department of Industrial Engineering, Istanbul, Türkiye

²(Prof. Dr.), Istanbul University-Cerrahpasa, Faculty of Engineering, Department of Industrial Engineering, Istanbul, Türkiye

³(Assoc. Prof. Dr.), Marmara University, Faculty of Engineering, Department of Computer Engineering, Istanbul, Türkiye

⁴(Dr.), Istanbul University-Cerrahpasa, Faculty of Engineering, Department of Computer Engineering, Istanbul, Türkiye

⁵Private company, Istanbul, Türkiye

Corresponding author : Ali ALSAÇ

E-mail : ali.alsac@ogr.iuc.edu.tr

ABSTRACT

Designing a prediction method with machine learning algorithms and increasing the prediction success is one of the most important research areas and aims of today. Models designed using classification algorithms are frequently used especially in problem types that require prediction. In this study, real life data is used to answer the question of which problem type should be included in the Information Technology Service Management (ITSM) system. An important step in the search for a solution is to examine the dataset with regularization methods. Experimental results have been obtained to establish the overfitting or underfitting balance of the dataset with L1 and L2 regularization methods. While the Root-Mean-Square Error (RMSE) value was approximately 0.13 in the regression model without regularization, this value was found to be approximately 0.083 after L1 regularization. With the regularized dataset, new results were obtained using Artificial Neural Network (ANN), Logistic Regression (LR), Support Vector Machine (SVM) classifier algorithms. SVM algorithm was the most successful model with a performance of approximately 0.73. It is followed by LR and ANN respectively. Accuracy, Precision, Recall and F1Score were used as evaluation metrics. It is seen that the use of regularization methods, especially in the preparation of real-life data for use in machine learning or other artificial intelligence research, will contribute to increasing the success level of the model.

Keywords: IT service management, regularization, prediction, classification

ÖZ

Matematik düzleminde bir tahmin yöntemi tasarlamak ve başarılı sonuçlarından faydalanmak günümüzün önemli araştırma alanlarından ve amaçlarından biri olarak öne çıkmaktadır. Sınıflandırma algoritmaları kullanılarak tasarlanan modeller özellikle tahmin gerektiren problem türlerinde sıklıkla kullanılmaktadır. Çalışmada gerçek hayat verileri kullanılarak bir gerçek hayat problemi olan müşteriden gelen çözüm talebinin Bilgi Teknolojisi Hizmet Yönetimi (BTHY) sistemi içinde hangi sorun tipine dahil edilmesi gerektiği sorusuna cevap aranmaktadır. Çözüm arayışının önemli bir aşamasında veri kümesinin Regülerizasyon yöntemleri ile incelenmesi yer almaktadır. L1 ve L2 regülerizasyon yöntemleri ile veri kümesinin overfitting ya da underfitting dengesinin kurulması için deneysel sonuçlar alınmıştır. Regülerizasyon uygulanmamış regresyon modelinde Kök Ortalama Kare Hatası (RMSE) değeri yaklaşık olarak 0,13 iken L1 regülerizasyonu sonucunda bu değer yaklaşık 0,083 olarak bulunmuştur. Düzenleştirilmiş veri kümesi ile Yapay Sinir Ağları (YSA), Lojistik Regresyon (LR), Destek Vektör Makinaları (DVM) sınıflandırıcı algoritmaları kullanılarak yeni sonuçlar elde edilmiştir. DVM algoritması yaklaşık 0,73 başarımla sonuçlu ile en başarılı model olmuştur. Sırasıyla LR ve YSA takip etmektedir. Değerlendirme metrikleri olarak Accuracy, Precision, Recall ve F1Score kullanılmıştır. Özellikle gerçek hayat verilerinin makina öğrenmesi ya da diğer yapay zeka araştırmalarında kullanımı için hazırlanması aşamasında Regülerizasyon yöntemlerinden faydalanmanın modelin başarı düzeyinin artmasında katkısı olacağı görülmektedir.

Anahtar Kelimeler: Bilgi işlem servis yönetimi, regülerizasyon, tahmin, sınıflandırma

Submitted : 22.11.2023

Revision Requested : 30.11.2023

Last Revision Received : 30.11.2023

Accepted : 01.12.2023

Published Online : 14.12.2023



This article is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0)

1. INTRODUCTION

The way for businesses, public institutions and organizations that produce goods or services to fulfill their activities faster has been opened with the participation of computers in business management. Production through computer-aided systems has facilitated the control of costs and increased the efficiency of production management. The increase in the transaction speed of service-producing enterprises has enabled them to work with more customers and increased their transaction volumes. According to TUIK 2021 data, 43.1% of the enterprises active in 2021 were in the service sector and 36.5% in the trade sector, while the service sector accounted for 37.4% of total employment in employment, while the employment share of the industrial sector was 28.9% (YENİSU, 2021). Information and Communication Technologies, which is one of the leading service sector areas, is the main area analyzed in this article.

In the study titled "Information and Communication Technologies Sector 2021 Market Data" conducted by Deloitte for TUBISAD, the predictions for the future of the IT sector are as follows (Deloitte & TUBISAD, 2022);

- Geopolitical and economic uncertainties and supply chain disruptions will increase spending on flexible and agile solutions
- As the focus on digitalization increases, companies position digital transformation as a strategic priority
- Increasing interest in emerging technologies such as analytics, cloud computing, digital customer experience and security solutions
- Remote working is becoming permanent and more than 50% of total employees are expected to switch to remote working in 2024

As can be understood from the items listed above, the importance of customer service will increase in many sectors. The permanence of remote working, which has become widespread with the Covid-19 pandemic, has also enabled the IT sector to expand its sphere of influence. Information technologies realized a growth of approximately 25% in the 2020-2021 period (Deloitte & TUBISAD, 2022).

Information technologies consist of hardware, software and service applications. Service applications include the following components (Deloitte & TUBISAD, 2022).

- Outsourcing services
- Consultancy services
- Development, integration, installation and operation services
- Support, care and training services

In this paper, we examine regularization models used to improve the success level of methods and models that seek solutions to the problem of assigning customer requests to the most appropriate expert. Assigning customer requests to the most appropriate consultant with the help desk application enables the control or reduction of time-related costs. The core of the study is to weight the attributes that affect the performance of the assignment problem in the model with machine learning methods on the data set to be explained in the following sections, or to penalize the attributes that do not affect the solution or cause the model to overlearn.

Customer satisfaction will increase if the customer's request reaches the right consultant in the fastest way possible. Currently, the customer service approach tries to increase service quality by categorizing customers in classical business models according to transaction volume, business nature, frequency or direct strategic importance for the business. However, this method of work allocation does not meet the quest to increase the speed and quality of service provided by finding the most appropriate consultant for the incoming demand.

The overlearning of the above-mentioned model is generally defined as "overfitting" in the literature. The concept of model is the broadest definition that covers all the methods, definitions, evaluations and explanations used to address and solve the problem. The concept of hyperparameterization also stands out as a very important concept in machine learning models. The negativities caused by overlearning of the model and hyperparameters in this context are discussed in the second section.

In the first part of the study, the literature is presented. In the second section, the methodology of the study is discussed and in the third section, the real-life data of a company operating in the IT sector are evaluated in an experimental environment using the machine learning algorithms described in the methodology and the success of the model is examined. The fourth section includes a discussion section where the results obtained from the model are evaluated. The fifth and final section presents the results obtained within the subject integrity of the article. The paper concludes with an area where the authors express their gratitude to the institutions and individuals who have supported their scientific work, followed by a bibliography.

2. RELATED WORK

In the IT sector, the use of criteria such as education, work experience, availability in working hours, age, language skills, salary group, etc. in the assignment of experts to customers constitutes the essence of the model designed for the problem. In order to increase the success level of the model, regularization is used as an experimental method. In the literature, the two cases mentioned above, namely expert assignments in the IT sector and regularization methods, are examined in academic databases. In Web of Science and Google Scholar databases, the keywords "expert recommendation", "issue classification", "regularization", "help desk" are searched in relation to the keywords "Machine Learning" and "Artificial Intelligence" and sample studies on the topics studied in the article are selected. When a search was made by selecting All Fields, 1665 publication results were obtained. When Open Access was selected, the number decreased to 663. When WoS Index is selected as SCI-EXPANDED and CPCI-S, publication date between 2008-2023, the number of publications is 512. Citation Topics Meso; Artificial Intelligence and Machine Learning, Knowledge Engineering, Software Engineering, Numerical Methods were selected to increase the precision of the search and thus the number of publications was simplified to 138. Web of Science Categories; Computer Science Artificial Intelligence, Computer Science Information Systems, Mathematics, Computer Science Software Engineering were selected and the number of publications was simplified to 85. In this paper, the use of regularization techniques to improve the performance of algorithms used in classification problems will be evaluated with experimental results. Learning the notations of regularization methods and examining their applications will be sufficient to understand the technique.

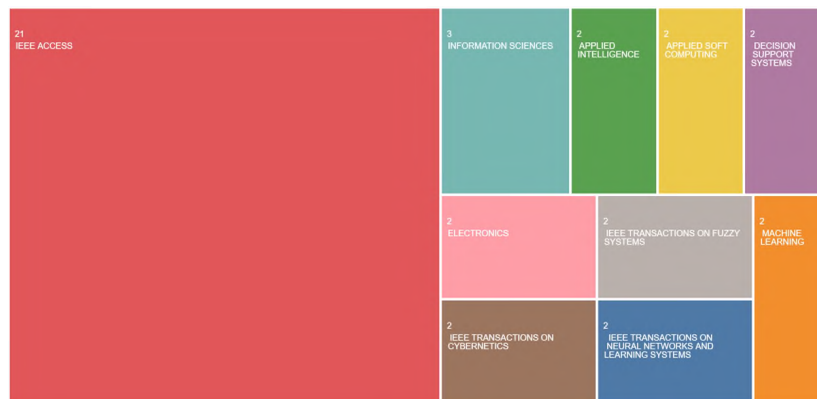


Figure 1. Sources selected for the article

As a result of the literature review on the classification of demands from customers, which is shown in Figure 1, some of the articles that are suitable for the content of the article are described below.

This paper is a follow-up to a conference paper published in 2022, which used data from a company operating in the ITSM sector to estimate the time it takes to resolve tickets to an expert. 16970 data sets were used in this study and Linear Regression, Decision Tree Regression, Random Forest Regression, Support Vector Machines Regression and Multiple Regression algorithms were used. Various metrics such as MAE, MSE and MAPE were used to evaluate these supervised models. The results show varying levels of success with different supervised machine learning algorithms for this challenging task. Among the trained models, Decision Trees and Random Forest Regression were evaluated as the algorithms with the best results (Yildiz et al., 2022).

Jonsson et al. (2016) investigated different classifiers for fault assignment problems and discovered that utilizing Stacked Generalization to combine the best classifiers enhances performance. An ensemble approach called stacked generalization (DG) uses the output of several classifiers as the input for a final classifier that determines the final class. According to the authors, the DG model's classification accuracy of four mistake datasets from a telecoms business ranged from 57% to 89%, which was comparable to the human approach in use at the time. Ultimately, the study demonstrated that using a topic's non-textual fields yields more encouraging outcomes.

Helming et al. (2011) suggests in their paper a novel model-based method that takes into account the connections between task items and system properties when assigning. They contrast this method with other approaches that investigate both textual content and structure information. Every technique is used on many kinds of work items, such as tasks and bug reports. They look through the model repositories of three distinct projects, including historical data, for our assessment to see how well they perform using various techniques. In this paper, they present a new model-based

method for semi-automatically assigning task items and compare other machine learning strategies that are currently in use. Every method is used on a single, unified model that is built within the UNICASE tool.

The fault classification method proposed by Zibran (2016) is based on topic modeling. The method used is the Labeled Latent Dirichlet Discrimination (EGDA) algorithm, named after the German mathematician Johann Peter Gustav Lejeune Dirichlet. In this study, we investigate the effectiveness of labeled EGDA in automatically classifying error reports into a predefined set of categories.

Bhattacharya et al. (2012) use a probabilistic graph-based model in this paper, which they propose to be a model that makes highly accurate predictions. This is the first study to look at how several machine learning dimensions (features, training history, and classifiers) affect prediction accuracy in fault assignment, as well as how fault discard graphs affect it. Using data from Eclipse and Mozilla that spans 21 combined development years and 856,259 bug reports, they objectively assess their methodology. They demonstrate how their method can greatly shorten throwing pathways and achieve up to 86.09% prediction accuracy in fault assignment. They contend that the greatest results for their dataset come from combining a Naive Bayes classifier with scatter plots, incremental learning, and product-component characteristics. They highlight optimization strategies that shorten training and prediction times while delivering excellent prediction accuracy.

In recent years, research on regularization methods has increased significantly over time due to the need to develop more accurate and reliable prediction models. Papers that examine various clustering and optimization problems using regularization methods include evaluations based on cost function results.

Li and Zhou (2009) addressed the group weight finding problem by employing a cost function that combined hinge loss and L1 regularization. They utilized Quadratic programming to minimize this cost function, conducting experiments with Decision tree classifiers and UCI datasets. Additionally, they proposed a semi-supervised version and found that the Regularized Selective Ensemble Algorithm (RSE) could generate ensembles with strong generalization ability while maintaining a small size.

Zhang and Zhou (2011) tackled the weight finding problem, formulating three distinct cost functions: LP1, which utilized only Hinge loss; LP2, incorporating Hinge loss and L1 adjustment; and LP3, allowing negative weights. Linear programming was employed to minimize these cost functions, and the experiments featured the K-Nearest Neighbor (KNN) algorithm as base classifiers along with UCI datasets.

Goldberg and Eckstein (2012) approached the weight finding problem using the indicator loss function and L0 regularization. They considered this problem NP-hard in specific cases and provided various relaxation strategies and bounds for solving it. Importantly, their work was primarily theoretical in nature, distinguishing it from other practical implementations.

Tinoco et al. (2013) combined MLP and SVM algorithms for classifying remote sensing images, employing genetic algorithms to find the weights. An improved version of their work utilized hinge loss and L1 regularization, with linear programming employed to minimize the cost function. Both versions classified remote sensing images using an ensemble of MLP and SVM classifiers.

Hautamaki et al. (2013) explored sparse ensembles in the speaker verification domain, modeling ensemble weight finding with a cross entropy loss function and three regularization functions: L1, L2, and L1+L2. The Nelder-Mead method was used to minimize these cost functions, and logistic regression classifiers were employed in the experiments.

Yin et al. (2012) addressed ensemble weight finding, incorporating a cost function with squared loss, L1 regularization, and diversity terms based on Yule's Q statistic. Their experiments featured neural network classifiers on six UCI datasets, and the proposed cost function was initially minimized using genetic algorithms.

Şen and Erdogan (2013) modeled ensemble weight finding using a cost function that included hinge loss and two regularization functions: L1 and group sparsity. Convex optimization techniques were employed to minimize this cost function, and experiments involved comparing 13 classifiers on 12 UCI datasets and three other datasets using CVX Toolbox.

Mao et al. (2013) tackled ensemble weight finding using a cost function consisting solely of absolute loss, minimizing it through a 0-1 matrix decomposition. In a subsequent work, they proposed a cost function with squared loss and L1 regularization, minimized using a quadratic form approach. Decision tree weak classifiers and UCI datasets were used in both studies.

Özgür et al. (2018) introduced a sparsity-driven weighted ensemble classifier (SDWEC) to enhance classification accuracy and minimize the number of classifiers. SDWEC formed ensembles with pre-trained classifiers, and the assigned weights determined how base classifiers voted. Efficiency tests on 11 datasets showed that SDWEC outperformed or matched state-of-the-art classifier ensemble methods, achieving similar accuracy levels with fewer classifiers and reducing testing time for the ensemble.

3. APPROACH

It is important to note that supervised learning is utilized when we wish to anticipate a certain outcome based on a specific input and we have instances of input/output pairings. Our training set consists of these input/output pairs, from which we construct a machine learning model. Making precise forecasts for fresh, never-before-seen data is our aim. Building the training set for supervised learning frequently takes human labor, but once done, it automates and frequently accelerates a tedious or impractical activity.

Linear Models (Linear Regression, SVM, etc.) stand out as the most widely used Machine Learning algorithms. However, they have an important drawback, they are very prone to Overfitting. In its simplest form, as seen in the 2-dimensional plane shown in Figure 2, the best fitting line (or curve) segment to the data points is tried to be found.

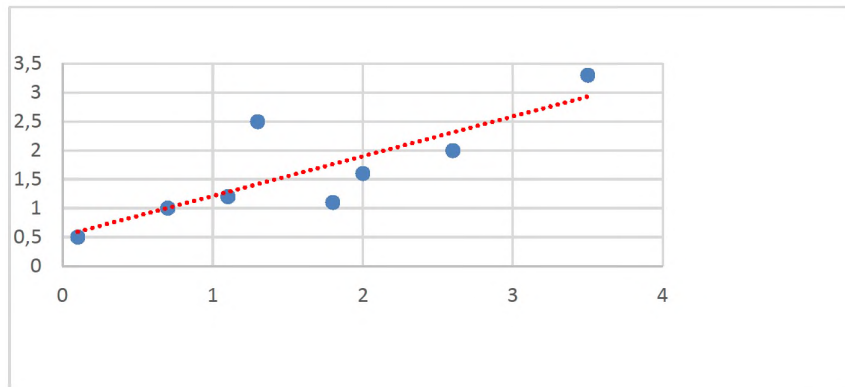


Figure 2. Linear Regression

The equation expressing linear regression models using coefficient weights is shown below.

$$y = w_0 + w_1x_1 \quad (1)$$

As the number of variables increases, the coefficients also increase in number. Increasing the number of variables also means increasing the attributes added to the model. This situation can be explained with the following equation.

$$y = w_0x_0 + w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4 + w_5x_5 + \dots + w_nx_n \quad (2)$$

Linear models are machine learning models that are prone to overfitting the training data. In case of overfitting the training data, the accuracy, reliability and generalization of the model in the testing phase are weakened. The concepts of overfitting, hyperparameters, variance and bias should also be mentioned.

Bias-variance decomposition is an important tool for understanding machine learning algorithms and its use in experimental studies has grown rapidly in recent years. The concepts of bias and variance help explain how very simple learners can outperform more complex learners and how groups of models can outperform single models (Domingos, 2000). In machine learning studies, bias is defined as the difference between the true value and the predicted value, and variance is defined as the amount by which the predictions deviate from the average prediction.

Every model has bias and variance error components. Bias and variance are inversely related; trying to reduce one component of the model will cause the other component to increase (Geman et al., 1992). Low bias and low variance are desirable characteristics in a model. The errors of the bias component are due to incorrect assumptions in the learning method. Figure 3 shows the relationship between bias and variance. At the point where the error is at its lowest, there is the necessary agreement between bias and variance for the model to obtain successful predictions. Complexity beyond this point means high variance. It is seen that the classification success will decrease.

In equation (2) above, as the coefficients $w_0, w_1, w_2, \dots, w_n$ increase, the variance will increase and a model that is difficult to generalize will emerge.

Since variance is sensitive to changes in the fit of the model, even a small change in the training data, it generates errors; therefore, high variance can lead to the problem of overfitting (Dangeti, 2017).

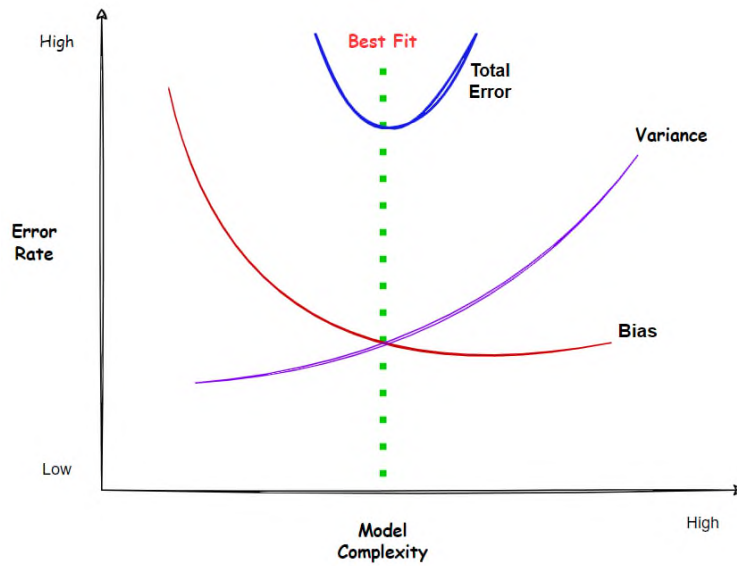


Figure 3. Bias Variance Correlation

The only measure of whether an algorithm will perform well on new data is the evaluation on the test set. However, intuitively, simple models are expected to generalize better when working with new data. Therefore, we always want to find the simplest model. Overfitting, or more widely, overfitting, is the process of creating a model that is too sophisticated for the data that we currently have. When a model is overfitted, it performs well on the training set but is unable to be applied to fresh data because it was fitted too tightly to the training set’s features. However, if your model is very simplistic, you might not be able to fully capture all of the nuances and variations in the data. Underfitting or underlearning is the process of selecting a model that is overly simplistic (Muller & Guido, 2017). As shown in Figure 4, the model fits all the outcome points. The distribution of the dataset is also very influential on overfitting. For example, if the class distribution is 90% to 10% in a two-class data set, a prediction model run on this data set will have a 90% success rate in the training data. This will also lead to overlearning, i.e. the model will be memorized. On the test data set or a new data set, this success cannot be achieved.

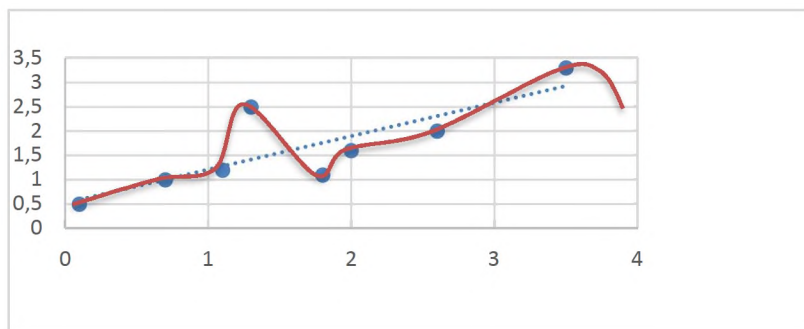


Figure 4. Extreme learning curve in regression models

The more the complexity of the model is allowed to increase, the better the training data is predicted. However, if the model becomes too complex, it starts to focus too much on each data point in the training set and the model cannot generalize well to new data. There is a sweet spot in between that will provide the best generalization performance. This is where the desired model tuning takes place. The variation in model performance between overfitting and underfitting is shown in Figure 5. Overfitting models have high variance and low bias. Where underfitting is observed, high bias is observed.

In order to prevent overfitting, multiple learning algorithms (ensemble), early stopping, cross-validation, feature engineering, expanding the volume of the dataset to create diversity and reducing the complexity of the model with the regularization method examined in this article are used. Linear models are prone to overfitting.

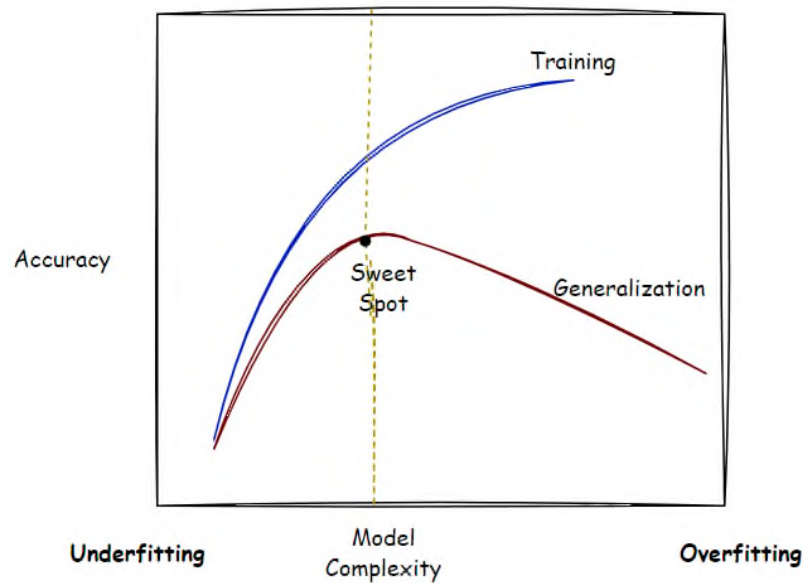


Figure 5. Impact of model complexity on training and test accuracy

The performance of a machine learning model can vary due to multiple factors, including data structure, size, the number of records related to classes, algorithms, performance verification methods, sampling techniques, and feature selection methods. Hyperparameters used in algorithms constitute a crucial factor influencing model performance, as they are parameters manipulated by the model developer. Consequently, hyperparameter tuning is a significant research area focused on optimizing these parameters to achieve the best possible model performance (Koçoğlu & Özcan, 2022).

Two types of parameters are encountered in machine learning models. These are called model parameters and hyperparameters. Model parameters are parameters that are included in the model and can be estimated from the data. They are not added to the model later by the expert during data analysis. Weights in a neural network, support vectors in a support vector machine, coefficients in linear regression or logistic regression are examples of model parameters (Tanyildizi & Demirtas, 2019).

Hyperparameters, unlike parameters, are not estimated from the data and depend on manual adjustment by the expert designing the model (Mantovani et al., 2017). Hyperparameters are adjustable parameters that can be selected by expert experience and trial and error methods. However, selecting the right hyperparameters requires an algorithmic process. Hyperparameter selection is seen as an optimization problem (Doğan, 2021). Kernel parameter (γ), epsilon value (ϵ) used in support vector machines; neighborhood value (k) in K-Nearest Neighbor algorithm; filter size, number of filters, number of neurons, number of layers, activation function, etc. used in deep neural network algorithms are among the examples of hyperparameters (Şipal et al., 2022).

Examples of hyperparameter solution methods considered as optimization problems are Grid Search, Random Search, Bayesian Optimization, Cross Validation and Hyperopt, Scikit Optimize and Optuna, which are included in the library of Python software language called alternative methods (Doğan, 2021).

3.1. Regularization

Overfitting is one of the most typical problems that any data scientist deals with. It is common for a machine learning model to perform well on training data, but not so well on testing data or new data sets. This suggests that the model is unable to predict the output or the target column of unseen data by introducing noise into the output. Noise is data points in the data set that do not really reflect the true qualities of our data but are there by chance (Kotsilieris et al., 2022).

Regularization greatly reduces the variance of the model without introducing a large bias. Consequently, the tuning parameter (α) used in regularization techniques limits the impact on variance and bias. As the value of (α) increases, the value of the coefficients decreases, reducing variance. This increase in (α) is useful to some extent because it only reduces variance without sacrificing any important features in the data, thus avoiding over-fitting. However, once a certain value is reached, the model starts to lose important features, leading to bias and Poor Fit. Consequently, the

value of (α) should be chosen carefully. It is a useful strategy to improve the accuracy of regression models (Friedrich et al., 2023).

Data collection and data preprocessing are the main causes of Overfitting. A data set with an uneven distribution of features, noises, random data fluctuations and variance can have an adverse effect on model training. The model learns these random errors and fluctuations so well during training that the accuracy of the training data model becomes extremely high, at which point the overfitting problem is encountered. A simple solution to overfitting is to update and penalize the weights. Table 1 shows the types of regularization and general approaches. In this paper, experimental studies with Ridge, Lasso and Elastic Net methods based on penalization type regularization are evaluated.

Table 1. Overview of types of regularization, general approaches and methods (Friedrich et al., 2023)

Regularization Type	Description	Solution Approach and Methods
Penalization	Add penalty term(s) to fitting criterion	Ridge regression, LASSO, elastic net
		Bayesian regularization priors
		Constraints for parameters
		Random effects
		Semiparametric regression
Early stopping	Early stopping of an iterative fitting procedure	Coefficient paths in penalization approaches
		Boosting
		Pruning of trees
		Learning rate in deep neural networks
Ensembling	Combine multiple base-procedures to an ensemble	Bagging
		Random forests
		(Bayesian) model averaging
		Boosting
Other approaches	-	Injecting noise
		Random probing in model selection
		Out-of-sample evaluation

In linear forecasting models, the Least Squares method aims to minimize the forecast error. The coefficients of the model tend to grow in (-) or (+) direction. On the other hand, regulatory extensions penalize the growth of the model's coefficients. Since the penalization prevents the model coefficients from growing, it prevents the model from producing extreme results. The sum of the squares of the error is the sum of the differences between the actual data point and the result point, i.e. the predicted value, formed by the data point taken into the function. Linear regression models are based on explaining this total value.

The cost (loss) function when applying regularization is shown in equation (3).

$$L_R = \Sigma(y_{gercek} - y_{tahmin})^2 + \alpha \Sigma w_i^2 \tag{3}$$

The equation $\alpha \Sigma w_i^2$ denotes the regularization part, while the coefficient α in the equation is the regularization coefficient and is a hyper parameter. As mentioned in section 2.c, the hyperparameters are given to the model externally. The hyper parameter α takes values between 0 and 1.

w is a parameter and is always a positive value. It is clear from this explanation that the model always wants to keep the parameters w small. The hyperparameter α is determined according to the parameter w. The operations performed for this purpose are called penalization (Tian & Zhang, 2022).

In this paper, experiments are carried out with Lasso, Ridge and Elastic Net regularization methods.

3.1.1. L1 Regularization (Lasso)

Lasso regularization has an approach that forces the coefficients to converge to zero (Emmert-Streib & Dehmer, 2019).

$$L(w) = \Sigma_{i=1}^n (y_i - \hat{y}_i)^2 + \alpha \Sigma_{j=1}^p |w_j^1| \tag{4}$$

When the equation is examined, it is seen that the exponent of the parameter w is 1. The Lasso regularization is called L1, referring to the exponent 1 of the parameter w (Bharambe et al., 2022). The parameters and hyperparameters in equation (4) are defined below:

α = penalty term (between 0 and 1)
 $|w_j^1|$ = absolute value of the coefficients (slope of the curve)
 y_i = actual result
 \hat{y}_i = prediction result

3.1.2. L2 Regularization (Ridge)

Ridge regression reduces the size of the regression coefficients so that the coefficients of the variables are close to zero. The penalty term L2, the sum of the squared coefficients, is used to penalize the regression model that causes the coefficients to shrink. The alpha (α) constant, the hyper parameter, is used to fine-tune the amount of penalty. It is very important to choose a perfect value for α . When α is set to 0, the penalty component has no effect and the OLS coefficients are calculated using Ridge regression. However, when α approaches infinity, the shrinkage penalty becomes more significant and the Ridge regression coefficients approach zero (Golam Kibria & Banik, 2016).

$$L(w) = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \alpha \sum_{j=1}^p w_j^2 \quad (5)$$

When the equation is examined, it is seen that the exponent of the parameter w is 2. Ridge regularization is called L2, referring to the exponent 2 of the parameter w , just as in the Lasso method (Bharambe et al., 2022).

3.1.3. Elastic Net Regularization

Elastic Net emerged in reaction to criticism of Lasso, which relies heavily on data for variable selection, making it unstable. Ridge regression and Lasso's penalties are combined to get the best of both approaches (Paper, 2019).

$$L(w) = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \alpha \sum_{j=1}^p |w_j^1| + \alpha_2 \sum_{j=1}^p w_j^2 \quad (6)$$

In the Elastic Net method, additional cost is added to the loss, cost function to form a hybrid of the L1 and L2 methods (Bharambe et al., 2022).

3.2. Support Vector Machines

The support vector machine can classify data into two or more classes with linear separation mechanisms in two-dimensional space, planar separation mechanisms in three-dimensional space and hyperplane separation mechanisms in multi-dimensional space (ÇELİK et al., 2021).

The case where a group of data can be separated by a line is the case where the group can be separated linearly. The idea is that the object separating the two classes is a corridor rather than a line, and that the width of this corridor is determined by some data vectors and is as wide as possible (Cortes & Vapnik, 1995).

In SVM literature, an attribute is termed a predictor variable, and a feature denotes a transformed symbol used to describe the hyperplane. Feature selection involves the task of choosing the most appropriate representation. A collection of features describing a case, such as a row of predictor values, is referred to as a vector. Therefore, the objective of SVM modeling is to identify the optimal hyperplane that separates sets of vectors, placing one-category cases of the target variable on one side of the plane and the other-category cases on the opposite side (Witten et al., 2016). The purpose of SVM in linear problems is to find a hyperplane passing through the features. This hyperplane consists of two lines where the features belonging to the classes are the furthest apart. Figure 6 shows the lines on this hyperplane.

In a non-linear dataset, SVMs cannot draw a linear hyperplane. Therefore, Kernel is used. The Kernel method greatly improves machine learning on nonlinear data. The operation of the SVM estimator (y) is expressed as follows (ARSLAN et al., 2020).

$$y = (K_{x_i} W_{j_k}) + b \quad (7)$$

The kernel function K_{x_i} is the bias term of the SVM network " b " and W_{j_k} is the weight vector. K and W denote Lagrange multipliers.

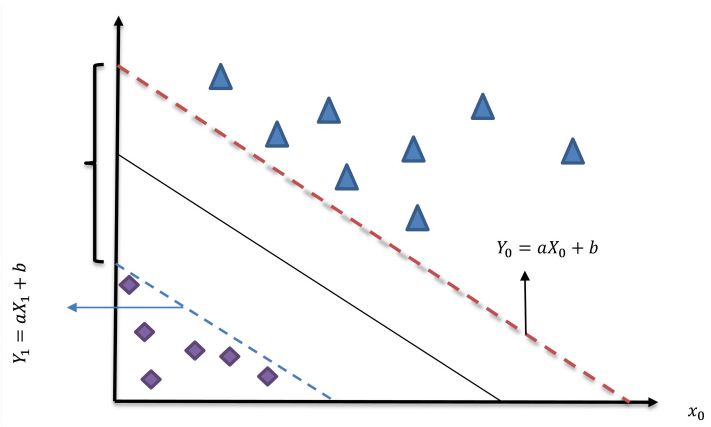


Figure 6. Hyperplane and support vectors

3.3. Logistic Regression

Logistic regression analysis is named after the logit transformation applied to the dependent variable (Hair et al., 2010). Logistic regression analysis is divided into three according to the type of scale on which the dependent variable is measured and the number of options of the dependent variable. If the dependent variable is a categorical variable with two options, it is called "Binary Logistic Regression Analysis". For example, binary logistic regression is applied when students are classified as successful and unsuccessful according to their completion of an academic program. If the dependent variable is a variable with more than two categories (levels), it is called "Multinomial Logistic Regression Analysis". For example, if there is a dependent variable consisting of students studying in three different academic programs, multinomial nominal logistic regression is applied. If the dependent variable is obtained with an ordinal scale, then "Ordinal Logistic Regression Analysis" is used. For example, ordinal logistic regression is applied when students' achievement in the academic program they are studying is grouped as "low", "medium" and "high" (Cook et al., 2001).

In logistic regression analysis, logit transformation is applied to the dependent variable and the logit of the dependent variable is estimated with the help of the independent variable. Logistic regression analysis, also called logit model, is a method used to determine the cause-and-effect relationship between independent variables and the dependent variable when the dependent variable has two, three or multiple categories and explains the effects of independent variables on the dependent variable with odds ratio (TAZEGÜL et al., 2016). Odds ratio is also called betting odds.

$$Odds\ Ratio = \frac{P_i}{1 - P_i} \tag{8}$$

According to Equation 8, P_i represents the probability of occurrence of an observed situation ($i=1,2,3,\dots,n$) and $1-P_i$ represents the probability of non-occurrence of an observed situation. In this case, the dependent variable takes the value 1 for P_i ($Y_i = 1$) and 0 for $1-P_i$ ($Y_i = 0$), making it bicategorical. Moreover, independent variables can be continuous, categorical or both. Odds ratio is defined as the ratio of the probability that a situation will occur to the probability that it will not occur.

The odds ratio ensures that the probability estimation takes a value between 0 and 1. However, in order to prevent the odds ratio from taking a value below zero, the logit value should be calculated by taking the natural logarithm of the value obtained with the odds ratio. As a result of the calculation of the logit value, a metric variable that can be converted into a probability between 0-1 is obtained (ŞENEL & ALATLI, 2014).

The model in which the logit value is obtained by taking the natural logarithm after the odds ratio is calculated is shown in Equation 9.

$$\text{logit}(Y) = \ln_e \left[\frac{P_i}{1 - P_i} \right] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n \tag{9}$$

If the odds ratio is less than 1, the logit value takes a negative value, while if it is greater than 1, it takes a positive value. As a result of logistic regression analysis, the intended model, which is a non-linear logarithmic function, is obtained and the model coefficients are shown as logarithmic values. In order to eliminate this situation that makes the

interpretation of the model coefficients difficult, the exponential logistic coefficient value obtained by taking the anti-logarithms of the coefficients and denoted by the symbol $\text{Exp}(\beta)$ is used. The model coefficients provide information about the direction of the relationship, while the exponential logistic coefficient provides information about how many times the change in the independent variable will decrease or increase the likelihood value.

3.4. Artificial Neural Network

ANN is a polycentric, parallel computational or rather modeling method inspired by the nerve cells in the human brain, first named as "neurons" by the German scientist Heinrich Wilhelm Gottfried von Waldeyer-Hartz in 1890 (Anderson D, 1992). The first concrete modeling technique based on neural cells was introduced by Frank Rosenblatt in 1958 as a simple perceptron (Yoon, 1989). Figure 7 shows the structure of an artificial neural network.

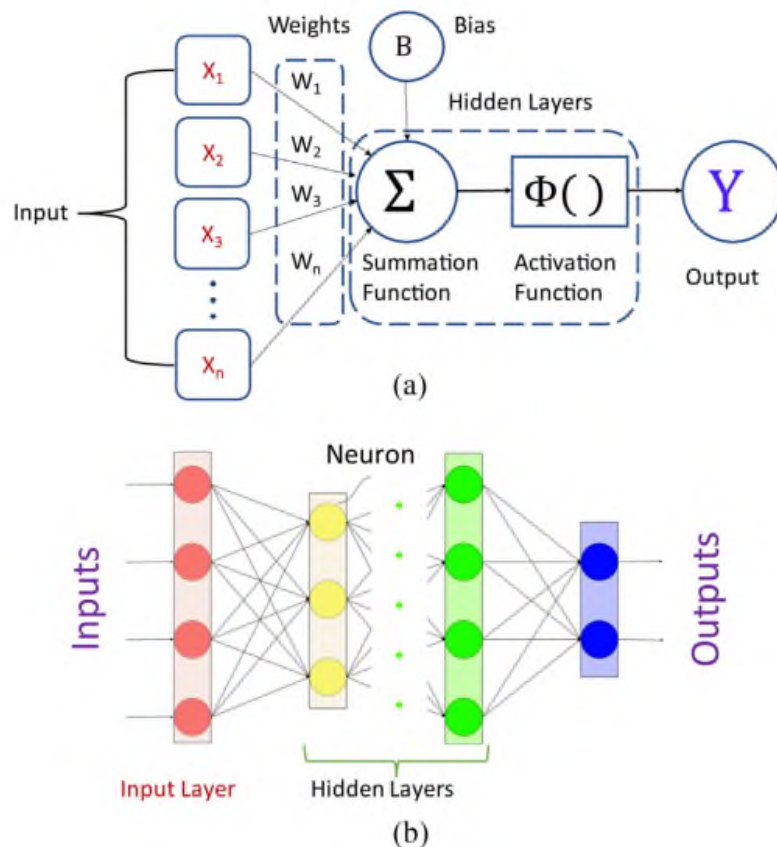


Figure 7. (a) An artificial neuron, and (b) Structure of an artificial neural network (Sinha et al., 2023)

The optimal architecture for a neural network should be sufficiently large to learn the problem yet small enough to generalize effectively. A network smaller than the optimal architecture struggles to learn the problem, while a larger network may overlearn the training data, resulting in poor generalization. Two primary approaches guide the determination of network structure: growing/constructive and pruning/destructive. The choice depends on whether the network's structure starts small and grows during learning (constructive) or begins large and shrinks during learning (destructive) (Aran et al., 2009).

There are basically three main layers in ANN, which are inspired by the information processing process of the brain. The names of these layers are input layer, hidden layer and output layer. The number of hidden layers can be one or more. The hidden layer between the input and output layer consists of structures called neurons. Each neuron in the layer is connected to all the neurons in the layer after it, but not to the neurons in the current layer. The input layer contains the parameters related to the state to be classified in the output layer. The hidden layer performs information processing and the output layer produces the class label or estimates the continuous time value (ÖZBİLGİN & KURNAZ, 2023).

Each relation between layers is assigned a weight value. As given in Equation 10, the values in the input layer are multiplied by the weights and given to a non-linear function by adding bias.

$$h_j = f(\sum_i w_{ij}x_i + b_j^1) \tag{10}$$

In the equation 10, x_i is the input parameters and w_{ij} is the weight value connecting input i to hidden neuron j . h_j is the output of hidden neuron j . b_j^1 bias and f is the activation function.

The mathematical expression of the sigmoid activation function, which is also widely used as a function, is as given in Equation 11.

$$f(x) = \frac{1}{1 + e^{-x}} \tag{11}$$

The mathematical expression of the output y_j in the output layer is as in the following equations 12 and 13.

$$y_j = \text{soft max}(z_j) = \frac{e^{z_j}}{\sum_j e^{z_j}} \tag{12}$$

$$z_j = \sum_i a_{ij}h_i + b_j^2 \tag{13}$$

Here, z_j represents the output units, b_j^2 the bias and a_{ij} the weight value between the j th output neuron and the i th hidden neuron.

Artificial Neural Networks (ANN) are commonly used in regression and clustering problems as well as in classification.

4. EXPERIMENT SETUP

4.1. Dataset

In this study, research is conducted on a problem of classification type. Classification is a technique that decomposes data in accordance with predetermined outputs. Since the outputs are known in advance, classification learns the dataset supervised (Giudici & Jiang, 2006).

Table 2. Jira Data

Jira Data			
1	ISSUEID	21	ORIGINAL_REPORTER
2	PKEY	22	ASSIGNEE1
3	ISSUE	23	ASSIGNEE2
4	ISSUETYPE	24	ASSIGNEE3
5	ISSUE_TYPE	25	ASSIGNEE4
6	PRIORITY	26	ASSIGNEE5
7	COMPONENT	27	WORKLOG_ASSIGNEE1
8	LABEL	28	WORKLOG_ASSIGNEE2
9	URGENCY	29	WORKLOG_ASSIGNEE3
10	IMPACT	30	WORKLOG_ASSIGNEE4
11	PLATFORM	31	WORKLOG_ASSIGNEE5
12	ISSUE_CATEGORY	32	LOG_HOURS_ASSIGNEE1
13	ISSUE_SUB_CATEGORY	33	LOG_HOURS_ASSIGNEE2
14	SUMMARY	34	LOG_HOURS_ASSIGNEE3
15	DESCRIPTION	35	LOG_HOURS_ASSIGNEE4
16	ASSIGNEE	36	LOG_HOURS_ASSIGNEE5
17	REPORTER	37	COMMENTOR_COUNT
18	CREATED	38	COMMENT_COUNT
19	RESOLUTIONDATE	39	WAITING_HOURS_AT_CUSTOMER
20	DUEDATE		

In our problem, it is aimed to assign an ITSM company to the expert who will answer the requests from customers in the fastest and most accurate way according to the characteristics of the request. In our study, which is supported

by experimental results to evaluate the effects of the regularization method on the solution sought with classification algorithms, the data of the Help Desk services provided through Oracle platforms in Table 3 and the personnel data of the human resources department of the company in Table 4 will be used. The data belongs to the year 2022.

Table 3. HR Data

HR Data			
1	PERSON_ID	9	IS_BILGISI
2	FIRST_NAME	10	KADEME
3	LAST_NAME	11	PERSONEL_TIPI
4	USER_PERSON_TYPE	12	EMAIL_ADDRESS
5	ORG_NAME	13	DATE_OF_BIRTH
6	TAKIM	14	CINSIYET
7	POZISYON	15	EFFECTIVE_START_DATE
8	SORUMLULUK	16	EFFECTIVE_END_DATE

The two datasets are combined into a single dataset based on the assignment problem that forms the core of the study. Experimental results are performed on the new merged dataset. There are 39 variables in the Jira dataset and 16 variables in the HR dataset.

4.2. Preprocessing

Since the "assignee" in Table 3 and the "email address" in Table 4 contain the same information, these two tables are used as reference for merging. What is meant by table merging is the merging of data sets. The merged data sets are then used as a single data set in the experimental processes. The model prepared for the defined problem will be studied with this new data set. The results of the combined data set in the evaluations are shown in Table 5.

Table 4. Attributes and Discriptions

Feature ID	Description	Feature ID	Description
1	ISSUEID	28	WORKLOG_ASSIGNEE2
2	PKEY	29	WORKLOG_ASSIGNEE3
3	ISSUE	30	WORKLOG_ASSIGNEE4
4	ISSUETYPE	31	WORKLOG_ASSIGNEE5
5	ISSUE_TYPE	32	LOG_HOURS_ASSIGNEE1
6	PRIORITY	33	LOG_HOURS_ASSIGNEE2
7	COMPONENT	34	LOG_HOURS_ASSIGNEE3
8	LABEL	35	LOG_HOURS_ASSIGNEE4
9	URGENCY	36	LOG_HOURS_ASSIGNEE5
10	IMPACT	37	COMMENTOR_COUNT
11	PLATFORM	38	COMMENT_COUNT
12	ISSUE_CATEGORY	39	WAITING_HOURS_AT_CUSTOMER
13	ISSUE_SUB_CATEGORY	40	PERSON_ID
14	SUMMARY	41	FIRST_NAME
15	DESCRIPTION	42	LAST_NAME
16	ASSIGNEE	43	USER_PERSON_TYPE
17	REPORTER	44	ORG_NAME
18	CREATED	45	TAKIM
19	RESOLUTIONDATE	46	POZISYON
20	DUEDATE	47	SORUMLULUK
21	ORIGINAL_REPORTER	48	IS_BILGISI
22	ASSIGNEE1	49	KADEME
23	ASSIGNEE2	50	PERSONEL_TIPI
24	ASSIGNEE3	51	EMAIL_ADDRESS
25	ASSIGNEE4	52	DATE_OF_BIRTH
26	ASSIGNEE5	53	CINSIYET
27	WORKLOG_ASSIGNEE1	54	EFFECTIVE_START_DATE
		55	EFFECTIVE_END_DATE

After the mentioned stages, the preparation process of the dataset is completed and it is made suitable for working with machine learning algorithms. By combining the 39-variable Jira dataset and the 16-variable HR dataset, a new 55-variable dataset was organized. Some of the variables in the expanded dataset, such as ISSUEID, may erroneously affect the experimental results of the model positively or negatively. ISSUEID only shows the registration order of the incoming request in the JIRA system. Therefore, it will not be useful to include it in the model. In the process of organizing the data set for the model, we tried to remove similar variables from the data set. As a result of the editing

Table 5. The data for classification analysis

	Column	Dtype
0	PRIORITY	object
1	URGENCY	object
2	IMPACT	object
3	ISSUE_CATEGORY	object
4	IS_BILGISI	object
5	Total_Assignee	int64
6	Total_Worklog_Assginee	int64
7	Total_Log_Hours_Assignee	int64
8	COMMENTOR_COUNT	int64
9	COMMENT_COUNT	int64
10	WAITING_HOURS_AT_CUSTOMER	int64
11	ISSUE_TYPE	object
12	Complition_Time	int64

process, a data set with 12 variables was obtained as shown in Table 6. In the next stages, data analysis is performed on the data set with Python programming language version 3.11.3.

As can be seen in Table 6, there are 7 properties of data type int64 and 6 properties of data type object. Int data type represents integer data. Object data type represents all other data types except specific data types. It is frequently encountered in data where objects such as letters, numbers and signs are used together. The dataset uses more than 6.4 MB of memory.

Another important point to consider when preparing the dataset is the distribution of the data. If the data input is concentrated in one data point, this will confuse the forecasting model. New data is predicted in the same way as it would be in the same data set. This is an example of overlearning and in this problem, we try to influence this situation with regularization models. The feature column named position was removed from the dataset with the drop function because it was skewed to a single class exceeding 90

Table 6. Data distributions for categorical data

ISSUE_CATEGORY				PRIORITY		IS_BILGISI	
IS1	0,300	WIP	0,002323	Major	0,62	IS2	0,27
PO	0,110	BI	0,002231	Minor	0,26	Junior	0,16
Custom	0,080	QA	0,001677	Critical	0,05	KD	0,14
HR	0,080	XTR	0,001600	Trivial	0,04	Senior	0,14
Salesforce	0,078	OIE	0,001569	Blocker	0,02	UZY	0,09
AP	0,075	PIM	0,000877			DU	0,09
INV	0,054	LINUX	0,000569	URGENCY		Uzman	0,03
GL	0,054	IPROC	0,000354	Medium	0,46	Consultant	0,02
Database	0,036	Training	0,000323	Low	0,43	Principal	0,02
AR	0,028	CE	0,000308	High	0,11	BİY	0,01
Sysadmin	0,025	OPMCosting	0,000308			Danışman	0,01
OE	0,020	IT	0,000292	IMPACT		DU	0,00
FA	0,014	Hyperion	0,000138	I1	0,58	Yönetici	0,00
Development	0,012	FAH	0,000062	SPSP	0,29	YAS	0,00
EAM	0,009	GRC	0,000046	OCWW	0,07	SY	0,00
PA	0,005	WMS	0,000046	NCA	0,04	Partner	0,00
ISUPPLIER	0,003	OrgPub	0,000031	ABSP	0,03	Müdür	0,00
CST	0,003					SUPC	0,00

By looking at the plot graph shared in Figure 8 for the columns containing *numerical* data, the distribution situations are conveyed. In the graphs, it is seen that the distributions of *Total Log Hours Assignee* and *Waiting Hours at Customer* are in an undesirable situation. It is natural to encounter such graphical shapes in real life data because the probabilistic state of life is reflected in the data. *Total Assignee* and *Completion Time* features show a favorable distribution.

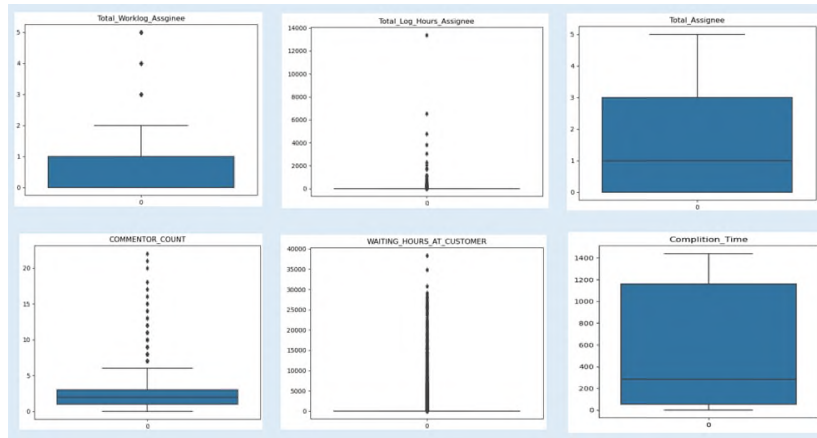


Figure 8. Plot graphs of numerical properties

Box Plot graphs of columns containing categorical data were analyzed according to Completion Time, which is the label class. Some of the inferences obtained from the graph shared in Figure 9 ;

- Tickets with Issue_type of Others and Incident take longer to complete
- Tickets with Is_information SY (Sales Manager) take longer to complete
- Tickets with GRC as Issue_category take longer to complete
- Tickets with Impact as OCWW take longer to complete
- Tickets with Priority as Trivial take longer to complete

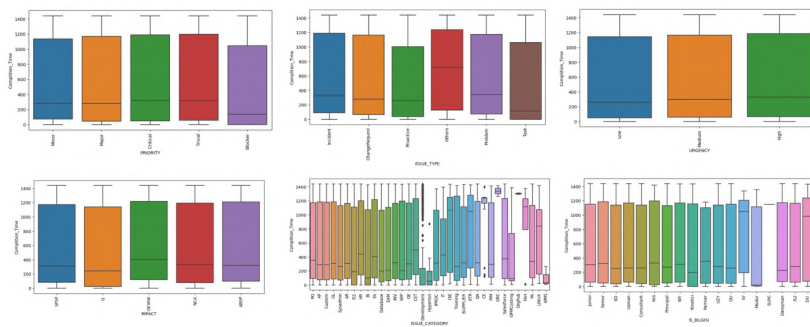


Figure 9. Box Plot Graphs for categorical data

A heat map was created to understand the correlation value between these variables. According to the heat map given in Figure 10;

- According to the heat map, the highest correlation is between Total_Log_Hours_Assignee, COMMENTOR_COUNT.
- The column most closely related to Completion_Time is COMMENTOR_COUNT.
- The weakest associated features/attribute is between Total_Log_Hours_Assignee and Completion Time.

Information about the dataset is described in this section, including data types, descriptions of real-life data and the steps involved in preparing the data for the algorithms. One hot encoding method was applied to categorical data with *get_dummies* function. This increased the number of columns, i.e. the number of features, to 77. After the outlier check, there were 64993 rows, which decreased to 64512 rows after this process, which means that 7 per thousand data was removed, which is an acceptable rate. Finally, the dataset was divided into 70% training 45158 and 30% test 19354 set.

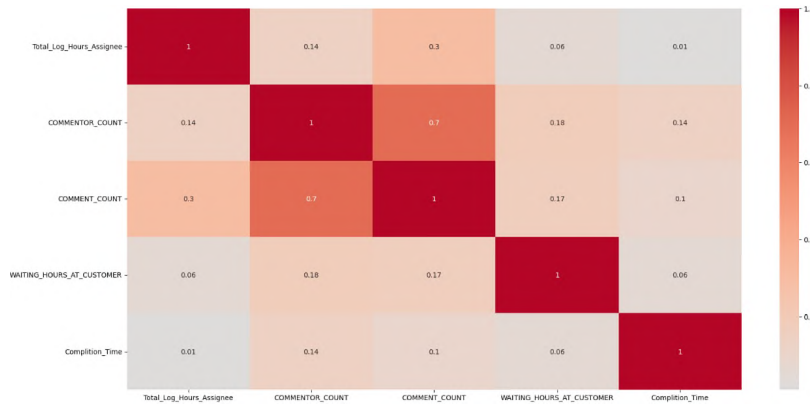


Figure 10. Heatmap to understand the correlation value between these variables

4.3. Evaluation Metrics

There are a number of evaluation metrics that are used to assess the models generated by classification algorithms and to compare which classification model yields better results. These metrics are usually based on a table structure called a Confusion Matrix. The Confusion Matrix is a table structure designed to illustrate how a classifier performs in machine learning and statistics classification problems (Ha et al., 2011).

There are four possible outcomes in dataset classification: a true positive (TP) when a truly positive example is correctly classified as positive, a false negative (FN) when a truly positive example is incorrectly classified as negative, a true negative (TN) when a truly negative example is correctly classified as negative, and a false positive (FP) when a truly negative example is incorrectly classified as positive (ALAN & KARABATAK, 2020).

Table 7. Confusion Matrix

Predicted Values		Actual Values	
		Positive	Negative
Positive	+	TP	FP
Negative	-	FN	TN

The performance metrics you use to measure how successful the model is are very important. If the evaluation is not done with the right metrics, a successful model can be characterized as unsuccessful and an unsuccessful model as successful. Machine Learning models are measured with the following metrics according to their types; Accuracy, Recall, Precision, F1 Score, ROC-AUC Curve, Log-Loss (Logarithmic Loss). The calculation methods of these criteria are given in the formulas below;

Accuracy is the rate at which the model created using the training set correctly classifies the data in the test set.

$$(Accuracy) = \frac{TP + TN}{TP + TN + FP + FN} \tag{14}$$

Recall is the rate at which the classifier correctly predicts data that belong to the positive class.

$$(Recall) = \frac{TP}{TP + FN} \tag{15}$$

Precision indicates the proportion of the positively predicted classifications that are correctly predicted.

$$Kesinlik(Precision) = \frac{TP}{TP + FP} \tag{16}$$

F Measure (F1 Score) is the weighted average of Precision and Recall. Therefore, it considers both FP (False Positive) and FN (False Negative) values.

$$(F1\ Score) = 2 * \frac{(Precision) * (Recall)}{(Precision) + (Recall)} \quad (17)$$

Apart from these criteria, another method used to evaluate classification performance is the ROC-AUC Curve (YETGINLER & ATACAK, 2020). It shows how successful the model is in separating the classes from each other on *Figure 11*. ROC stands for Receiver Operating Characteristic Curve or Probability Curve; AUC stands for Area Under the Curve or Area Under the Probability Curve. It takes a value between 0 and 1 and the closer it is to 1, the more successful the model is.

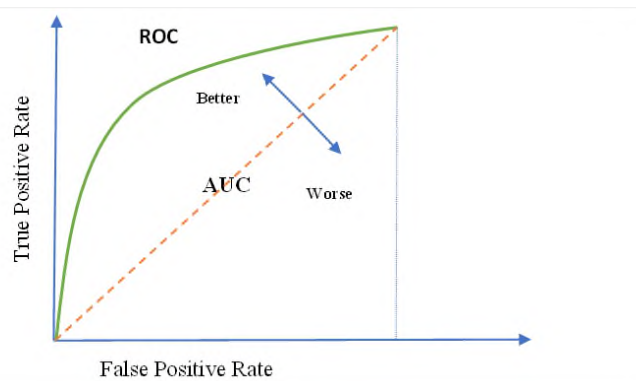


Figure 11. ROC-AUC Curve

In ROC curves, the x-axis is the FPO (False Positive Odds Ratio), while the y-axis is the TPO (True Positive Odds Ratio). For different threshold values, TPO and FPO values, i.e. sensitivity and precision values, are calculated. TPO and FPO pairs form the ROC curve. The ROC Curve is an increasing function between (0,0) and (1,1). After ROC analysis, it can be decided whether a test whose diagnostic success is evaluated is useless or a very good test (Orynbassar et al., 2022).

Root Mean Squared Error (RMSE) and r^2 also used to measure the performance of our model and k-fold cross validation is applied.

5. EXPERIMENTAL RESULTS AND DISCUSSION

The study adheres to the Cross Industry Standard Process for Data Mining (CRISP-DM) steps utilized in both Data Mining and Machine Learning. This process involves six steps: understanding the problem, understanding the data, preparing the data for analysis, developing models, evaluating model performances, and selecting the best model for application (Koçoğlu & Esnaf, 2022). In these stages, data analysis is performed on the data set with Python programming language version 3.11.3. The main libraries used are NumPy, SciPy, Matplotlib, Seaborn and SciKit-Learn.4

The data set was first fed into the linear regression model to find the prediction values of the completion time of the solution request from the customer called ticket. Then, the prediction success of the linear regression model was found and the effect of applying L2 (Ridge) and L1 (Lasso) regularization methods on this success value was observed. Figure 12 shows the mentioned processes on the diagram.

For the experimental application, the specified data set needs to be loaded and data preprocessing operations need to be performed. In data preprocessing, outlier detection and conversion of numerical data that may cause deviations such as year and month into string data are prioritized. Categorical data is corrected by encoding (LabelEncoder and OneHotEncoder). The next stage is Normalization, scaling and missing data management. Normalization rescales the data between 0 and 1. Feature scaling is based on the normal distribution and is calculated as the ratio of the distance of each value from the mean to the standard deviation. These are the preparations for using the dataset in experimental studies according to machine learning methods.

The problem is defined as correcting the success of the regression model established for predicting the completion time of a customer request with regularization and examining the classification success of the customer requests sent to the system according to the ISSUE TYPE class label through the classification method of the corrected model.

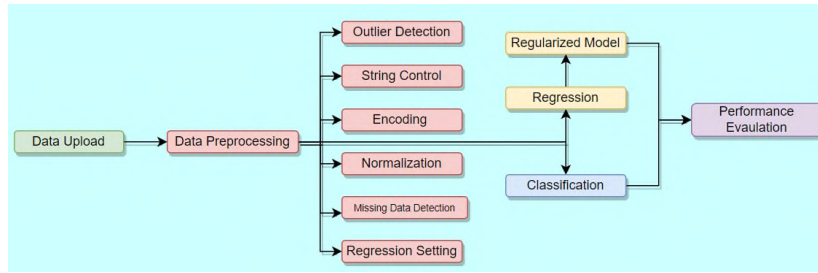


Figure 12. Diagram of the proposed model

In order to investigate the effect of regularization methods on the success of the model, an application is made on the dataset. The process starts with an unregularized OLS model and then a regularization model is constructed with L1 and L2 approaches.

The basic approach when reading the OLS table is to ensure that the p value is less than 0.05 *Confidence Interval*. Variables greater than this value do not have a significant effect on the result. Therefore, variables that are greater than this value are identified and removed from the model. Table 8 shows the results of the first iteration of the OLS method. According to the values under the p>|t| column in the list, variables such as PRIORITY_Blocker, 'IMPACT_NCA', 'ISSUE_CATEGORY_AP', etc. are dropped from the model with the code block 'X_train.drop('PRIORITY_Blocker',axis=1, inplace=True)'. The same process is continued until there is no variable with a p value greater than 0.005 in the list. In the model, this conclusion is reached at the end of the 3rd iteration. Score was calculated as R-squared: 0.753

Table 8. OLS Regression results for first iteration

	coef	std err	t	P> t	[0.025	0.975]		coef	std err	t	P> t	[0.025	0.975]
Total_Assignee	392.672	3.032	12.951	0.000	33.324	45.210	ISSUE_CATEGORY_IT	-425.168	130.730	-0.325	0.745	-298.750	213.716
Total_Worklog_Assignee	-108.761	4.396	-2.474	0.013	-19.492	-2.260	ISSUE_CATEGORY_LINUX	2.148.940	107.443	2.000	0.045	4.304	425.484
Total_Log_Hours_Assignee	-0.0726	0.033	-2.227	0.026	-0.136	-0.009	ISSUE_CATEGORY_OE	156.124	27.505	0.568	0.570	-38.298	69.523
COMMENTOR_COUNT	520.994	2.619	19.890	0.000	46.965	57.233	ISSUE_CATEGORY_OIE	1.703.153	62.245	2.736	0.006	48.315	292.316
COMMENT_COUNT	-28.274	0.494	-5.720	0.000	-3.796	-1.859	ISSUE_CATEGORY_OPMCosting	-2.101.062	134.870	-1.558	0.119	-474.453	54.241
WAITING_HOURS_AT_CUSTOMER	0.0039	0.002	2.057	0.040	0.000	0.008	ISSUE_CATEGORY_OrgPub	1.26E-09	6.16e-13	2.046	0.041	5.28e-14	2.47e-12
PRIORITY_Blocker	-0.5575	17.609	-0.032	0.975	-35.071	33.956	ISSUE_CATEGORY_PA	-260.093	41.579	-0.626	0.532	-107.505	55.486
PRIORITY_Critical	595.563	12.322	4.833	0.000	35.405	83.708	ISSUE_CATEGORY_PIM	0.8541	82.525	0.010	0.992	-160.896	162.604
PRIORITY_Major	525.489	8.352	6.292	0.000	36.179	68.918	ISSUE_CATEGORY_PO	1.280.628	23.673	5.410	0.000	81.664	174.462
PRIORITY_Minor	354.272	8.918	3.973	0.000	17.948	52.906	ISSUE_CATEGORY_Salesforce	-225.475	26.330	-0.856	0.392	-74.154	29.059
PRIORITY_Trivial	571.923	17.623	3.245	0.001	22.650	91.734	ISSUE_CATEGORY_Sysadmin	1.332.519	54.681	1.738	0.082	-202.191	12.162
URGENCY_High	657.044	10.047	6.540	0.000	46.012	85.396	ISSUE_CATEGORY_Training	-950.147	139.459	-1.655	0.098	-140.091	406.595
URGENCY_Low	557.391	9.455	5.895	0.000	37.208	74.271	ISSUE_CATEGORY_WIP	-6.046.066	365.248	-1.655	0.098	-1.320.499	111.286
URGENCY_Medium	827.236	9.076	9.115	0.000	64.934	100.513	ISSUE_CATEGORY_WMS	2.317.465	65.252	3.552	0.000	103.851	359.642
IMPACT_ABSP	358.105	15.959	2.244	0.025	4.530	67.091	ISSUE_CATEGORY_XTR	32.622	44.163	0.074	0.941	-83.297	89.821
IMPACT_I1	519.183	16.278	3.189	0.001	20.013	83.824	IS_BILGISI_BIY	-182.579	44.968	-0.435	0.664	-100.515	63.999
IMPACT_NCA	252.520	18.489	1.366	0.172	-10.987	61.491	IS_BILGISI_Consultant	-92.548	39.860	-0.232	0.816	-87.381	68.871
IMPACT_OCWW	659.721	12.105	5.450	0.000	42.246	89.698	IS_BILGISI_DU	2.007.261	63.014	3.185	0.001	77.217	324.235
IMPACT_SPSP	252.143	9.700	2.599	0.009	6.202	44.226	IS_BILGISI_Danişman	-364.473	45.141	-0.807	0.419	-124.925	52.030
ISSUE_CATEGORY_AP	-24.818	23.137	-1.07	0.915	-47.830	42.867	IS_BILGISI_IS2	53.188	39.332	0.135	0.892	-71.773	82.411
ISSUE_CATEGORY_AR	245.116	25.860	0.948	0.343	-26.175	75.198	IS_BILGISI_ISD	60.753	39.517	0.154	0.878	-71.378	83.529
ISSUE_CATEGORY_BI	-798.816	55.688	-1.434	0.151	-189.032	29.269	IS_BILGISI_Junior	-172.307	39.568	-0.435	0.663	-94.785	60.324
ISSUE_CATEGORY_CE	4.459.041	140.326	3.178	0.001	170.862	720.946	IS_BILGISI_KD	-6.798.101	356.890	-1.905	0.057	-1.379.321	19.701
ISSUE_CATEGORY_CST	628.408	47.624	1.320	0.187	-30.503	156.185	IS_BILGISI_Müdür	-273.108	253.948	-0.108	0.914	-525.053	470.431
ISSUE_CATEGORY_Custom	143.265	23.131	0.619	0.536	-31.011	59.664	IS_BILGISI_Partner	147.473	42.809	0.344	0.730	-69.158	98.653
ISSUE_CATEGORY_Database	-428.091	24.977	-1.714	0.087	-91.765	6.147	IS_BILGISI_Principal	4.731.995	503.157	0.940	0.347	-512.996	1.459.395
ISSUE_CATEGORY_Development	-2.228.992	31.378	-7.104	0.000	-284.400	-161.399	IS_BILGISI_SUPC	1.493.105	102.654	1.455	0.146	-51.893	350.514
ISSUE_CATEGORY_EAM	-42.952	33.791	-0.127	0.899	-70.527	61.936	IS_BILGISI_SY	586.655	39.561	1.483	0.138	-18.875	136.206
ISSUE_CATEGORY_FA	523.978	29.420	1.781	0.075	-5.266	110.062	IS_BILGISI_Senior	91.607	39.863	0.230	0.818	-68.972	87.293
ISSUE_CATEGORY_FAH	2.702.357	258.644	1.045	0.296	-236.711	777.182	IS_BILGISI_UZY	-80.007	41.051	-0.195	0.845	-88.462	72.460
ISSUE_CATEGORY_GL	475.042	23.890	1.988	0.047	0.679	94.329	IS_BILGISI_Uzman	1.339.403	99.567	1.345	0.179	-61.214	329.094
ISSUE_CATEGORY_GRC	893.393	365.407	0.244	0.807	-626.864	805.543	IS_BILGISI_Yönetici	-539.267	67.779	-0.796	0.426	-186.774	78.921
ISSUE_CATEGORY_HR	1.310.836	23.211	5.647	0.000	85.590	176.578	ISSUE_TYPE_ChangeRequest	-264.431	10.036	-2.635	0.008	-46.113	-6.773
ISSUE_CATEGORY_Hyperion	-5.930.275	211.631	-2.802	0.005	-1.007.828	-178.227	ISSUE_TYPE_Incident	39.533	17.168	0.230	0.818	-29.697	37.603
ISSUE_CATEGORY_INV	262.815	23.800	1.104	0.269	-20.367	72.930	ISSUE_TYPE_Others	2.811.207	13.729	20.477	0.000	254.212	308.029
ISSUE_CATEGORY_IPROC	187.429	139.665	0.134	0.893	-255.003	292.489	ISSUE_TYPE_Proactive	-545.683	11.530	-4.733	0.000	-77.167	-31.969
ISSUE_CATEGORY_IS1	213.409	25.708	0.830	0.406	-29.047	71.728	ISSUE_TYPE_Problem	250.827	14.633	1.714	0.087	-3.598	53.764
ISSUE_CATEGORY_ISUPPLIER	-292.476	47.796	-0.612	0.541	-122.929	64.434	ISSUE_TYPE_Task	-249.781	11.034	-2.264	0.024	-46.605	-3.352

We examine the results obtained for the OLS regression by applying L1 and L2 regularization methods. The results of the regularized model are compared. The purpose of regularization models is to add a penalty term to Linear Regression to prevent the coefficients from growing too large. But they work a little differently; Ridge penalizes high coefficient values but does not force them to zero. Lasso forces as many coefficients as possible to zero.

In the Ridge method model, the model is built separately with all combinations for the hyperparameters and their values to be tested and the most successful hyperparameter set is determined according to the specified metric. Grid-SearchCV method is used for this in the model. Alpha parameter values for GridSearchCV are 0.0001, 0.001, 0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0, 9.0, 10.0, 20, 50, 100, 500, 1000. As a result of the experiments, the most appropriate param value was found to be 6 and cross-validation was defined as 10-fold.

R2 Score (Train): 0.75054674060912334

R2 Score (Test): 0.74791859089008832

RMSE (Train): 0.083440

RMSE (Test): 0.084280

In Lasso regularization, like Ridge, the initial value of the alpha parameter is given as 0.0001. An important point at this stage is that 10-fold fitting is performed for each of the 28 candidates, for a total of 280 fits.

R2 Score (Train): 0.759693391808405374

R2 Score (Test): 0.748265086687702795

RMSE (Train): 0.082810

RMSE (Test): 0.082721

Table 9 is provided to evaluate the results of the Ridge and Lasso methods collectively. Since both models have the same R2 score around 0.75 and RMSE score around 0.08, it is better to choose the simpler model. In this respect, the Lasso model does a better job as it does feature selection resulting in 45 features while the Ridge model has 76 features which is 31 features more than the Lasso model.

Table 9. Ridge and Lasso results

	Metric	Ridge	Lasso
0	R2 Score (Train)	0.750547	0.759693
1	R2 Score (Test)	0.747919	0.748265
2	RMSE (Train)	0.083440	0.082810
3	RMSE (Test)	0.084280	0.082721

The mean RMSE obtained in the regression analysis before regularization was calculated as 0.126. After regularization, the RMSE error metric is around 0.08. It is stated that the Lasso, i.e. L1 regularization obtained a more successful result. Therefore, Lasso has actually done Feature Elimination (Variable Reduction). One of the most common uses of Lasso is this Feature Elimination process.

Also, all the above features have a positive correlation with the ticket's completion time data. A zero coefficient indicates that the variable indicated by that coefficient is insignificant for the outcome. Therefore, the model is simplified by reducing the number of variables. Simplicity is a good thing for Machine Learning. Variance is reduced, Bias is increased and better generalization is possible.

A new situation has emerged by examining the data set with regularization methods. It is seen that some attributes have no effect on the model and higher prediction success can be achieved by excluding these attributes from the model. At this stage, it is seen that the dataset to be used in the classification experiments will now be corrected and more reliable.

A common approach to running classification algorithms is to use an exploratory data analysis approach. Exploratory Data Analysis (EDA) is an approach to summarizing data by taking its key features and visualizing them with appropriate representations (Sahoo et al., 2019).

One of the issues analyzed in the problem is the classification of the type of customer demand by using a classification method. This information is defined in the dataset by an attribute called ISSUE_TYPE. Therefore, the dependent parameter labeled in classification algorithms is categorical. LR, SVM and ANN classifier algorithms are generally

used in models where categorical dependent variables are class labels. Figure 13 shows the initialization screen of the LR algorithm for Python coding. The mathematical flow of the algorithm is coded by calling the suitable libraries.

```
[2]: import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as seabornInstance
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression # import Logistic regression
from sklearn import metrics

%matplotlib inline

[3]: dataset = pd.read_excel('omdSVM.xlsx')
dataset.describe()
```

	Completion_Time	Total_Assignee	Total_Worklog_Assignee	Total_Log_Hours_Assignee	COMMENTOR_COUNT	CO
count	64995.000000	64995.000000	64995.000000	64995.000000	64995.000000	
mean	556.486453	1.516809	0.691761	3.683068	1.985122	
std	546.193641	1.394241	0.725491	69.228870	1.485799	
min	0.000000	0.000000	0.000000	0.000000	0.000000	
25%	53.000000	0.000000	0.000000	0.000000	1.000000	
50%	281.000000	1.000000	1.000000	1.000000	2.000000	
75%	1160.000000	3.000000	1.000000	2.000000	3.000000	
max	1440.000000	5.000000	5.000000	13380.000000	22.000000	

Figure 13. Python Coding Screenshot for Logistic Regression Algorithm

During the preparation of the data set for classification algorithms according to the EDA approach, various library and function updates are encountered depending on the software language used in the coding phase. As can be seen in Figure 14, one of these updates was encountered during the histogram graph display phase. *histplot* function is used instead of *distplot* function.

```
plt.figure(figsize=(15,10))
plt.tight_layout()
sns.distplot(dataset['ISSUE_TYPE'])
```

C:\Users\ali.alsac\AppData\Local\Temp\ipykernel_22852\2263707846.py:3: UserWarning:
`distplot` is a deprecated function and will be removed in seaborn v0.14.0.
Please adapt your code to use either `displot` (a figure-level function with similar flexibility) or `histplot` (an axes-level function for histograms).
For a guide to updating your code to use the new functions, please see
<https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751>

```
sns.distplot(dataset['ISSUE_TYPE'])
```

```
plt.figure(figsize=(15,10))
plt.tight_layout()
sns.histplot(dataset["ISSUE_TYPE"], kde=True)
```

<Axes: xlabel='ISSUE_TYPE', ylabel='Count'>

Figure 14. Histogram graph plot display update warning

The code block of the ROC curve of the LR algorithm is given in Figure 15.

```

# Plot the ROC curve
plt.figure(figsize=(8, 6))
plt.plot(fpr, tpr, color='darkorange', lw=2, label=f'AUC = {roc_auc:.4f}')
plt.plot([0, 1], [0, 1], color='navy', lw=2, linestyle='--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Receiver Operating Characteristic (ROC) Curve - Logistic Regression')
plt.legend(loc='lower right')
plt.show()

```

Figure 15. The code block of the ROC curve of the LR algorithm

As mentioned in the Evaluation Metrics section, ROC analysis is used to decide whether a test whose results are evaluated for predictive accuracy is useless or a very good test. According to the graph shared in Figure 16, the tests show that the model's approach to generating problem-specific solutions is positive. The area under the ROC curve, i.e. the AUC ratio, was calculated as 0.8622. It is a ratio that shows that the LR algorithm model is good at separating classes from each other.



Figure 16. LR ROC Curve

The code block given in Figure 17 was used to display all the classification algorithm results in a table. With this code, Model, Accuracy, Precision, Recall and F1 Score results can be seen in Table 10.

```

predictions_base = pd.concat(predictions, ignore_index=True, sort=False)
predictions_base = predictions_base.sort_values(by=['Recall'], ascending=False).style.hide(axis='index')
predictions_base

```

Figure 17. Coding block for Prediction Results

The data set used in the application consists of real life data. As seen in the data preparation section, the data is not balanced, that is, homogeneously distributed. Therefore, only accuracy is not used in the evaluation.

As seen in Table 10, SVM algorithm has the highest prediction success among the classification algorithms. LR and ANN algorithms gave very close results. In fact, all three algorithms have close values. Based on the principles mentioned in the evaluation metrics section, it is seen that the SVM algorithm has the best result when the accuracy is high and the F1 score is low. While creating the ANN architecture, the architectural structure of the model has a great impact on the accuracy rates. In this study, the architecture established for the ANN (single hidden layer with

Table 10. Prediction Results

Model	Accuracy	Precision	Recall	F1 Score
LR	0.728881	0.144476	0.761194	0.242857
SVM	0.732825	0.194476	0.732172	0.182833
ANN	0.724444	0.144111	0.771484	0.246841

ten neurons and 10000 iterations, sigmoid function as activation function and ten repetitions) was the model that gave accuracy rates of 70%. Optimizing the parameters of the ANN architecture also takes days.

In addition, compared to ANN, SVM method, after determining the kernel function (the most used kernel function is the radial basis function), results can be obtained in a single move. In this point of view, it is a much easier method to implement compared to ANN.

In this study, the ANN and SVM algorithms produced better results than the Logistic Regression algorithm. However, it should be considered that methods such as logistic regression can get results very quickly using a simple background, while methods such as ANNs are difficult to mature their architecture and reach iteration numbers such as 50000-100000 on the computer.

6. CONCLUSIONS AND FUTURE WORK

ITSM services have a very important place in the modernization processes of organizations. ITSM enables teams within organizations to produce value faster as software-centric services increase. It is now a necessity to add Artificial Intelligence-supported models to the systems used for collaboration, ease of use and a faster, quality value-producing ITSM service. In our study, data sets obtained through Jira tools used for ITSM services were used. It is imperative to improve existing systems and utilize up-to-date techniques and technologies in order to assign customer requests to the most accurate expert and to provide the most appropriate service in accordance with the service conditions included in the SLA agreements between the service providers and service recipients.

As a result of the experimental investigations, it has been revealed that the regularization approach has a positive effect on improving model performance as an important elimination tool in feature engineering. The most important problem encountered especially in the analysis of *real life* datasets is the lack of understanding of the suitability of the features to the designed machine learning models. In this study, the examination of the dataset with $L1$ and $L2$ regularization methods and the resulting regularization of the dataset provided a more suitable dataset for the next stage, classification.

This study shows that the classification method can be used to learn the problem type of customer requests. It is observed that ANN, SVM and LR algorithms are suitable algorithms for classification. As a result of this study, it is concluded that ITSM companies should not delay in establishing their own neural networks and quickly incorporating artificial intelligence into their business processes.

In summary, using classification and regularization methods in demand type (Issue-Type) forecasting allows for more accurate, interpretable, and robust models. This approach leverages the strengths of both techniques, resulting in better predictions and more informed decision-making.

In future studies, it is expected to make a positive contribution to the literature by understanding that regularization is an advantageous method, especially in examining data sets consisting of real-life data according to the feature engineering approach, and that classification studies carried out on data sets arranged according to the logic of regularization can give better results.

Peer Review: Externally peer-reviewed.

Author Contributions: Conception/Design of Study- M.M.Y., M.C.G., M.D., A.A., T.U.; Data Acquisition- T.U., A.A., M.C.G.; Data Analysis/Interpretation- M.M.Y., M.C.G., M.D., A.A., T.U.; Drafting Manuscript- A.A., T.U.; Critical Revision of Manuscript- M.M.Y., M.C.G.; Final Approval and Accountability- M.M.Y., M.C.G., M.D., A.A., T.U.; Material and Technical Support- M.D., A.A., T.U.; Supervision- M.M.Y., M.C.G.

Conflict of Interest: The authors have no conflict of interest to declare.

Grant Support: This work is supported by TUBITAK 1509 (program number 9210017) and TUBITAK 2244 (program number 119C056).

Acknowledgements: I would like to greatly acknowledge Experteam which is a trademark of Uzman Bilişim A.Ş.

ORCID IDs of the authors / Yazarların ORCID ID'leri

Ali Alsaç	0000-0002-8585-4501
Mehmet Mutlu Yenisey	0000-0002-4532-344X
Murat Can Ganiz	0000-0001-8338-991X
Mustafa Dağtekin	0000-0002-0797-9392
Taner Ulusinan	0009-0000-3647-0408

REFERENCES

- ALAN, A., & KARABATAK, M. (2020). Veri Seti - Sınıflandırma İlişkisinde Performansa Etki Eden Faktörlerin Değerlendirilmesi. *Fırat Üniversitesi Mühendislik Bilimleri Dergisi*, 32(2). <https://doi.org/10.35234/fumbd.738007>
- Anderson D, M. G. (1992). Artificial Neural Networks Technology. *Kaman Sciences Corporation*, 258(6).
- Aran, O., Yildiz, O. T., & Alpaydin, E. (2009). An incremental framework based on cross-validation for estimating the architecture of a multilayer perceptron. *International Journal of Pattern Recognition and Artificial Intelligence*, 23(2). <https://doi.org/10.1142/S0218001409007132>
- ARSLAN, H., ÜNEŞ, F., DEMİRCİ, M., TAŞAR, B., & YILMAZ, A. (2020). Keban Baraj Gölü Seviye Değişiminin ANFIS ve Destek Vektör Makineleri ile Tahmini. *Osmaniye Korkut Ata Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, 3(2). <https://doi.org/10.47495/okufbed.748018>
- Bharambe, Prof. P., Bagul, B., Dandekar, S., & Ingle, P. (2022). Used Car Price Prediction using Different Machine Learning Algorithms. *International Journal for Research in Applied Science and Engineering Technology*, 10(4). <https://doi.org/10.22214/ijrasnet.2022.41300>
- Bhattacharya, P., Neamtiu, I., & Shelton, C. R. (2012). Automated, highly-accurate, bug assignment using machine learning and tossing graphs. *Journal of Systems and Software*, 85(10). <https://doi.org/10.1016/j.jss.2012.04.053>
- ÇELİK, E., DAL, D., & AYDİN, T. (2021). Duygu Analizi İçin Veri Madenciliği Sınıflandırma Algoritmalarının Karşılaştırılması. *European Journal of Science and Technology*. <https://doi.org/10.31590/ejosat.905259>
- Cook, D., Dixon, P., Duckworth, W. M., Kaiser, M. S., Koehler, K., Meeker, W. Q., & Stephenson, W. R. (2001). Binary Response and Logistic Regression Analysis. *Project Beyond Traditional Statistical Methods, ML*.
- Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning*, 20(3). <https://doi.org/10.1023/A:1022627411411>
- Dangeti, P. (2017). Statistics for Machine Learning: Techniques for exploring supervised, unsupervised, and reinforcement learning models with Python and R. In *Packt Publishing*.
- Deloitte, & TUBISAD. (2022). *Bilgi ve İletişim Teknolojileri Sektörü 2021 Pazar Verileri*.
- Doğan, C. (2021). *İstatistiksel ve Makine Öğrenme ile Derin Sinir Ağlarında Hiper-Parametre Seçimi İçin Melez Yaklaşım* [Yüksek Lisans]. Hacettepe Üniversitesi.
- Domingos, P. (2000). A Unified Bias-Variance Decomposition. *Aaai/Iaai*.
- Emmert-Streib, F., & Dehmer, M. (2019). High-Dimensional LASSO-Based Computational Regression Models: Regularization, Shrinkage, and Selection. In *Machine Learning and Knowledge Extraction* (Vol. 1, Issue 1). <https://doi.org/10.3390/make1010021>
- Friedrich, S., Groll, A., Ickstadt, K., Kneib, T., Pauly, M., Rahnenführer, J., & Friede, T. (2023). Regularization approaches in clinical biostatistics: A review of methods and their applications. In *Statistical Methods in Medical Research* (Vol. 32, Issue 2). <https://doi.org/10.1177/09622802221133557>
- Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural Networks and the Bias/Variance Dilemma. *Neural Computation*, 4(1). <https://doi.org/10.1162/neco.1992.4.1.1>
- Golam Kibria, B. M., & Banik, S. (2016). Some ridge regression estimators and their performances. *Journal of Modern Applied Statistical Methods*, 15(1). <https://doi.org/10.22237/jmasm/1462075860>
- Goldberg, N., & Eckstein, J. (2012). Sparse weighted voting classifier selection and its linear programming relaxations. *Information Processing Letters*, 112(12). <https://doi.org/10.1016/j.ipl.2012.03.004>
- Ha, J., Kambe, M., & Pe, J. (2011). *Data Mining: Concepts and Techniques*. In *Data Mining: Concepts and Techniques*. <https://doi.org/10.1016/C2009-0-61819-5>
- Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). Multivariate Data Analysis. In *Vectors*. <https://doi.org/10.1016/j.ijpharm.2011.02.019>
- Hautamaki, V., Kinnunen, T., Sedlak, F., Lee, K. A., Ma, B., & Li, H. (2013). Sparse classifier fusion for speaker verification. *IEEE Transactions on Audio, Speech and Language Processing*, 21(8). <https://doi.org/10.1109/TASL.2013.2256895>
- Helming, J., Arndt, H., Hodaie, Z., Koegel, M., & Narayan, N. (2011). Automatic Assignment of Work Items. *Communications in Computer and Information Science*, 230. https://doi.org/10.1007/978-3-642-23391-3_17
- Jonsson, L., Borg, M., Broman, D., Sandahl, K., Eldh, S., & Runeson, P. (2016). Automated bug assignment: Ensemble-based machine learning in large scale industrial contexts. *Empirical Software Engineering*, 21(4). <https://doi.org/10.1007/s10664-015-9401-9>
- Koçoğlu, F. Ö., & Esnaf, Ş. (2022). Machine Learning Approach and Model Performance Evaluation for Tele-Marketing Success Classification. *International Journal of Business Analytics*, 9(5). <https://doi.org/10.4018/ijban.298014>
- Koçoğlu, F. Ö., & Özcan, T. (2022). A grid search optimized extreme learning machine approach for customer churn prediction. *Journal of Engineering Research*.
- Kotsilieris, T., Anagnostopoulos, I., & Livieris, I. E. (2022). Special Issue: Regularization Techniques for Machine Learning and Their Applications. In *Electronics (Switzerland)* (Vol. 11, Issue 4). <https://doi.org/10.3390/electronics11040521>

- Li, N., & Zhou, Z. H. (2009). Selective ensemble under regularization framework. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 5519 LNCS. https://doi.org/10.1007/978-3-642-02326-2_30
- Mantovani, R. G., Horvath, T., Cerri, R., Vanschoren, J., & De Carvalho, A. C. P. L. F. (2017). Hyper-Parameter Tuning of a Decision Tree Induction Algorithm. *Proceedings - 2016 5th Brazilian Conference on Intelligent Systems, BRACIS 2016*. <https://doi.org/10.1109/BRACIS.2016.018>
- Mao, S., Xiong, L., Jiao, L. C., Zhang, S., & Chen, B. (2013). Weighted ensemble based on 0-1 matrix decomposition. *Electronics Letters*, 49(2). <https://doi.org/10.1049/el.2012.3528>
- Muller, A. C., & Guido, S. (2017). Introduction to Machine Learning with Python: a guide for data scientist. In *O'Reilly Media, Inc.*
- Orynassar, A., Sapazhanov, Y., Kadyrov, S., & Lyublinskaya, I. (2022). Application of ROC Curve Analysis for Predicting Students' Passing Grade in a Course Based on Prerequisite Grades. *Mathematics*, 10(12). <https://doi.org/10.3390/math10122084>
- ÖZBİLGİN, F., & KURNAZ, Ç. (2023). Koroner Arter Hastalığının İris Görüntülerinden Yerel İkili Örüntüler ve Yapay Sinir Ağı Kullanılarak Tahmini. *Karadeniz Fen Bilimleri Dergisi*, 13(2). <https://doi.org/10.31466/kfbd.1266996>
- Özgür, A., Nar, F., & Erdem, H. (2018). Sparsity-driven weighted ensemble classifier. *International Journal of Computational Intelligence Systems*, 11(1). <https://doi.org/10.2991/ijcis.11.1.73>
- Paper, D. (2019). Hands-on Scikit-Learn for Machine Learning Applications: Data Science Fundamentals with Python. In *Hands-on Scikit-Learn for Machine Learning Applications: Data Science Fundamentals with Python*. <https://doi.org/10.1007/978-1-4842-5373-1>
- Sahoo, K., Samal, A. K., Pramanik, J., & Pani, S. K. (2019). Exploratory data analysis using python. *International Journal of Innovative Technology and Exploring Engineering*, 8(12), 4727–4735. <https://doi.org/10.35940/ijitee.L3591.1081219>
- Şen, M. U., & Erdogan, H. (2013). Linear classifier combination and selection using group sparse regularization and hinge loss. *Pattern Recognition Letters*, 34(3). <https://doi.org/10.1016/j.patrec.2012.10.008>
- ŞENEL, S., & ALATLI, B. (2014). Lojistik Regresyon Analizinin Kullanıldığı Makaleler Üzerine Bir İnceleme. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, 5(1). <https://doi.org/10.21031/epod.67169>
- Sinha, K., Uddin, Z., Kawsar, H. I., Islam, S., Deen, M. J., & Howlader, M. M. R. (2023). Analyzing chronic disease biomarkers using electrochemical sensors and artificial neural networks. In *TrAC - Trends in Analytical Chemistry* (Vol. 158). <https://doi.org/10.1016/j.trac.2022.116861>
- Şipal, B., Ormancı, B. B., & Altınel, A. B. (2022). KELİME ANLAM BULANIKLIĞINI GİDERMEK İÇİN DİFÜZYON REGÜLARİZASYON VE NORMALİZASYON TEKNİKLERİNİN KULLANILMASI. In *MÜHENDİSLİK ALANINDA ULUSLARARASI ARAŞTIRMALAR VI* (pp. 75–85).
- Tanyıldızı, E., & Demirtas, F. (2019). Hiper Parametre Optimizasyonu Hyper Parameter Optimization. *1st International Informatics and Software Engineering Conference: Innovative Technologies for Digital Transformation, IISEC 2019 - Proceedings*. <https://doi.org/10.1109/UBMYK48245.2019.8965609>
- TAZEGÜL, A., YAZARKAN, H., & YERDELEN, C. (2016). İşletmelerin Finansal Başarılı ve Başarısız Olma Durumlarının Veri Madenciliği ve Lojistik Regresyon Analizi İle Tahmin Edilebilirliği. *Ege Akademik Bakis (Ege Academic Review)*, 16(1). <https://doi.org/10.21121/eab.2016119960>
- Tian, Y., & Zhang, Y. (2022). A comprehensive survey on regularization strategies in machine learning. In *Information Fusion* (Vol. 80). <https://doi.org/10.1016/j.inffus.2021.11.005>
- Tinoco, S. L. J. L., Santos, H. G., Menotti, D., Santos, A. B., & Dos Santos, J. A. (2013). Ensemble of classifiers for remote sensed hyperspectral land cover analysis: An approach based on Linear Programming and Weighted Linear Combination. *International Geoscience and Remote Sensing Symposium (IGARSS)*. <https://doi.org/10.1109/IGARSS.2013.6723730>
- Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). Data Mining: Practical Machine Learning Tools and Techniques. In *Data Mining: Practical Machine Learning Tools and Techniques*.
- YENİSU, E. (2021). Ekonomiye Harekete Geçiren Kilit Sektörler Nelerdir? Türkiye Üzerine Bir Girdi-Çıktı Analizi. *İzmir İktisat Dergisi*, 36(4). <https://doi.org/10.24988/ije.721302>
- YETGINLER, B., & ATACAK, İ. (2020). Sentiment Analyses on Movie Reviews using Machine Learning-Based Methods. *Artificial Intelligence Studies*, 3(2). <https://doi.org/10.30855/ais.2020.03.02.01>
- Yildiz, M., Alsac, A., Ulusinan, T., Ganiz, M. C., & Yenisey, M. M. (2022). IT Support Ticket Completion Time Prediction. *Proceedings - 7th International Conference on Computer Science and Engineering, UBMK 2022*. <https://doi.org/10.1109/UBMK55850.2022.9919591>
- Yin, X. C., Huang, K., Hao, H. W., Iqbal, K., & Wang, Z. Bin. (2012). Classifier ensemble using a heuristic learning with sparsity and diversity. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 7664 LNCS(PART 2). https://doi.org/10.1007/978-3-642-34481-7_13
- Yoon, B. L. (1989). Artificial neural network technology. *ACM SIGSMALL/PC Notes*, 15(3), 3–16. <https://doi.org/10.1145/74657.74658>
- Zhang, L., & Zhou, W. Da. (2011). Sparse ensembles using weighted combination methods based on linear programming. *Pattern Recognition*, 44(1). <https://doi.org/10.1016/j.patcog.2010.07.021>
- Zibran, M. F. (2016). On the effectiveness of labeled latent dirichlet allocation in automatic bug-report categorization. *Proceedings - International Conference on Software Engineering*. <https://doi.org/10.1145/2889160.2892646>

How cite this article

Alsac, A., Yenisey, M.M., Ganiz, M.C., Dagtekin, M, Ulusinan, T. (2023). The efficiency of regularization method on model success in issue type prediction problem. *Acta Infologica*, 7(2), 360-383. <https://doi.org/10.26650/acin.1394019>

The Future of Smart Campuses: Combining Digital Twin and Green Metrics

Akıllı Kampüslerin Geleceği: Dijital İkiz ve Yeşil Metriklerin Birleştirilmesi

İlknur Teke¹ , Orkun Teke² , Murat Kılıncı¹ 

¹(Lec. Dr.) Manisa Celal Bayar University, Computer Research and Application Center, Manisa, Türkiye
²(Lec. Dr.) Manisa Celal Bayar University, Manisa Vocational School of Technical Sciences, Manisa, Türkiye

Corresponding author : İlknur Teke
E-mail : ilknur.teke@cbu.edu.tr

*The paper was presented at the 10th International Management Information Systems Conference, 18-20 October 2023, Istanbul, Türkiye

ABSTRACT

The “Smart Campus” concept combines environmental sustainability and technological innovation and plays an important role in educational institutions. In this context the contribution of “Digital Twins” and “Green metrics (GM)” to smart campus is one of the important areas of research. In this study, a digital twin architecture is proposed, including energy and climate change, waste and water issues, which account for 48% in GM criteria. Digital simulation and communication protocols, predictive analysis and dynamic decision support can be synthesized between the physical world and the sensor-based framework. This synthesis reveals the potential to reduce environmental damage through effective waste management, efficient use of water resources, identification of efficiency gaps through real-time analysis of energy consumption, and reduction of carbon footprint through energy savings. This study aims that the combined approach presented with the proposed framework according to GM criteria will contribute to future educational environments by ensuring smart campus sustainability.

Keywords: Digital twin, green metrics, smart campus

ÖZ

Çevresel sürdürülebilirlik ile teknolojik yeniliğin birleşimini temsil eden “Akıllı Kampüs” kavramı, geleceğin eğitim kurumlarında önemli bir rol oynamaktadır. Bu açıdan, “Dijital İkiz” teknolojisinin ve “Green Metrics (GM)” ölçümlerinin akıllı kampüslere nasıl katkı sağlayacağı önemli araştırma alanlarından biridir. Bu çalışma kapsamında, GM kriterleri arasında etki değeri %48 olan enerji ve iklim değişikliği, atık ve su konuları dikkate alınarak bir dijital ikiz mimarisi önerilmiştir. Bu kriterler kullanılarak oluşturulan sensör tabanlı çerçeve ile dijital simülasyon ve fiziksel dünya arası iletişim protokolü, tahmine dayalı analiz ve dinamik karar desteği sentezlenebilmektedir. Bu sentez ile kampüslerdeki atıkların etkili bir şekilde yönetilerek çevresel zararın azaltılması, su kaynaklarının verimli kullanımı, enerji tüketiminin gerçek zamanlı analizi ile verimlilik açıklarının belirlenmesi, enerji tasarrufu sağlayarak karbon ayak izinin azaltılması potansiyelleri ortaya çıkmaktadır. Bu çalışmada GM kriterleri doğrultusunda önerilen çerçeve ile birlikte sunulan birleşik yaklaşımla beraber, akıllı kampüslerin sürdürülebilir olması sağlanarak geleceğin eğitim ortamlarına katkıda sağlanması amaçlanmaktadır.

Anahtar Kelimeler: Akıllı kampüs, dijital ikiz, yeşil metrikler

Submitted : 04.11.2023
Revision Requested : 08.12.2023
Last Revision Received : 10.12.2023
Accepted : 11.12.2023
Published Online : 21.12.2023



This article is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0)

1. INTRODUCTION

In recent years of rapid technological developments, the concept of "Digital Twin", which is created as a copy of the physical world, has attracted great attention in various sectors from manufacturing to urban planning. Digital twin technology, which aims to improve business processes and increase efficiency, has gained importance primarily for organizations operating in areas where test and defective product costs are quite high by using real-time data from the physical world (Kumaş & Erol, 2021). However, digital twin technology, which is used to support digital transformation and decision-making processes in many sectors and based on data for desired results, is a concept that expands and continues to develop according to the areas in which it is used (VanDerHorn & Mahadevan, 2021). Three features stand out for the success of the digital twin concept, which creates a virtual mirror by modeling the behavior of the physical world. These are the design of the digital twin as a dynamic representation of the physical world, the bidirectional data flow between them, and the connection of the digital twin to include all processes in the physical world (Trauer, Schweigert-Recksiek, Engel, Spreitzer & Zimmermann, 2020). The conceptual diagram for digital twin technology is presented in Figure 1.

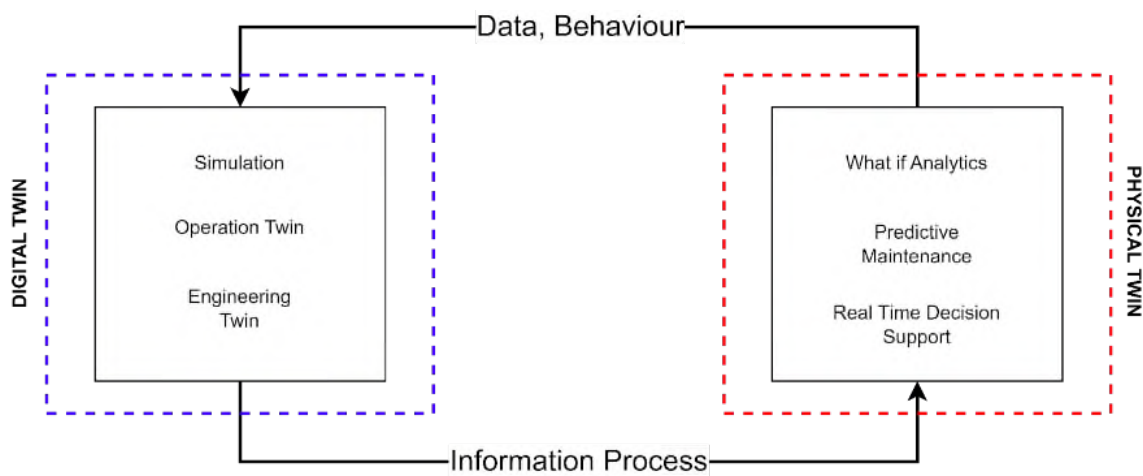


Figure 1. The Conceptual Diagram for Digital Twin

The Internet of Things (IoT), which is among the technologies that have gained importance with the Industry 4.0 revolution, is considered as a dynamic structure that connects the physical world and the digital world using any network or service (Georgios, Kerstin & Theofylaktos, 2019). Today, the concept of IoT has an important place in the design and acquisition of smart structures designed with data flow from physical structures. IoT-based technologies can be used to create smart campuses aimed at improving service quality and improving campus business operations. In this way, an infrastructure can be created for the presentation of applications in which variables such as people, spaces, vehicles, etc. in the campus ecosystem are in connection with each other (Abuarqoub et al., 2017).

The smart campus concept, which has an important place within the scope of digitalization in educational institutions, can be designed not only to improve campus activities, but also to optimize corporate business processes and prioritize sustainability. The driving force behind this formation is the adoption of "GreenMetric" (GM) measurements. GreenMetric is a sustainability ranking launched by the University of Indonesia in 2010 to see the green space rating on campuses using 6 criteria and 39 indicators. These criteria are indicated in Figure 2 as Setting and Infrastructure (SI), Energy and Climate Change (EC), Waste (WS), Water (WR), Transport (TR) and Education and Research (ED) (UI GreenMetric, 2023). These measurements are used to assess the environmental impact of campus activities, from energy consumption and waste generation to transport modes and water use.

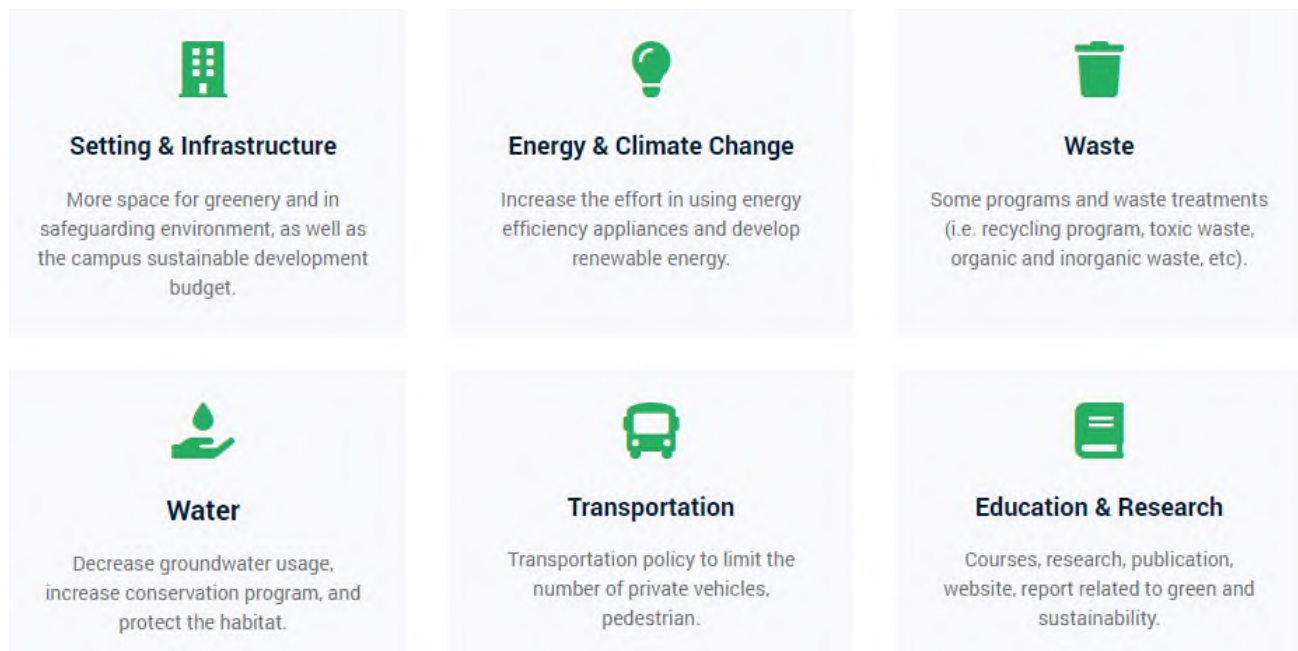


Figure 2. GreenMetric Criteria

The conceptual framework presented within the scope of digital twin in this study is based on the application of digital twin technology to the smart campus concept by blending it with GreenMetric criteria. The study focused on three GM criteria: energy and climate change, waste and water. The digital twin concept, created with the data obtained from the IoT-based smart campus structure, has been prepared to obtain real-time information about the GM criteria and to support the decision-making processes of this information. In addition, the other aim of the study is to estimate the sustainability plan of the University by combining the digital twin concept and GM measurements, which are created using the data flow from the smart campus system. Using smart campus formation and digital twin and GreenMetric criteria, this study is expected to contribute the way in promoting a culture of sustainability through campus communities, reducing ecological footprints and encouraging "eco-conscious" behaviors.

2. LITERATURE REVIEW

Within the scope of GreenMetric (GM), digital twin and smart campuses, many studies are found in the literature. In particular, GM evaluation criteria, which are used to understand and develop the sustainability potential of smart campuses, can be used in smart campuses since they can produce predictive simulations when considered together with digital twin architectures. In this direction, GM criteria such as energy and climate change, waste and water categories were evaluated and discussed for the digital twin architecture created in the smart campus by conducting literature research.

First of all, in the study of Corrado, DeLong, Holt, Hua & Tolk, in 2022, a model proposal is presented by examining the existing research in the field of GreenMetric and digital twin on smart cities. The model proposal, which evaluates a city as a sociotechnical complex system based on current research, enables sustainability planning at various levels related to city planning and governance and provides computational decision support for complex challenges (Corrado et al., 2022). In another study by Pexyeon, Saraubon, and Nilsook in 2022, the use of digital twins obtained with IoT output data is discussed in order to better understand the energy management potential in campuses. One of the highlights of the study is the emphasis on new energy models and management that can solve future emergencies by externally controlling the virtual energy data obtained with the digital twin in smart campuses (Pexyeon et al., 2022). Considered from this aspect, the architecture proposed in this study is similar to the results obtained by the authors in this direction.

Han et al. in 2022, it is recommended to use digital twin technology to digitally create the physical campus environment at the university, detect the physical campus in real time, accurately map the virtual campus to the virtual campus, and provide reverse control of the twin virtual campus to the physical campus. The results obtained within the scope of the study show that the virtual-real campus system can improve school management and teaching, and important implications can be obtained by promoting campus smart systems within the scope of the implementation

of development processes (Han et al., 2022). Wang's research in 2022 highlights that traditional campus management faces problems such as cost, maintenance, low efficiency and energy waste. With the findings obtained in this direction, a control cloud platform was created in order to reduce costs, reduce energy consumption, and improve the use of assets and equipment by using smart IoT research, and a dynamic campus management was provided (Wang, 2022).

In another study conducted by Suwartha and Sari in 2013, the ranking and application results of GreenMetric, which provides the development of the current situation and policies regarding green campus and sustainability in universities all over the world, were evaluated. In the study, it is emphasized that the most important criteria achieved by many universities are energy and climate change (Suwartha & Sari, 2013). With this aspect, in the study carried out on the axis of digital twin, smart campuses and GreenMetric; energy and climate change, waste and water criteria were discussed, and the data set and digital twin architecture design were carried out based on the relevant criteria.

Celebi et al. in 2020, campuses defined as a component of cities were discussed and a study was conducted in the example of Aksaray University. In the study, which was carried out using the "Smart City Circle" proposed by Boyd Cohen, sustainable environment-oriented approaches were discussed through university examples (Celebi et al., 2020). Finally, Zaballos et al. in 2020, a smart campus concept is proposed to explore the integration of building information modeling tools with IoT-based wireless sensor networks in the fields of environmental monitoring and emotion sensing and to give an idea about the level of comfort. The preliminary results obtained highlight the importance of monitoring workspaces as it has been proven that productivity is directly affected by environmental parameters. Comfort monitoring infrastructure can also be used to monitor physical parameters in educational buildings to improve energy efficiency (Zaballos et al., 2020). For this reason, since energy and climate change are in the first place within the scope of GreenMetric criteria, campus building and facility data are also included in the data layer within the scope of architecture.

Within the model proposed within the scope of the study, machine learning techniques to be selected according to the data size will also differ. As the data size increases and the number of features increases, different and more complex machine learning methods may be needed. Three important perspectives emerge here; model simplification to reduce computational complexity, optimization approach to increase computational efficiency, and computational parallelism to increase computational capacity (Wang et al. 2020). Regarding the choice of algorithms, Sala et al. 2018, emphasizes that statistics and simple ML models are useful and sufficient for low-dimensional data sets, while for complex data and high-dimensional data sets, tree-based algorithms and advanced artificial neural networks are recommended. As for another important step, hyperparameter optimization, Gambella et al. 2018, stated that an important part of most machine learning approaches is the selection of the hyperparameters of the learning model. They emphasized that Hyperparameter Optimization is usually driven by the experience of the data scientist and the characteristics of the dataset and typically follows heuristic rules or cross-validation approaches. Furthermore, Li's research in 2017 stated that gradient descent algorithms are the most important and popular techniques for optimization of models related to deep learning. Also, that study emphasized that "adagrad" converges faster than "adam" and other optimization methods on various tasks in different neural network structures.

Looking at the literature review, it is seen that the GreenMetric evaluation criteria used to understand and develop the sustainability potential of smart campuses, when combined with digital twin architectures, have the potential to enable the development of predictive simulations in smart campuses. In this context, a conceptual framework is presented in the study and the layers in this proposal are detailed in the following sections.

3. CONCEPTUAL FRAMEWORK

A conclusion may review the main points of the paper, do not replicate the abstract as the conclusion. A conclusion might elaborate on the importance of the work or suggest applications and extensions.

In this study, a conceptual framework that combines the digital twin concept with the data obtained from the smart campus formation using IoT technology is presented. A conceptual framework is a visual or written representation of the relationships between key concepts or variables in a study. It is used to guide the research process and help the researcher make sense of the data and set of relationships. The conceptual framework approach establishes a link between the facts describing the subject under investigation and research practice (Leshem & Trafford, 2007). The conceptual framework approach, which can be specific to the research topic and present the study as a logical master plan, encompasses the theoretical framework concept, which can be summarized as the interpretation of other researchers' ideas about the study (Kivunja, 2018). Within the scope of this study a diagram representing the conceptual framework is created and presented in Figure 3.

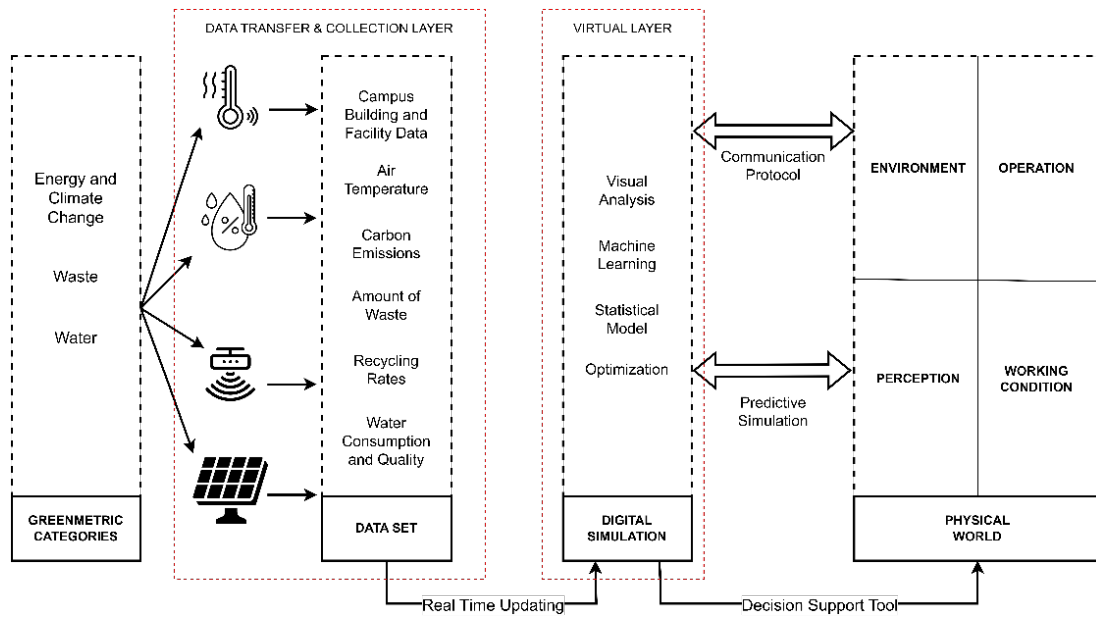


Figure 3. Conceptual Framework Design of Digital Twin for GM Criteria in Smart Campuses

In the architecture presented in Figure 3, it is aimed to integrate the data collected continuously and in real time from the physical world with various simulation models for the GM criteria. With this integration, a real-time model is obtained by creating a data-driven digital twin that represents the physical process. This aims to allow the digital twin to simulate the behavior of the physical object and predict how it performs or will perform under different conditions. Detailed explanations of the structures in the conceptual framework created are presented in the following sections.

3.1. GreenMetric Criteria

The UI GreenMetric World Universities Ranking is a ranking of universities’ environmental commitments and initiatives, launched by Universitas Indonesia in 2010. The ranking is based on 39 indicators across 6 criteria, including campus infrastructure, energy and climate change actions, water management, waste management, transport, green education and research. The distributions of the criteria, which have the degree of importance in percentages, affect the ranking in different degrees. The percentage distributions of the impacts of the criteria on the ranking are presented in Table 1 (Lauder et al., 2015; UI GreenMetric, 2023).

Table 1. Impact of GM criteria on ranking

Framework	Criteria	Weighting (%)
UI GreenMetric	Setting and Infrastructure	15
	Energy and Climate Change	21
	Waste	18
	Water	10
	Transportation	18
	Education and Research	18

Universities can carry out studies for all criteria at the same time or give priority to studies for criteria with higher importance. One of the important parameters here is to make a planning according to the criteria that can be solved in the short term and to take beneficial actions.

Sonetti et al. in the study they carried out in 2016, they conducted an analysis study by examining the GreenMetric criteria of different universities. In this study, it was revealed that the need for improvement was most intense in the 3 areas selected in the study (Sonetti et al., 2016). Demiroğlu et al., in their study in 2017, examined 5 campuses in Turkey on the basis of Green Campus approaches. In the findings of the study, it was seen that the investments in the selected campuses focused on the use of renewable energy in the campus, improving waste management, reducing the carbon footprint and using water resources more efficiently (Demiroğlu et al., 2017).

Lukman et al. in 2009, they conducted a study based on the use and functioning (construction and maintenance, heating, lighting and water consumption) and daily consumption of various items (paper and plastic bottles). In the improvements presented in this study, it has been revealed that improvements such as heating-cooling, changing energy habits, making waste management more sustainable contribute much more than construction improvements and are the right approach for campuses (Lukman et al., 2009). Finley et al., 2012 and Register, 2006 determined that the "Eco City" concept, which was developed by integrating smart energy, waste management and water management with new building construction, is the best method for university campuses (Finlay & Massey, 2012; Register, 2005).

Massuco and others. with a study carried out by Genoa University in 2023, the energy performance of the campus was improved. Various measures have been taken, including the implementation of a system for real-time monitoring of the electricity consumption of the buildings that consume the most energy. The results allowed the reduction of energy consumption and provided important and practical guidelines for energy saving, highlighting the domino effect of the changes made in this chapter (Massucco et al., 2023).

In this study, not all GreenMetric criteria are chosen based on research objectives and focus. A strategic decision-making process is carried out. Based on the findings of the literature review, "Energy and Climate Change, Waste and Water" criteria were selected for resource management and cost efficiency, easier to measure, quickly implementable, increasing public awareness and perception by appealing to large audiences, open to innovation and technology, integrating interdisciplinary features and maximizing the contribution of staff. Basically, one of the most important reasons for selecting these criteria is that they are in line with the United Nations Sustainable Development Objectives, especially Objective 7 "Affordable Clean Energy", Objective 12 "Responsible Consumption and Production" and Objective 13 "Climate Action".

These criteria correspond to 49% in order of importance, that is, almost half of the total criteria, and the improvements to be made at this stage will be stimulating and exemplary for other categories as well. Another prospective goal of the study will be a holistic approach that includes all criteria.

3.2. Dataset Collection with IoT

Collecting data with IoT-based sensors for measurements in the field of "Energy and Climate Change, Waste and Water" in smart campuses plays a very important role in understanding and improving the sustainability performance of the campus. Sensor-based data collection provides real-time and accurate information on energy production, energy consumption, waste generation and water use, if available, enabling data-based decision making and targeted sustainability initiatives. Considering the criteria used in this study, it is aimed to ensure that the measurements are taken as instant and fluent data by using sensors with different skills in different ways for each criterion (Table 2).

Table 2. Data collected from sensors according to GM criteria

GM Criteria	Sensor	Data Collected
Energy and Climate Change	- Air Conditioning System Related Sensors (Temperature, Humidity, Carbon Dioxide, Movement, etc.) - Energy Meters and Submeters - Lighting Sensors	- Building and facility data on campus - Heat - Carbon Emission Amount - Energy Consumption
Waste	- Weight Sensors (Load Cell) - Occupancy Sensors - RFID Tags - Image Sensors	- Waste Amount - Recycle Rate
Water	- Water Flow Sensors (Flowmeter etc.) - Water Quality Sensors - Sensors for rainwater harvesting (If equipped) - pH Sensors	- Water Consumption - Water Quality

By connecting a collection of sensors or devices to a cloud-based platform and processing the data collected by these devices, data collected by simply framed IoT systems trigger processes that may require user input or automated system responses. A holistic IoT system consists of four main components: sensors or devices, data transmission, data processing, and user interface.

Sensors/Devices collect data from the environment. A device can have multiple sensors; For example, a smartphone has many features such as GPS, camera, accelerometer. Basically, the sensor or sensors collect data from the environment for a specific purpose.

In the Data Transmission Phase, after data is collected from the devices, this data must be transmitted to the cloud. This can be accomplished with a variety of technologies; these technologies may include connecting to the Internet via Wi-Fi, Bluetooth, satellite, low-power wide area networks (LPWAN), or direct Ethernet. The type of connection to use depends on the respective IoT application.

In the Data Processing Phase, when the data reaches the cloud, the related software processes this data and decides to apply the most accurate data preprocessing methods. This may include sending an alert or automatically adjusting sensors or device without requiring user input. However, in some cases user input may be required and that's where the UI comes in.

The User Interface is actually the part where the software communicates with the user. All operations performed by the user are carried out in the opposite direction through the system. In other words, actions are sent from the UI to the cloud and then back to the sensors or devices to make the necessary change.

The exact connectivity, networking, and communication protocols used by web-enabled devices will vary depending on specific IoT applications. IoT is increasingly using artificial intelligence (AI) and machine learning technologies to make data collection processes both simpler and faster. This significantly increases the IoT's ability to collect, analyze and transform data into meaningful information.

The Internet of Things (IoT) represents a network of embedded, interconnected and communicating devices containing sensors and software to collect, exchange and process data. In the context of smart campus sustainability covering selected GM criteria, IoT provides a powerful data collection platform to collect high-resolution, real-time data for in-depth analysis and informed decision-making in the digital twin creation process.

Sharma and Suryakanthi conducted a field study on IoT applications for university campuses in their study in 2015. In this study, the development of a Smart System that optimizes power consumption by analyzing the use of all electrical and electronic devices for the campus is discussed. The proposed system has been developed using a combination of IoT, Wireless Sensors, Network Security, Green information technologies, Big Data Analysis (Sharma & Suryakanthi, 2015). In a study by Khan and Naseer in 2020, a solution was proposed to improve wastebasket management, which is a small and important component of the university waste management system. The basic idea of this project is to provide a healthier and cleaner university environment by using the Internet of Things (IoT) protocol (Khan & Naseer, 2020). Anh Khoa et al. in 2020, a new method is proposed that performs waste management powerfully and efficiently by estimating the probability of filling the waste levels in the bins (Anh Khoa et al., 2020). With the smart green campus vision proposed in the study conducted by Abdulmouti et al., it is aimed to provide people with different innovative systems and to support the development of the country. It has been observed that 63.7% of electricity is saved when electrical energy is obtained from solar energy and innovative applications are applied in the smart green campus, and 0.02 percent of the emissions released into the air and carbon dioxide (Abdulmouti et al., 2022).

3.3. Data Collection for Energy and Climate Change Criteria

Within the scope of Energy and Climate Change criteria, the campus needs to be transformed in areas such as adaptation to climate change, reduction of carbon footprint and smart energy consumption. Some materials are also needed to collect the data set required for transformation in this area. Sensors or devices that provide important data for this criterion are mainly energy meters and sub-meters, heating, ventilation and air conditioning (HVAC) and lighting sensors, SCADA tools and climate sensors (Temperature, humidity and other air parameters).

Energy meters and sub-meters are specified as key components of the energy monitoring infrastructure of the smart campus. By providing precise measurement and monitoring of electricity consumption at various levels, they provide valuable insight into energy usage patterns and identifies opportunities for energy optimization. Energy meters are devices used to measure the total electrical energy consumption of a particular area or the entire building. They are typically installed at the main electrical supply point of the campus or individual buildings. Energy meters provide cumulative energy consumption data, usually measured in kilowatt-hours (kWh) or megawatt-hours (MWh) over a period such as daily, weekly, or monthly. Submeters, also known as branch circuit meters or individual meters, are devices installed below energy meters to monitor energy consumption in more detail. They measure energy use for specific areas, sections, floors or individual equipment within a building. Like energy meters, sub-meters can be integrated with data communication networks for real-time data transmission and analysis.

HVAC and lighting sensors, another source of monitoring and instantaneous data, are an important element of a smart campus' energy monitoring and management system. These sensors enable real-time data collection and analysis

from heating, ventilation, air conditioning and lighting systems, providing valuable insight into energy use patterns and opportunities to optimize energy efficiency. Sensors such as "temperature sensor, humidity sensor, occupancy and motion sensors, carbon dioxide sensors" are used in this system, which presents the flowing data to the user with the cooperation of different sensor groups. Thus, in order to keep it in the desired comfort range, operations such as arranging the operation of the instruments according to the indoor temperature, humidity level in the interior, occupancy or movement in the rooms and evaluating the indoor air quality can be performed. In addition to the sensors in this group, which can also be called Climate Sensors, special sensors such as solar radiation sensors and precipitation sensors can be added. These sensors can help campuses adapt their HVAC systems to prepare for extreme weather events and maintain indoor comfort and safety in extreme temperatures, laying the groundwork for applications such as mold prevention, better control of HVAC systems, energy demand forecasts, and climate resilience.

Recently, clean energy investments created by universities on campus have become very popular. It is aimed to build cleaner and more sustainable campuses for the future, both with educational presentations to students and with studies carried out within the scope of sustainable campus goals. By integrating sensors to measure power output from renewable energy sources such as solar panels or wind turbines, universities can gain valuable insight into the performance of these renewable energy systems and their contribution to the overall energy mix of the campus. Thus, added value is provided in areas such as data-driven decision making, reduction of carbon footprint, resource planning and energy management with maximum efficiency.

3.4. Data Collection Water Criteria

Water is getting more and more important in today's world. With the effect of global climate change, it becomes imperative to take precautions against possible problems in access to clean water. In this context, university campuses can contribute to the goal of sustainability and a more accessible future by making water management smarter and more efficient. At this stage, more efficient taps for water use and sinks with sensitive sensors can be preferred in university campuses. In addition, data providers and water flow sensors, rainwater harvesting sensors and water quality sensors can be used to measure how well the measures taken are working.

Detailed monitoring of water flow is crucial for early detection of potential leaks and excessive water consumption. Also, the management of water distribution and allocation between campus buildings and facilities can be done more efficiently based on this data. Detection of sudden changes in water flow is critical for monitoring leaks in the plumbing system. Real-time water flow data enables campus managers to make informed decisions about water usage, budgeting and sustainability initiatives and facilitates resource tracking. Such practices not only provide significant added value, but also help the campus move up the Green Metric rankings. Therefore, it is of critical importance that the data obtained as a result of measurement and monitoring practices is collected in a healthy and regular manner.

3.5. Data Collection Waste Criteria

In smart waste management systems, various sensors are used to optimize waste collection, encourage recycling and reduce waste in a smart campus, and the data from these sensors provide important outputs for decision support systems.

Smart waste bins can be equipped with fill level sensors that continuously monitor the amount of waste in the bin. These sensors use various technologies such as ultrasonic, infrared or weight-based sensors to measure the fill level. Real-time fill level data allows waste collection teams to optimize their routes and prioritize bins that need to be emptied, reducing unnecessary trips and fuel consumption. Smart waste bins provide more efficient waste collection practices, resulting in cost savings in labor, fuel and vehicle maintenance. Ensuring that the trash cans are emptied at the right time reduces the possibility of overflow.

Separation and recycling sensors are used to analyze waste streams and identify recyclable materials in waste. These sensors can also help identify and remove non-recyclable materials or contaminants that may hinder the recycling process.

Waste generation tracking with weight sensors contributes to waste reduction initiatives by helping to measure the amount of waste produced by specific buildings or areas. It can also evaluate the effectiveness of recycling programs by comparing the weights of recyclable and non-recyclable waste.

3.6. Digital Simulation

Using the data collected from the sensors mentioned in the data set collection section, it is aimed to create insightful and interactive visualizations with the help of visual analytical techniques. This step of the approach plays an important role in creating a closer to the original digital twin, providing important ideas for next steps, and making the right decisions by making data-related transactions healthier.

3.7. Visual Analysis

Data visualization is an important step to give insight into the data collected and to make the preprocessing of various data healthier and better for decision processes.

The purpose of visual analysis for the energy and climate change criterion is to gain insight into energy consumption patterns and their relationship to climate factors. It is aimed to follow the trend with the visualizations, to help identify the peak energy demand periods as well as potential opportunities for energy efficiency and renewable energy integration. The main purpose is to enable decision support mechanisms to work properly by being informed about energy consumption patterns and their relationship with climate factors. Data visualization tools such as different types of graphs, heat maps and scatter plots will be used to show energy consumption trends over time. In order to understand how weather affects energy demand, performing correlation analyzes with climate parameters such as temperature and humidity will help create a healthy digital twin by facilitating the system's response to current and possible extreme situations. In addition, markings on the map to display the distribution of renewable energy sources such as solar panels and wind turbines throughout the campus play an important role in the analysis of data in order to follow these mini power plants and make their future forecasts more reliable.

Visualizations will provide a comprehensive understanding of water consumption patterns in campus buildings and facilities. Visualizations help identify high water use areas, detect leaks and optimize water conservation efforts. We use a variety of graphical and data visualization techniques, such as water flow maps, to show water use patterns over time and spatially. Interactive visualizations allow users to explore water consumption trends in different campus areas and buildings. Comparative visualizations of water use before and after the implementation of conservation measures help evaluate the effectiveness of water conservation initiatives.

As for waste, which is one of the GreenMetric criteria, added value can be provided in understanding waste accumulation patterns, optimizing waste collection routes, and encouraging waste reduction and recycling efforts. Visualization of waste composition and recycling rates is important to give an idea about the effectiveness of recycling programs. Interactive waste collection route maps help optimize collection schedules and visualize data is essential to reduce unnecessary travel and associated emissions.

Visualizations enable data-driven decisions to improve energy efficiency, water conservation and waste reduction strategies, while resource optimization and understanding of consumption and production patterns enable campuses to allocate resources more efficiently, resulting in cost savings and environmental benefits. Additionally, Interactive visualizations will engage the campus community in sustainability efforts, foster awareness and participation, and enable stakeholder engagement.

3.8. Physical World

The definition of the Physical World refers to the campus area where data for all criteria specified in the above sections are collected. Within the scope of the study, it refers to the tangible, observable and measurable aspects of a campus that can be characterized as smart, especially in relation to energy and climate change, water management and waste management. The physical world, which includes the real infrastructure, resources and behaviors in the campus environment, which is subjected to data collection, analysis and optimization through smart technologies, sensors and similar tools, is a real-world representation of the digital twin created.

In order to create a healthy digital twin in this area and to offer more accurate solutions, the infrastructure that will provide the most appropriate data collection should be established. A communication protocol that will contribute to decision support processes between processes such as data analysis and machine learning in the operation processes within the created environment will enable the digital twin process to sit on a stronger infrastructure.

One of the most important purposes of the digital twin within the scope of predictive simulation is to imitate the main system, to bring the perceived results with the analysis of the created working conditions to the system integration and to ensure that the physical world is fully complete. For the physical world, which represents all the physical components of the existing infrastructure in the campuses, and the targeted digital twin, creating the most suitable working conditions

for the hardware integrated for data collection and monitoring is a necessity for the healthy functioning of the proposed approach.

4. DISCUSSION AND CONCLUSIONS

Contribution to environmental sustainability requires a more comprehensive and detailed understanding of the ecological impacts of various activities, especially within large institutional structures such as universities. Given their scope, the diversity of their activities and their overall impact, universities have a critical role to play in promoting and implementing sustainable practices. However, the complexity and scale of activities on a university campus make it difficult to monitor and manage the environmental footprint. In addition, the specific focus of this study on blending digital twin technology with GM metrics is a crucial step towards assessing and improving sustainability practices on university campuses. It also creates an opportunity for world universities to be ranked and gain prestige. In other words, the university gains environmental, economic and social benefits by contributing to environmental sustainability policies while at the same time increasing its own prestige.

Establishing such a framework will also have a full benefit-oriented and value-adding effect, given the global focus on achieving the United Nations' Sustainable Development Goals and the Green Deal. With universities, which are very important institutions in society, leading these initiatives, a data-driven, technologically advanced approach can serve as a guide for other institutions and sectors seeking sustainability.

In order for the proposed model to be sustainable and efficient, it should be taken into consideration that the machine learning and optimization techniques to be used in the model should be selected according to the type and size of the collected data. For low-dimensional data studies, statistical models (regression, etc.), which are more efficient in terms of both time and cost, are foreseen to be used. For more complex and larger data sets, depending on the problem, decision tree-based algorithms (XGBosst etc.) or classical artificial neural networks, which have been proven to work well in the literature, can be used. In order to obtain more differentiated and advanced results, advanced models of these networks (LSTM etc.) can be selected depending on the problem content. In addition, scalability and efficiency should be considered in case the size of the data collected in line with GM criteria increases by using IoT technologies. In this case, it is foreseen to use cloud solutions that provide fast and secure access to data, storage of high-dimensional data sets, better management of risks and synchronization advantages.

In conclusion, this study proposes a digital twin architecture to be built within the smart campus concept, combining GM criteria with selected energy and climate change, waste and water issues. It makes a valuable contribution to the growing literature on the application of digital twin technology and IoT in university campuses, with a particular focus on sustainability. This integrated approach has significant potential in the sustainability context by providing universities with the opportunity to assess, plan and monitor their sustainability efforts in a more effective and data-driven manner. The proposed architecture aims to improve the environmental performance of campuses by combining a sensor-based framework and digital simulation with predictive analysis and dynamic decision support protocols, including communication with the physical world. In this context, the potentials of making waste management more efficient, ensuring efficient use of water resources, and identifying efficiency gaps by analyzing energy consumption in real time are highlighted. The proposed framework provides a holistic and data-driven approach to assessing the environmental performance of smart campuses. With its effective solutions on critical issues such as energy and climate change, waste and water management, it is thought to provide an important step towards the construction of environmentally friendly and sustainable educational environments in future educational institutions. The combination of "Digital Twin" technology and the concept of "Green Metrics" will allow smart campuses to make significant progress in environmental sustainability and green consensus. With the spread of such smart campuses in educational institutions in the future, it will be possible to build an environmentally friendly, efficient and innovative educational environment. This study should be considered as an important step in shaping the educational environments of the future.

Peer Review: Externally peer-reviewed.

Author Contributions: Conception/Design of Study- İ.T., O.T., M.K.; Data Acquisition- O.T.; Data Analysis/Interpretation- M.K.; Drafting Manuscript- İ.T.; Critical Revision of Manuscript- İ.T., O.T., M.K.; Final Approval and Accountability- İ.T., O.T., 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.

ORCID IDs of the authors / Yazarların ORCID ID'leri

İlknur Teke 0000-0002-6383-4067
Orkun Teke 0000-0003-4390-263X
Murat Kılınç 0000-0003-4092-5967

REFERENCES

- Abdulgouti, H., Skaf, Z., & Alblooshi, S. (2022). Smart Green Campus: The Campus of Tomorrow. *Advances in Science and Engineering Technology International Conferences (ASET)*, 1–8. <https://doi.org/10.1109/ASET53988.2022.9735087>
- Abuarqoub, A., Abusaimh, H., Hammoudeh, M., Uliyan, D., Abu-Hashem, M. A., Murad, S., Al-Jarrah, M., & Al-Fayez, F. (2017). A survey on internet of things enabled smart campus applications. *ACM International Conference Proceeding Series*, Part F130522. <https://doi.org/10.1145/3102304.3109810>
- Anh Khoa, T., Phuc, C. H., Lam, P. D., Nhu, L. M. B., Trong, N. . . . Duc, D. N. M. (2020). Waste Management System Using IoT-Based Machine Learning in University. *Wireless Communications and Mobile Computing*, 2020(1), 1-13. <https://doi.org/10.1155/2020/6138637>
- Çelebi, H., Bahadır, T., Şimşek, İ., & Tulun, Ş. (2020). The Importance of Smart Campuses in the Context of Boyd Cohen Wheel and Sustainable Environmental Dimension. *Journal of Engineering Sciences and Design*, 8(3), 952–960. <https://doi.org/10.21923/JESD.703431>
- Corrado, C. R., DeLong, S. M., Holt, E. G., Hua, E. Y., & Tolk, A. (2022). Combining green metrics and digital twins for sustainability planning and governance of smart buildings and cities. *Sustainability*, 14(20), 12988. <https://doi.org/10.3390/SU142012988>
- Demiroğlu, D., Karadağ, A., & Cengiz, A. E. (2017). Evaluation of the green campus approach on the campuses in Turkey. *Eurasian Journal of Civil Engineering and Architecture Journal*, 1(1), 53–65.
- Finlay, J., & Massey, J. (2012). Eco-campus: Applying the ecocity model to develop green university and college campuses. *International Journal of Sustainability in Higher Education*, 13(2), 150-165. <https://doi.org/10.1108/14676371211211836>
- Gambella, C., Ghaddar, B., & Naoum-Sawaya, J. (2019). Optimization Models for Machine Learning: A Survey. *European Journal of Operational Research*, 290, 807-828.
- Georgios, L., Kerstin, S., & Theofylaktos, A. (2019). Internet of Things in the Context of Industry 4.0: An Overview. *International Journal of Entrepreneurial Knowledge Issue*, 1(1), 4–19. <https://doi.org/10.2478/IJEK-2019-0001>
- Han, X., Yu, H., You, W., Huang, C., Tan, B., Zhou, X., & Xiong, N. N. (2022). *Intelligent Campus System Design Based on Digital Twin. Electronics*, 11(21), 3437. <https://doi.org/10.3390/ELECTRONICS11213437>
- Khan, M. N., & Naseer, F. (2020). IoT Based University Garbage Monitoring System for Healthy Environment for Students. *IEEE 14th International Conference on Semantic Computing (ICSC)*, 354–358. <https://doi.org/10.1109/ICSC.2020.00071>
- Kivunja, C. (2018). Distinguishing between Theory, Theoretical Framework, and Conceptual Framework: A Systematic Review of Lessons from the Field. *International Journal of Higher Education*, 7(6), 44–53. <https://doi.org/10.5430/ijhe.v7n6p44>
- Kumaş, E., & Erol, S. (2021). Endüstri 4.0'da Anahtar Teknoloji Olarak Dijital İkizler. *Politeknik Dergisi*, 24(2), 691-701. <https://doi.org/10.2339/politeknik.778934>
- Lauder, A., Sari, R., Suwartha, N., & Tjahjono, G. (2015). Critical review of a global campus sustainability ranking: GreenMetric. *Journal of Cleaner Production*, 108, 852-863. <https://doi.org/10.1016/j.jclepro.2015.02.080>
- Leshem, S., & Trafford, V. (2007). Overlooking the conceptual framework, *Innovations in Education and Teaching International*, 44(1), 93–105. <https://doi.org/10.1080/14703290601081407>
- Li, P. (2017). Optimization Algorithms for Deep Learning. *University of Hong Kong Department of Systems Engineering and Engineering Management Journal*. 30(3), 1-12. Retrieved from <http://lipiji.com/docs/li2017optdl.pdf>.
- Lukman, R., Tiwary, A., & Azapagic, A. (2009). Towards greening a university campus: The case of the University of Maribor, Slovenia. *Resources, Conservation & Recycling*, 53, 639–644. <https://doi.org/10.1016/j.resconrec.2009.04.014>
- Massucco, S., Del Borghi, A., Delfino, F., Laiolo, P., Marin, V., Moreschi, L., & Vinci, A. (2023). University of Genoa best practices in managing Energy and Climate Change. *IOP Conference Series: Earth and Environmental Science*, 1194, 012001. <https://doi.org/10.1088/1755-1315/1194/1/012001>
- Pexyeau, T., Saraubon, K., & Nilsook, P. (2022). IoT, AI and Digital Twin For Smart Campus. *Invention, and Innovation Congress: Innovative Electricals and Electronics, RI2C*, 160–164. <https://doi.org/10.1109/RI2C56397.2022.9910286>
- Register, R. (2005). *Ecocities: Rebuilding Cities in Balance with Nature*. New York, NY: New Society Publishers.
- Sala, R., Zambetti, M., Pirola, F., & Pinto, R. (2018). How to select a suitable machine learning algorithm: A feature-based, scope-oriented selection framework. *Proceedings of the Summer School Francesco Turco, 2018(1)*, 87-93.

- Sharma, K., & Suryakanthi, T. (2015). Smart System: IoT for University. *International Conference on Green Computing and Internet of Things (ICGCIoT)*, 2015(1), 1586–1593. <https://doi.org/10.1109/ICGCIoT.2015.7380720>
- Sonetti, G., Lombardi, P., & Chelleri, L. (2016). True Green and Sustainable University Campuses? Toward a Clusters Approach. *Sustainability*, 8, 83. <https://doi.org/10.3390/su8010083>
- Suwartha, N., & Sari, R. F. (2013). Evaluating UI GreenMetric as a tool to support green universities development: assessment of the year 2011 ranking. *Journal of Cleaner Production*, 61, 46–53. <https://doi.org/10.1016/J.JCLEPRO.2013.02.034>
- Trauer, J., Schweigert-Recksiek, S., Engel, C., Spreitzer, K., & Zimmermann, M. (2020). What Is a Digital Twin? – Definitions and Insights From an Industrial Case Study In Technical Product Development. *Proceedings of the Design Society: Design Conference*, 1, 757–766. <https://doi.org/10.1017/DSD.2020.15>
- UI GreenMetric. (2023, July 21). What is UI Green Metric?[<https://Greenmetric.Ui.Ac.Id/about/Welcome>]. Retrieved from <https://greenmetric.ui.ac.id/>
- VanDerHorn, E., & Mahadevan, S. (2021). Digital Twin: Generalization, characterization and implementation. *Decision Support Systems*, 145, 113524. <https://doi.org/10.1016/J.DSS.2021.113524>
- Wang, Y. (2022). Research on smart campus construction based on digital twins. *Fifth International Conference on Mechatronics and Computer Technology Engineering (MCTE 2022)*. 12500Q3, 914–920. <https://doi.org/10.1117/12.2662569>
- Wang, M., Fu, W., He, X., Hao, S., & Wu, X. (2020). A Survey on Large-Scale Machine Learning. *IEEE Transactions on Knowledge and Data Engineering*, 34, 2574-2594.
- Zaballos, A., Briones, A., Massa, A., Centelles, P., & Caballero, V. (2020). A Smart Campus' Digital Twin for Sustainable Comfort Monitoring. *Sustainability*, 12(21), 9196. <https://doi.org/10.3390/SU12219196>

How cite this article

Teke, I., Teke, O., Kilinc, M. (2023). The Future of Smart Campuses: Combining Digital Twin and Green Metrics. *Acta Infologica*, 7(2), 384-395. <https://doi.org/10.26650/acin.1386072>

A Systematic Literature Review for New Technologies in IT Audit

Bilgi Teknolojileri Denetiminde Yeni Teknolojiler Üzerine Bir Sistemantik Literatür Taraması

Nur Sena Tanrıverdi¹ , Nazım Taşkın² 

¹(PhD Candidate), Bogazici University, Faculty of Managerial Sciences, Department of Management Information Systems, Istanbul, Türkiye

²(Assist. Prof.), Bogazici University, Faculty of Managerial Sciences, Department of Management Information Systems, Istanbul, Türkiye

Corresponding author : Nur Sena Tanrıverdi

E-mail : nur.tanriverdi@boun.edu.tr

ABSTRACT

Information technology (IT) audit focuses on auditing companies' IT systems and processes. The systems companies use are getting more complicated and better integrated. This means more data also needs to be audited. An IT audit often requires performing repetitive manual tasks, which makes IT audits more labor-intensive and costly. Current technological advancements have immense potential for improving an IT audit's performance, quality, and accuracy. Therefore, by leveraging advanced data processing and analysis technology, this workload can be lowered, allowing the auditing process to be performed effectively and efficiently with higher-quality outcomes. To achieve this objective, a systematic literature review (SLR) has been conducted to identify studies that use artificial intelligence (AI), machine learning (ML), predictive analytics, process mining, and natural language processing (NLP) techniques applied within IT auditing. Process mining is seen to have emerged as the most commonly used technique among the analyzed studies. The studies also reveal that combining techniques such as process mining and data mining, natural language processing, and machine learning enables effective and efficient audit processes by conducting continuous, automated, or online auditing work. The application of these new techniques in the examined studies are seen to generally provide solutions regarding the audit's testing stage. Overall, the study reveals a limited number of academic studies to have examined how these techniques are implemented into IT audits.

Keywords: Information technology audit, application audit, IT process audit, emerging technologies, systematic literature review

ÖZ

Bilgi teknolojisi (BT) denetimi, şirketlerin BT sistemlerini ve süreçlerini denetlemeye odaklanır. Bir şirkette kullanılan sistemler daha karmaşık ve entegre hale gelmektedir. Bu durum, denetlenmesi gereken verilerin artması ile sonuçlanır. BT denetimi genellikle manuel ve tekrarlanan görevlerin yapılmasını gerektirir, bu da BT denetimlerini daha emek yoğun ve maliyetli hale getirir. Mevcut teknolojik gelişmeler, BT denetim sürecinin performansını, kalitesini ve doğruluğunu iyileştirmek için büyük bir potansiyele sahiptir. Bu nedenle, gelişmiş veri işleme ve analiz teknolojisi kullanılarak bahsedilen iş yükü azaltılabilir ve denetim süreci daha kaliteli sonuçlarla etkin ve verimli bir şekilde gerçekleştirilebilir. Söz konusu amaca ulaşmak için, BT denetimi içinde uygulanan yapay zeka, makine öğrenimi, tahmine dayalı analitik, süreç madenciliği ve doğal dil işleme tekniklerini kullanan çalışmaları belirlemek için sistemantik literatür taraması yapılmıştır. İncelenen çalışmalarda süreç madenciliğinin en çok kullanılan teknik olduğu görülmüştür. Çalışmalar ayrıca, süreç madenciliği ve veri madenciliği, doğal dil işleme ve makine öğrenimi gibi teknikleri birleştirmenin, sürekli, otomatik veya çevrimiçi denetim çalışması yürüterek etkin ve verimli denetim süreçlerine sahip olmayı mümkün kıldığını ortaya koydu. İncelenen çalışmalarda bu yeni tekniklerin uygulanması genellikle denetimin test aşamasına ilişkin çözümler sunmaktadır. Genel olarak çalışma, bu tekniklerin BT denetimlerine uygulanmasını inceleyen sınırlı sayıda akademik çalışma olduğunu ortaya koymaktadır.

Anahtar Kelimeler: Bilgi teknolojileri denetimi, bilgi teknolojileri genel kontrolleri, uygulama denetimi, BT süreç denetimi, gelişen teknolojiler, sistemantik literatür taraması

Submitted : 12.07.2022
Revision Requested : 10.02.2023
Last Revision Received : 07.06.2023
Accepted : 10.07.2023
Published Online : 21.12.2023



This article is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0)

1. INTRODUCTION

Information technology (IT) audit is an audit that focuses on companies' IT systems and processes. The audit process follows specific steps and tasks to determine critical issues and incompatible events within processes. At the end of the audit work, an audit opinion is provided to companies.

IT audit work is repetitive in nature (Manhanga, 2020), and performing repetitive manual tasks takes a serious amount of time and requires high labor costs. Moreover, the volume of data has increased as a result of the increasing number of integrated systems and digitalized business processes (Dzurani & Mălăescu, 2016). Obtaining a high volume of data from systems and getting a sample of the data to audit are both challenges for an audit (Alexiou, 2019). Alongside sampling high volumes of data, another significant issue is examining all the data in a system.

However, systems have become capable of storing and processing high volumes of data, and analytical methods have also evolved. Therefore, current technology is also pushing changes in IT auditing. At the same time, new technologies can increase an IT audit's performance and lower the associated transaction costs while boosting the audit's quality and accuracy (Khan, Adi, & Hussain, 2021; Flores & Riquenes, 2020). Therefore, current top issues in auditing involve reducing the amount of time and costs required for audit work and increasing efficiency, effectiveness, and quality. These issues make the usage of new technologies in IT auditing crucial.

The Information Systems Audit and Control Association (ISACA, 2019a) conducted a survey among IT auditors about future predictions and directions of IT audits. This survey asked IT auditors about current technology usage in IT audit and predictions about technologies that will be used in the future. According to the survey results, 26% of IT auditors claimed their firms to use process mining, and 23% stated predictive analytics to be in use. In the report, 23% and 26% of the auditor participants respectively predicted that process mining and predictive analytics will still be used in the next 1-2 years. Additionally, 25% of the participants mentioned planning to use artificial intelligence (AI) within their audit work in the next 3-5 years, while 24% said they will use machine learning (ML) in the future. As explained above, IT audit work creates a substantial workload that is projected to grow even further in the future. However, this workload can be mitigated by using advanced data processing and analysis technologies, resulting in an efficient and effective auditing process with enhanced accuracy. Therefore, this literature review aims to identify studies that employ AI, ML, predictive analytics, process mining, and natural language processing (NLP) techniques. These identified techniques are currently being utilized or are anticipated to gain more prominence within the IT auditing domain. Their application facilitates more efficient and effective IT auditing, ultimately elevating the quality of audits. Additionally, this study is the first attempt to gather different approaches concerning the adoption of new technologies in IT auditing, an area limited by current techniques.

The primary aim of this study is to describe technologies that have already been integrated or that have the potential to be integrated into IT audits. By providing key insights from existing studies in the field of IT auditing, this review study aims to provide a comprehensive overview of the topic. With this aim, the study has created the following research questions:

RQ-1: What are the common approaches proposed for the use of new technologies/techniques in IT auditing?

RQ-2: In which step of IT auditing will the new technologies/techniques be used?

RQ-3: What data analytics algorithms are commonly used in IT auditing?

2. BACKGROUND INFORMATION

This section will provide background information about the key areas of this study, IT audits, and new technologies before applying the literature review methodology and presenting the related literature. After explaining a solid understanding of these key areas, the literature review methodology will be constructed and implemented.

2.1. IT Audit

IT auditing implies assessing and investigating all IT processes and the systems that process and store a company's critical data, such as financial data, customer data, and employees' personal data. An IT audit tests the defined control environment of a company and evaluates its compatibility with the company's policies (Carlin & Gallegos, 2007). An IT audit may also require an assessment of the organization's usage of IT in supporting the efficiency, effectiveness, and economics of its business processes (Carlin & Gallegos, 2007).

An IT audit is performed based on the collected audit evidence. ISACA (2019b) defined audit evidence as set of data employed to substantiate the audit assessment. Reports, documents, and system logs are audit evidence obtained based on the audit scope and period. IT auditors examine the audit evidence in detail and can conduct investigations directly on the systems in addition to these collected documents and logs. Information can be extracted from a system

during auditing. Auditors also perform many inquiries in an audit on conducting and monitoring specific systems and processes with the people who are responsible in order to understand the processes. Although some of the audit evidence within IT audits are extracted from systems, IT auditors also perform manual audits. According to Mendez (2020), an IT auditor conducts the IT audit by following certain steps, such as planning, fieldwork or documentation, reporting, and follow up.

ISACA is a global professional association focused on IT governance, audits, risk, and compliance (ISACA, n.d.). The association creates a framework for IT governance, audits, and risk areas by collecting ideas from practitioners and academicians within the global network. They also publish reports and articles on their areas of activity.

An IT audit can be performed for different purposes. For example, an IT general controls audit aims to evaluate whether the controls of financial systems have been sufficiently set and are being efficiently run within the auditing period of a financial system and also evaluate whether financial systems are reliable (Barta, 2018; Chen, Hsu, & Hu, 2021). This audit is performed mainly under financial audits (Krieger, Drews, & Velte, 2021). Application checks can also be run within this type of audit; however, an IT application controls audit is subjected to evaluating the application itself (Barta, 2018; ISACA, 2019b). All IT components and systems are regulated and monitored under definitive process. Therefore, these processes (i.e., IT processes) are generally subject to an audit (ISACA, 2019b).

ISACA (2020) conducted a survey to provide a benchmark for IT audits. According to the frequencies of the answers the survey participants gave, the survey results show IT audits to be responsible for conducting IT general control audits, application audits, and IT process audits. Therefore, these audit types have been identified as the main focus of IT audits in this literature review study.

2.2. New Technologies

Based on the research ISACA (2019a) conducted, how IT audits are currently used, and what is expected to be used regarding IT audits in the future, this study has identified four technologies and techniques: AI, ML, predictive analytics, and process mining. These technologies and techniques mostly require structured data. However, a lot of unstructured data is also found to be audited as audit evidence in an IT audit. Therefore, natural language processing is also involved in the technologies on which this study focuses in order to cover the unstructured textual data analysis based on Schumann and Marx Gómez's (2021) approach.

3. RESEARCH METHODOLOGY

This paper uses a systematic literature review (SLR) methodology to comprehensively review new technology usage in the IT audit literature. SLR facilitates evidence-based guidelines for researchers and gives more rigorous results than ad-hoc literature review (Kitchenham et al., 2009; Tranfield, Denyer, & Smart, 2003). This article follows three stages within the SLR similar to how Tranfield et al. (2003) conducted their study: planning the SLR, conducting it, and reporting on it.

The planning stage should determine the needs of the review and develop the literature review protocol. The aim and needs of the SLR have already been explained in the Introduction. Additionally, the paper has adopted Schumann and Marx Gómez's (2021) structured literature review protocol due to the focus of this study resembling their focus on using natural language processing in internal auditing. The review protocol is given in Table 1. They conducted searches on six databases: Science Direct, Scopus, Web of Science, ACM Digital Library, IEEE Xplore Digital Library, and American Accounting Association. Databases with broader coverage than the others were accessed through the university library website. Thus, all databases were used to search papers except for the American Accounting Association database.

Table 1. Review Protocol

Unit of analysis	Journal articles
Type of analysis	Qualitative
Time period	2000-2021
Search fields	Title, Abstract, Keywords
Databases	Science Direct, Scopus, Web of Science, ACM Digital Library, and IEEE Xplore Digital Library
Total number of articles used in this study	8

The journal articles were evaluated for their effects and contributions to their fields (Hult, Reimann, & Schicke,

2009). In addition, the time interval as a search criterion was chosen based on Schumann and Marx Gómez (2021)'s study. Therefore, articles published between 2000-2021 were selected.

When conducting the review stage, the determined literature review protocol was executed. Keywords should be determined to select the relevant literature, and the keywords were selected based on concepts clarified in Section 2 (Background Information; see Table 2). Each concept has been expressed using different words and word groups in order to have the most generic terms as keywords written in different formats.

Table 2. List of Keywords and References

Perspective	Keywords
Technology	"artificial intelligence", "AI"
	"predictive analytics"
	"process mining"
	"machine learning", "ML"
	"natural language processing", "NLP"
Audit	"IT audit", "information technology audit"
	"IT general control audit", "information technology general control audit", "ITGC"
	"application audit"
	"IT process audit", "information technology process audit"

The search query given in Table 3 was run on the selected databases using the determined criteria.

Table 3. Database Search Query

In Title-Abstract-Keywords: ("artificial intelligence" OR "AI" OR "predictive analytics" OR "process mining" OR "machine learning" OR "ML" OR "natural language processing" OR "NLP") AND ("IT audit" OR "information technology audit" OR "IT general control audit" OR "information technology general control audit" OR "ITGC" OR "application audit" OR "IT process audit" OR "information technology process audit")

After the search was completed, 636 journal articles were collected from five databases. Of these articles, 317 are found on Science Direct, eight on Scopus, 289 on Web of Science, three on ACM Digital Library, and 19 on IEEE Xplore Digital Library (Figure 1). Only papers written in English were considered in this review, which then left 600 papers. The initial set was checked to identify duplicate papers. After removing duplicates, 530 unique journal articles were obtained in the initial data set.

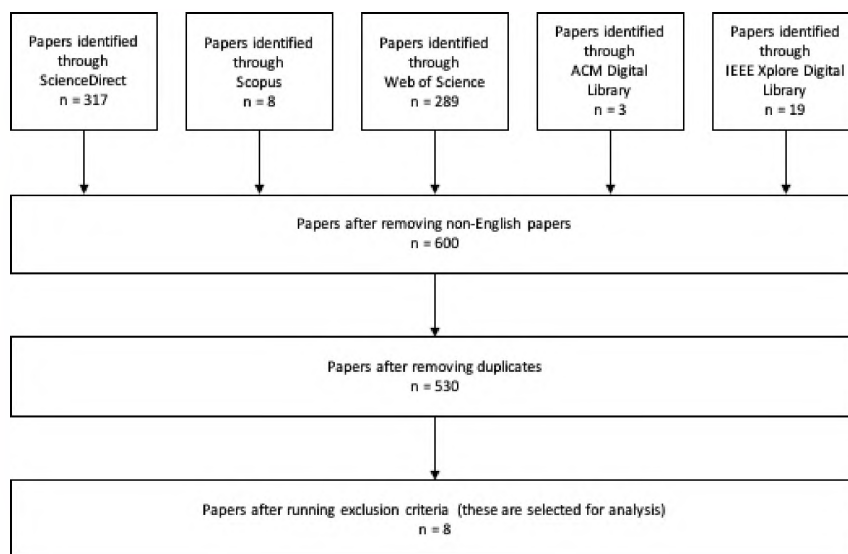


Figure 1. Flowchart of the literature Review process.

The aim of this literature review is to find studies that have applied such techniques as AI, ML, predictive analytics, process mining, and/or NLP with regard to IT auditing. This paper has identified studies that focus on tasks within any step of the IT auditing process. Therefore, the exclusion criteria given in Table 4 were formed and applied to the initial set of studies. The number of papers excluded from the initial set after applying the criteria are given in Table 4.

Applying the first exclusion criterion (C1) excluded 407 papers for containing only the searched keywords without the context of corporate auditing. The second exclusion criterion (C2) eliminated papers related to auditing that did not refer to conducting an audit or that did not contain such things as auditing tasks or auditing methodologies. In addition, papers that focused on the auditability of a proposed system, model, or new technology without giving an auditing approach or methodology were eliminated. This second criterion eliminated another 25 papers. The third exclusion criterion (C3) was run to eliminate papers that were related to auditing but that only handled the topic at a conceptual level. Based on this third exclusion criterion, another 16 papers were excluded. The main focuses of these eliminated papers involved such topics as listing the advantages of new approaches in auditing, their acceptance factors, auditors' qualifications, and changes in the audit profession.

After running the exclusion criteria C1, C2, and C3, the remaining papers are seen to include tasks and approaches within the auditing process. Starting from this point, the papers were analyzed with a greater focus on IT auditing. However, the remaining 74 papers did not include any studies implementing new techniques and technologies into application auditing, IT general controls (ITGC), IT process audit types in IT auditing, or other types of IT auditing.

Meanwhile, the final set of articles had some papers that focused on providing a method to audiences rather than just applying methods and techniques to specific auditing subjects. Jans and Hosseinpour (2019) emphasized the independence of the method from any specific auditing subject. When considering the potential applicability of the method papers to other audit subjects, the remaining set of articles was decided to be evaluated accordingly. Therefore, a fourth exclusion criterion was developed with this aim. According to the fourth exclusion criterion (C4), papers were excluded that provided specific solutions for certain audit subject data (e.g., tax, finance, fraud), because applying these solutions to IT auditing was unsuitable or evaluating their applicability to the IT audit area was ineffective. Meanwhile, papers related to business process auditing or internal auditing were not excluded because they can be applied to the field of IT audits. This criterion eliminated another 57 papers.

The fifth exclusion criterion (C5) filtered out an additional nine papers. These papers had no applications focused on AI, NLP, ML, predictive analytics, or process mining, which are the new technologies that have been determined in this research. Concepts covering such technologies as advanced data analytics or big data analytics have been included. For example, many papers provided blockchain applications within auditing as a new approach. The final exclusion criterion (C6) eliminated four papers that were inaccessible through the university library's access rights. Eight papers remained from the initial set after applying these exclusion criteria.

The criteria C1, C2, and C3 were adapted from Schumann and Marx Gómez's (2021) study, whereas the criteria C4, C5, and C6 were created to capture research suitable for this study. The exclusion criteria and the number of excluded papers are given in Table 4.

Table 4. *The Exclusion Criteria and Number of Excluded Papers*

ID	Exclusion Criteria	Number of Excluded Papers	ID	Exclusion Criteria	Number of Excluded Papers
C1	No actual auditing reference	407	C4	No approaches that are adaptable to the IT audit engagement process	57
C2	No task within the audit engagement process	25	C5	No reference to the relevant technology	9
C3	No technical approach	16	C6	Article is inaccessible	8

Eight papers were determined for the analysis after applying the exclusion criteria. The papers were then evaluated based on the categories given in Table 5. These categories were created deductively. Eight categories (i.e., audit type, IT audit stage, new techniques/technologies, algorithm, data type, methodology and framework, validation, and evaluation) were created by following Schumann and Marx Gómez's (2021) approach. The full text of the eight papers were qualitatively analyzed based on these categories in order to fulfill the research questions. Thus, planning and conducting the SLR stages have been performed so far.

Table 5. *Categories for Literature Analysis*

Category	Description of Category
Audit Type	Type of audit (application audit, IT process audit, ITGC, others)
IT Audit Stage	Stage of IT audit (planning, fieldwork/documentation, report/follow up)
New Techniques/Technologies	Determined new techniques/technologies (AI, NLP, ML, predictive analytics, process mining)
Algorithm	Applied data analytics algorithm (e.g. clustering, support vector machines, etc.)
Data Type	Type of inputted data of the model (unstructured, structured)
Methodology and Framework	Whether methodology or framework is provided in the paper
Validation	Whether validation is performed in the paper
Evaluation	Whether performance of techniques is evaluated in the paper

4. RESULTS

This section reports on the literature review in stages similar to those Tranfield et al. (2003) suggested for a literature review. The following paragraphs and sections give descriptive and thematic analyses of the articles.

Among the eight articles in the final set, the earliest publication occurred in 2008, while the most recent articles were published in 2021 (Table 6). As seen in Table 6, three recent papers published in 2020 and 2021 employed AI, ML, and natural language processing. While four of the papers published before 2020 employed process mining, one employed data mining.

Table 6. *The Papers' Publication Year, Audit Type, and New Techniques/Technologies*

Reference	Year	Audit Type	New Techniques/Technologies
Rozinat & van der Aalst (2008)	2008	Other	Process mining
Caron, Vanthienen, & Baesens (2013)	2013	Other	Process mining
Kuna, Gartia-Martinez, & Villatoro (2014)	2014	Other	Other (Data mining)
Zerbino, Aloini, Dulmin, & Mininno (2018)	2018	Other	Process mining
Jans & Hosseinpour (2019)	2019	Other	Process mining & Other (Data mining)
Yesmin & Carter (2020)	2020	Other	ML
Khan et al. (2021)	2021	Other	ML & NLP
Chen et al. (2021)	2021	Other	AI

The following subsections provide thematic analyses of the eight papers. These eight papers have been analyzed based on the eight categories given in Table 5. The literature analysis results are presented based on categories and in separate subsections titled the same as the category names except for the categories of validation and evaluation, which have been grouped together under the subsection titled Technique Validation and Evaluation.

4.1. Audit Types

Analysis of the papers shows none of the articles to focus on implementing new techniques into IT audits. The aim was to find the most frequently conducted IT audit types in the literature identified by the keywords of IT process audit, application audit, and ITGC. Other IT audit types and the specified IT audit types were not found in the literature. However, audit types other than IT audits (e.g., internal audit, business process audit) were found in the literature and identified in Table 6 as "Other;" these were then analyzed to determine their audit subjects. As a result, business process audits are dominant in the final set.

In some papers, the authors named the subject of the audit work as an internal audit (Chen et al., 2021; Jans &

Hosseinpour, 2019), information systems audit (Kuna et al., 2014; Zerbino et al., 2018), or business process audit (Caron et al., 2013; Khan et al., 2021; Rozinat & van der Aals, 2008). In these articles, internal audits and information systems refer to business process audits. These articles focused on different business processes such as innovation management (Khan et al., 2021), claim-handling processes (Caron et al., 2013) and freight export port processes (Zerbino et al., 2018).

One article focused on a privacy audit (Yesmin & Carter, 2020). Privacy audits can be evaluated as a business process audit. However, the article's auditing process involved technical auditing, not business process auditing. The methods and techniques studied in these papers, which had been predominantly subjected to business process audits, were evaluated, and these methods and techniques are considered suitable for adapting to IT auditing.

4.2. IT Audit Stages

The papers were analyzed based on the audit stages on which they were focused. Except for Chen et al.'s (2021) study, all papers focused on the testing stage. Chen et al. (2021) focused on the planning stage of an IT audit with the aim of identifying critical risks in an audited company as planning for an internal control audit. They developed a hybrid decision-making model that involved data exploration.

4.3. New Techniques

Several techniques were used in the papers, including process mining, ML, NLP, and AI. Most papers developed an approach with process mining (Caron et al., 2013; Jans & Hosseinpour, 2019; Rozinat & van der Aalst, 2008; Zerbino et al., 2018). Jans and Hosseinpour (2019) combined process mining and data mining. AI (Chen et al., 2021), ML techniques (Khan et al., 2021; Yesmin & Carter, 2020), NLP techniques (Khan et al., 2021), and data mining techniques (Kuna et al., 2014) were employed within the testing stages of audits. The testing stage constitutes the main work and mostly entails repetitive labor-intensive work.

The four process mining articles (Caron et al., 2013; Jans & Hosseinpour, 2019; Rozinat & van der Aalst, 2008; Zerbino et al., 2018) were method articles that developed a method using process mining. These papers employed conformance checking algorithms to determine the derivations of processes from as-is processes. The techniques performed in these papers also correspond to the testing stage.

Kuna et al. (2014) employed data mining techniques to detect anomalous actions within processes of an academic management system, purchase management system, and inventory control system. Yesmin and Carter (2020) employed supervised ML algorithms in order to determine inappropriate access to a hospital system. Khan et al. (2021) applied NLP and ML algorithms to text-based evidence to automatically review innovation management systems' compliance with the ISO56002 innovation management system standard. Therefore, anomalous data detection, inappropriate access identification, and compliance check of text-based audit evidence are the techniques that were performed within the testing step.

4.4. Algorithms

The studies employed different data analytics algorithms and combinations of algorithms. Because four process mining articles are found in the final set, the common approach among these studies can be more clearly identified. Conformance checking algorithms specially developed for compliance analysis were commonly used in the process mining articles (Caron et al., 2013; Jans & Hosseinpour, 2019; Rozinat & van der Aalst, 2008; Zerbino et al., 2018). Process discovery algorithms (e.g., Alpha algorithm, heuristic miner, fuzzy miner) are first used in process mining algorithms to determine the executed processes in the event log. A conformance checking algorithm is then used to test the compliance of the process executions to an as-is process model; this was determined using the process discovery algorithms.

Zerbino et al. (2018) analyzed data in two steps. Before conducting compliance tests, they first created a control-flow model. The control flow model was developed with process discovery techniques. They used a fuzzy miner algorithm for the process discovery. Once they had the control-flow model, they employed a conformance checking algorithm to identify non-conformances within the process. Rozinat and van der Aalst (2008) employed two different approaches to the conformance checking algorithm in their method that overcomes specific problems regarding log replay analysis. The first algorithm solves the problem of not being fired from an invisible task due to no related log event. The second algorithm was created to solve duplicate tasks resulting from a model task and log event mapping. Jans and Hosseinpour (2019) presented a different approach. They proposed applying both process mining and data mining algorithms within

their framework. For their first step, they used the most suitable process discovery algorithm from among all the process discovery algorithms (e.g., alpha algorithm, heuristic miner, fuzzy miner, genetic miner, inductive miner) to identify executed processes. Then they ran a compliance checking algorithm to determine deviations in the process execution compared to the as-is process. After completing the process mining step and determining the deviations, they grouped the deviations using data mining classification algorithms. Their study combined two analytic approaches, while Caron et al. (2013) did not specify the algorithm in their study.

Additionally, no specific algorithm was seen to have been frequently used in the articles that use ML, NLP, AI, or data mining algorithms. Each article employed a different algorithm. However, two articles (Kuna et al., 2014; Yesmin & Carter, 2020) used algorithms to classify identified deficiencies. Firstly, Yesmin and Carter (2020) implemented supervised ML algorithms to proactively identify unauthorized employee access to the hospital system. Their approach also enabled the categorization of detected access into three groups: appropriate access, unexplained access, and flagged access. They also provide customized explanations for detected access. Kuna et al.'s (2014) study implemented various data mining algorithms to determine anomalous data in audit logs. They first employed LOF and DBSCAN algorithms for outlier detection within audit logs and then applied the C4.5, Bayesian Network (BN), and PART algorithms to classify the detected outliers.

Khan et al.'s (2021) study is the only one to have analyzed textual data while employing NLP. They applied NLP techniques such as lemmatization and stop-word removal for data cleansing, the clustering method for closeness and associations between words, and word frequency for visualizing textual data.

Meanwhile, Chen et al. (2021) applied a hybrid model that included various algorithms for identifying risks in internal auditing. They mentioned their integrated hybrid model to contain rule generation techniques based on AI algorithms such as support vector machines, ensemble learning, decision trees, and multiple rule-based decision-making techniques. However, they did not explain how they implemented the AI algorithms.

4.5. Methodology and Framework

This category presents the methodologies and frameworks provided in the articles. The authors provided their methodology, framework, or only the steps they followed while implementing the algorithm in their articles (Caron et al., 2013; Jans & Hosseinpour, 2019; Kuna et al., 2014; Rozinat & van der Aalst, 2008; Zerbino et al., 2018). Some authors developed an audit tool in their study (Khan et al., 2021; Yesmin & Carter, 2020). Only one study (Chen et al., 2021) did not explain their methodology.

Zerbino et al. (2018) proposed a process mining-enabled methodology for an information systems audit containing five steps: justification and planning, data extraction, control-flow model construction, model enrichment, and conformance checking. They used their methodology in an attempt to find a solution for ineffectiveness at detecting non-conformances, frauds, and abuses. Caron et al. (2013) proposed a technique for compliance checking due to the inadequacy of existing techniques in process discovery and visualization, conformance checking and delta analysis, and logic-based property verification with regard to compliance checking. Their proposed technique involved comprehensive rule-based compliance checking using process mining. The technique also contained the architecture for compliance checking, two dimensions of business rule taxonomy, rule restriction, formal specification of rule patterns, and definition of the scope of rules. Jans and Hosseinpour (2019) provided a transactional verification framework for continuous auditing. Their framework contained six phases from building an event log to identifying deviations in transactions by combining process mining and data mining techniques. Rozinat and van der Aalst (2008) proposed a conformance checker technique to extend current techniques. Their conformance checker is an incremental technique for validating compliance in a process model and an event log. The conformance checker provides new metrics with precise definitions on how to implement these metrics. Meanwhile, Kuna et al. (2014) presented no framework for the method they applied. However, they did provide their procedure clearly, including how to read data from the database, preprocess the data, apply outlier detection rules, and classify detected outliers.

Yesmin and Carter (2020) provide neither a defined methodology or framework for applying their technique. Nevertheless, they built an automatic privacy auditing tool, including their solution. They also presented a framework for evaluating the outputs of the employed technique. Khan et al. (2021) developed an AI-based audit tool with a fuzzy front-end of the ISO56002 Innovation Management System Standard. They also presented no framework but did provide their procedure, which involved the specific tasks of preprocessing input data sets, cluster mapping and visualizing the standard, visualizing the standard's cluster map with audit transcripts, normalizing Map 2, and automated scoring by creating a reverse map.

4.6. Validation and Evaluation of the Techniques

This section analyzes whether the studies included any validation or performance evaluation for their proposed models. The articles had different approaches for validating and evaluating their techniques. First, Zerbino et al. (2018) applied the proposed process mining-enabled methodology to the freight export port process managed by a port community system (PCS). They did not perform a numerical evaluation of the technique's performance but did verbally assess the technique, mentioning the drawbacks of their technique in the paper. These drawbacks are related to being familiar with the process and the information systems from which the input data for the process mining methods come with regard to determining deviations in processes.

Two application examples were provided in Rozinat and van der Aalst's (2008) study. The first was applied to the administrative processes of a municipality in the Netherlands. The other application analyzed the conformance of web service behavior. The technique provided in Rozinat and van der Aalst's study contained metrics for evaluating conformance checks, and they also approached how to evaluate the metrics.

Yesmin and Carter's (2020) approach was an application of a method rather than the creation of one. As such, they did not require validation for any method. However, they demonstrated the benefit of their method's results. In order to accomplish this, they provided a validation framework that included various scenarios about user behavior on the privacy auditing tool they had created. All scenarios were created from different groups of employees' perspectives and corrected by experts. Their framework enabled them to test their results with different use cases. Yesmin and Carter's evaluated how well their model determined categories and how it had evolved itself to make more accurate classifications. They explained that they had achieved a statistically significant learning curve as a result.

Kuna et al. (2014) implied the solution to three different systems: the academic management system of a university, the purchase management system of a local government, and the inventory control system of a wholesaler. They also provided an evaluation that compared false positives with the efficacy values of the results from the three applications. Moreover, to demonstrate their developed method's effectiveness and efficiency, Khan et al. (2021) compared the compliance scores of audits performed by the developed AI tool and manually conducted audits of five organizations. Khan et al. evaluated the efficiency and effectiveness metrics to compare the automated and manual audits.

Meanwhile, Caron et al. (2013) mentioned testing the proposed technique in claims-handling processes in the insurance industry. However, they did not provide any evaluation results. Additionally, Jans and Hosseinpour (2019) and Chen et al. (2021) neither evaluated nor conducted any validation for their proposed techniques.

4.7. Data Type

The structure and source of input data varied in the articles based on the employed techniques. Apart from Khan et al.'s (2021) study, all the others (Caron et al., 2013; Chen et al., 2021; Jans & Hosseinpour, 2019; Kuna et al., 2014; Rozinat & van der Aalst, 2008; Yesmin & Carter, 2020; Zerbino et al., 2018) used structured data.

Khan et al. (2021) used unstructured data, which was also text-based. They also used two different sets of data as inputs for the audit. The first were transcripts from the interviews about the conducted innovation management process in organizations. The other were the requirements for fuzzy front-end with regard to the ISO56002 Innovation Management System Standard in text format.

The articles that employed process mining as a technique used structured data. These data included event log data belonging to the executed process. The fundamental attributes of event log data are case ID as a unique identifier, event name as activity, and timestamp defining when the event was executed. Process mining studies in the final set used event log data belonging to different processes such as the administrative processes of a municipality in the Netherlands and processes on a web service (Rozinat & van der Aalst, 2008), claim-handling process (Caron et al., 2013), and freight export port process (Zerbino et al., 2018). Event log data were generally obtained from the system on which the processes run and on which event logs are simultaneously stored.

The other papers that employed ML and AI used structured data in their analysis (Chen et al., 2021; Kuna et al., 2014; Yesmin & Carter, 2020). Yesmin and Carter (2020) collected user and patient data from different systems such as human resources and scheduling software. Kuna et al. (2014) used audit logs from three different systems: academic management system, purchase management system, and inventory control system. Chen et al.'s (2021) data sources were the most varied among all the papers. They collected data using a questionnaire for risk identification, in which the participants were domain experts.

5. DISCUSSION

This study has performed an SLR to present an overview of papers that had implemented new technologies and techniques (i.e., AI, ML, NLP, predictive analytics, and process mining) into IT audits. Even if some papers had IT audits as a subject, no paper apparently focused on using new technology and techniques, especially in IT audits. However, the literature had some papers that provided a method for applying certain techniques to related audit areas, such as business process and information systems audits. Techniques applied in similar areas have the potential to be implemented in IT audits. Therefore, this study aimed at presenting an overview of these papers. Additionally, due to the study being the first one in the context of IT auditing, no earlier studies conducted in the literature could be referred to in this context.

The first finding from this literature review study is the lack of academic studies focused on implementing techniques into IT audits. In the research conducted among IT auditors by ISACA (2019a), however, 26% of the IT auditors claimed to currently use process mining and 23% to use predictive analytics. As such, the gap between practice and academic studies in terms of the technologies used in IT audit can be revealed.

As was given in the results, eight papers were identified in total. These papers included techniques adaptable to IT audits. One paper each contained the following techniques: artificial intelligence (Chen et al., 2021), machine learning (Yesmin & Carter, 2020), natural language processing (Khan et al., 2021), and data mining (Kuna et al., 2014). Meanwhile, four papers employed process mining (Caron et al., 2013; Jans & Hosseinpour, 2019; Rozinat & van der Aalst, 2008; Zerbino et al., 2018). Just as the IT auditors had mentioned their frequent usage of process mining in ISACA's (2019a) study, the most studied technique was process mining in the studies analyzed here. These papers also employed a combination of these techniques, such as process mining with data mining (Jans & Hosseinpour, 2019) or NLP with ML (Khan et al., 2021).

Seven papers (Caron et al., 2013; Jans & Hosseinpour, 2019; Khan et al., 2021; Kuna et al., 2014; Rozinat & van der Aalst, 2008; Yesmin & Carter, 2020; Zerbino et al., 2018) provided a methodology, framework, or procedure to explain what they had performed or proposed to apply as a technique. These papers are helpful for practitioners and academicians. Most of the papers had validated their proposed techniques with either actual data (Caron et al., 2013; Khan et al., 2021; Rozinat & van der Aalst, 2008; Yesmin & Carter, 2020; Zerbino et al., 2018) or artificial data (Kuna et al., 2014; Rozinat & van der Aalst, 2008). However, the authors from two studies did not validate their proposed methods (Chen et al., 2021; Jans & Hosseinpour, 2019). Additionally, providing performance evaluations in a paper, as had occurred in five of them (Khan et al., 2021; Kuna et al., 2014; Rozinat & van der Aalst, 2008; Yesmin & Carter, 2020; Zerbino et al., 2018), makes assessing the suggested technique more understandable for audiences.

From the audit perspective, seven papers (Caron et al., 2013; Jans & Hosseinpour, 2019; Khan et al., 2021; Kuna et al., 2014; Rozinat & van der Aalst, 2008; Yesmin & Carter, 2020; Zerbino et al., 2018) out of eight had clearly provided a solution for the testing stage of the audit. Only one study (Chen et al., 2021) handled the planning stage. However, considerable potential is found for applying different techniques to other audit stages. Whether the input data of the technique is structured or unstructured changed based on the technique. The input data for the techniques corresponded to the audit evidence. Thus, any kind of audit evidence can be used in the various techniques.

This paper can support researchers studying new technologies in IT auditing by providing insights. The results from this study may also help audit teams when they implement and develop new technological solutions for their audit processes. In addition, the main reasons are understood as to why studies apply new audit techniques. According to the authors, these reasons are to provide a continuous audit approach (Caron et al., 2013; Jans & Hosseinpour, 2019), to automate the audit work (Khan et al., 2021; Kuna et al., 2014; Zerbino et al., 2018), and to perform online or real-time audits (Zerbino et al., 2018) in organizations. These reasons justify the need to apply new techniques that add effectiveness and efficiency to the audit processes of today's increased audit workload. These reasons also support the aim of this literature review work. The studies can be concluded to have developed new approaches, mainly process mining, data mining, ML, and NLP techniques or some combination of these techniques and to have achieved effective and efficient audits by conducting continuous, automated, or online audit works. Meanwhile, this study has identified the gap between practice and academic studies in the field of IT audits. Closing this gap would expand the IT audit literature and also provide an example of how the related technologies are used. Both academicians and practitioners can benefit from closing the gap. One method developed by any company specific to their needs can be transformed into an academic study by providing a performance evaluation or even the insights of the involved parties who use these new approaches or have develop these techniques to understand the adaptation process through their experiences.

As mentioned before, after not finding studies related to implementing new technologies into IT audits, the articles that used these technologies for other audit areas were examined according to their applicability to IT audits. The

existing literature may have more studies with the potential for use in IT audits. These articles were not fully covered in this study. Therefore, future systematic literature review studies can expand their search using additional keywords that can be selected according to the relevance of the area to IT audits.

Additionally, future literature review studies could specify research questions to specific IT domains or processes. Specific evidence types could also be determined, such as third-party contracts, previous years' audit reports, deployment logs of applications, or request records from request management systems.

Moreover, conference proceedings can also be evaluated. The current study could be enhanced with other emerging technologies such as blockchain and big data analytics. New technologies could also be enhanced with data mining because of its relevance.

This study also encountered other limitations with regard to the searched databases. Conceptualization was performed based on the articles published in ISACA's database; however, any advanced searches were not conducted on this database, because ISACA's publication database does not provide an advanced search function that would enable a systematic search, unlike the other databases. A review of the ISACA database would be beneficial.

6. FUTURE STUDIES AND RECOMMENDATIONS for IT AUDITS

Studies on IT auditing enriched with AI, ML, NLP, predictive analytics, and process mining are not available in the literature. Studies and real-life applications should focus on having continuous and automated auditing while closing the gap between practice and academic studies in the IT audit area.

Based on the reviewed literature, this study suggests that, in order to emphasize the value of implementing new techniques into the auditing process, validation and evaluation of proposed methods should occur, as many papers did to show the applicability and effectiveness of their methods. In addition, the method, framework, or steps a proposed technique follows should be given to enlighten people who are considering adopting the technique. Providing information regarding input data type (i.e., audit evidence) is also significant in clarifying technique details.

IT processes subject to audits such as access management, project management, or processes within the software development life cycle are suitable for assessment using process mining. A log of users' activities within these processes can be extracted from related systems. An IT process that starts at one system and ends up at another can be also analyzed with process mining techniques. Process mining can also be combined with machine learning or data mining techniques, such as the ones Jans and Hosseinpour (2019) performed for more sophisticated evaluations of IT audits.

As Khan et al. (2021) proposed, an NLP technique can be employed for conducting IT audits based on standards, frameworks, or regulations such as ISO27001, Cobit 4.1, SOX, or SSAE.

Most of the studies had provided a solution for the testing stage of an IT audit. As such, the potential could exist for focusing on the automatization of areas in the planning, reporting, and follow-up stages of IT audits rather than just focusing on the testing stage.

Additional suggestions have been intuitively made as follows. NLP techniques, especially in the planning stage, could be employed because textual data are generally used in that stage for determining the scope of audits and controls. For example, summarizing the previous year's audit reports, being one of the inputs of the planning stage, can be performed using NLP techniques. An audit's scope, risk, and controls can be identified for auditing an IT service provider using third-party agreements belonging to the provider with regard to outsourced IT services, and NLP can be used for this analysis. While concluding an audit, machine learning algorithms can also help determine the classification of deficiencies and create an audit opinion based on the audit results.

Peer Review: Externally peer-reviewed.

Author Contributions: Conception/Design of Study- N.S.T.; N.T.; Data Acquisition- N.S.T.; Data Analysis/Interpretation- N.S.T.; N.T.; Drafting Manuscript- N.S.T.; Critical Revision of Manuscript- N.S.T.; N.T.; Final Approval and Accountability- N.S.T.; N.T.

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.

ORCID IDs of the authors / Yazarların ORCID ID'leri

Nur Sena Tanrıverdi 0000-0002-4728-988X

Nazım Taşkın 0000-0002-5327-9012

REFERENCES

- Alexiou, S. (2019). Trends, challenges and strategies for effective audit in a rapidly changing landscape. *ISACA Journal*, 6. Retrieved from <https://www.isaca.org/resources/isaca-journal/issues/2019/volume-6/trends-challenges-and-strategies-for-effective-audit-in-a-rapidly-changing-landscape>
- Barta, G. (2018). The increasing role of IT auditors in financial audit: risks and intelligent answers. *Business, Management and Education*, 16(1), 81-93. <https://doi.org/10.3846/bme.2018.2142>
- Carlin, A., & Gallegos, F. (2007). IT audit: A critical business process. *Computer*, 40(7), 87-89. <https://doi.org/10.1109/MC.2007.246>
- Caron, F., Vanthienen, J., & Baesens, B. (2013). Comprehensive rule-based compliance checking and risk management with process mining. *Decision Support Systems*, 54(3), 1357-1369. <https://doi.org/10.1016/j.dss.2012.12.012>
- Chen, F. H., Hsu, M. F., & Hu, K. H. (2021). Enterprise's internal control for knowledge discovery in a big data environment by an integrated hybrid model. *Information Technology and Management*, 23, 213-231. <https://doi.org/10.1007/s10799-021-00342-8>
- Dzurani, A. C., & Mălăescu, I. (2016). The current state and future direction of IT audit: challenges and opportunities. *Journal of Information Systems*, 30(1), 7-20. <https://doi.org/10.2308/isisys-51315>
- Flores, I. G., & Riquenes, J. R. (2020). Audit 2.0, a perspective for its execution in the business environment using process mining techniques. *Vivat Academia*, 23(150), 47-57. <http://doi.org/10.15178/va.2020.150.47-57>
- Hult, G. T. M., Reimann, M., & Schilke, O. (2009). Worldwide faculty perceptions of marketing journals: rankings, trends, comparisons, and segmentations. *Global Edge Business Review*, 3(3), 1-23.
- ISACA (2019a). Future of IT Audit Report. Schaumburg, USA.
- ISACA (2019b). CISA Review Manual. 27th Edition. Schaumburg, USA.
- ISACA, & Protiviti Inc. (2019). 2019 IT Audit Benchmarking Study. Today's toughest challenges in IT audit: Tech Partnerships, Talent, Transformation.
- ISACA. (n.d). History [Web page]. Retrieved from <https://www.isaca.org/why-isaca/about-us/history>
- Jans, M., & Hosseinpour, M. (2019). How active learning and process mining can act as continuous auditing catalyst. *International Journal of Accounting Information Systems*, 32, 44-58. <https://doi.org/10.1016/j.accinf.2018.11.002>
- Kitchenham, B., Brereton, O. P., Budgen, D., Turner, M., Bailey, J., & Linkman, S. (2009). Systematic literature reviews in software engineering - A systematic literature review. *Information and Software Technology*, 51(1), 7-15. <https://doi.org/10.1016/j.infsof.2008.09.009>
- Khan, R., Adi, E., & Hussain, O. (2021). AI-based audit of fuzzy front end innovation using ISO56002. *Managerial Auditing Journal*, 36(4), 564-590. <https://doi.org/10.1108/MAJ-03-2020-2588>
- Krieger, F., Drews, P., & Velte, P. (2021). Explaining the (non-) adoption of advanced data analytics in auditing: A process theory. *International Journal of Accounting Information Systems*, 41, 100511. <https://doi.org/10.1016/j.accinf.2021.100511>
- Kuna, H. D., García-Martínez, R., & Villatoro, F. R. (2014). Outlier detection in audit logs for application systems. *Information Systems*, 44, 22-33. <http://dx.doi.org/10.1016/j.is.2014.03.001>
- Manhanga, L. (2020). *Data analytics in information systems (IS) auditing: An examination of the cost-effectiveness of the use of data analytics in information systems auditing* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 27997995)
- Mendez, H. (2020). Innovation in the IT audit process: the IT semantic audit models audit segments using semantic graphs. *ISACA Journal*, 1. Retrieved from <https://www.isaca.org/resources/isaca-journal/issues/2020/volume-1/innovation-in-the-it-audit-process>
- Rozinat, A., & Van der Aalst, W. M. (2008). Conformance checking of processes based on monitoring real behavior. *Information Systems*, 33(1), 64-95. <https://doi.org/10.1016/j.is.2007.07.001>
- Schumann, G., & Gómez, J. M. (2021). Natural language processing in internal auditing - a structured literature review. *AMCIS 2021 Proceedings*. https://aisel.aisnet.org/amcis2021/sig_acctinfosystem_asys/sig_acctinfosystem_asys/1
- Tranfield, D., Denyer, D., & Smart, P. (2003). Towards a methodology for developing evidence-informed management knowledge by means of systematic review. *British Journal of Management*, 14(3), 207-222. <https://doi.org/10.1111/1467-8551.00375>
- Yesmin, T., & Carter, M. W. (2020). Evaluation framework for automatic privacy auditing tools for hospital data breach detections: A case study. *International Journal of Medical Informatics*, 138, 104123. <https://doi.org/10.1016/j.ijmedinf.2020.104123>

Zerbino, P., Aloini, D., Dulmin, R., & Mininno, V. (2018). Process-mining-enabled audit of information systems: Methodology and an application. *Expert Systems with Applications*, 110, 80-92. <https://doi.org/10.1016/j.eswa.2018.05.030>

How cite this article

Tanriverdi, N.S., Taskin, N. (2023). A systematic literature review for new technologies in IT audit. *Acta Infologica*, 7(2), 396-408. <https://doi.org/10.26650/acin.1142281>

DESCRIPTION

Acta Infologica (ACIN) is the publication of Informatics Department of the Istanbul University. It is an open access, scholarly, peerreviewed journal published biannually in June and December. The journal was founded in 2017.

AIM AND SCOPE

ACIN aims to contribute to the scientific community interested in the field of informatics and aims to provide a platform for researchers exploring issues based on the concepts of data-information-knowledge, information and communication technologies and applications. The journal welcomes multidisciplinary studies regarding the field as well.

The areas of study covered in the scope of ACIN are in below;

Intelligent Systems

Information Security and Law

Knowledge Management

Computer Networks

Computer Architecture

Information Systems

Bioinformatics

Geographic Information Systems

E-Applications

Internet Technologies

Decision Support Systems and Business Intelligence

Microcontroller and Applications

Mobile Systems

Modeling and Optimization

Project Management

Social and Digital Media

Data Mining

Database Systems

Artificial Intelligence and Machine Learning

Software Engineering

EDITORIAL POLICIES AND PEER REVIEW PROCESS

Publication Policy

The subjects covered in the manuscripts submitted to the Journal for publication must be in accordance with the aim and scope of the journal. The journal gives priority to original research papers submitted for publication.

General Principles

Only those manuscripts approved by its every individual author and that were not published before in or sent to another journal, are accepted for evaluation.

Submitted manuscripts that pass preliminary control are scanned for plagiarism using iThenticate software. After plagiarism check, the eligible ones are evaluated by editor-in-chief for their originality, methodology, the importance of the subject covered and compliance with the journal scope.

Short presentations that took place in scientific meetings can be referred if indicated in the article. The editor hands over the papers matching the formal rules to at least two national/international referees for evaluation and gives green light for publication upon modification by the authors in accordance with the referees' claims. Changing the name of an author (omission, addition or order) in papers submitted to the Journal requires written permission of all declared authors. Refused manuscripts and graphics are not returned to the author.

Open Access Statement

The journal is an open access journal and all content is freely available without charge to the user or his/her institution. Except for commercial purposes, users are allowed to read, download, copy, print, search, or link to the full texts of the articles in this journal without asking prior permission from the publisher or the author. This is in accordance with the BOAI definition of open access.

The open access articles in the journal are licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) license.

Article Processing Charge

All expenses of the journal are covered by the Istanbul University. Processing and publication are free of charge with the journal. There is no article processing charges or submission fees for any submitted or accepted articles.

Peer Review Process

Only those manuscripts approved by its every individual author and that were not published before in or sent to another journal, are accepted for evaluation.

Submitted manuscripts that pass preliminary control are scanned for plagiarism using iThenticate software. After plagiarism check, the eligible ones are evaluated by Editor-in-Chief for their originality, methodology, the importance of the subject covered and compliance with the journal scope. Editor-in-Chief evaluates manuscripts for their scientific content without regard to ethnic origin, gender, sexual orientation, citizenship, religious belief or political philosophy of the authors and ensures a fair double-blind peer review of the selected manuscripts.

The selected manuscripts are sent to at least two national/international external referees for evaluation and publication decision is given by Editor-in-Chief upon modification by the authors in accordance with the referees' claims.

Editor-in-Chief does not allow any conflicts of interest between the authors, editors and reviewers and is responsible for final decision for publication of the manuscripts in the Journal.

Reviewers' judgments must be objective. Reviewers' comments on the following aspects are expected while conducting the review.

- Does the manuscript contain new and significant information?
- Does the abstract clearly and accurately describe the content of the manuscript?
- Is the problem significant and concisely stated?
- Are the methods described comprehensively?
- Are the interpretations and conclusions justified by the results?
- Is adequate references made to other Works in the field?
- Is the language acceptable?

Reviewers must ensure that all the information related to submitted manuscripts is kept as confidential and must report to the editor if they are aware of copyright infringement and plagiarism on the author's side.

A reviewer who feels unqualified to review the topic of a manuscript or knows that its prompt review will be impossible should notify the editor and excuse himself from the review process.

The editor informs the reviewers that the manuscripts are confidential information and that this is a privileged interaction. The reviewers and editorial board cannot discuss the manuscripts with other persons. The anonymity of the referees is important.

COPYRIGHT NOTICE

Authors publishing with the journal retain the copyright to their work licensed under the Creative Commons Attribution-NonCommercial 4.0 International license (CC BY-NC 4.0) (<https://creativecommons.org/licenses/by-nc/4.0/>) and grant the Publisher non-exclusive commercial right to publish the work. CC BY-NC 4.0 license permits unrestricted, non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

PUBLICATION ETHICS AND PUBLICATION MALPRACTICE STATEMENT

Acta Infologica (ACIN) is committed to upholding the highest standards of publication ethics and pays regard to Principles of Transparency and Best Practice in Scholarly Publishing published by the Committee on Publication Ethics (COPE), the Directory of Open Access Journals (DOAJ), to access the Open Access Scholarly Publishers Association (OASPA), and the World Association of Medical Editors (WAME) on <https://publicationethics.org/resources/guidelines-new/principles-transparency-and-best-practice-scholarly-publishing> All parties involved in the publishing process (Editors, Reviewers, Authors and Publishers) are expected to agree on the following ethical principles.

All submissions must be original, unpublished (including as full text in conference proceedings), and not under the review of any other publication synchronously. Each manuscript is reviewed by one of the editors and at least two referees under double-blind peer review process. Plagiarism, duplication, fraud authorship/denied authorship, research/data fabrication, salami slicing/salami publication, breaching of copyrights, prevailing conflict of interest are unethical behaviors.

All manuscripts not in accordance with the accepted ethical standards will be removed from the publication. This also contains any possible malpractice discovered after the publication. In accordance with the code of conduct we will report any cases of suspected plagiarism or duplicate publishing.

RESEARCH ETHICS

Acta Infologica (ACIN) adheres to the highest standards in research ethics and follows the principles of international research ethics as defined below. The authors are responsible for the compliance of the manuscripts with the ethical rules.

- Principles of integrity, quality and transparency should be sustained in designing the research, reviewing the design and conducting the research.
- The research team and participants should be fully informed about the aim, methods, possible uses and requirements of the research and risks of participation in research.
- The confidentiality of the information provided by the research participants and the confidentiality of the respondents should be ensured. The research should be designed to protect the autonomy and dignity of the participants.
- Research participants should participate in the research voluntarily, not under any coercion.
- Any possible harm to participants must be avoided. The research should be planned in such a way that the participants are not at risk.
- The independence of research must be clear; and any conflict of interest or must be disclosed.
- In experimental studies with human subjects, written informed consent of the participants who decide to participate in the research must be obtained. In the case of children and those under wardship or with confirmed insanity, legal custodian's assent must be obtained.
- If the study is to be carried out in any institution or organization, approval must be obtained from this institution or organization.
- In studies with human subject, it must be noted in the method's section of the manuscript that the informed consent of the participants and ethics committee approval from the institution where the study has been conducted have been obtained.

AUTHOR RESPONSIBILITIES

It is authors' responsibility to ensure that the article is in accordance with scientific and ethical standards and rules. And authors must ensure that submitted work is original. They must certify that the manuscript has not previously been published elsewhere or is not currently being considered for publication elsewhere, in any language. Applicable copyright laws and conventions must be followed. Copyright material (e.g. tables, figures or extensive quotations) must be reproduced only with appropriate permission and acknowledgement. Any work or words of other authors, contributors, or sources must be appropriately credited and referenced.

All the authors of a submitted manuscript must have direct scientific and academic contribution to the manuscript. The

author(s) of the original research articles is defined as a person who is significantly involved in “conceptualization and design of the study”, “collecting the data”, “analyzing the data”, “writing the manuscript”, “reviewing the manuscript with a critical perspective” and “planning/conducting the study of the manuscript and/or revising it”. Fund raising, data collection or supervision of the research group are not sufficient roles to be accepted as an author. The author(s) must meet all these criteria described above. The order of names in the author list of an article must be a co-decision and it must be indicated in the Copyright Agreement Form. The individuals who do not meet the authorship criteria but contributed to the study must take place in the acknowledgement section. Individuals providing technical support, assisting writing, providing a general support, providing material or financial support are examples to be indicated in acknowledgement section.

All authors must disclose all issues concerning financial relationship, conflict of interest, and competing interest that may potentially influence the results of the research or scientific judgment.

When an author discovers a significant error or inaccuracy in his/her own published paper, it is the author’s obligation to promptly cooperate with the Editor to provide retractions or corrections of mistakes.

RESPONSIBILITY FOR THE EDITOR AND REVIEWERS

Editor-in-Chief evaluates manuscripts for their scientific content without regard to ethnic origin, gender, sexual orientation, citizenship, religious belief or political philosophy of the authors. He/She provides a fair double-blind peer review of the submitted articles for publication and ensures that all the information related to submitted manuscripts is kept as confidential before publishing.

Editor-in-Chief is responsible for the contents and overall quality of the publication. He/She must publish errata pages or make corrections when needed.

Editor-in-Chief does not allow any conflicts of interest between the authors, editors and reviewers. Only he has the full authority to assign a reviewer and is responsible for final decision for publication of the manuscripts in the Journal.

Reviewers must have no conflict of interest with respect to the research, the authors and/or the research funders. Their judgments must be objective.

Reviewers must ensure that all the information related to submitted manuscripts is kept as confidential and must report to the editor if they are aware of copyright infringement and plagiarism on the author’s side.

A reviewer who feels unqualified to review the topic of a manuscript or knows that its prompt review will be impossible should notify the editor and excuse himself from the review process.

The editor informs the reviewers that the manuscripts are confidential information and that this is a privileged interaction. The reviewers and editorial board cannot discuss the manuscripts with other persons. The anonymity of the referees must be ensured. In particular situations, the editor may share the review of one reviewer with other reviewers to clarify a particular point.

MANUSCRIPT ORGANIZATION

LANGUAGE

The publication language of the journal is English.

Manuscript Organization and Submission

All correspondence will be sent to the first-named author unless otherwise specified. Manuscript is to be submitted online via dergipark.org.tr/login that can be accessed at <http://acin.istanbul.edu.tr> and it must be accompanied by a title page specifying the article category (i.e. research article, review etc.) and including information about the manuscript (see the Submission Checklist) and cover letter to the editor. Manuscripts should be prepared in Microsoft Word 2003 and upper versions. In addition, Copyright Agreement Form that has to be signed by all authors must be submitted.

1. Use ACIN article document as a template if you are using Microsoft Word 6.0 or upper versions. Otherwise, use this document as an instruction set.

2. The first letters of words in the article title should be written in uppercase; the entire title should not be capitalized. Avoid writing formulas in the title. Do not write “(Invited)” or similar expressions in the title.
3. The abstract must be between 150–250 words and written as one paragraph. It should not contain displayed mathematical equations or tabular material. The abstract should include three to five different keywords or phrases, as this will help readers to find it. It is important to avoid over-repetition of such phrases as this can result in a page being rejected by search engines. Ensure that your abstract reads well and is grammatically correct.
4. Underneath the abstracts, a minimum of 3 and a maximum of 5 keywords that inform the reader about the content of the study should be specified. Keywords must be defined by taking into consideration authorities like “TR Dizin Anahtar Terimler Listesi”, “Medical Subject Headings”, “CAB Theasarus”, “JISCT”, “ERIC”, etc.
5. The manuscripts should contain mainly these components: title, abstract and keywords; sections, references, tables and figures.
6. A title page including author information must be submitted together with the manuscript. The title page is to include fully descriptive title of the manuscript and, affiliation, title, e-mail address, ORCID, postal address, phone, mobile phone and fax number of the author(s) (see The Submission Checklist).
7. References should be prepared as APA 6th edition.

REFERENCES

Reference Style and Format

Acta Infologica (ACIN) complies with APA (American Psychological Association) style 6th Edition for referencing and quoting. For more information:

- American Psychological Association. (2010). Publication manual of the American Psychological Association (6th ed.). Washington, DC: APA.
- <http://www.apastyle.org>

Accuracy of citation is the author’s responsibility. All references should be cited in text. Reference list must be in alphabetical order. Type references in the style shown below.

Citations in the Text

Citations must be indicated with the author surname and publication year within the parenthesis. If more than one citation is made within the same paranthesis, separate them with (;).

Samples:

More than one citation;

(Esin, et al., 2002; Karasar, 1995)

Citation with one author;

(Akyolcu, 2007)

Citation with two authors;

(Saymer & Demirci, 2007)

Citation with three, four, five authors;

First citation in the text: (Ailen, Ciembrune, & Welch, 2000) Subsequent citations in the text: (Ailen, et al., 2000)

Citations with more than six authors;

(Çavdar, et al., 2003)

Citations in the Reference

All the citations done in the text should be listed in the References section in alphabetical order of author surname without numbering. Below given examples should be considered in citing the references.

Basic Reference Types

Book

a) Turkish Book

Karasar, N. (1995). *Araştırmalarda rapor hazırlama* (8th ed.) [Preparing research reports]. Ankara, Turkey: 3A Eğitim Danışmanlık Ltd.

b) Book Translated into Turkish

Mucchielli, A. (1991). *Zihniyetler* [Mindsets] (A. Kotil, Trans.). İstanbul, Turkey: İletişim Yayınları.

c) Edited Book

Ören, T., Üney, T., & Çölkesen, R. (Eds.). (2006). *Türkiye bilişim ansiklopedisi* [Turkish Encyclopedia of Informatics]. İstanbul, Turkey: Papatya Yayıncılık.

d) Turkish Book with Multiple Authors

Tonta, Y., Bitirim, Y., & Sever, H. (2002). *Türkçe arama motorlarında performans değerlendirme* [Performance evaluation in Turkish search engines]. Ankara, Turkey: Total Bilişim.

e) Book in English

Kamien R., & Kamien A. (2014). *Music: An appreciation*. New York, NY: McGraw-Hill Education.

f) Chapter in an Edited Book

Bassett, C. (2006). Cultural studies and new media. In G. Hall & C. Birchall (Eds.), *New cultural studies: Adventures in theory* (pp. 220–237). Edinburgh, UK: Edinburgh University Press.

g) Chapter in an Edited Book in Turkish

Erkmen, T. (2012). Örgüt kültürü: Fonksiyonları, öğeleri, işletme yönetimi ve liderlikteki önemi [Organization culture: Its functions, elements and importance in leadership and business management]. In M. Zencirkıran (Ed.), *Örgüt sosyolojisi* [Organization sociology] (pp. 233–263). Bursa, Turkey: Dora Basım Yayın.

h) Book with the same organization as author and publisher

American Psychological Association. (2009). *Publication manual of the American psychological association* (6th ed.). Washington, DC: Author.

Article

a) Turkish Article

Mutlu, B., & Savaşer, S. (2007). Çocuğu ameliyat sonrası yoğun bakımda olan ebeveynlerde stres nedenleri ve azaltma girişimleri [Source and intervention reduction of stress for parents whose children are in intensive care unit after surgery]. *Istanbul University Florence Nightingale Journal of Nursing*, 15(60), 179–182.

b) English Article

de Cillia, R., Reisigl, M., & Wodak, R. (1999). The discursive construction of national identity. *Discourse and Society*, 10(2), 149–173. <http://dx.doi.org/10.1177/0957926599010002002>

c) Journal Article with DOI and More Than Seven Authors Lal, H., Cunningham, A. L., Godeaux, O., Chlibek, R., Diez-Domingo, J., Hwang, S.-J. ... Heineman, T. C. (2015). Efficacy of an adjuvanted herpes zoster subunit vaccine in older adults. *New England Journal of Medicine*, 372, 2087–2096. <http://dx.doi.org/10.1056/NEJMoa1501184>

d) Journal Article from Web, without DOI

Sidani, S. (2003). Enhancing the evaluation of nursing care effectiveness. *Canadian Journal of Nursing Research*, 35(3), 26–38. Retrieved from <http://cjr.mcgill.ca>

e) Journal Article with DOI

Turner, S. J. (2010). Website statistics 2.0: Using Google Analytics to measure library website effectiveness. *Technical Services Quarterly*, 27, 261–278. <http://dx.doi.org/10.1080/07317131003765910>

f) Advance Online Publication

Smith, J. A. (2010). Citing advance online publication: A review. *Journal of Psychology*. Advance online publication. <http://dx.doi.org/10.1037/a45d7867>

g) Article in a Magazine Henry, W. A., III. (1990, April 9). Making the grade in today's schools. *Time*, 135, 28–31.

Doctoral Dissertation, Master's Thesis, Presentation, Proceeding

a) Dissertation/Thesis from a Commercial Database Van Brunt, D. (1997). *Networked consumer health information systems* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 9943436)

b) Dissertation/Thesis from an Institutional Database

Yaylalı-Yıldız, B. (2014). *University campuses as places of potential publicness: Exploring the political, social and cultural practices in Ege University* (Doctoral dissertation). Retrieved from <http://library.iyte.edu.tr/tr/hizli-erisim/iyte-tez-portali>

c) Dissertation/Thesis from Web

Tonta, Y. A. (1992). *An analysis of search failures in online library catalogs* (Doctoral dissertation, University of California, Berkeley). Retrieved from <http://yunus.hacettepe.edu.tr/tonta/yayinlar/phd/ickapak.html>

d) Dissertation/Thesis abstracted in Dissertations Abstracts International

Appelbaum, L. G. (2005). Three studies of human information processing: Texture amplification, motion representation, and figureground segregation. *Dissertation Abstracts International: Section B. Sciences and Engineering*, 65(10), 5428.

e) Symposium Contribution

Krinsky-McHale, S. J., Zigman, W. B., & Silverman, W. (2012, August). Are neuropsychiatric symptoms markers of prodromal Alzheimer's disease in adults with Down syndrome? In W. B. Zigman (Chair), *Predictors of mild cognitive impairment, dementia, and mortality in adults with Down syndrome*. Symposium conducted at the meeting of the American Psychological Association, Orlando, FL.

f) Conference Paper Abstract Retrieved Online

Liu, S. (2005, May). *Defending against business crises with the help of intelligent agent based early warning solutions*. Paper presented at the Seventh International Conference on Enterprise Information Systems, Miami, FL. Abstract retrieved from http://www.iceis.org/iceis2005/abstracts_2005.htm

g) Conference Paper - In Regularly Published Proceedings and Retrieved Online

Herculano-Houzel, S., Collins, C. E., Wong, P., Kaas, J. H., & Lent, R. (2008). The basic nonuniformity of the cerebral cortex. *Proceedings of the National Academy of Sciences*, 105, 12593–12598. <http://dx.doi.org/10.1073/pnas.0805417105>

h) Proceeding in Book Form

Parsons, O. A., Pryzwansky, W. B., Weinstein, D. J., & Wiens, A. N. (1995). Taxonomy for psychology. In J. N. Reich, H. Sands, & A. N. Wiens (Eds.), *Education and training beyond the doctoral degree: Proceedings of the American Psychological Association National Conference on Postdoctoral Education and Training in Psychology* (pp. 45–50). Washington, DC: American Psychological Association.

i) Paper Presentation Nguyen, C. A. (2012, August). *Humor and deception in advertising: When laughter may not be the best medicine*. Paper presented at the meeting of the American Psychological Association, Orlando, FL.

Other Sources

a) Newspaper Article

Browne, R. (2010, March 21). This brainless patient is no dummy. *Sydney Morning Herald*, 45.

b) Newspaper Article with no Author

New drug appears to sharply cut risk of death from heart failure. (1993, July 15). *The Washington Post*, p. A12.

c) Web Page/Blog Post Bordwell, D. (2013, June 18). David Koepp: Making the world movie-sized [Web log post].

Retrieved from <http://www.davidbordwell.net/blog/page/27/>

d) Online Encyclopedia/Dictionary Ignition. (1989). In *Oxford English online dictionary* (2nd ed.). Retrieved from <http://dictionary.oed.com>

Marcoux, A. (2008). Business ethics. In E. N. Zalta (Ed.). *The Stanford encyclopedia of philosophy*. Retrieved from <http://plato.stanford.edu/entries/ethics-business/>

e) Podcast

Dunning, B. (Producer). (2011, January 12). *inFact: Conspiracy theories* [Video podcast]. Retrieved from <http://itunes.apple.com/>

f) Single Episode in a Television Series

Egan, D. (Writer), & Alexander, J. (Director). (2005). Failure to communicate. [Television series episode]. In D. Shore (Executive producer), *House*; New York, NY: Fox Broadcasting.

g) Music

Fuchs, G. (2004). Light the menorah. *On Eight nights of Hanukkah* [CD]. Brick, NJ: Kid Kosher.

SUBMISSION CHECKLIST

Ensure that the following items are present:

- Cover letter to the editor
 - The category of the manuscript
 - Confirming that “the paper is not under consideration for publication in another journal”.
 - Including disclosure of any commercial or financial involvement.
 - Confirming that last control for fluent English was done.
 - Confirming that journal policies detailed in Information for Authors have been reviewed.
 - Confirming that the references cited in the text and listed in the references section are in line with APA 6.
- Copyright Agreement Form
- Permission of previous published material if used in the present manuscript
- Title page
 - The category of the manuscript
 - The title of the manuscript
 - All authors’ names and affiliations (institution, faculty/department, city, country),
 - e-mail addresses
 - Corresponding author’s email address, full postal address, telephone and fax number
 - ORCIDs of all authors.
- Main Manuscript Document
 - The title of the manuscript
 - Abstract (150-250 words)
 - Key words: 3-5 words
 - Grant support (if exists)
 - Conflict of interest (if exists)
 - Acknowledgement (if exists)
 - References
 - All tables, illustrations (figures) (including title, explanation, captions)

TELİF HAKKI ANLAŞMASI FORMU / COPYRIGHT AGREEMENT FORM

İstanbul Üniversitesi
 Istanbul University



Dergi Adı: Acta INFOLOGICA (ACIN)
 Journal name: Acta INFOLOGICA (ACIN)

Telif Hakkı Anlaşması Formu
 Copyright Agreement Form

Sorumlu Yazar Responsible/Corresponding Author	
Makalenin Başlığı Title of Manuscript	
Kabul Tarihi Acceptance Date	
Yazarların Listesi List of Authors	

Sıra No	Adı-Soyadı Name - Surname	E-Posta E-Mail	İmza Signature	Tarih Date
1				
2				
3				
4				
5				

Makalenin türü (Araştırma makalesi, Derleme, v.b.) Manuscript Type (Research Article, Review, etc.)

Sorumlu Yazar:
Responsible/Corresponding Author:

Çalıştığı kurum	University/company/institution
Posta adresi	Address
E-posta	E-mail
Telefon no; GSM no	Phone; mobile phone

Yazar(lar) aşağıdaki hususları kabul eder:
 Sunulan makalenin yazar(lar)ın orijinal çalışması olduğunu ve intihal yapmadıklarını,
 Tüm yazarların bu çalışmaya aslı olarak katılmış olduklarını ve bu çalışma için her türlü sorumluluğu aldıklarını,
 Tüm yazarların sunulan makalenin son halini gördüklerini ve onayladıklarını,
 Makalenin başka bir yerde basılmadığını veya basılmak için sunulmadığını,
 Makalede bulunan metnin, şekillerin ve dokümanların diğer şahıslara ait olan Telif Haklarını ihlal etmediğini kabul ve taahhüt ederler.
 İSTANBUL ÜNİVERSİTESİ'nin bu fikri eseri, Creative Commons Atıf-GayriTicari 4.0 Uluslararası (CC BY-NC 4.0) lisansı ile yayınlamasına izin verirler. Creative Commons Atıf-GayriTicari 4.0 Uluslararası (CC BY-NC 4.0) lisansı, eserin ticari kullanım dışında her boyut ve formatta paylaşılmasına, kopyalanmasına, çoğaltılmasına ve orijinal esere uygun şekilde atıfta bulunmak kaydıyla yeniden düzenleme, dönüştürme ve eserin üzerine inşa etme dâhil adapte edilmesine izin verir.
 Yazar(lar)ın veya varsa yazar(lar)ın işverenin telif dâhil patent hakları, fikri mülkiyet hakları saklıdır.
 Ben/Biz, telif hakkı ihlali nedeniyle üçüncü şahıslara vuku bulacak hak talebi veya açılacak davalarda İSTANBUL ÜNİVERSİTESİ ve Dergi Editörlerinin hiçbir sorumluluğunun olmadığını, tüm sorumluluğun yazarlara ait olduğunu taahhüt ederim/ederiz.
 Ayrıca Ben/Biz makalede hiçbir suç unsuru veya kanuna aykırı ifade bulunmadığını, araştırma yapılırken kanuna aykırı herhangi bir malzeme ve yöntem kullanılmadığını taahhüt ederim/ederiz.
 Bu Telif Hakkı Anlaşması Formu tüm yazarlar tarafından imzalanmalıdır/onaylanmalıdır. Form farklı kurumlarda bulunan yazarlar tarafından ayrı kopyalar halinde doldurularak sunulabilir. Aneak, tüm imzaların orijinal veya kanıtlanabilir şekilde onaylı olması gerekir.

The author(s) agrees that:
 The manuscript submitted is his/her/their own original work and has not been plagiarized from any prior work,
 all authors participated in the work in a substantive way and are prepared to take public responsibility for the work,
 all authors have seen and approved the manuscript as submitted,
 the manuscript has not been published and is not being submitted or considered for publication elsewhere,
 the text, illustrations, and any other materials included in the manuscript do not infringe upon any existing copyright or other rights of anyone.
 İSTANBUL UNIVERSITY will publish the content under Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) license that gives permission to copy and redistribute the material in any medium or format other than commercial purposes as well as remix, transform and build upon the material by providing appropriate credit to the original work.
 The Contributor(s) or, if applicable the Contributor's Employer, retain(s) all proprietary rights in addition to copyright, patent rights.
 I/We indemnify İSTANBUL UNIVERSITY and the Editors of the Journals, and hold them harmless from any loss, expense or damage occasioned by a claim or suit by a third party for copyright infringement, or any suit arising out of any breach of the foregoing warranties as a result of publication of my/our article. I/We also warrant that the article contains no libelous or unlawful statements and does not contain material or instructions that might cause harm or injury.
 This Copyright Agreement Form must be signed/ratified by all authors. Separate copies of the form (completed in full) may be submitted by authors located at different institutions; however, all signatures must be original and authenticated.

Sorumlu Yazar; Responsible/Corresponding Author	İmza / Signature	Tarih / Date
	/...../.....