

# Journal of Data Analytics and Artificial Intelligence Applications

---

Volume 1 • Issue 1 • Year 2025

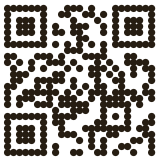
Owner Şadi Evren Şeker  
Istanbul University, Faculty of Computer and Information Technologies,  
Istanbul, Türkiye

Responsible Manager Şadi Evren Şeker  
Istanbul University, Faculty of Computer and Information Technologies,  
Istanbul, Türkiye

Correspondence Istanbul University, Faculty of Computer and Information Technologies  
🏠 Bozdoğan Kemerli Cad. No:1 Vezneciler, Fatih, 34134 İstanbul, Türkiye  
☎ +90 (212) 440 00 00 - 12450  
✉ d3ai@istanbul.edu.tr  
🌐 <https://d3ai.istanbul.edu.tr/>

Publisher Istanbul Universtiy Press  
🏠 İstanbul Üniversitesi Merkez Kampüsü, 34452 Beyazıt/Fatih, İstanbul,  
Türkiye  
☎ +90 (212) 440 00 00  
✉ iupress@istanbul.edu.tr  
🌐 <https://iupress.istanbul.edu.tr/>

Publication Type Periodical



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

## Editorial Management

**Editor-in-Chief** **Şadi Evren Şeker**  
*İstanbul University, İstanbul, Türkiye*

**Co-Editors-in-Chief** **İsmail Önden**  
*İstanbul University, İstanbul, Türkiye*  
**Ezgi Zorarpacı**  
*İstanbul University, İstanbul, Türkiye*

**Promotion Manager** **Bora Çalışkan**  
*İstanbul University, İstanbul, Türkiye*

**Editorial Assistant** **Mert Sülük**  
*İstanbul University, İstanbul, Türkiye*

**Editorial Management Board Members** **Fatma Önay Koçoğlu,**  
*İstanbul University, İstanbul, Türkiye*  
**Pınar Sarısaray Bölük,**  
*İstanbul University, İstanbul, Türkiye*  
**Mehmet Ali Ertürk**  
*İstanbul University, İstanbul, Türkiye*

**Language Editor** **Elizabeth Mary Earl**  
*İstanbul University, İstanbul, Türkiye*

## Editorial Board Members

- Ayse Tüysüz Erman** *Istanbul University, İstanbul, Türkiye*  
aysegul.erman@istanbul.edu.tr
- Abbas Memiş** *Istanbul University, İstanbul, Türkiye*  
abbas.memis@istanbul.edu.tr
- Sezin Güleriyüz Ergül** *Istanbul University, İstanbul, Türkiye*  
sezing@istanbul.edu.tr
- Ezgi Zorarpacı** *Istanbul University, İstanbul, Türkiye*  
ezgi.zorarpaci@istanbul.edu.tr
- Hatice Nizam Özoğur** *Istanbul University, İstanbul, Türkiye*  
hatice.nizamozogur@istanbul.edu.tr
- A. Halim Zaim** *Istanbul Technical University, İstanbul, Türkiye*  
azaim@itu.edu.tr
- Ahmet Kaplan** *Istanbul Sabahattin Zaim University, İstanbul, Türkiye*  
ahmet.kaplan@izu.edu.tr
- Banu Öntürk Diri** *Yıldız Technical University, İstanbul, Türkiye*  
diri@yildiz.edu.tr
- Şebnem Baydere** *Yeditepe University, İstanbul, Türkiye*  
sbaydere@cse.yeditepe.edu.tr
- Enes Eryarsoy** *Sabancı University, İstanbul, Türkiye*  
email: eneseryarsoy@sabanciuniv.edu
- Dursun Delen** *Oklahoma State University, Oklahoma, United States of America (USA)*  
dursun.delen@okstate.edu
- Ender Özcan** *University of Nottingham, Nottingham, England*  
ender.ozcan@nottingham.ac.uk
- Dragan Pamucar** *University of Belgrade, Belgrade, Serbia*  
dragan.pamucar@fon.bg.ac.rs
- Vladimir Simic** *University of Belgrade, Belgrade, Serbia*  
vsima@sf.bg.ac.rs
- Taymaz Rakhar Farsi** *Louisiana State University, Baton Rouge, United States of America (USA)*  
taymaz.akan@lsuhs.edu
- Serkan Varol** *University of Tennessee at Chattanooga, Chattanooga, United States of America (USA)*  
Serkan-Varol@utc.edu
- Özlem Durmaz İncel** *University of Twente, Enschede, The Netherlands*  
ozlem.durmaz@utwente.nl
- Atakan Aral** *Umea University, Umea, Sweden*  
atakan.aral@umu.se
- Sercan Aygün** *University of Louisiana at Lafayette, Lafayette, United States of America (USA)*  
sercan.aygun@louisiana.edu
- Berk Canberk** *Edinburgh Napier University, United Kingdom*  
B.Canberk@napier.ac.uk
- Luca Vollero** *Università Campus Bio-Medico di Roma, Italy*  
l.vollero@unicampus.it



## Table of Contents

Research Article	
Modern AI Models for Text Analysis: A Comparison of Chatgpt and Rag	<b>01</b>
Aslan Nurzhanov, Altynbek Sharipbay	
Research Article	
Children of the Tree: Optimised Rule Extraction from Machine Learning Models	<b>14</b>
Hilal Meydan, Mert Bal	
Research Article	
Financial Performance Evaluation of Companies Listed in Corporate Governance and Sustainability Indices: Application of the IVSF-RBNAR Method	<b>36</b>
Karahan Kara, Galip Cihan Yalçın, Hamide Özyürek	
Research Article	
Enhancing SME Operations with Machine Learning and Business Intelligence: A Case Study of Kolay.ai	<b>61</b>
Rabia Yörük	
Research Article	
Analysis of Word Similarities in Tax Laws Using the Word2Vec Method	<b>84</b>
Ali İhsan Özgür Çilingir	
Review Article	
Machine Learning Implementation in Automated Software Testing: A Review	<b>110</b>
Normi Sham Awang Abu Bakar	

# Journal of Data Analytics and Artificial Intelligence Applications

Research Article

 Open Access

## Modern AI Models for Text Analysis: A Comparison of Chatgpt and Rag



Aslan Nurzhanov<sup>1</sup>   & Altynbek Sharipbay<sup>2</sup> 

<sup>1</sup> L.N.Gumilyov Eurasian National University, Faculty of Information Technology, Department of Information Security, Astana, Kazakhstan

<sup>2</sup> L.N.Gumilyov Eurasian National University, Faculty of Information Technology, Department of Artificial Intelligence Technology, Astana, Kazakhstan

### Abstract

This study presents a comparative analysis of two text-processing models: ChatGPT and Retrieval-Augmented Generation (RAG).

ChatGPT, built on the Generative Pre-trained Transformer (GPT) architecture, excels at generating coherent and contextually appropriate texts, making it widely applicable in fields such as education, healthcare, and business. However, it has a significant limitation—it relies solely on pre-trained data, lacking the ability to access real-time information, which can affect the relevance of its responses in dynamic contexts.

In contrast, RAG integrates text generation with external data retrieval, offering a substantial advantage in terms of real-time data relevance. This feature enhances both the accuracy and completeness of the generated responses, especially for tasks that require up-to-date information. The study evaluates both models based on several key performance indicators, including accuracy, completeness, processing time, and scalability.


The conclusion highlights the strengths and weaknesses of each model and suggests potential improvements for their future application across various domains. By offering a deeper understanding of the capabilities and limitations of these technologies, this research contributes to their optimal use and further development.

### Keywords


Artificial intelligence (AI) · machine learning · natural language processing (NLP) · ChatGPT · RAG



Citation: Aslan Nurzhanov, and Altynbek Sharipbay. 2025. Modern AI Models for Text Analysis: A Comparison of Chatgpt and Rag. *Journal of Data Analytics and Artificial Intelligence Applications* 1, 1 (January 2025), 1-13. <https://doi.org/10.26650/d3ai.002>

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

© 2025. Nurzhanov, A. & Sharipbay, A.

 Corresponding author: Aslan Nurzhanov [as777an@gmail.com](mailto:as777an@gmail.com)



## 1. INTRODUCTION

In recent years, AI models have become essential in text processing, including automatic text generation and data analysis. Popular solutions include the generative transformer-based model ChatGPT [1] and hybrid models like RAG [2].

ChatGPT, developed by OpenAI, generates coherent text by training on extensive datasets, allowing it to handle diverse queries. However, it is limited by its reliance on an internal knowledge base, affecting accuracy and relevance [3, 4].

RAG enhances language models by combining retrieval and generation, achieving state-of-the-art performance in many NLP tasks [5, 6].

This study offers a comparative analysis of ChatGPT and RAG, highlighting their advantages, limitations, and potential applications in various tasks.

## 2. ANALYSIS OF MODERN TEXT PROCESSING METHODS

In this study, over 40 research papers were reviewed, focusing on the application of modern models, including ChatGPT and RAG, in data processing, particularly in text analysis. These works cover a broad range of topics related to the use of these technologies across various fields—from education and healthcare to science and engineering [7]. The articles examine both the advantages of these models, such as high flexibility and improved accuracy through the use of external data, as well as their limitations, including issues with bias, inaccuracies in responses, and computational costs. The comparative analysis of the studies highlighted the key features of applying ChatGPT and RAG and developed a benchmark methodology for their effective use in various text processing scenarios [8, 9].

### 2.1. ChatGPT

ChatGPT is a large language model developed by OpenAI based on the GPT architecture. It has been trained on vast amounts of textual data to generate coherent and meaningful responses to text-based queries. The model can engage in conversations, answer questions, generate text, and even solve tasks that require complex contextual understanding and logic [1].

#### 2.1.1. Advantages and disadvantages of using RAG for text processing

A comparative review of studies on ChatGPT highlights its diverse applications and varying strengths and weaknesses. In education, research indicates that ChatGPT enhances learning by providing quick answers, assisting with essay writing, and explaining complex concepts, leading to high user satisfaction [6, 10].

In medicine, ChatGPT aids in drafting reports and making recommendations, thus improving clinical efficiency [7, 11].

However, limitations exist, particularly regarding accuracy and reliability in technical fields, as some studies note that the model can generate plausible but incorrect responses [8, 12]. In addition, concerns about political bias and objectivity have been raised [9].

In summary, while ChatGPT shows significant potential, caution is necessary in fields where accuracy and impartiality are critical.



### 2.1.2. 2.1.2. Benchmark methodology for using ChatGPT

The benchmark methodology for ChatGPT in text generation assesses its ability to produce coherent, accurate, and contextually relevant responses. Key performance metrics include accuracy, recall, precision, and F1-Score, along with processing time to evaluate efficiency.

Testing involves large-scale datasets, such as SQuAD, and various response formats (e.g., structured answers and free-form text) to gauge performance flexibility. Since ChatGPT relies solely on internal knowledge, it generates high-quality responses with minimal latency, supporting scalable response generation for numerous queries [13].

## 2.2. 2.2. RAG

RAG is a powerful approach for text processing that combines the strengths of generative models with the ability to retrieve information from external data sources [14]. This approach offers several key advantages that make it appealing for various tasks, although it also comes with certain limitations.

### 2.2.1. 2.2.1. Advantages and disadvantages of using RAG for text processing

RAG excels in retrieving relevant data from external databases, enhancing the accuracy and reliability of responses, particularly in critical fields such as medicine and law [15]. It effectively handles long texts and complex queries, enriching responses with the necessary context, which is beneficial in multi-layered tasks such as legal analysis [16].

Additionally, RAG's adaptability to various domains, leveraging specialised databases, provides versatility compared to traditional models trained on limited datasets [17].

However, RAG's high computational complexity poses a disadvantage, especially with large datasets, leading to slower response times in time-sensitive tasks [18]. Setting up an RAG also requires substantial effort to integrate external databases efficiently, with potential performance issues if not carefully configured [19]. Furthermore, the effectiveness of the RAG relies on the quality of the external data; outdated or incorrect information can undermine its accuracy [20].

In summary, while RAG is a powerful tool for text processing in complex tasks requiring high accuracy, its implementation necessitates considerable computational resources, careful configuration, and access to high-quality data.

### 2.2.2. 2.2.3. Benchmark methodology for using the RAG

The benchmark methodology for RAG in text processing utilises objective metrics to evaluate performance, including Precision (relevance of retrieved documents), Recall (ability to retrieve all relevant documents), and Accuracy (overall correctness of predictions).

Processing Time measures the speed of document retrieval and response generation, while Scalability assesses the model's capability to handle increased queries or dataset sizes without performance loss. These metrics provide a quantifiable basis for optimising RAG models in real-world applications, ensuring effective performance in complex text processing tasks [21].

## 3. RESEARCH METHODOLOGY

The methodology of this research is focused on evaluating the performance of the ChatGPT and RAG models using key metrics. For a systematic analysis, the study is divided into several critical aspects.



### 3.1. Literature review of the evaluation methods

The evaluation methodologies for the ChatGPT and RAG models incorporate various metrics, including accuracy, recall, and F1-score for ChatGPT, as well as precision and recall for RAG. A detailed overview of the evaluation methods is presented in [Table 1 \[22-24\]](#).

**Table 1.** Overview of the evaluation methods for the ChatGPT and RAG models

Study title	Model evaluated	Key metrics	Summary of the findings
Evaluating ChatGPT as a Question-Answering System: A Comprehensive Analysis and Comparison with Existing Models [22]	ChatGPT	Accuracy, Recall, and F1-score	Compared ChatGPT with traditional QA systems, testing various interaction modes and evaluation methods.
Evaluation of Retrieval-Augmented Generation: A Survey [23]	RAG	Precision, Recall	Discusses metrics for assessing RAG's retrieval capabilities and generated text quality, including answer relevance and faithfulness.
CRAG—Comprehensive RAG Benchmark [24]	RAG	Context Precision and Answer Relevance	Outlines the key evaluation metrics for the RAG, providing insights into the evaluation methodologies.

### 3.2. Experimental conditions for testing the ChatGPT and RAG

To test ChatGPT, QA datasets were used, including popular test sets like SQuAD, which evaluate the model's ability to generate accurate answers to complex questions. Various datasets and query types were employed to assess the performance in different generation modes [22, 25, 26].

For the RAG, testing required integrating external databases, using complex queries that necessitated information retrieval. The conditions included working with large datasets and evaluating the relevance of the retrieved data [23, 27].

### 3.3. The abilities of the ChatGPT and RAG models in processing and classifying extremist texts

To evaluate the effectiveness of ChatGPT and RAG in processing and classifying extremist texts, a specialised dataset was developed. It consisted of examples of extremist content across eight categories: political, religious, racial, national (ethnic), economic, social, youth, and environmental extremism. Additionally, the dataset included materials related to extremism, such as articles, reports, and publications, addressing various aspects of extremist activities, their consequences, and methods of prevention.

Experimental setup for ChatGPT: Input texts were processed directly without external data retrieval. The model's responses were evaluated based on binary classification (extremist or non-extremist) and type classification (correct identification of the specific type of extremism).

Experimental setup for the RAG: The model retrieved contextual data from an external database. Similar evaluation metrics were used, with additional emphasis on the relevance of the retrieved documents.

Key Evaluation Metrics:

- True Positives by type of extremism (TPv): The number of texts correctly classified not only as extremist but also by the correct type of extremism.

- False Positives by type of extremism (FPv): The number of texts with extremist content correctly identified as extremist but misclassified regarding the specific type of extremism.
- False Negatives by type of extremism (FNv): The number of texts containing extremist content of a specific type that the model either failed to classify correctly or failed to recognise as extremist (missed classification).

In the implementation of the RAG model, the external knowledge base was stored in MD (Markdown) files, allowing for a simple and structured format that facilitated efficient processing. These MD files contained texts organised into thematic segments, simplifying the retrieval of relevant information.

Vector representations of the data, stored in the Chroma database, were generated based on the content of these MD files. This setup ensured efficient data management and reduced the system load during query execution. Texts were split into chunks of 300 characters, ensuring consistent representation in the vector database and improving the accuracy of context retrieval.

Additionally, the LangChain library was used to orchestrate the processes of information retrieval and response generation. LangChain enabled seamless integration between the knowledge base, MD files, and vector search operations. During response generation, the RAG model used ChatGPT, leveraging its generative capabilities to analyse retrieved information and produce coherent and contextually relevant outputs. This approach ensured high accuracy and relevance in the tasks related to text classification and processing.

## 4. EXPERIMENTAL RESULTS

Two models, ChatGPT and RAG, were used in the experiments. Each model was tested in conditions as close as possible to real-world text processing scenarios, including tasks involving short, long, and implicit texts. ChatGPT was tested on various question-answer tasks using datasets like SQuAD, generating answers based on pre-trained data without external search [22]. In contrast, the RAG included a data retrieval component, enabling the system to find information in real databases before generating a response. This ensured the integration of additional sources to improve the accuracy and relevance of the answers [14].

### 4.1. Performance comparison of ChatGPT and RAG by criteria

*Accuracy:* ChatGPT performs well in providing contextually correct answers but struggles with complex, factually precise questions due to a lack of external data support [22]. In contrast, RAG shows higher accuracy, particularly in tasks requiring factual retrieval from external sources [25].

*Recall:* ChatGPT often fails to deliver complete answers, especially with implicit texts [28]. RAG demonstrates higher recall by effectively retrieving and integrating information from multiple sources [25].

*Processing Time:* ChatGPT is faster for tasks without information retrieval, while RAG takes longer due to its need to search external sources [29].

*Scalability:* ChatGPT handles numerous queries efficiently [3], whereas RAG faces scalability challenges with large datasets [30].

*Relevance:* ChatGPT's responses can be general or contextually limited due to reliance on pre-trained data [3]. RAG provides more relevant and accurate responses, especially for complex queries requiring current information [31].

*Processing Long Texts:* ChatGPT manages long texts but may lose critical information due to its context window limitations [32]. RAG effectively processes long texts by breaking them into chunks and retrieving relevant data as needed.

In the study of the performance of the ChatGPT and RAG models, the results of the comparative analysis are presented in Table 2 [18, 33-37].

**Table 2.** Comparison of the ChatGPT and RAG performance

Criteria	ChatGPT	RAG	Comments
<b>Accuracy</b>	50.5% (PubMedQA), 15.06% (HotpotQA)	56.42% (PubMedQA), 12.07% (HotpotQA)	RAG shows a slight improvement in accuracy compared to ChatGPT, especially when accessing external data [33, 34]
<b>Recall</b>	1.09% (PubMedQA), 22.63% (HotpotQA)	3.05% (PubMedQA), 25.05% (HotpotQA)	RAG provides better extraction of relevant information, especially on complex question-and-answer tasks [33]
<b>Processing time</b>	~0.3–0.5 sec	~1.5–2 sec	ChatGPT is faster because it does not require access to external data sources [35]
<b>Scalability</b>	High	Average	ChatGPT scales better due to the lower computational cost of extracting information [36]
<b>Relevance of the response</b>	Low for complex queries	Higher thanks to the external data	RAG is more relevant for tasks that require searching for relevant information [37]
<b>Processing long texts</b>	Moderate	High	RAG handles long texts better by extracting information from external sources [18]

These results provide a clear comparison of the performance of ChatGPT and RAG across different criteria, highlighting the strengths and limitations of each model in various text processing tasks.

## 4.2. Advantages of the RAG in handling long texts

RAG demonstrated clear advantages in working with long texts and texts with implicit content. Thanks to its data retrieval component, the model could locate and use relevant information from external sources, significantly improving the quality of its generated responses. In tasks where the text requires detailed processing and contextual understanding, RAG outperforms ChatGPT, which relies solely on internal model data. This was confirmed in studies where long documents and complex texts requiring in-depth analysis were tested [16].

Thus, RAG proved to be highly effective in complex scenarios that require the integration of external information, while ChatGPT remains more suitable for quicker, less complex tasks involving shorter texts [38].

### 4.2.1. Evaluation results of the ChatGPT and RAG models' capabilities in processing and classifying extremist texts

As part of the conducted scientific study, 160 extremist texts were tested, evenly distributed across eight types of extremism with 20 texts for each category. The analysis results are presented in Table 3.

**Table 3.** Comparative table of extremist text classification results by the ChatGPT and RAG models

Type of extremism	ChatGPT			RAG		
	TPv	FPv	FNv	TPv	FPv	FNv
Political	20	1	0	15	1	5
Religious	17	1	3	10	4	10
Racial	16	2	4	13	3	7
National (ethnic)	20	2	0	15	3	5
Economic	15	1	5	12	2	8
Social	16	1	4	11	3	9
Youth	16	0	4	12	3	8
Environmental	15	0	5	11	4	9

The ChatGPT model demonstrated high classification accuracy across most categories of extremism, particularly in tasks where the texts contained explicit content. It achieved maximum TPv values for political and national (ethnic) extremism (20 out of 20), while maintaining low FPv and FNv values. However, in some cases, such as religious and economic extremism, the model made more errors, with FNv reaching up to 5.

The RAG model, leveraging its ability to extract additional data from external sources, showed stable performance in classifying complex and veiled texts. However, its TPv values were generally lower than those of ChatGPT across almost all categories, especially for religious and social extremism, where FNv reached 10 and 9, respectively.

The classification results highlight the differences in the effectiveness of ChatGPT and RAG depending on the type of extremism.

For political extremism, ChatGPT correctly classified all texts (TPv = 20), with no omissions (FNv = 0) and only one FPv error. In contrast, RAG delivered a lower performance, correctly classifying 15 texts (TPv = 15), while missing 5 texts (FNv = 5) and maintaining a similar number of FPv errors.

In religious extremism, ChatGPT performed better, correctly classifying 17 texts (TPv = 17) with one FPv error and three FNv omissions. The RAG model was less accurate, correctly classifying only 10 texts (TPv = 10), making more FPv errors (4), and missing 10 texts (FNv = 10).

For racial extremism, ChatGPT achieved slightly better results, correctly classifying 16 texts (TPv = 16) with 4 FNv omissions. RAG performed worse, correctly classifying 13 texts (TPv = 13) and missing 7 (FNv = 7). Both models had comparable FPv errors (2–3).

In the category of national (ethnic) extremism, ChatGPT again demonstrated maximum accuracy, correctly classifying all texts (TPv = 20) with no omissions (FNv = 0). RAG underperformed, missing 5 texts (FNv = 5) and correctly classifying 15 texts (TPv = 15).

For economic extremism, ChatGPT correctly classified 15 texts (TPv = 15) with 5 FNv omissions, while RAG showed lower accuracy, correctly classifying 12 texts (TPv = 12) and missing 8 (FNv = 8).

The analysis of social extremism also highlighted ChatGPT's superiority, correctly classifying 16 texts (TPv = 16) with 4 FNv omissions. RAG demonstrated lower accuracy, correctly classifying 11 texts (TPv = 11) and missing 9 (FNv = 9).

For youth extremism, ChatGPT showed higher accuracy, correctly classifying 16 texts (TPv = 16) with 4 FNv omissions, while RAG correctly classified 12 texts (TPv = 12) and missed 8 (FNv = 8).



Finally, in the category of environmental extremism, ChatGPT correctly classified 15 texts (TPv = 15) with 5 FNv omissions. In comparison, RAG correctly classified 11 texts (TPv = 11), missed 9 texts (FNv = 9), and had more FPv errors (4 compared to 0 for ChatGPT).

These results demonstrate that ChatGPT achieves higher accuracy in classifying most types of extremism, particularly in cases involving explicit text content. In contrast, the RAG encounters challenges when classifying veiled or complex texts, which is reflected in its higher FPv and FNv error rates.

## 5. DISCUSSION

The discussion section focuses on a detailed comparison between the strengths and weaknesses of the two models used in this research, ChatGPT and RAG. This analysis is essential for understanding how each model performs in different contexts, highlighting their advantages and limitations in real-world applications. By evaluating these models, we can identify areas where improvements could further enhance their performance. The following subsections provide a breakdown of the key features and shortcomings of both models, followed by potential strategies for enhancing their capabilities in the field of text processing.

### 5.1. Strengths and weaknesses of each model

ChatGPT's key strengths include fast text generation, scalability, and versatility, enabling real-time query processing without the need for external database access [36]. However, it has significant drawbacks in terms of accuracy and reliability, as it relies on static training data, making it unsuitable for critical fields like medicine or law where current information is essential. Additionally, ChatGPT may produce plausible-sounding but factually incorrect responses [39].

In contrast, RAG improves the accuracy and relevance of responses by integrating with external databases, allowing access to real-time information, which is crucial for complex tasks in medicine and law [14]. Nonetheless, RAG has its own disadvantages, including the high computational costs associated with real-time information retrieval, which can slow down the performance. The complexity of setting up an RAG for integration with various databases also complicates its use across tasks [40].

### 5.2. Potential improvements in the text processing models

Improving ChatGPT's performance could involve integrating an information retrieval component like RAG to access real-time data, enhancing its accuracy in fields such as medicine and finance. Fine-tuning on specialised datasets for tasks such as medical or legal consultations could also increase relevance and reliability.

For RAG, optimising the data retrieval process with more efficient ranking algorithms could reduce search times and computational costs. In addition, incorporating automatic fact-checking and data verification systems would enhance accuracy and minimise irrelevant information.

In summary, both models possess unique strengths, and targeted improvements could enhance their effectiveness across various tasks.



### 5.3. Evaluation of the ChatGPT and RAG models' capabilities in processing and classifying extremist texts

The results of the comparative analysis revealed that the ChatGPT and RAG models exhibit varying levels of effectiveness in classifying extremist texts depending on the category and complexity of the content.

ChatGPT demonstrated high classification accuracy, particularly in tasks involving explicit content. The model achieved maximum TPv values for political and national (ethnic) extremism, correctly classifying all texts in these categories. Low FPv and FNv values across most categories confirm the model's ability to handle texts effectively that do not require deep analysis or the extraction of an additional context. However, in categories with veiled content, such as religious and economic extremism, ChatGPT delivered less accurate results, with more omissions (FNv up to 5).

RAG, on the other hand, showcased stable performance when working with veiled and complex texts due to its ability to retrieve additional information from external data sources. However, this capability did not always result in higher accuracy. In categories like religious and social extremism, the model exhibited a significant number of omissions (FNv up to 10) and classification errors related to extremist types (FPv up to 4). This may be attributed to the model's reliance on the quality of external data and the challenges associated with integrating these data into the classification process.

## 6. CONCLUSION

The comparison of the ChatGPT and RAG models is crucial for advancing artificial intelligence in specialised domains, such as processing extremist texts. ChatGPT serves as a powerful tool for text generation based on pre-trained knowledge, making it effective for quick analysis and tasks that do not require access to external data. However, its limitations become clear in situations where handling current or domain-specific information is essential.

In contrast, RAG, by integrating mechanisms for retrieving data from external sources, shows significant potential for processing texts that require deep analysis and contextual understanding. This capability is particularly important in fields such as medicine, law, or the analysis of veiled extremist content.

The comparison of these models helps identify the key aspects of their applications and determine the directions for further improvements. For instance, integrating the strengths of both models could lead to the development of hybrid systems that combine the accuracy and speed of ChatGPT with the data retrieval capabilities of RAG, offering significant potential for use in critically important areas.

Future research should focus on optimising each model to enhance its effectiveness in specialised applications. For ChatGPT, it would be beneficial to integrate data retrieval mechanisms, enabling the model to access real-time information. This could involve developing hybrid architectures that combine the model's pre-trained knowledge with contextual search capabilities. Additionally, fine-tuning ChatGPT on domain-specific datasets, such as those in medicine, law, or extremist content analysis, would significantly improve its accuracy for specific tasks. Developing algorithms to better analyse veiled texts, leveraging methods of deep context analysis to uncover nuanced meanings, is also essential. To improve the handling of long texts, the model's context window should be expanded, and mechanisms for segmenting text into chunks with subsequent interpretation integration should be implemented.



For the RAG, improving the quality of data retrieval is a priority. This can be achieved by employing more accurate ranking algorithms and implementing automatic verification mechanisms to minimise the impact of outdated or irrelevant information. Reducing the computational complexity of the model is also essential, which can be accomplished by optimising its architecture to accelerate search processes without compromising accuracy and employing dynamic selection of relevant sources. Additionally, RAG should be tailored to specialised tasks through fine-tuning on relevant databases, such as medical or legal sources, while strengthening its contextual retrieval mechanisms to better handle veiled texts.

A combined optimisation of these models could involve developing hybrid systems that merge the speed and accuracy of ChatGPT with RAG's external data retrieval capabilities. Such systems could dynamically adapt to various tasks, automatically switching between pre-trained knowledge generation and external data retrieval based on contextual demands. To enhance effectiveness, it is recommended to test these systems on real-world datasets that reflect practical application conditions and to develop new evaluation metrics that account for requirements in accuracy, speed, and scalability. Ethical and legal considerations are also critical, including measures to prevent the generation of biased content and to ensure data confidentiality, particularly in sensitive areas such as medicine or counter-extremism efforts. These steps will help tailor the models to specialised applications and ensure their successful deployment across diverse contexts.

The scalability of the AI models and the reduction of computational costs are critical areas for future improvement. This study identifies several strategies that can be implemented to enhance these aspects for both the ChatGPT and RAG models.

For ChatGPT, computational efficiency can be improved through techniques such as parameter reduction or weight quantisation, which reduce resource demands without significantly impacting performance. Similarly, dynamic sampling strategies can be employed in RAG to minimise the number of external data queries, ensuring that only the most relevant information is retrieved.

Caching mechanisms present another promising avenue for optimisation. In the case of RAG, frequently accessed data can be cached, thereby reducing the retrieval times and computational overhead. For ChatGPT, the use of precomputed contexts for common queries could accelerate processing while maintaining accuracy. Additionally, parallel processing on high-performance computing platforms offers potential scalability improvements for both models, allowing them to handle larger datasets and more complex tasks efficiently.

This study makes a substantial contribution to the existing literature by highlighting the strengths and limitations of ChatGPT and RAG in text classification tasks. Specifically, it demonstrates that ChatGPT excels in tasks involving explicit content, while RAG is better suited for handling veiled or contextually complex texts. These findings enrich current knowledge by providing a clearer understanding of the contexts in which each model performs optimally.

Moreover, the results underscore the practical applicability of these models in real-world scenarios, such as healthcare and legal domains, which require high levels of accuracy and reliability. By bridging the gap between academic research and practical deployment, this study provides a valuable foundation for developing hybrid systems that combine the strengths of both ChatGPT and RAG. Such systems could enhance the precision and adaptability of AI models, making them more suitable for diverse and critical applications.

Finally, the findings of this research lay the groundwork for further exploration into computationally efficient hybrid architectures. The evaluation metrics developed and applied in this study, such as TPv, FPv, and

FNv, could serve as benchmarks for analysing the performance of AI models in other specialised tasks. These insights reinforce the importance of integrating computational efficiency, scalability, and practical adaptability into the development of next-generation AI systems.



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


#### Author Details **Aslan Nurzhanov**

<sup>1</sup> L.N.Gumilyov Eurasian National University, Faculty of Information Technology, Department of Information Security, Astana, Kazakhstan

 0009-0001-4617-7798

#### **Altynbek Sharipbay**

<sup>2</sup> L.N.Gumilyov Eurasian National University, Faculty of Information Technology, Department of Artificial Intelligence Technology, Astana, Kazakhstan

 0009-0000-5511-7466

## References

- [1] Yuntao Wang, Yanghe Pan, Miao Yan, Zhou Su1, and Tom H. Luan. 2023. A Survey on ChatGPT: AI-Generated Contents, Challenges, and Solutions. arXiv:2305.18339. Retrieved from <https://arxiv.org/abs/2305.18339>
- [2] Siyun Zhao, Yuqing Yang, Zilong Wang, Zhiyuan He, Luna K. Qiu, Lili Qiu. 2024. Retrieval Augmented Generation (RAG) and Beyond: A Comprehensive Survey on How to Make your LLMs use External Data More Wisely. arXiv:2409.14924. Retrieved from <https://arxiv.org/abs/2409.14924>
- [3] Walid Harir. 2024. Unlocking the Potential of ChatGPT: A Comprehensive Exploration of its Applications, Advantages, Limitations, and Future Directions in Natural Language Processing. arXiv:2304.02017. Retrieved from <https://arxiv.org/abs/2304.02017>
- [4] Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, Yupeng Wu. 2023. How Close is ChatGPT to Human Experts? Comparison Corpus, Evaluation, and Detection. arXiv:2301.07597. Retrieved from <https://arxiv.org/abs/2301.07597>
- [5] Shayekh Bin Islam, Md Asib Rahman, K S M Tozammel Hossain, Enamul Hoque, Shafiq Joty, Md Rizwan Parvez. 2024. Open-RAG: Enhanced Retrieval-Augmented Reasoning with Open-Source Large Language Models. arXiv:2410.01782. Retrieved from <https://arxiv.org/abs/2410.01782>
- [6] Chengcheng Yu, Jinzhe Yan, Na Cai. 2024. ChatGPT in higher education: factors influencing ChatGPT user satisfaction and continued use intention. 2024. *Frontiers in Education*, 9, Article 2 (May 2024), 11 pages. <https://doi.org/10.3389/educ.2024.1354929>
- [7] Tirth Dave, Sai Anirudh Athaluri, Satyam Singh. 2023. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. *Frontiers in Artificial Intelligence*, 6, Article 64 (May 2023), 5 pages. <https://doi.org/10.3389/frai.2023.1169595>
- [8] Karen D. Wang, Eric Burkholder, Carl Wieman, Shima Salehi, Nick Haber. 2024. Examining the potential and pitfalls of ChatGPT in science and engineering problem-solving. *Frontiers in Education*, 8, Article 5 (January 2024), 11 pages. <https://doi.org/10.3389/educ.2023.1330486>
- [9] Sasuke Fujimoto, Kazuhiro Takemoto. 2023. Revisiting the political biases of ChatGPT. *Frontiers in Artificial Intelligence*, 6, Article 174 (October 2023), 6 pages. <https://doi.org/10.3389/frai.2023.1232003>





- [10] Anissa M. Bettayeb, Manar Abu Talib, Al Zahraa Sobhe Altayasinah, Fatima Dakalbab. 2024. Exploring the impact of ChatGPT: conversational AI in education. *Frontiers in Education*, 9, Article 1 (July 2024), 16 pages. <https://doi.org/10.3389/feduc.2024.1379796>
- [11] Maria Grazia Maggio, Gennaro Tartarisco, Davide Cardile, Mirjam Bonanno, Roberta Bruschetta, Loris Pignolo, Giovanni Pioggia, Rocco Salvatore Calabrò, Antonio Cerasa. 2024. Exploring ChatGPT's potential in the clinical stream of neurorehabilitation. *Frontiers in Artificial Intelligence*, 7, Article 1 (January 2024), 15 pages. <https://doi.org/10.3389/frai.2024.1407905>
- [12] Rex Bringula. 2024. ChatGPT in a programming course: benefits and limitations. *Frontiers in Education*, 9, Article 73 (February 2024), 6 pages. <https://doi.org/10.3389/feduc.2024.1248705>
- [13] Zeyan Liu, Zijun Yao, Fengjun Li, and Bo Luo. 2024. Check Me If You Can: Detecting ChatGPT-Generated Academic Writing using CheckGPT. arXiv:2306.05524. Retrieved from <https://arxiv.org/abs/2306.05524>
- [14] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, Haofen Wang. 2024. Retrieval-Augmented Generation for Large Language Models: A Survey. arXiv:2312.10997v5. Retrieved from <https://arxiv.org/abs/2312.10997v5>
- [15] Guangzhi Xiong, Qiao Jin, Xiao Wang, Minjia Zhang, Zhiyong Lu, Aidong Zhang. 2024. Improving Retrieval-Augmented Generation in Medicine with Iterative Follow-up Questions. arXiv:2408.00727. Retrieved from <https://arxiv.org/abs/2408.00727>
- [16] Ziyang Jiang, Xueguang Ma, Wenhua Chen. 2024. LongRAG: Enhancing Retrieval-Augmented Generation with Long-context LLMs. arXiv:2406.15319. Retrieved from <https://arxiv.org/abs/2406.15319>
- [17] Xun Xian, Ganghua Wang, Xuan Bi, Jayanth Srinivasa, Ashish Kundu, Charles Fleming, Mingyi Hong, Jie Ding. 2024. On the Vulnerability of Applying Retrieval-Augmented Generation within Knowledge-Intensive Application Domains. arXiv:2409.17275. Retrieved from <https://arxiv.org/abs/2409.17275>
- [18] Zhuowan Li, Cheng Li, Mingyang Zhang, Qiaozhu Mei, Michael Bendersky. 2024. Retrieval Augmented Generation or Long-Context LLMs? A Comprehensive Study and Hybrid Approach. arXiv:2407.16833. Retrieved from <https://arxiv.org/abs/2407.16833>
- [19] Ye Yuan, Chengwu Liu, Jingyang Yuan, Gongbo Sun, Siqi Li, Ming Zhang. 2024. A Hybrid RAG System with Comprehensive Enhancement on Complex Reasoning. arXiv:2408.05141. Retrieved from <https://arxiv.org/abs/2408.05141>
- [20] Fei Wang, Xingchen Wan, Ruoxi Sun, Jiefeng Chen, Sercan Ö. Arık. 2024. Astute RAG: Overcoming Imperfect Retrieval Augmentation and Knowledge Conflicts for Large Language Models. arXiv:2410.07176. Retrieved from <https://arxiv.org/abs/2410.07176>
- [21] Siyun Zhao, Yuqing Yang, Zilong Wang, Zhiyuan He, Luna K. Qiu, Lili Qiu. 2024. Retrieval Augmented Generation (RAG) and Beyond: A Comprehensive Survey on How to Make your LLMs use External Data More Wisely. arXiv:2409.14924. Retrieved from <https://arxiv.org/abs/2409.14924>
- [22] Hossein Bahak, Farzaneh Taheri, Zahra Zojaji, Arefeh Kazemi. 2023. Evaluating ChatGPT as a Question Answering System: A Comprehensive Analysis and Comparison with Existing Models. arXiv:2312.07592. Retrieved from <https://arxiv.org/abs/2312.07592>
- [23] Hao Yu, Aoran Gan, Kai Zhang, Shiwei Tong, Qi Liu, Zhaofeng Liu. 2024. Evaluation of Retrieval-Augmented Generation: A Survey. arXiv:2405.07437. Retrieved from <https://arxiv.org/abs/2405.07437>
- [24] Xiao Yang, Kai Sun, Hao Xin, Yushi Sun, Nikita Bhalla, Xiangsen Chen, Sajal Choudhary, Rongze Daniel Gui, Ziran Will Jiang, Ziyu Jiang, Lingkun Kong, Brian Moran, Jiaqi Wang, Yifan Ethan Xu, An Yan, Chenyu Yang, Eting Yuan, Hanwen Zha, Nan Tang, Lei Chen, Nicolas Scheffer, Yue Liu, Nirav Shah, Rakesh Wang, Anuj Kumar, Wen-tau Yih, Xin Luna Dong. 2024. CRAG – Comprehensive RAG Benchmark. arXiv:2406.04744. Retrieved from <https://arxiv.org/abs/2406.04744>
- [25] Kunal Sawarkar, Abhilasha Mangal, Shivam Raj Solanki. 2024. Blended RAG: Improving RAG (Retriever-Augmented Generation) Accuracy with Semantic Search and Hybrid Query-Based Retrievers. arXiv:2404.07220v1. Retrieved from <https://arxiv.org/abs/2404.07220v1>
- [26] Yizheng Huang, Jimmy Huang. 2024. Exploring ChatGPT for Next-generation Information Retrieval: Opportunities and Challenges. arXiv:2402.11203. Retrieved from <https://arxiv.org/abs/2402.11203>
- [27] Yihang Zheng, Bo Li, Zhenghao Lin, Yi Luo, Xuanhe Zhou, Chen Lin, Jinsong Su, Guoliang Li, Shifu Li. 2024. Revolutionizing Database Q&A with Large Language Models: Comprehensive Benchmark and Evaluation. arXiv:2409.04475. Retrieved from <https://arxiv.org/abs/2409.04475>
- [28] Yuntao Wang, Yanghe Pan, Miao Yan, Zhou Su, Tom H. Luan. 2023. A Survey on ChatGPT: AI-Generated Contents, Challenges, and Solutions. arXiv:2305.18339. Retrieved from <https://arxiv.org/abs/2305.18339>
- [29] Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, Zhaochun Ren. 2023. Is ChatGPT Good at Search? Investigating Large Language Models as Re-Ranking Agents. arXiv:2304.09542. Retrieved from <https://arxiv.org/abs/2304.09542>
- [30] Yu Bai, Yukai Miao, Li Chen, Dan Li, Yanyu Ren, Hongtao Xie, Ce Yang, Xuhui Cai. 2024. Pistis-RAG: A Scalable Cascading Framework Towards Trustworthy Retrieval-Augmented Generation. arXiv:2407.00072. Retrieved from <https://arxiv.org/abs/2407.00072>



- [31] Xiaohua Wang, Zhenghua Wang, Xuan Gao, Feiran Zhang, Yixin Wu, Zhibo Xu, Tianyuan Shi, Zhengyuan Wang, Shizheng Li, Qi Qian, Ruicheng Yin, Changze Lv, Xiaoqing Zheng, Xuanjing Huang. 2024. Searching for Best Practices in Retrieval-Augmented Generation. arXiv:2407.01219. Retrieved from <https://arxiv.org/abs/2407.01219>
- [32] Weizhi Fei, Xueyan Niu, Pingyi Zhou, Lu Hou, Bo Bai, Lei Deng, Wei Han. 2023. Extending Context Window of Large Language Models via Semantic Compression. arXiv:2312.09571. Retrieved from <https://arxiv.org/abs/2312.09571>
- [33] Yuetong Zhao, Hongyu Cao, Xianyu Zhao, Zhijian Ou. 2024. An Empirical Study of Retrieval Augmented Generation with Chain-of-Thought. arXiv:2407.15569. Retrieved from <https://arxiv.org/abs/2407.15569>
- [34] Yizheng Huang, Jimmy Huang. 2024. A Survey on Retrieval-Augmented Text Generation for Large Language Models. arXiv: 2404.10981v1. Retrieved from <https://arxiv.org/html/2404.10981v1>
- [35] Harry Guinness. 2024. What is RAG (retrieval augmented generation)? (August 2024). Retrieved October 15, 2024 from <https://zapier.com/blog/retrieval-augmented-generation/>.
- [36] Frank Joublin, Antonello Ceravola, Joerg Deigmoeller, Michael Gienger, Mathias Franzius, Julian Eggert. 2023. A Glimpse in ChatGPT Capabilities and its impact for AI research. arXiv:2305.06087. Retrieved from <https://arxiv.org/abs/2305.06087>
- [37] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, Douwe Kiela. 2021. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. arXiv:2005.11401. Retrieved from <https://arxiv.org/abs/2005.11401>
- [38] Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, Zhaochun Ren. 2023. Is ChatGPT Good at Search? Investigating Large Language Models as Re-Ranking Agents. arXiv:2304.09542. Retrieved from <https://arxiv.org/abs/2304.09542>
- [39] Christoph Leiter, Ran Zhang, Yanran Chen, Jonas Belouadi, Daniil Larionov, Vivian Fresen, Steffen Eger. 2023. ChatGPT: A Meta-Analysis after 2.5 Months. arXiv:2302.13795. Retrieved from <https://arxiv.org/abs/2302.13795>
- [40] Ruiyang Qin, Zheyu Yan, Dewen Zeng, Zhenge Jia, Dancheng Liu, Jianbo Liu, Zhi Zheng, Ningyuan Cao, Kai Ni, Jinjun Xiong, Yiyu Shi. 2024. Robust Implementation of Retrieval-Augmented Generation on Edge-based Computing-in-Memory Architectures. arXiv:2405.04700. Retrieved from <https://arxiv.org/abs/2405.04700>



# Journal of Data Analytics and Artificial Intelligence Applications

Research Article

 Open Access

## Children of the Tree: Optimised Rule Extraction from Machine Learning Models



Hilal Meydan<sup>1</sup>   & Mert Bal<sup>1</sup> 

<sup>1</sup> Yıldız Technical University, Department of Mathematical Engineering, İstanbul, Türkiye

### Abstract



The “Children of the Tree” algorithm provides a strong understanding of how the imbalanced dataset is classified by extracting rules from each tree of the Random Forest (RF) model. Basically, it converts the divisions created at each node of the trees into “if-then” rules and extracts individual rules for each tree by differentiating the general “community model” perception in the RF. Thus, the algorithm finds the “Children of the Tree” by converting the forest into a rule set. This study, developed on the “German Credit Data Set”, which is one of the banking data sets on which many studies have been conducted in the literature; determines the rules that cause to fall into that class(class good or class bad) for candidate customers. In this way, the bank would see the rules for potential customers belonging to the risky class and have the chance to recommend the alternative plans/products that are suitable for their risk strategy to their potential customers. The study evaluates rule validity and reliability using association rule mining metrics—support, confidence, lift, leverage, conviction - calculates “Minimum Description Length” (MDL), and ranks rules by “support” and “MDL cost” to extract the simplest rules for each class. It addresses risk management in banking and marketing needs, using MDL cost and SMOTE to handle imbalanced datasets, setting it apart from other algorithms.

### Keywords

Children of the Tree · Machine Learning · Rule Extraction · Random Forest · Minimum Description Length



Citation: Hilal Meydan, and Mert Bal. 2025. Children of the Tree: Optimised Rule Extraction from Machine Learning Models. *Journal of Data Analytics and Artificial Intelligence Applications* 1, 1 (January 2025), 14-35. <https://doi.org/10.26650/d3ai.1606958>

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

© 2025. Meydan, H. & Bal, M.

 Corresponding author: Hilal Meydan [hilalmydan@gmail.com](mailto:hilalmydan@gmail.com)



## 1. Introduction

This article will discuss the “Children of the Tree” algorithm, which we introduce as a novel algorithm to the Machine Learning (ML) literature. “*Children of the Tree*” provides rule extraction for candidate bank customers, which we explain through Random Forest (RF) as the tree-based machine learning model. This study helps banks and non-bank financial institutions accurately detect potential customer risks and make decisions aligned with their risk strategies. The rules needed are extracted accurately and in a less complex manner. The algorithm can work on imbalanced data sets.

In this study, we examine how the “*Children of the Tree*” algorithm can efficiently generate rules for customer classification in the banking sector. This approach helps us understand the dataset classification by extracting decisions from each node of the tree-based models. These decisions are then converted into rules. It provides a powerful method for analysing the risk status of ongoing customers and gaining meaningful insights for candidate customers. In this way, the factors causing customers or groups to fall into certain classes are revealed. This enables banks to make strategic decisions in marketing, risk management, and customer relations while developing better strategies for potential customers.

There are studies in the literature on the use of rule extraction and tree-based models in customer classification. However, these studies often overlook the simultaneous evaluation of the reliability, simplicity, and validity of the extracted rules. Additionally, studies that focus on these aspects do not incorporate 'Minimum Description Length' (MDL)-based simplicity research. The “*Children of the Tree*” approach addresses this gap by measuring the confidence and validity of each rule using quantitative metrics for interestingness. Additionally, it selects simpler rules by performing an MDL cost analysis. Thus, it reduces complexity and ensures that the obtained rules are based on a more meaningful and simple basis. This is essential in high-risk sectors such as banking because the accuracy of the rules used in customer classification and risk analysis can directly affect the business strategies of banks.

In addition, generating insights for candidate customers by seeing the current risk status of the bank through existing customers will be a meaningful and valid analysis. In this respect, the “*Children of the Tree*” algorithm will provide a unique solution in customer classification and risk analysis in the banking industry and will make significant contributions. This study effectively addresses imbalanced datasets by incorporating the SMOTE technique, demonstrating its strength in this area.

In this study, Chapter 1 reviews the existing rule extraction algorithms. Mathematical programming-based algorithms and machine learning-based algorithms were investigated and verbally compared with our algorithm “Children of the Tree”. In addition, Chapter 1 briefly touches on the background of the association rule mining metrics and the Minimum Description Length (MDL) method, which is used to evaluate the inferred rules that are the basis of our algorithm. In addition, in this section, the SMOTE method is briefly explained and the dataset used in the study is introduced. Chapter 2 explains the basic principles, working steps, and mathematical background of our algorithm. In addition, it is discussed in detail how the metrics used in the evaluation of the rules extracted in our algorithm are calculated. Chapter 3 covers the applications of our algorithm to the dataset. In this section, the practical results and performance of our algorithm are conveyed. Chapter 4 is the discussion and conclusion section. In this section, our algorithm is compared with RuleFit [10] and Anchors [18], which are widely known in the literature and can be used as open source, on the German Credit Data dataset. According to the benchmark results, our algorithm generally produces a higher F1 score than these rule extraction algorithms. In addition, our algorithm could provide high and



close F1 scores for both classes, considering the minority class as well as the majority class. This shows that our study exhibits superior performance on unbalanced datasets.

According to the benchmark results, the Children of the Tree algorithm demonstrates superior performance, achieving an overall F1 score of 0.80 compared to 0.74 for RuleFit. Specifically, the F1 score for Class 1 (good class) is 0.80 for Children of the Tree and 0.73 for RuleFit, while for Class 2 (bad class), the F1 score is 0.81 for Children of the Tree and 0.75 for RuleFit. Additionally, when compared to Anchors, the Children of the Tree algorithm stands out by evaluating rules across broader metrics. Anchors provide high precision values, but these are often associated with very small subsets of the dataset, as seen in its low coverage values (e.g., 0.0104 and 0.0147 for Class 1 and Class 2, respectively). In contrast, Children of the Tree achieved confidence values analogous to precision, such as 0.78 for Class 2 and 0.69 for Class 1, while maintaining significantly higher support values (up to 0.48 for Class 2 and 0.40 for Class 1). This balance between confidence and support ensures that the generated rules are both meaningful and scalable. By addressing the challenges of imbalanced datasets and generating reliable, interpretable rules, Children of the Tree proves to be a robust and practical alternative to existing methods like RuleFit and Anchors.

### 1.1. Rule Extraction in Literature

The issue of rule extraction from ML models has been discussed and used in many areas (e.g., finance, healthcare, etc.). Until now, many researchers have conducted studies ranging from the interpretation of models, explainable/interpretable models, to the detection of model tendencies through rule extraction, and even the evolution of these rules into use by business units/owners. Rule extraction from tree-based machine learning models is also an important approach to make the decision processes of the model more interpretable. In ensemble models such as Random Forest, the analysis of rules from multiple decision trees is used to increase the explainability of the model and to obtain reliable decisions [1]. In this context, various algorithms that work in the form of "if-then" rules stand out in the literature.

*inTrees* Framework [2] ranks the rules extracted from the RF according to metrics such as frequency, error rate, and length. It reduces the complexity of the rules and makes them simpler and more understandable. However, it does not use the MDL technique, which is the way to reduce the complexity and provide the simplicity by analysing the model's error and rule length. In addition, it has some limitations in the area of rule overlap and explainability. *ExtractingRuleRF* [3] extracts and ranks the rules from RF using a greedy algorithm. This method prioritises predictive accuracy while limiting the interpretability coverage. *SIRUS* [4] is a method derived from RF and based on the rule frequency. It focuses on creating shorter and more stable rules. *RF+HC* [5] reduces the number of rules extracted from RF using the Hill-Climbing algorithm. This method applies optimisation to create small and meaningful rule sets. *defragTrees* [6] simplifies RF rules using a Bayesian model selection algorithm and optimises predictive performance. It preserves explainability while reducing complexity. *ForEx++* [7] generates high-quality rule ensembles from decision forests. It presents a framework based on average metrics and focuses on predictive performance. *MIRCO* [8] uses mathematical programming to minimize rule heterogeneity and complexity. This method produces rules that better represent the minority class.

*OptExplain* [9] extracts rules using logical inference, sampling and optimisation techniques. It offers the ability to explain particularly large data coverage. *RuleFit* [10] combines rules extracted from RF nodes with sparse linear regression. However, it may experience instability when working with highly correlated rules. *Node Harvest* [11] uses the rules extracted from RF as a weighted prediction model. It works with non-negative weights and provides more explainable models. *Forest-ORE* (Optimal Rule Ensemble) [12] is a method that

generates an explainable rule ensemble from Random Forest models. This method uses Mixed-Integer Programming (MIP) to optimise the balance between the predictive performance, rule coverage, and rule complexity.

Our algorithm “*Children of the Tree*” optimises rules using MDL (Minimum Description Length) [13], which offers a different evaluation mechanism than most other algorithms in the literature. The inTrees algorithm also evaluates rules in terms of length calculation to simplify rule sets, but it does not conduct an MDL-based optimisation study. For example, while algorithms such as *CN2* and *RIPPER* usually focus on metrics such as support and confidence, our approach evaluates the complexity and accuracy of the rules together [14,15]. This distinct evaluation approach makes direct comparison with algorithms like *CN2* and *RIPPER* challenging, as they do not incorporate complexity into their assessments. In addition, MDL-based optimisation allows the rules to be more compact and explainable. This unique feature places our algorithm in a distinct category, emphasising explainability and balance between complexity and accuracy. The fact that our algorithm can successfully work on imbalanced data sets is a significant advantage. Classical rule extraction algorithms such as *CN2* and *RIPPER* generally tend to overfit the majority class in such data sets [14,15].

On the other hand, our algorithm can effectively target the minority class in imbalanced datasets and extract meaningful rules for this class. This makes it unnecessary to compare our algorithm in the same context with others.

These studies have essentially set us a benchmark. Although they did not directly use the MDL principle in terms of rule extraction from the model, some of these studies aimed to balance model fit and complexity, which indirectly resembled the basic ideas of MDL.

Table 1 includes the basic properties, advantages, and disadvantages of mathematical programming-based algorithms in the literature that create benchmarks, and the comparison of the application domain in which they are suited.

**Table 1.** Mathematical programming-based algorithms

Algorithm Name	Basic Properties	Advantages	Disadvantages	Application Domain/ Fields
Forest-ORE [12]	It extracts rule sets from RF models that can be explained by the mixed-integer optimisation.	It strikes a balance between predictive performance and explainability.	It has high computational cost and is time consuming on large datasets.	Global model explainability and minority classes.
MIRCO [8]	It minimizes the total rule complexity and heterogeneity using mathematical programming.	Creates rules that better represent the minority class.	The computational cost is high.	Risk analysis and data mining.
OptExplain [9]	It creates rules through logical inference, sampling and optimisation techniques.	Explains the broad scope of data.	Logical operations and optimisation processes can be complex.	Optimisation and logical inference.
defragTrees [6]	It simplifies and optimises RF rules through Bayesian model selection.	Predictive simplifies the rules while maintaining accuracy.	Bayesian processes can be slow on large datasets.	Bayesian modelling, finance, and healthcare.

On the other hand, Table 2 includes the comparison of the basic properties, advantages, disadvantages and application domain of rule extraction algorithms that can work on machine learning models in the literature that creates benchmarks for us.

**Table 2.** Machine Learning-Based Algorithms

Algorithm Name	Basic Properties	Advantages	Disadvantages	Application Domain/Fields
RuleFit [10]	It derives if-then based rules, learns rules from complex models and combines them with linear models.	L1 regularisation selects important rules and provides a balance between accuracy and explainability.	Requires SMOTE or optimisation on unbalanced datasets.	Classification and regression, financial analysis.
RIPPER [15]	It is if-then based and uses incremental pruning to reduce errors in rules.	Creates simple, fast, and explainable rules.	Performance may degrade on large data sets.	Medicine, finance, and small data sets.
CN2 [14]	It is if-then based and produces rules with an implicit (covering) algorithm.	It is powerful in unbalanced data sets and generates meaningful rules.	Accuracy may decrease in complex data sets.	Biology, medicine, classification.
PART [16]	It is if-then based and extracts partial rules from the decision trees.	It creates explainable and simple rules and is effective in multi-class problems.	Performance in complex relationships is limited.	Education, classification.
Bayesian Rule Lists (BRL) [17]	It is if-then based, sorts and optimises the rules according to the Bayesian probability model.	It offers a balance of explainability and accuracy and is robust on small datasets.	The computational cost is high for large data sets.	Healthcare/Medicine, law, finance.
Anchors [18]	It is if-then based, creating local rules that explain each predicted situation.	It is powerful in making sense of complex patterns and produces explanatory and intuitive rules.	Scalability may be limited to large datasets.	Model explainability and engineering.
C4.5 ve CART [19,20]	It is if-then based and creates rules with the decision tree algorithm.	Easy to apply, fast and explainable.	Is prone to overfitting.	Classification and regression, training.
Slipper [21]	It is if-then based and increases the accuracy of the rules with boosting.	It improves performance and can be effective on imbalanced datasets.	The computational cost may increase due to boosting.	Binary classification.
Scalable Rule-Based Learner (SRL) [22]	It is if-then based and creates scalable rules on large datasets.	It is fast, explainable and optimisable on large datasets.	May produce oversimplified results on small data sets.	Large data sets and real-time applications.
Interpretable Decision Sets (IDS) [23]	It is if-then based and produces non-overlapping and low-complexity rules.	Explainability is at the forefront, and the overlap between rules is minimized.	The computational cost is high for large data sets.	Healthcare, law, and sectors requiring high reliability.

Algorithm Name	Basic Properties	Advantages	Disadvantages	Application Domain/Fields
EBM (Explainable Boosting Machines) [24]	It is based on Gradient Boosting and creates explainable rules by modelling each feature independently.	Near-Gradient Boosting accuracy, meaningful explanations.	Performance in complex relationships may be limited.	Healthcare, finance, and critical decision-making processes.
TE2Rules [25]	It optimises the rules extracted from the tree ensemble models with a balance of fidelity and explainability.	High balance of fidelity and explainability; covers all decision paths.	It has high computational cost and is time consuming on large datasets.	Machine learning, explainable models.
SIRUS [4]	Creates stable and explainable rule sets; derived from RF models.	It creates short and decisive rules and provides stability.	May overlook rare but important rules.	Regression and classification, stable models.
inTrees [2]	It derives rules from all the decision paths in the RF and optimises these rules.	Optimises by considering the frequency of rules.	The rules are highly expressive, but complexity can increase in large data sets.	Machine learning, predictive models.
ExtractingRuleRF [3]	It extracts rules from the RF and weights them with the greedy algorithm.	Optimises the accuracy and coverage of the rules.	Predictive accuracy is prioritised over explainability.	Predictive performance, financial analysis.
RF + HC (Hill-Climbing) [5]	It uses hill-climbing to optimise the rules within the RF.	Creates small and meaningful rule sets.	The optimisation process can be lengthy.	Optimised small datasets.
Node Harvest [11]	It combines the rules obtained from the RF nodes with a weighted prediction model.	Creates a simple rule set with non-negative weights.	Predictive performance may be limited.	Machine learning, rule-based analysis.
ForEx++ [7]	Generates high-quality rule populations, improving the predictive performance.	Optimises predictive performance.	Optimised rule size may limit explainability.	Risk analysis and data mining.
<b>Children of the Tree</b> <sup>1</sup>	It is if-then based, uses RF models, is suitable for working on imbalanced data sets, creates rules by balancing with SMOTE and ranks according to MDL cost.	It achieves high accuracy and F1 scores in both classes and finds the least cost rules.	Since computational costs can be high in numerous data sets, feature-based filtering should be added for such data sets.	Healthcare, finance, data-intensive sectors, imbalanced data sets, Machine Learning classification problems.

<sup>1</sup>This article describes an algorithm for extracting rules from a new model. The algorithm details are available in Section 2. The Application results are available in Section 3.





## 1.2. Quantitative Association Rule Mining Measures in the Literature

Association rule mining is a technique frequently used in data mining to discover dependencies and patterns between elements in large data sets. It is also referred to in the literature as “interestingness metrics”. In particular, metrics such as “support”, “confidence”, and “lift” are among the most used metrics in association rule mining. The development of these metrics provides valuable information to the user by determining meaningful relationships between elements in the data. This approach enables the analysis of past associations to inform future studies and decision-making processes [26, 27].

*Support* and *confidence* metrics, first introduced by Agrawal and Srikant<sup>2</sup>, form the basis of association rule mining and express the probability of the co-occurrence of elements in a dataset [26]. Later, additional metrics such as *lift* were developed to help determine the degree of dependency of the rules, expressing positive or negative dependencies. Interestingness metrics such as *the certainty factor* and *netconf* provide more meaningful results, especially by eliminating misleading or independent rules<sup>3</sup> [27].

## 1.3. Minimum Description Length in the Literature

The principle of Minimum Description Length (MDL) was developed through a series of papers, primarily by Jorma Rissanen [28,30,31]. Its roots lie in the Kolmogorov or algorithmic complexity theory developed by Solomonoff, Kolmogorov, and Chaitin in the 1960s<sup>4</sup> [32].

The Minimum Description Length (MDL) principle is a formalisation of Occam’s Razor in machine learning and statistics. In model selection, MDL seeks to balance model complexity and data adaptability [29].

### 1.3.1. Concept of MDL

MDL suggests that the best model for a dataset is the one that compresses the data most effectively. This approach consists of two main components [13,33]:

- **Model Complexity ( $L(h)$ ):** Refers to the definition length of the hypothesis or model, i.e., the number of bits necessary to represent the model. A simple model usually has a shorter definition length.
- **Data Adaptation Cost ( $L(D | h)$ ):** Refers to the length required to describe the data based on the model or hypothesis. A well-fitting model requires fewer bits to encode its errors or deviations from the data.

### 1.3.2. MDL Formulation

Mathematically, the total definition length  $L(D, h)$  is given by [33]:

$$L(D, h) = L(h) + L(D | h) \quad (1)$$

Here,

$L(h)$ : It is the definition length (complexity) of the model itself,

$L(D | h)$ : The definition length of the model relative to the data (error or redundancy encoding).

In MDL, the goal is to minimize this total definition length. This strikes a balance between model simplicity and data fidelity.

<sup>2</sup>See: References Section, source number 26.

<sup>3</sup>For a detailed analysis, please see the “Measures” section under the title “QUANTITATIVE ASSOCIATION RULES” in the 2nd Chapter of the References Section, reference number 27.

<sup>4</sup>For details, the source numbered 32 in the Bibliography Section can be examined.



## 1.4. SMOTE Method

SMOTE (Synthetic Minority Oversampling Technique) was introduced by Chawla et al. in 2002. It aims to enhance classification models by increasing the number of minority class samples in imbalanced datasets, enabling better predictions for the minority class. SMOTE produces synthetic data points by interpolating between an example in the minority class and one of its k-nearest neighbours, providing a wider decision boundary for the minority class. The effectiveness of SMOTE is usually evaluated by metrics such as AUC [34].

## 1.5. Introduction of the Dataset

In this study, we used the Statlog (German Credit Data)<sup>5</sup> dataset, a widely recognised resource in the literature for credit risk analysis. This dataset effectively captures the characteristics of bank customers and is suitable for extracting rules related to customer risk levels. This dataset is used to determine the risk level of prospective customers in the bank's marketing and risk management departments. This dataset is valuable because it combines the demographic, financial, and behavioural characteristics of the applicants. In assessing credit risk, multidimensional data such as a customer's age, employment status, past credit payments, credit period, and requested credit amount provide detailed insights into potential risks. In such imbalanced datasets, it is critical to develop models with high predictive accuracy and derive statistically significant rules to ensure effective decision-making.

Talking about the nature of the dataset is important to understand the area that the study serves. The dataset includes 20 features and 1 target variable, which are used to evaluate loan applications. These features include demographic, financial, and behavioural information about each applicant. Age indicates the age of the applicant, while Personal Status and Sex refers to the applicant's marital status and gender. Housing represents the ownership status of the applicant's residence and is classified as "own house," "rent," or "free accommodation". Number of Dependents represents how many people rely on the applicant financially [35].

Among the financial characteristics in the data set, the loan amount (Credit Amount) refers to the amount of credit requested; Duration in Months indicates the repayment period of the loan in months. The amount of deposits (Savings) refers to the amount of the applicant's savings and is categorically divided into different ranges from low to high. The Existing Property attribute classifies the type and value of properties owned by the applicant. Other Installment Plans indicate whether the applicant has additional loan agreements. The Other Debtors attribute indicates whether the applicant has a guarantor or other debtors in the loan application. Employment status and occupation (Job) is a characteristic that categorically expresses the employment status and occupation of the applicant. Employment Duration is the time worked in the current workplace, and the length of this period can be a criterion for measuring financial stability.

Among behavioural characteristics, the number of existing credits that the applicant has is an important factor in evaluating the loan application. On the other hand, the purpose of the loan indicates the purpose for which the loan is taken and is divided into categories such as "car", "furniture", "education". Credit History provides a summary of the applicant's past loan payments, and regularity in payments plays a fundamental role in understanding credit risk.

Telephone ownership indicates whether the applicant has a phone, which is particularly important for communication. Foreign Worker indicates whether the applicant is a foreign employee. It is predicted that

---

<sup>5</sup>The dataset is sourced from the UCI Machine Learning Repository; It classifies people defined by a set of characteristics as having good or bad credit risks. For details, the source numbered 35 in the Bibliography Section can be examined.

these two features may offer indirect effects to the model in evaluating a person's loan application. The Credit Risk field, on the other hand, refers to our target variable, which consists of two different classes. Class 1 refers to applications in a good (no risk) condition, while class 2 refers to applications in a poor (risky) condition.

While all these features come together in different aspects in the credit risk analysis and provide information about the loan repayment capacity of the applicant, in our study, it was estimated whether the person was risky in terms of granting loans on the classification model and it was tried to provide meaningful rules for bank prospective customers based on the situation of the customers in the bank with different rule extractions according to this risk class.

Table 3 shows the features in the dataset can be summarised in tabular form as follows.

**Table 3.** Original German Loan Dataset Specifications and Descriptions<sup>6</sup>

Features Name	Data type	Demographics	Description
Checking Account	Categorical		Existing "checking account" information
Duration	Numeric		Loan Term (Term)
Credit History	Categorical		Credit History
Purpose	Categorical		Purpose of Obtaining Loans
Credit Amount	Numeric		Loan Amount
Saving Account	Categorical		Saving account information
Employment Duration	Categorical	Other	Length of work in the employee's current job (time interval)
Installment rate	Numeric		Installment rate as a percentage of disposable income
Personal Status and Sex	Categorical	Marital status	Marital status and gender
Other Debtors	Categorical		Other debtors/guarantors
Present Residence	Numeric		Where he currently resides
Property	Categorical		Properties
Age	Numeric	Age	Age
Other Installment Plans	Categorical		Other payment plans
Housing	Categorical	Other	Housing
Number of Credits	Numeric		Number of other loans in this bank
Job	Categorical	Profession	Work
Dependents	Numeric		Number of people responsible for providing care
Telephone	Binary		Phone
Foreign Worker	Binary	Other	Foreign Employee
Credit Risk	Binary	Target	Risk class (good/bad)

The input of the dataset to our model is as shown in Figure 1 below before the label encoder is made.

<sup>6</sup>For the details of the data set, source number 35 in the Bibliography Section can be examined.

checking	duration	credit_hi	purpose	credit_an	savings	employ	installme	personal	other_de	present	property	age	other_ins	housing	number	job	depende	telephon	foreign_v
A11	6	A34	A43	1169	A65	A75	4	A93	A101	4	A121	67	A143	A152	2	A173	1	A192	A201
A12	48	A32	A43	5951	A61	A73	2	A92	A101	2	A121	22	A143	A152	1	A173	1	A191	A201
A14	12	A34	A46	2096	A61	A74	2	A93	A101	3	A121	49	A143	A152	1	A172	2	A191	A201
A11	42	A32	A42	7882	A61	A74	2	A93	A103	4	A122	45	A143	A153	1	A173	2	A191	A201
A11	24	A33	A40	4870	A61	A73	3	A93	A101	4	A124	53	A143	A153	2	A173	2	A191	A201

**Figure 1.** German Credit Data-Input Data

The digitisation result of the value contained in each categorical variable is shown in Table 4 below.

**Table 4.** Encoding values based on decision variables and categories

Columns	Category	Descriptions/Values	Encoded
checking_account	A11	X < 0 DM (Deutsche Mark)	0
checking_account	A12	0 <= X < 200 DM	1
checking_account	A13	X >= 200 DM/salary assignments for at least 1 year	2
checking_account	A14	no checking account	3
credit_history	A30	no credits taken/all credits paid back duly	0
credit_history	A31	all credits at this bank paid back duly	1
credit_history	A32	existing credits paid back duly till now	2
credit_history	A33	delay in paying off in the past	3
credit_history	A34	critical account/other credits existing (not at this bank)	4
purpose	A40	Car (new)	0
purpose	A41	Car (used)	1
purpose	A410	others	2
purpose	A42	furniture/equipment	3
purpose	A43	radio/television	4
purpose	A44	domestic appliances	5
purpose	A45	repairs	6
purpose	A46	education	7
purpose	A48	retraining	8
purpose	A49	business	9
savings_account	A61	X < 100 DM	0
savings_account	A62	100 <= X < 500 DM	1
savings_account	A63	500 <= X < 1000 DM	2
savings_account	A64	X >= 1000 DM	3
savings_account	A65	unknown/no saving account	4
employment_duration	A71	unemployed	0
employment_duration	A72	X < 1 year	1
employment_duration	A73	1 <= X < 4 years	2
employment_duration	A74	4 <= X < 7 years	3
employment_duration	A75	X >= 7 years	4
personal_status_sex	A91	Male: divorced/separated	0
personal_status_sex	A92	female : divorced/separated/married	1
personal_status_sex	A93	male: single	2
personal_status_sex	A94	Male: married/widowed	3
other_debtors	A101	none	0



Columns	Category	Descriptions/Values	Encoded
other_debtors	A102	co-applicant	1
other_debtors	A103	guarantor	2
property	A121	real estate	0
property	A122	Building a society savings agreement/life insurance	1
property	A123	car or other, not in attribute 6	2
property	A124	unknown/no property	3
other_installment_plans	A141	bank	0
other_installment_plans	A142	stores	1
other_installment_plans	A143	none	2
housing	A151	rent	0
housing	A152	own	1
housing	A153	for free	2
job	A171	unemployed/ unskilled - non-resident	0
job	A172	unskilled-resident	1
job	A173	skilled employee/official	2
job	A174	management/ self-employed/highly qualified employee/officer	3
telephone	A191	none	0
telephone	A192	yes, registered under the customer's name	1
foreign_worker	A201	yes	0
foreign_worker	A202	no	1
credit_risk	1	Good	1
credit_risk	0	Bad	0

## 2. Children of the Tree

This article introduces a new algorithm for rule extraction from tree-based classification models such as Random Forest. In domains where rule-based predictions are crucial, such as banking, meaningful rules for prospective customers play a key role in the development of the algorithm. This algorithm is expected to provide a simple yet effective solution for the risk management and marketing departments of banks compared to existing methods in the literature.

Thanks to the "Minimum Description Length (MDL)" method, the algorithm reduces the rules that lead to unnecessary complexity and thus allows the creation of a more understandable and optimised rule set.

### 2.1. The Foundation of the "Children of the Tree"

A random forest model is an ensemble learning method that consists of multiple decision trees. Each decision tree classifies or predicts the  $T_i$  ( $i = 1, 2, \dots, N$ ) dataset through specific rules.

Decision trees typically start with a root node ( $r$ ). This root node represents the starting point of the dataset.

If a node  $d$  is a leaf node, then no distinction is made on  $d$ , and that node represents a class label or estimated value.

Leaf node metrics,  $M(d)$ , are calculated after all rule extraction is complete and the significance and explainability of each rule are evaluated. These metrics include criteria such as support, confidence, and MDL (Minimum Description Length).

$$M(d) = f(d)$$

Here,  $f(d)$  represents the metrics that indicate the significance of the rule inferred at the leaf node  $d$ .

In our study, these metrics are support, confidence, lift, leverage, conviction and mdl values.

If the inner node is  $d$ , it can have two child nodes ( $d_L$  and  $d_R$ ), and this node differentiates on the data using a property  $X_j$  and a threshold value  $\theta_j$ .

The decision rule is created by using the property and threshold value on the inner node. Mathematically, the rule for the inner node  $d$   $\varphi(d)$  is expressed as follows:

$$\varphi(d) = \begin{cases} X_j \leq \theta_j & \text{If } d \text{ switches to the left child node} \\ X_j > \theta_j & \text{If } d \text{ switches to the right child node} \end{cases}$$

This rule splits the pieces of data that are separated from the inner node into two. The left and right child nodes represent the subsets that this rule creates.

## 2.2. Algorithm Iterations

Below are the steps of "Children of the Tree", our rule-extraction algorithm from a tree-based machine learning model.

### ALGORITHM 1: Children of the Tree Algorithm Iterations

```

current_node root
current_node_type internal
current_node is not null
while current_node is not null, do
  if current_node is leaf, do
    add rule for current_node to rule_list
  else
    identify the left_child and right_child of the current_node
    split_data using feature  $X_j$  and threshold  $\theta_j$ 
    Generate rules:
      if  $(X_j \leq \theta_j)$  move to the left_child
      if  $(X_j > \theta_j)$  move to the right_child
    end
  Move to the next node in the tree (left_child or right_child)
end

```

- **Beginning:**

The decision tree is started from the root node ( $r$ ).

- **Decision Function:**

Any node  $d$  in a decision tree uses a decision function  $f(d)$  to classify the dataset. This function determines how the separation in the node and the results are obtained. In general, the decision function is defined as

$$f(d) = \begin{cases} \text{Class } C1 & \text{If } d \text{ is the leaf node} \\ \text{Decision } (X_j \leq \theta_j \text{ ve } d_L) \text{ veya } (X_j > \theta_j \text{ ve } d_R) & \text{If } d \text{ is the inner node} \end{cases}$$



Here,  $C_1$  is the class label on the leaf node, or it represents the predicted value. In the inner node, the decision function represents the rules that divide the dataset into two.

- **Leaf Node Control:**

If node  $d$  is a leaf node, the extracted rule is added to the list of rules:

Rule List  $\leftarrow$  Rule List  $\cup \{(D)\}$

- **Internal Node Processing:**

If node  $d$  is the inner node:

The left ( $d_L$ ) and right ( $d_R$ ) child nodes are passed.

Using the property  $X_j$  and the threshold value  $\theta_j$ ,

Rules( $d$ )  $\leftarrow \{(X_j \leq \theta_j \text{ and } d_L), (X_j > \theta_j \text{ and } d_R)\}$

Rules are created and the child nodes are passed.

Switching to the left and right child nodes allows similar operations to be performed on each child node. After the rules are issued, metrics such as support, confidence, lift, leverage, and conviction are calculated. The MDL (Minimum Description Length) value is calculated by evaluating the complexity and error costs of each rule. These metrics are evaluated after the entire rule extraction process is completed and the optimal rules are selected.

### 2.3. Evaluation of the Rules Metrics

"Children of the Tree" uses support, confidence, lift, leverage, and conviction metrics to determine how meaningful a rule or association is in the dataset and how interesting it is.

**Support:** Indicates how often the rule is passed in the dataset.

$$\text{Support} = \frac{N_{\text{rule}}}{N_{\text{total}}}$$

$N_{\text{rule}}$ : The number of instances to which the rule applies.

$N_{\text{total}}$ : Total number of instances.

**Confidence:** Confidence measures the probability that a rule is true. It refers to how often the rule is true, especially when given a property or condition.

$$\text{Confidence} = \frac{N_{\text{correct}}}{N_{\text{rule}}}$$

$N_{\text{correct}}$ : The number of instances where the rule is true.

$N_{\text{rule}}$ : The number of instances to which the rule applies.

**Lift:** Measures how well the rule is relative to the expected accuracy.

The upgrade shows how good the rule is compared to the expected accuracy rate.

$$\text{Lift} = \frac{\text{Confidence}}{\frac{N_{\text{class}}}{N_{\text{total}}}}$$

$\text{Confidence}$ : The confidence value of the rule.

$\frac{N_{\text{class}}}{N_{\text{total}}}$ : The proportion of the class that exists as a result of the rule in the dataset.

**Leverage:** Subtracts the support value of the rule from its expected support in the case of independence.

$$\text{Leverage} = \text{Support} - \frac{N_{\text{antecedent}}}{N_{\text{total}}} \times \frac{N_{\text{class}}}{N_{\text{total}}}$$

**Support:** The support value of the rule.

$\frac{N_{\text{antecedent}}}{N_{\text{total}}}$ : The expected accuracy of the feature (or condition).

$\frac{N_{\text{class}}}{N_{\text{total}}}$ : The proportion of the class that exists as a result of the rule in the dataset.

**Conviction:** It measures how persuasive the rule is to ensure its accuracy.

$$\text{Conviction} = \frac{1 - \frac{N_{\text{class}}}{N_{\text{total}}}}{1 - \text{Confidence}}$$

**Confidence:** The confidence value of the rule.

$\frac{N_{\text{class}}}{N_{\text{total}}}$ : The proportion of the class that exists as a result of the rule in the dataset.

## 2.4. Application of MDL in Rule Extraction

In the context of rule extraction, MDL can be used to evaluate the "cost" of a set of rules. Section 1.3 "Minimum Description Length in Literature" describes the way MDL is calculated in the literature. Accordingly, a rule that is too complex (with too many conditions) will have a high  $L(h)$  value, while a rule that adapts poorly to the data will have a high  $L(D|h)$  value. The best rule set would be the one that minimizes this unified definition length. Our algorithm implements MDL calculations based on the principles outlined in the literature.

### 1. Sample Calculation for MDL Cost

In our rule extraction algorithm, the MDL value is calculated as follows, alongside other metrics commonly used in the literature:

**Rule Complexity ( $L(h)$ ):** If a rule  $r$  has  $k$  conditions, each condition contributes to the complexity. For example, if we assume that each condition requires a certain number of bits, the total complexity of the rule can be expressed approximately as follows:

$$L(h) = k \cdot \log_2(N) \quad (2)$$

Here,  $N$  refers to the number of samples in the dataset and can be thought of as a "resolution" that determines the complexity of the rule.

**Cost of Error ( $L(D|h)$ ):** This represents the number of samples that were misclassified or incorrectly estimated by the rule. Let  $E$  be the number of misclassifications:

$$L(D|h) = E \cdot \log_2(N) \quad (3)$$

**Total MDL Cost:** By combining the two components, the MDL cost of an  $r$  rule can be written as:

$$\text{MDL}(r) = k \cdot \log_2(N) + E \cdot \log_2(N) \quad (4)$$

This cost function promotes simple rules (low  $k$ ) and correct rules (low  $E$ ).

### MDL's Interpretation

- A low MDL value indicates a good balance between model complexity and accuracy.
- A high MDL value indicates that the rule is either too complex (high  $k$ ) or contains too many errors (high  $E$ ), which makes it less preferable.

The interpretation of low or high MDL values should be considered relative to all other calculated MDL values in the dataset.





### 3. Application

This section presents the application of our algorithm. The functionality of "*Children of the Tree*" is explained in detail in Chapter 2. In this section, the "*Children of the Tree*" algorithm was applied on the dataset introduced in Section 1.5 "Introduction of the Dataset" and the results were obtained as shown in Table 5.

**Table 5.** Children of the Tree Algorithm Results: Rules, Metrics, and Related Classes

Rule	Class	Support	Confidence	Lift	Leverage	Conviction	MDL Cost
duration > 6.50 and savings_account <= 2.50 and checking_account <= 2.50 and other_debtors <= 1.50 and credit_history <= 3.50.	2	0,48	0,78	1,56	0,24	2,28	9176,16
savings_account <= 2.50 and other_debtors <= 1.50 and checking_account <= 2.50 and credit_history <= 3.50 and duration > 10.50.	2	0,45	0,79	1,59	0,22	2,42	9479,25
savings_account <= 1.50 and credit_history <= 3.50 and checking_account <= 2.50 and other_debtors <= 1.50 and personal_status_sex <= 2.50.	2	0,49	0,79	1,57	0,22	2,34	9521,05
present_residence <= 3.50 and checking_account <= 2.50 and foreign_worker <= 0.50 and other_debtors <= 1.50 and purpose <= 8.50.	2	0,43	0,76	1,52	0,22	2,07	9865,94
checking_account <= 2.50 and other_debtors <= 1.50 and credit_history <= 3.50 and duration > 12.50 and savings_account <= 3.50.	2	0,39	0,80	1,62	0,19	2,62	10116,77
checking_account <= 2.50 and other_debtors <= 1.50 and credit_amount > 1044.50 and number_credits <= 1.50 and credit_amount > 1173.00.	2	0,43	0,70	1,41	0,22	1,70	10189,93
credit_history <= 3.50 and checking_account <= 2.50 and duration > 7.50 and credit_history > 1.50 and other_debtors <= 1.50.	2	0,43	0,70	1,39	0,21	1,65	10325,80
credit_amount <= 10918.00 and telephone <= 0.50 and checking_account <= 2.50 and employment_duration <= 3.50 and dependents <= 1.50	2	0,40	0,74	1,48	0,20	1,92	10398,96
checking_account <= 2.50 and duration > 10.50 and	2	0,37	0,78	1,56	0,19	2,29	10409,41

Rule	Class	Support	Confidence	Lift	Leverage	Conviction	MDL Cost
other_debtors <= 1.50 and employment_duration <= 3.50 and telephone <= 0.50.							
employment_duration > 1.50 and purpose <= 8.50 and duration <= 25.50 and credit_history > 1.50 and credit_amount <= 7452.00.	1	0,40	0,69	1,38	0,20	1,61	10628,88
duration <= 27.50 and housing > 0.50 and employment_duration > 1.50 and credit_history > 1.50 and other_installment_plans > 1.50.	1	0,29	0,78	1,55	0,15	2,24	11339,56
duration <= 24.50 and savings_account <= 3.50 and employment_duration > 1.50 and credit_amount <= 7881.00 and duration > 7.00.	1	0,39	0,59	1,17	0,19	1,21	11381,37
other_installment_plans > 1.50 and employment_duration > 1.50 and credit_history > 1.50 and credit_amount <= 3897.50 and housing > 0.50.	1	0,28	0,79	1,58	0,14	2,37	11444,08
personal_status_sex > 1.50 and duration > 8.50 and duration <= 24.50 and credit_amount > 1081.00 and credit_amount <= 7521.00.	1	0,3	0,66	1,33	0,15	1,49	11768,06
credit_history > 1.50 and foreign_worker <= 0.50 and other_installment_plans > 1.50 and credit_amount <= 3916.00 and number_credits <= 1.50.	1	0,33	0,59	1,18	0,17	1,23	11788,97
savings_account <= 2.50 and duration <= 24.50 and employment_duration > 1.50 and duration > 7.50 and credit_history > 1.50	1	0,31	0,63	1,27	0,15	1,37	11830,77
checking_account > 2.50 and age > 24.50 and duration <= 45.00 and credit_amount <= 9569.00 and other_installment_plans > 1.50.	1	0,20	0,93	1,86	0,10	7,13	11914,38
other_installment_plans > 1.50 and personal_status_sex	1	0,27	0,70	1,40	0,13	1,66	11966,64



Rule	Class	Support	Confidence	Lift	Leverage	Conviction	MDL Cost
> 1.50 and credit_history > 1.50 and property <= 2.50 and savings_account <= 3.50.							

The "*Children of the Tree*" algorithm was produced with the original ideas of the authors, since no similar solution was found in the literature. In particular, the fact that the MDL interpretation is used in the step of rule extraction from machine learning models makes our algorithm valuable in terms of making a unique contribution to the literature.

The "*Children of the Tree*" algorithm is an optimisation algorithm for extracting rules from the machine learning model, based on Random Forest. In the study, the SMOTE technique was used to make the algorithm work on unbalanced data sets.

The study focuses on two main components. First, the imbalanced dataset was balanced using the SMOTE technique. Various hyperparameter optimizations were applied to the Random Forest model, achieving an overall F1 score of approximately 0.81. The F1 scores were 0.80 for Class 1 (good class) and 0.81 for Class 2 (bad class). Based on these results, the rule extraction mechanism detailed in Section 2 of the "*Children of the Tree*" algorithm was implemented on the resulting model. After extracting the rules, metrics such as support, confidence, lift, leverage, and conviction were calculated. The MDL value was determined by assessing the complexity and error costs of each rule. After that, all these metrics were evaluated. The optimal rules were selected. Table 5 presents the optimal rules, the classes they represent, and the corresponding metric values, including validity and prevalence. Additionally, the table highlights the MDL cost values, reflecting the simplicity of these rules.

In Table 5 above, our "*Children of the Tree*" algorithm is applied to "German Credit Data" to extract rules for potential customers through bank customers. The first column lists the extracted rules, while the second column indicates the classes to which the rules belong. In other words, if a rule applies to the bank's prospective customer, it indicates whether the customer is potentially risky or has a low risk level. The support values in the table indicate the frequency of each rule in the dataset, while the confidence values represent their validity. As is known, the MDL Cost field is a reflection of the mathematical calculation of the simplicity value of the rule in the table.

MDL costs are problem-specific and may vary significantly across different datasets and studies, sometimes being higher or lower than the values observed here. However, the lowest of the costs given in this study was calculated for class 2 and this cost belongs to the rule "duration > 6.50 and savings\_account <= 2.50 and checking\_account <= 2.50 and other\_debtors <= 1.50 and credit\_history <= 3.50" with a value of 9176.16<sup>7</sup>. The rule with the lowest MDL cost for Class 1 is "employment\_duration > 1.50 and purpose <= 8.50 and duration <= 25.50 and credit\_history > 1.50 and credit\_amount <= 7452.00" with an MDL Cost value of 10628.88<sup>8</sup>.

<sup>7</sup>The decision limit determined for each variable in these rules can be read with the numerical equivalents of the values described in Section 1.5 Introduction to the Dataset in Table 4 Table of Encoding Values Based on Decision Variables and Categories.

<sup>8</sup>Example rule reading: Decision class 1 for loans of 25.50 months or less when the employment\_duration is A73, A74 or A75, and for those with a credit history of less than 7452, the credit history is A32, A33 or A34. See Section 1.5 for a variable-based explanation of each categorical statement.



## 4. Conclusion of the Article and Future Works

In our study, the MDL principle is used to enhance the explainability and simplicity of the rules. MDL is useful in minimizing the definition length of the rules by considering the complexity of the rule (number of conditions) and the cost of error (number of misclassifications). This encourages rules that are not only true but also relatively less complex and explainable with fewer conditions. Our study introduces a new and effective algorithm that prioritises less complex rules, evaluates simplicity using MDL cost, and generates more general rules by accounting for the minority class in imbalanced datasets. This approach distinguishes it from alternative methods in rule optimisation.

In the future, we aim to enhance the algorithm's ability to generate more stable rules, similar to its competitors, while producing more general rules with higher accuracy and simplicity. Additionally, calculations can be performed based on the extraction and importance of the rules for the selected features. This will enable the generation of more focused rules for specific classes. In future studies, it is aimed to enable users to identify the most important features and obtain rules focused on them.

Below, [Table 6](#) presents a comparison of our algorithm with RuleFit, as discussed in Section 1.1 "Rule Extraction in Literature" in terms of model performance. The results for RuleFit were obtained by applying it within its framework, with the `tree_generator` parameter set to `random_forest_classifier`. The same preprocessing steps were performed for both algorithms, and the same features were used as input to the models. Additionally, both datasets were balanced using the SMOTE technique and both random forest models were configured with the same hyperparameter values.

**Table 6.** Comparison of Model Performances of RuleFit and Children of the Tree

Algorithm	Accuracy	F1 Score for Class 2	F1 Score for Class 1
Rule Fit	0.74	0.75	0.73
Children of the Tree	0.80	0.81	0.80

The results for "Children of the Tree" stand out against RuleFit for Class 2 (bad) and Class 1 (good). The superiority of our algorithm is evident from [Table 6](#), as it enables the model to make more reliable predictions and better distinguish between classes, which contributes to the extraction of more trustworthy rules being extracted. The reason for comparing model performance is that RuleFit, like our algorithm, can train on a random forest model within its framework. Therefore, comparing the performance of the two algorithms on datasets processed with identical preprocessing steps and feature selection allows for a fairer evaluation of their classification success.

Let us continue our comparison with another algorithm, Anchors, and this time analyse the rules extracted by the Anchors algorithm when it is applied to the same random forest model used with Children of the Tree. [Table 7](#) presents some rules extracted by Anchors. Anchors evaluate the rules it extracts based on precision and coverage. Precision reflects the accuracy of a rule, while coverage represents the proportion of the dataset to which the rule applies.

**Table 7.** Anchor Results Based on the Highest Precision Values

Class	Anchors	Precision	Coverage
2	checking_account = 0 AND personal_status_sex = 1 AND duration > 28.00	1.0	0.04
2	checking_account = 1 AND credit_amount > 4176.25 AND telephone = 0 AND property = 2 AND other_installment_plans = 2	1.0	0.01

Class	Anchors	Precision	Coverage
1	employment_duration = 4 AND duration <= 19.00 AND age > 39.25 AND other_installment_plans = 2 AND 1361.75 < credit_amount <= 2319.50	1.0	0.01
1	duration <= 12.00 AND present_residence > 3.00 AND housing = 1 AND property = 0 AND credit_amount <= 2319.50	0.99	0.03

In Table 7, rules with these rules with high precision apply to negligible subsets of the data. In contrast, the Children of the Tree algorithm evaluates rules across broader metrics. While they may not be fully comparable, the coverage metric in Anchors, which indicates how much of the dataset a rule covers, can be compared to the support values of Children of the Tree. Similarly, Anchors’ precision, which indicates how often a rule is correct, can be compared to the confidence values of Children of the Tree.

When analysed by class, the highest support and confidence values observed in Children of the Tree were 0.48 and 0.78 for Class 2 and 0.40 and 0.69 for Class 1, respectively. For Anchors, when we examine a rule for Class 1 that is comparable in length to the rules extracted by Children of the Tree, specifically the rule “employment\_duration = 4 AND duration <= 19.00 AND age > 39.25 AND other\_installment\_plans = 2 AND 1361.75 < credit\_amount <= 2319.50,” its precision is 1, but its coverage is only 0.0104. This indicates that its applicability across the dataset is extremely low. Similarly, for Class 2, the rule in Anchors that is most similar in length and has the best combination of precision and coverage, “checking\_account = 1 AND credit\_amount > 4176.25 AND telephone = 0 AND property = 2 AND other\_installment\_plans = 2,” has a precision of 1 but a coverage of just 0.0147.

Table 8 presents the results of our Children of the Tree algorithm, ranked by the highest confidence values. When compared to Table 7 above, if we consider confidence as analogous to precision and support as analogous to coverage, it is clear that the Children of the Tree algorithm stands out compared to Anchors, especially when evaluated based on similar rules.

**Table 8.** Children of the Tree Algorithm's Results Ranked by the Highest Confidence

Rule	Class	Confidence	Support
age > 33.50 and purpose <= 4.50 and credit_amount <= 4814.00 and checking_account > 2.50 and duration <= 16.50	1	1.0	0.05
checking_account > 2.50 and duration <= 16.50 and other_installment_plans > 1.50 and credit_history > 3.50	1	1.0	0.04
housing <= 0.50 and checking_account <= 2.50 and present_residence <= 3.50 and dependents <= 1.50 and other_installment_plans <= 1.50	2	1.0	0.04
credit_history <= 3.50 and personal_status_sex <= 1.50 and credit_history <= 1.50 and savings_account <= 1.50 and installment_rate > 2.50	2	1.0	0.04

The comparison of the Children of the Tree algorithm with both Anchors and RuleFit highlights the strengths of the proposed approach. While Anchors provides highly precise rules, these rules often apply to extremely small subsets of the dataset, limiting their practical utility. On the other hand, RuleFit exhibited lower model performance under the same hyperparameters and preprocessing steps, resulting in reduced classification success compared to Children of the Tree. The Children of the Tree algorithm effectively balances confidence (analogous to precision) and support (analogous to coverage) while achieving higher F1 scores for both classes. This makes Children of the Tree a superior alternative, particularly for unbalanced datasets, where generating meaningful and scalable rules is critical for decision-making processes, such as credit risk analysis. By leveraging metrics such as MDL cost and support, Children of the Tree demonstrates its ability



to produce rules that are both interpretable and practical, distinguishing itself as a robust and effective alternative to existing algorithms like Anchors and RuleFit.

#### 4.1. Limitations of the Study

The "Children of the Tree" algorithm presents notable contributions to rule extraction, yet certain limitations remain that can guide future enhancements. While the algorithm demonstrates strong performance on the German Credit Data dataset, its generalizability to datasets with different characteristics has not been fully explored. Additionally, the current approach lacks the ability to prioritise rules based on user-specified features or domain knowledge, which could increase its applicability in real-world scenarios. Finally, although comparisons were made with Anchors and RuleFit, future work evaluating the algorithm against other state-of-the-art rule extraction and explainable AI techniques would provide deeper insights into its relative advantages and areas for improvement.




Peer Review	Externally peer-reviewed.
Author Contributions	Conception/Design of Study- H.M., M.B.; Data Acquisition – H.M.; Data Analysis/Interpretation- H.M., M.B.; Drafting Manuscript- H.M., M.B.; Critical Revision of Manuscript- M.B.; Final Approval and Accountability- H.M., M.B.; Technical or Material Support- H.M., M.B.; Supervision- M.B.
Acknowledgements	We acknowledge GTech for their support, as the preliminary version of this study was used in the “AI-Powered Decision Support and Early Warning System in the Finance and Banking Sector” project. This project provided the foundation and inspiration for the development of the current study. GTech’s domain expertise and contributions in machine learning applications enriched the study by guiding feature selection.
Conflict of Interest	The authors have no conflict of interest to declare
Grant Support	The authors declared that this study has received no financial support.

#### Author Details

##### Hilal Meydan

<sup>1</sup> Yıldız Technical University, Department of Mathematical Engineering, İstanbul, Türkiye

 0009-0000-6145-4418

##### Mert Bal

<sup>1</sup> Yıldız Technical University, Department of Mathematical Engineering, İstanbul, Türkiye

 0000-0001-6250-929X

## References

- [1] Leo Breiman. 2001. Random Forests. *Machine Learning*. 45, 1 (Oct. 2001), 5-32. <https://doi.org/10.1023/A:1010933404324>.
- [2] Houtao Deng. 2019. Interpreting Tree Ensembles with inTrees. *International Journal of Data Science and Analytics*. 7, 4 (Dec. 2019), 277-287. <https://doi.org/10.1007/s41060-018-0144-8>.
- [3] Kim Phung Lu Thi, Ngoc Chau Vo Thi, and Nguyen Hua Phung. 2015. Extracting Rule RF in Educational Data Classification: From a Random Forest to Interpretable Refined Rules. In *Proceedings of the 2015 International Conference on Advanced Computing and Applications (ACOMP)*. 20-27. <https://doi.org/10.1109/ACOMP.2015.13>.
- [4] Clément Bénard, Gérard Biau, Sébastien da Veiga, and Erwan Scornet. 2019. SIRUS: Stable and Interpretable Rule Set for Classification. arXiv:1908.06852. Retrieved from <https://arxiv.org/abs/1908.06852>.



- [5] Morteza Mashayekhi and Robin Gras. 2015. Rule Extraction from Random Forest: The C Methods. In Kanade, T., Kittler, J., Kleinberg, J.M., et al. (Eds.). *Advances in Artificial Intelligence*. Vol. 3060. Springer Berlin Heidelberg, Berlin, Heidelberg, 223–237. [10.1007/978-3-319-18356-5\\_20](https://doi.org/10.1007/978-3-319-18356-5_20)
- [6] Satoshi Hara and Kohei Hayashi. 2017. Making Tree Ensembles Interpretable: A Bayesian Model Selection Approach. arXiv:1606.09066. Retrieved from <https://arxiv.org/abs/1606.09066>.
- [7] Md Nasim Adnan and Md Zahidul Islam. 2017. For A New Framework for Knowledge Discovery from Decision Forests. *Australasian Journal of Information Systems*. 21, (Nov. 2017). <https://doi.org/10.3127/ajis.v21i0.1539>.
- [8] S. Ilker Birbil, Mert Edali, and Birol Yucesoglu. 2020. Rule Covering for Interpretation and Boosting. arXiv:2007.06379. Retrieved from <https://arxiv.org/abs/2007.06379>.
- [9] Gelin Zhang, Zhe Hou, Yanhong Huang, Jianqi Shi, Hadrien Bride, Jin Song Dong, and Yongsheng Gao. 2021. Extracting Optimal Explanations for Ensemble Trees via Logical Reasoning. arXiv:2103.02191. Retrieved from <https://arxiv.org/abs/2103.02191>.
- [10] Jerome H. Friedman and Bogdan E. Popescu. 2008. Predictive Learning via Rule Ensembles. *The Annals of Applied Statistics*. 2, 3 (September 2008), 916–954. <https://doi.org/10.1214/07-AOAS148>.
- [11] Nicolai Meinshausen. 2010. Node Harvest. *The Annals of Applied Statistics*. 4, 4 (December 2010), 2049–2072. DOI:<https://doi.org/10.1214/10-AOAS367>. arXiv:0910.2145. Retrieved from <https://arxiv.org/abs/0910.2145>.
- [12] Haddouchi Maissae and Berrado Abdelaziz. 2024. Forest-ORE: Mining Optimal Rule Ensemble to Interpret Random Forest Models. arXiv:2403.17588. Retrieved from <https://doi.org/10.48550/arXiv.2403.17588>.
- [13] Peter D. Grünwald. 2007. The Minimum Description Length Principle. *Adaptive Computation and Machine Learning series*. The MIT Press. <https://doi.org/10.7551/mitpress/4643.001.0001>.
- [14] Peter Clark and Tim Niblett. 1989. The CN2 Induction Algorithm. *Machine Learning*. 3, (1989), 261–283. <https://doi.org/10.1007/BF00116835>.
- [15] Mlungisi Duma, Bhekisipho Twala, and Tshilidzi Marwala. Improving the Performance of the RIPPER in Insurance Risk Classification: A Comparative Study Using Feature Selection. arXiv:1108.4551. Retrieved from <https://doi.org/10.48550/arXiv.1108.4551>.
- [16] Eibe Frank and Ian H. Witten. 1998. Generating Accurate Rule Sets Without Global Optimization. In *Proceedings of the Fifteenth International Conference on Machine Learning (ICML '98)*. 144–151. Published: 24 July 1998.
- [17] Benjamin Letham, Cynthia Rudin, Tyler H. McCormick, and David Madigan. 2015. Interpretable Classifiers Using Rules and Bayesian Analysis: Building a Better Stroke Prediction Model. *The Annals of Applied Statistics*. 9, 3 (2015), 1350–1371. <https://doi.org/10.1214/15-AOAS848>.
- [18] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: high-precision model-agnostic explanations. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence (AAAI'18/IAAI'18/EAAI'18)*. AAAI Press, Article 187, 1527–1535.
- [19] Steven L. Salzberg. C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993. *Mach Learn* 16, 235–240 (1994). <https://doi.org/10.1007/BF00993309>.
- [20] Leo Breiman, Jerome Friedman, Charles J. Stone, and Richard A. Olshen. 1984. Classification and regression trees. Wadsworth International Group. <https://doi.org/10.1201/9781315139470>.
- [21] William W. Cohen and Yoram Singer. 1999. A simple, fast, and effective rule learner. In *Proceedings of the sixteenth national conference on Artificial intelligence and the eleventh Innovative applications of artificial intelligence conference innovative applications of artificial intelligence (AAAI '99/IAAI '99)*. American Association for Artificial Intelligence, USA, 335–342.
- [22] Zhuo Wang, Wei Zhang, Ning Liu, Jianyong Wang. 2021. Scalable Rule-Based Representation Learning for Interpretable Classification. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- [23] Himabindu Lakkaraju, Stephen H. Bach, and Jure Leskovec. 2016. Interpretable Decision Sets: A Joint Framework for Description and Prediction. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. Association for Computing Machinery, New York, NY, USA, 1675–1684. <https://doi.org/10.1145/2939672.2939874>
- [24] Harsha Nori, Samuel Jenkins, Paul Koch, and Rich Caruana. 2019. Interpretml: A Unified Framework for Machine Learning Interpretability. arXiv:1909.09223. Retrieved from <https://doi.org/10.48550/arXiv.1909.09223>.
- [25] G Roshan Lal, Xiaotong Chen, and Varun Mithal. 2022. TE2Rules: Explaining Tree Ensembles using Rules. arXiv:2206.14359. Retrieved from <https://doi.org/10.48550/arXiv.2206.14359>.
- [26] Agrawal, R., Imieliński, T., and Swami, A. 1993. Mining Association Rules Between Sets of Items in Large Databases. *ACM SIGMOD Record*. 22, 2 (June 1993), 207–216. <https://doi.org/10.1145/170036.170072>.



- [27] Elif VAROL ALTAY and BİLAL ALATAŞ. 2020. Nicel Birliktelik Kural Madenciliği İçin Baskın Olmayan Sıralama Genetik Algoritma-II'nin Duyarlılık Analizi. BİLİŞİM TEKNOLOJİLERİ DERGİSİ, Cilt 13, (Ocak 2020).
- [28] Jorma Rissanen, Modeling by the shortest data description, *Automatica* 14 (1978) 465–471.
- [29] Teemu Roos. 2017. Minimum Description Length Principle. In *Encyclopedia of Machine Learning and Data Mining*(editors:Sammut, C., Webb, G.I.). Springer, 823–827. [https://doi.org/10.1007/978-1-4899-7687-1\\_894](https://doi.org/10.1007/978-1-4899-7687-1_894).
- [30] Jorma Rissanen. 1989. *Stochastic Complexity and Statistical Inquiry*. World Scientific.
- [31] Andrew R. Barron, Jorma Rissanen and Bin Yu. 1998. The minimum description length principle in coding and modeling. *IEEE Trans. Inform. Theory* 44(6), 2743–2760. <https://doi.org/10.1109/18.720554>.
- [32] Ming Li and Paul Vitányi. 2019. *An Introduction to Kolmogorov Complexity and Its Applications*. Springer. ISBN: 978-3-030-11297-4.
- [33] Peter Grünwald and Teemu Roos. 2019. Minimum Description Length Revisited. *International Journal of Mathematics for Industry*. 11, 01 (2019), 1930001. <https://doi.org/10.1142/S2661335219300018>.
- [34] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall and W. Philip Kegelmeyer. 2002. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research* 16 (2002) 321–357. <https://doi.org/10.48550/arXiv.1106.1813>
- [35] Hans Hofmann. 1994. Statlog (German Credit Data) [Dataset]. UCI Machine Learning Repository. <https://doi.org/10.24432/C5NC77>.









# Journal of Data Analytics and Artificial Intelligence Applications

Research Article

 Open Access

## Financial Performance Evaluation of Companies Listed in Corporate Governance and Sustainability Indices: Application of the IVSF-RBNAR Method



Karahan Kara<sup>1</sup>  , Galip Cihan Yalçın<sup>2</sup>  & Hamide Özyürek<sup>3</sup> 

<sup>1</sup> İzmir Demokrasi University, Faculty of Economics and Administrative Sciences, Department of Business, İzmir, Türkiye

<sup>2</sup> Independent Researcher, Ankara, Türkiye

<sup>3</sup> OSTİM Technical University, Faculty of Economics and Administrative Sciences, Department of Business, Ankara, Türkiye

### Abstract

Companies strive to improve their financial conditions not only to gain a competitive advantage but also to be listed on corporate governance and sustainability indices. Companies listed in the Corporate Governance Index are understood to have high levels of corporate governance success, whereas those listed in the Sustainability 25 Index are recognised for being long-term environmentally sustainable organisations. It is evident that companies listed in both indices not only implement successful governance practices but also adopt sustainability principles. The primary motivation of this research is to evaluate the financial performance of companies listed on both indices. Accordingly, the main objective of this study is to develop a method for calculating the financial performance of these companies and to demonstrate its applicability. In this study, which is approached as a decision problem using the multi-criterion decision-making (MCDM) approach, the IVSF-RBNAR (Interval-Valued Spherical Fuzzy - Reference-Based Normalised Assessment Ranking) method is proposed for calculating financial performance. This method allows criterion weighting based on expert opinions and performance ranking based on reference distance. In the application phase of this study, the financial performance levels of the ten companies listed in The Corporate Governance Index and The Sustainability 25 Index were determined by considering seven financial ratio indicators. As a result, Doğan Companies Group Holding Inc. was identified as having the highest financial performance. The most significant financial ratio was determined to be Return on Assets (ROA). The study also presents research implications and suggests future research directions.

### Keywords

Financial Performance Analysis · Corporate Governance Index · Sustainability 25 Index · Interval-Valued Spherical Fuzzy Sets · Reference-Based Normalised Assessment Ranking Method

### Author Note

This research was presented as a Turkish abstract at the VII. Congress on Critical Debates in Social Sciences held at İzmir Democracy University on November 22, 2024.



Citation: Karahan Kara, Galip Cihan Yalçın, and Hamide Özyürek. 2025. Financial performance evaluation of companies listed in corporate governance and sustainability indices: application of the IVSF-RBNAR method. *Journal of Data Analytics and Artificial Intelligence Applications* 1, 1 (January 2025), 36-60. <https://doi.org/10.26650/d3ai.1607081>

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

© 2025. Kara, K., Yalçın, G. C. & Özyürek, H.



✉ Corresponding author: Karahan Kara [karahan.kara@idu.edu.tr](mailto:karahan.kara@idu.edu.tr)

## 1. Introduction

Corporate governance and sustainability are critical factors that determine the long-term success of companies and form the foundation of modern business practices [1]. Corporate governance consists of a set of principles and procedures that regulate a company's management structure to ensure transparency, accountability and the protection of stakeholder rights. In contrast, sustainability is an approach that comprehensively addresses environmental, social, and governance (ESG) parameters [2]. The interaction between these two concepts ensures that businesses are managed not only according to financial performance but also in consideration of their societal responsibilities. Consequently, companies adhering to both corporate governance and sustainability principles can achieve sustainable success in their internal management processes and in their relationships with external stakeholders.

Corporate governance and sustainability have become not only ethical obligations but also key areas that define a company's strategic competitive advantages and help it achieve sustainable growth objectives [3]. In this context, corporate governance and sustainability indices have emerged as important tools for measuring and evaluating the level of implementation in these areas. These indices, which are particularly critical for investors, provide a comprehensive assessment of a company's corporate governance practices and sustainability performance, enabling investment decisions to be based on a more informed and sound foundation [4].

In Türkiye, the BIST Corporate Governance Index and the BIST Sustainability Index are important indicators used to identify companies that demonstrate high performance in corporate governance and sustainability. The Sustainability Index is a tool that measures companies' ESG performance. This index objectively evaluates how well companies adhere to sustainability principles and practices in this area. It assesses sustainability practices based on various criteria, including environmental factors (such as carbon footprint, energy consumption, and water usage), social factors (such as labour rights, occupational health and safety, and community relations), and governance factors (such as ethical standards, corporate governance practices, and transparency) [5].

The Sustainability Index provides investors and other stakeholders comprehensive information about companies' sustainability performance. This enables investors to make investments in companies that align with sustainability goals. The index encourages companies to act responsibly towards the environment and society, enhancing their long-term success and reputation. Sustainability indices strengthen companies' competitive advantages by increasing transparency and accountability. They promote sustainable investments and encourage companies to adopt more sustainable business practices. Launched on November 21, 2022, the BIST Sustainability 25 Index comprises companies with high sustainability performance as well as large and liquid companies. This index is an important indicator that brings together companies in Türkiye based on sustainability principles (It identifies and evaluates companies that adopt best practices in sustainability [6]. Commitment to sustainability principles also influences a company's financial performance. While measuring companies' financial performance, the index also reveals how well these companies align with their sustainability goals. It supports companies' strategic environmental, social, and governance decisions, helping them better plan their sustainability objectives to achieve long-term success.



The Corporate Governance Index is a tool that objectively measures the extent to which companies adhere to corporate governance principles. This index provides an indicator of companies' corporate governance performance. Since the index assesses corporate governance practices based on specific criteria (such as shareholders, transparency, stakeholders, and board of directors), it allows for the analysis of how successful companies are in meeting these criteria [7]. An analysis of corporate governance practices enhances the trust of investors and other stakeholders in companies. By highlighting well-managed companies, the index offers investors safer investment options. The Corporate Governance Index serves as an appropriate benchmark for understanding the relationship between financial performance and corporate governance practices.

The index provides investors and companies with insights into areas for improvement in corporate governance practices. This helps in making strategic decisions. The Corporate Governance Index is a comprehensive measurement tool that evaluates companies' corporate governance levels across multiple criteria. This enables a more detailed analysis of company management quality.

Companies included in these indices have a broad impact on the business world, not only through their financial success but also by fulfilling their environmental and social responsibilities. Therefore, companies' performance in these indices goes beyond financial indicators, highlighting their potential for sustainable growth and their contributions to society.

The importance of corporate governance and sustainability indices in the business world and financial analysis lies in the fact that these indices provide a comprehensive evaluation that reflects not only a company's corporate governance practices but also its capacity for long-term value creation [8]. Analysing the performance of companies included in these indices is crucial because it helps to understand not only financial outcomes but also the impact on companies' strategies for sustainable growth and risk management.

The relationship between corporate governance and sustainability is becoming increasingly important. Good corporate governance supports a company's long-term sustainability, while strong practices in both corporate governance and sustainability create trust among investors and stakeholders, positively impacting a company's overall performance. In this context, examining the performance of companies included in the Corporate Governance Index and the Sustainability Index offers a comprehensive approach to understanding how companies are managed in terms of both financial and social/environmental sustainability. Corporate governance principles and sustainability practices are critical factors that influence investors' decisions, and companies with high scores are perceived as more transparent, ethical, and trustworthy. This enhances investor confidence and supports the company's long-term success.

The Corporate Governance Index and the Sustainability Index are important tools that objectively demonstrate companies' performance in terms of transparency and accountability. Companies included in these indices are subject to greater scrutiny and evaluation, both in terms of their internal governance mechanisms and in terms of their external environmental impact and social responsibility [9]. This allows for more reliable analysis of performance. Companies that exhibit strong performance in both corporate governance and sustainability are more likely to achieve long-term success and not just short-term gains. Well-managed companies that focus on sustainability principles gain a stronger competitive advantage in their markets, improve operational efficiency, and enhance their social prestige by fulfilling their social responsibilities. These factors directly influence companies' financial performance, contributing to more sustainable long-term success.

Companies included in the Corporate Governance Index and the Sustainability Index not only achieve financial goals but also prioritise fulfilling their environmental and social responsibilities. These companies emphasise the importance of environmental and social contributions while achieving financial success in alignment with societal values.

This study adopts a multi-criterion decision-making (MCDM) approach to identify the financial performance levels of selected companies as a decision-making problem. The research also incorporates expert opinions into the analysis. The primary motivation of this study is to propose a method for evaluating the financial performance of companies in decision-making models and to demonstrate its applicability.

In this context, the IVSF-RBNAR (*Interval-Valued Spherical Fuzzy - Reference-Based Normalised Assessment Ranking*) Method is proposed and applied. In this method, expert opinions are collected using linguistic expressions, which are then transformed into IVSF numbers [10]. The influence of experts on the decision-making process is assessed using IVSF sets, and the weights of the criteria are determined using the IVSWAM (*Interval-Valued Spherical Weighted Arithmetic Mean*) aggregation operator [10]. The ranking of companies' performance is performed using the RBNAR method [11]. The primary reason for choosing this methodology is its ability to rank alternatives based on their distances to reference points, providing a robust framework for performance evaluation. This study highlights the effectiveness of the IVSF-RBNAR method in addressing the challenges of decision-making in corporate financial performance analysis.

The primary objectives of this research are as follows:

- *Financial Performance Assessment*: To calculate the financial performance of companies listed on the corporate governance and sustainability index using financial ratio values derived from their financial reports for 2022.
- *Identification and Analysis*: To identify companies included in both indices and analyse their financial reports to compute relevant financial ratio metrics.
- *Decision-Making Problem*: This study treats financial performance evaluation as a decision-making problem by applying an MCDM approach.
- *Methodology Application*: To apply the IVSF-RBNAR Method for financial performance assessment. This involves: (i) Utilising expert opinions to determine the importance of criteria. (ii) Employing linguistic expressions to assess expert expertise and calculate their weights. (iii) Calculating criteria weights using the IVSWAM aggregation operator. (iv) The financial performance of companies is ranked based on their distance from the reference points using the RBNAR method.
- *Case Study Implementation*: To implement the IVSF-RBNAR method on a sample of 10 companies listed in the indices, involving: 7 experts, 7 criteria, and 10 alternatives (companies).

This research makes the following key contributions to the field of financial performance assessment and decision-making methodologies:

- *Methodological Advancement*: Introduces and demonstrates the applicability of the IVSF-RBNAR Method for financial performance evaluation by integrating IVSF and the RBNAR approach.
- *Expert-driven decision-making*: This approach develops a robust framework to incorporate expert opinions into the financial performance evaluation process. This includes: (i) The use of linguistic expressions to assess and quantify the expertise of decision-makers. (ii) The application of the IVSWAM aggregation operator to calculate the importance weights of criteria based on expert inputs.

- *Corporate Financial Analysis*: Provides a structured approach for analysing the financial performance of companies listed on the corporate governance and sustainability index using financial ratios derived from company reports.
- *Identification of Key Performance Drivers*: Highlights Return on Assets (ROA) as the most critical criterion for evaluating financial performance, providing actionable insights for stakeholders.
- *Best-Performing Company Recognition*: Identifies Doğan Companies Group Holding Inc. as the company with the highest financial performance among the evaluated entities.
- *Support for Decision-Making Models*: The IVSF-RBNAR method is validated as a reliable and effective tool for ranking corporate financial performance, making a valuable contribution to decision-making processes in corporate governance and sustainability contexts.

This study is organised into seven sections: *Section 2-Literature Review*: Provides an overview of relevant studies and theoretical foundations related to financial performance evaluation, corporate governance and sustainability indices. *Section 3 – Methodology*: The methodological framework is outlined, detailing the IVSF-RBNAR method, expert judgement integration, and criterion weighting processes. *Section 4 - Application*: This section demonstrates the practical implementation of the proposed methodology, including data collection, calculations, and performance rankings for the selected companies. *Section 5 – Results*: This section presents the findings of the study, highlighting company rankings, key criteria, and the significance of the applied method. *Section 6-Research Implications*: This section discusses the theoretical, practical, and methodological contributions of the study to financial performance evaluation and decision-making processes. *Section 7 - Conclusion*: The study's key outcomes, limitations, and recommendations for future research.

## 2. Literature Review

Financial analysis is a fundamental tool for understanding the dynamics of financial markets and making informed decisions. As markets become increasingly complex and interconnected, the ability to accurately assess market trends, forecast future movements, and effectively analyse investment opportunities is critical for investors, policymakers, and businesses. Accurate financial analysis enables stakeholders to minimise risks, optimise returns, and allocate resources more efficiently, thus supporting global economic stability and contributing to sustainable growth.

In this context, financial ratios play a paramount role. Financial ratios provide critical indicators for understanding a company's financial health, performance, and efficiency. These ratios allow investors and analysts to assess a company's profitability, debt levels, and operational efficiency. In particular, key metrics, such as profitability ratios, offer valuable insights into company financial performance. These indicators enable in-depth analysis of financial statements, helping decision-makers to take more accurate and strategic actions. Additionally, the use of financial ratios in combination allows for multiple perspectives to be evaluated, making the decision-making process more robust and reliable.

The application of advanced analytical methods, such as MCDM, further enhances financial analysis by evaluating market factors and performance indicators more comprehensively. These methods do not rely solely on traditional financial metrics; instead, they consider other factors in the market, enabling the generation of more accurate and predictable results. Consequently, the adoption of more sophisticated multi-criteria

decision models has increased in financial research, making financial decision-making processes more reliable and effective.

Building upon the significance of financial ratios and advanced analytical techniques, the integration of MCDM methods has emerged as a critical component of modern financial market analysis. Over recent years, academic and practical research has increasingly focused on harnessing these methods to optimise decision-making processes, empowering investors, financial analysts, and decision-makers to make more informed, precise, and effective choices across various indices and sectors. The application of MCDM methods allows for a more comprehensive evaluation of complex financial data, facilitating the consideration of multiple criteria, and enhancing the quality of financial decisions. This section provides an overview of key studies in the literature, focusing on the methods used and their impact on sectoral and financial decision-making. Kara et al. [11] conducted a performance analysis of the technology sector on the Istanbul Stock Exchange, utilising SVN-CIMAS-CRITIC-RBNAR (*Single-Valued Neutrosophic - Criteria Importance Assessment - Criteria Importance Through Intercriteria Correlation - Reference-Based Normalisation Alternative Ranking*) methods. This study aimed to develop decision support mechanisms for investments in the technology sector. Kaya et al. [12] examined the sustainability index using a combination of FUCOM (*Full Consistency Method*), GRA (*Grey Relational Analysis*), MABAC (*Multi-Attributive Border Approximation Area Comparison*), and TOPSIS techniques. For Order Performance By Similarity To Ideal Solution method, evaluating the impact of sustainability factors on sectoral performance. Isik et al. [13] employed the DEMATEL (*Decision-Making Trial and Evaluation Laboratory*), CRITIC, EDAS (*Evaluation Based on Distance from Average Solution*), and WASPAS (*Weighted Aggregated Sum Product Assessment*) methods to analyse the food and beverage sector on the Istanbul Stock Exchange, providing insights into the opportunities and challenges within the sector. Alsanousi et al. [14] analysed five sectors of the Saudi Arabian stock market in 2022 using BWM and TOPSIS methods, highlighting the effectiveness of these techniques in sectoral performance analysis. Biswas [15] focused on the energy sector by using the ERUNS (*Evaluation Based on Relative Utility and Nonlinear Standardisation*) methodology to evaluate energy sector performance and provide recommendations for efficient resource use. Elma [16] examined the Bosa İstanbul Sustainability Index on the Istanbul Stock Exchange using various methods, including FUCA (*Faire Un Choix Adéqua*), VIKOR (*Vlekriterijumsko KOMpromisno Rangiranje*), TOPSIS, and others, offering insights into sustainability factors and their influence on financial markets. Işık et al. [17] studied the insurance sector, applying Pythagorean fuzzy AHP (*Analytic Hierarchy Process*) and MAIRCA (*MultiAttributive Ideal-Real Comparative Analysis*) methods to assess risks and opportunities within the sector.

Hoang et al. [18] explored the performance of electronic enterprises globally using the spherical fuzzy AHP and WASPAS methods. This study aims to provide strategic recommendations for firms in the electronics sector. Nguyen et al. [19] analysed the retail sector in Vietnam using Pythagorean fuzzy AHP and CoCoSo (*Combined Compromise Solution*) methods to uncover key decision-making factors. Kara et al. [20] conducted a study on the BIST Sustainability Index, employing MEREC (*Method Based on The Removal Effects Of Criteria*) - RBNAR methods to evaluate sector performance. Güçlü and Muzac [21] focused on the iron and steel sector in Türkiye, using the Extended Grey MULTIMOORA (*Multi-Objective Optimisation by Ratio Analysis Plus Full Multiplicative Form*) method to analyse sectoral risks and opportunities.

Yüksel and Uncu [22] investigated the railway transportation sector in Türkiye using the EDAS method and provided recommendations for increasing sector efficiency. Lam et al. [23] utilised Fuzzy TOPSIS methods to analyse the performance of firms on the Dow Jones Stock Exchange in the United States and identify the

key factors influencing investment decisions. Miguez et al. [24] used AHP and TOPSIS methods to study the tourism sector in Spain and develop decision support tools for the tourism industry. Makki and Alqahtani [25] examined the energy sector in Saudi Arabia using AHP and TOPSIS methods to evaluate sector performance over the period 2019-2021.

Liew et al. [26] employed the entropy-DEMATEL-TOPSIS methods to analyse firms listed on the Dow Jones Stock Exchange, providing valuable insights into their future performance. Ghosh and Bhattacharya [27] used MEREC and Grey-based CoCoSo methods to investigate the hospitality and tourism sectors in India, offering strategic decision-making support to firms in these industries. Bae et al. [28] utilised Fuzzy AHP and TOPSIS methods to analyse the airline sector in the United States, assisting industry stakeholders in making more informed decisions. Katrancı et al. [47] utilised the Indifference Threshold-Based Attribute Ratio Analysis (ITARA) and the Cost Estimation, Benchmarking, and Risk Assessment (COBRA) methods to evaluate the financial performance of 25 companies listed on Borsa Istanbul. These studies highlight the applicability of various MCDM methods across different sectors and their contribution to improving decision-making in financial markets. The integration of different MCDM techniques allows for a more in-depth analysis of sectoral dynamics, leading to more informed and strategic decision-making. Overall, the literature provides significant insights into the decision-making processes within financial markets, enhancing the quality and efficiency of decision support systems. Table 1 presents a summary of the literature review, showcasing key studies relevant to the stock exchange, sector/industry, methods, and years.

**Table 1.** Literature review of financial performance analysis using MCDM

Authors	Stock Exchange (SE)	Sector/Industry	Years	Methods
Kara et al. [11]	İstanbul SE	Technology	2023	SVN-CIMAS-CRITIC-RBNAR
Kaya et al. [12]	İstanbul SE	Sustainability index	2021	FUCOM-GRA-MABAC-MOOSRA-OCRA-TOPSIS-TODIM-VIKOR
Isik et al. [13]	İstanbul SE	Food/Beverage	2021	DEMATEL-CRITIC-EDAS-WASPAS-TOPSIS
Alsanousi et al. [14]	Saudi Stock Market	5 sectors	2022	BWM and TOPSIS
Biswas et al. [15]	-	Energy	-	ERUNS
Elma [16]	İstanbul SE	BIST Sustainability Index	2022	FUCA, VIKOR, TOPSIS, SAW, CODAS, RAFSI and GRA
Işık et al. [17]	-	Insurance	-	Pythagorean fuzzy AHP and MAIRCA
Hoang et al. [18]	-	10 electronic enterprises	-	Spherical fuzzy AHP and WASPAS
Nguyen et al. [19]	-	Retailing industry	-	Pythagorean fuzzy AHP and CoCoSo
Kara et al. [20]	İstanbul SE	BIST Sustainability Index	2022	MEREC-RBNAR
Güçlü and Muzac [21]	İstanbul SE	Iron and Steel	-	Extended Grey MULTIMOORA
Yüksel and Uncu [22]	-	Railway Transportation	2015-2021	EDAS
Lam et al. [23]	Dow Jones, Inc.	-	-	Fuzzy TOPSIS
Miguez et al. [24]	-	Tourism	2019	AHP- TOPSIS



Authors	Stock Exchange (SE)	Sector/Industry	Years	Methods
Makki and Alqahtani [25]	Saudi	Energy	2019-2021	AHP-TOPSIS
Liew et al. [26]	Dow Jones, Inc.	-	2015-2020	Entropy-DEMATEL-TOPSIS
Ghosh and Bhattacharya [27]	Indian	hospitality and tourism	2019-2021	MEREC-Grey CoCoSo
Bae et al. [28]	-	Airline	2018	Fuzzy AHP TOPSIS

### 3. Methodological Framework

In this study, the IVSF-RBNAR method is proposed and applied to calculate companies’ financial performance using the MCDM approach. This method comprises three key stages: *Stage 1*: Expert weights are determined by considering the expertise levels of the specialists consulted for criteria weighting. *Stage 2*: Weights of criteria are identified using IVSF sets and the IVSWAM aggregation operator. *Stage 3*: The financial performance of the companies is determined using the RBNAR method. The stages of this hybrid method are sequentially interconnected. The decision-maker weights obtained in Stage 1 are utilised in Stage 2. Similarly, the criterion weights calculated in Stage 2 are employed in Stage 3. The methodology of the study is illustrated in Figure 1. In the methodology section, basic IVSF set calculations are provided. The steps of the IVSF-RBNAR method are then demonstrated in detail.

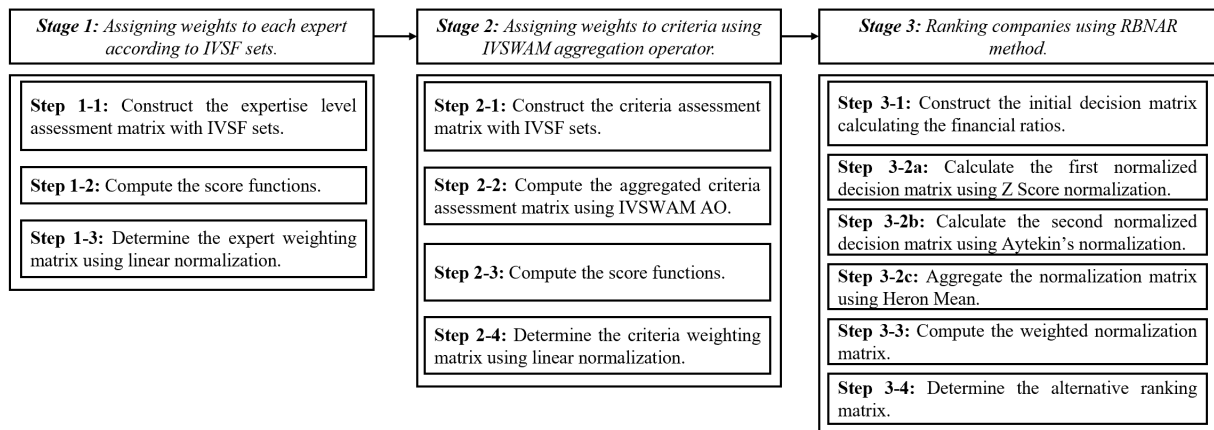


Figure 1. Methodology framework

#### 3.1. Fundamentals of Interval Valued Spherical Fuzzy (IVSF) Sets

**Definition 1.** In the specified domain of discourse, denoted as  $\mathcal{L}$ , the symbol  $\tilde{\mathcal{X}}$  as introduced  $\tilde{\mathcal{X}} = \{ \langle \ell (\chi_{\tilde{\mathcal{X}}}^l(\ell), \chi_{\tilde{\mathcal{X}}}^u(\ell)), (\phi_{\tilde{\mathcal{X}}}^l(\ell), \phi_{\tilde{\mathcal{X}}}^u(\ell)), (\varphi_{\tilde{\mathcal{X}}}^l(\ell), \varphi_{\tilde{\mathcal{X}}}^u(\ell)) \mid \ell \in \mathcal{L} \rangle \}$ , represents the presence of IVSF sets operating within the broader framework of set  $\mathcal{L}$  [10]. Here, the constraints are  $0 \leq \chi_{\tilde{\mathcal{X}}}^l(\ell) \leq \chi_{\tilde{\mathcal{X}}}^u(\ell) \leq 1, 0 \leq \phi_{\tilde{\mathcal{X}}}^l(\ell) \leq \phi_{\tilde{\mathcal{X}}}^u(\ell) \leq 1$ , and  $0 \leq \varphi_{\tilde{\mathcal{X}}}^l(\ell) \leq \varphi_{\tilde{\mathcal{X}}}^u(\ell) \leq 1$ .

Within the given context, the functions  $\chi_{\tilde{\mathcal{X}}}^l(\ell)$ ,  $\phi_{\tilde{\mathcal{X}}}^l(\ell)$ , and  $\varphi_{\tilde{\mathcal{X}}}^l(\ell)$  are interpreted as abstract representations of the lower degree of membership, the lower degree of non-membership, and the lower degree of hesitancy, respectively. The functions  $\chi_{\tilde{\mathcal{X}}}^u(\ell)$ ,  $\phi_{\tilde{\mathcal{X}}}^u(\ell)$ , and  $\varphi_{\tilde{\mathcal{X}}}^u(\ell)$  represent abstract concepts corresponding to the upper degree of membership, non-membership, and hesitancy, respectively. These functions are formally defined such that the inequality  $0 \leq (\chi_{\tilde{\mathcal{X}}}^u(\ell))^2 + (\phi_{\tilde{\mathcal{X}}}^u(\ell))^2 + (\varphi_{\tilde{\mathcal{X}}}^u(\ell))^2 \leq 1$  holds for all elements  $\ell$  belonging to the set  $\mathcal{L}$ .







$$\tilde{\mathcal{X}}_1^\Theta = \left\{ \left( \begin{array}{c} \left( \chi_{\tilde{\mathcal{X}}_1}^l(\ell) \right)^\Theta \\ \left( \chi_{\tilde{\mathcal{X}}_1}^u(\ell) \right)^\Theta \\ \left( 1 - \left( 1 - \left( \phi_{\tilde{\mathcal{X}}_1}^l(\ell) \right)^2 \right)^\Theta \right)^{\frac{1}{2}} \\ \left( 1 - \left( 1 - \left( \phi_{\tilde{\mathcal{X}}_1}^u(\ell) \right)^2 \right)^\Theta \right)^{\frac{1}{2}} \\ \left( \left( 1 - \left( \phi_{\tilde{\mathcal{X}}_1}^l(\ell) \right)^2 \right)^\Theta - \left( 1 - \left( \phi_{\tilde{\mathcal{X}}_1}^l(\ell) \right)^2 - \left( \phi_{\tilde{\mathcal{X}}_1}^l(\ell) \right)^2 \right)^\Theta \right)^{\frac{1}{2}} \\ \left( \left( 1 - \left( \phi_{\tilde{\mathcal{X}}_1}^u(\ell) \right)^2 \right)^\Theta - \left( 1 - \left( \phi_{\tilde{\mathcal{X}}_1}^u(\ell) \right)^2 - \left( \phi_{\tilde{\mathcal{X}}_1}^u(\ell) \right)^2 \right)^\Theta \right)^{\frac{1}{2}} \end{array} \right\} | \ell \text{ in } \mathcal{L} \text{ for } \Theta > 0.$$

Definition 2 is expected to adhere to the following criteria [10]:

- (i)  $\tilde{\mathcal{X}}_1 \oplus \tilde{\mathcal{X}}_2 = \tilde{\mathcal{X}}_2 \oplus \tilde{\mathcal{X}}_1,$
- (ii)  $\tilde{\mathcal{X}}_1 \otimes \tilde{\mathcal{X}}_2 = \tilde{\mathcal{X}}_2 \otimes \tilde{\mathcal{X}}_1,$
- (iii)  $\Theta(\tilde{\mathcal{X}}_1 \oplus \tilde{\mathcal{X}}_2) = \Theta\tilde{\mathcal{X}}_2 \oplus \Theta\tilde{\mathcal{X}}_1 \text{ for } \Theta > 0,$
- (iv)  $(\tilde{\mathcal{X}}_1 \otimes \tilde{\mathcal{X}}_2)^\Theta = \tilde{\mathcal{X}}_1^\Theta \otimes \tilde{\mathcal{X}}_2^\Theta \text{ for } \Theta > 0,$
- (v)  $\Theta_1\tilde{\mathcal{X}}_1 \oplus \Theta_2\tilde{\mathcal{X}}_1 = (\Theta_1 + \Theta_2)\tilde{\mathcal{X}}_1 \text{ for } \Theta_1, \Theta_2 > 0,$
- (vi)  $\tilde{\mathcal{X}}_1^{\Theta_1} \otimes \tilde{\mathcal{X}}_1^{\Theta_2} = \tilde{\mathcal{X}}_1^{(\Theta_1 + \Theta_2)} \text{ for } \Theta_1, \Theta_2 > 0.$

**Definition 3.** In scenario in which  $\tilde{\mathcal{X}}_1 = \left( \left( \chi_{\tilde{\mathcal{X}}_1}^l(\ell), \chi_{\tilde{\mathcal{X}}_1}^u(\ell) \right), \left( \phi_{\tilde{\mathcal{X}}_1}^l(\ell), \phi_{\tilde{\mathcal{X}}_1}^u(\ell) \right), \left( \varphi_{\tilde{\mathcal{X}}_1}^l(\ell), \varphi_{\tilde{\mathcal{X}}_1}^u(\ell) \right) \right)$  presents an IVSF numbers within the set  $\mathcal{L}$ , the score function, denoted as  $S(\tilde{\mathcal{X}}_1)$ , is computed using Eq. (1).

$$S(\tilde{\mathcal{X}}_1) = \frac{\left( \chi_{\tilde{\mathcal{X}}_1}^l(\ell) \right)^2 + \left( \chi_{\tilde{\mathcal{X}}_1}^u(\ell) \right)^2 - \left( \phi_{\tilde{\mathcal{X}}_1}^l(\ell) \right)^2 - \left( \phi_{\tilde{\mathcal{X}}_1}^u(\ell) \right)^2 - \left( \frac{\varphi_{\tilde{\mathcal{X}}_1}^l(\ell)}{2} \right)^2 - \left( \frac{\varphi_{\tilde{\mathcal{X}}_1}^u(\ell)}{2} \right)^2}{2} + 1; \quad (1)$$

$$S(\tilde{\mathcal{X}}_1) \in [0, 2]$$

**Definition 4.** Consider  $\tilde{\mathcal{X}}_a = \left( \left( \chi_{\tilde{\mathcal{X}}_a}^l(\ell), \chi_{\tilde{\mathcal{X}}_a}^u(\ell) \right), \left( \phi_{\tilde{\mathcal{X}}_a}^l(\ell), \phi_{\tilde{\mathcal{X}}_a}^u(\ell) \right), \left( \varphi_{\tilde{\mathcal{X}}_a}^l(\ell), \varphi_{\tilde{\mathcal{X}}_a}^u(\ell) \right) \right)$  presents an IVSF sets ( $\tilde{\mathcal{X}}_a = (\tilde{\mathcal{X}}_1, \tilde{\mathcal{X}}_2, \dots, \tilde{\mathcal{X}}_A)$ ). The formulation of the IVSWAM aggregation operator is shown in Eq. (2):

$$IVSWAM(\tilde{\mathcal{X}}_1, \tilde{\mathcal{X}}_2, \dots, \tilde{\mathcal{X}}_A) = \oplus_{a=1}^A \Theta_a \tilde{\mathcal{X}}_a =$$



$$\left\{ \left( \begin{array}{c} \left( 1 - \prod_{a=1}^A \left( 1 - \left( \chi_{\tilde{x}_a}^l(\ell) \right)^2 \right)^{\Theta_a} \right)^{\frac{1}{2}} \\ \left( 1 - \prod_{a=1}^A \left( 1 - \left( \chi_{\tilde{x}_a}^u(\ell) \right)^2 \right)^{\Theta_a} \right)^{\frac{1}{2}} \\ \prod_{a=1}^A \left( \phi_{\tilde{x}_a}^l(\ell) \right)^{\Theta_a} \\ \prod_{a=1}^A \left( \phi_{\tilde{x}_a}^u(\ell) \right)^{\Theta_a} \\ \left( \prod_{a=1}^A \left( 1 - \left( \chi_{\tilde{x}_a}^l(\ell) \right)^2 \right)^{\Theta_a} - \prod_{a=1}^A \left( 1 - \left( \chi_{\tilde{x}_a}^l(\ell) \right)^2 - \left( \phi_{\tilde{x}_a}^l(\ell) \right)^2 \right)^{\Theta_a} \right)^{\frac{1}{2}} \\ \left( \prod_{a=1}^A \left( 1 - \left( \chi_{\tilde{x}_a}^u(\ell) \right)^2 \right)^{\Theta_a} - \prod_{a=1}^A \left( 1 - \left( \chi_{\tilde{x}_a}^u(\ell) \right)^2 - \left( \phi_{\tilde{x}_a}^u(\ell) \right)^2 \right)^{\Theta_a} \right)^{\frac{1}{2}} \end{array} \right) \mid \ell \in \mathcal{L} \right\}. \quad (2)$$

here, we introduce the associated weight vector  $\Theta_a = (\Theta_1, \Theta_2, \dots, \Theta_A)$  where  $\sum_{a=1}^A \Theta_a = 1$  and  $\Theta \in [0, 1]$ .

### 3.2. The IVSF-RBNAR Method using IVSWAM Aggregation operator

The IVSF-RBNAR method is used to evaluate financial performance. Let consider  $B = \{B_1, B_2, \dots, B_z, \dots, B_Z\}$  ( $z = 1, 2, \dots, Z$ ) presents companies,  $C = \{C_1, C_2, \dots, C_v, \dots, C_V\}$  ( $v = 1, 2, \dots, V$ ) presents criteria,  $E = \{E_1, E_2, \dots, E_f, \dots, E_F\}$  ( $f = 1, 2, \dots, F$ ) represent the decision makers. The procedural steps of the IVSF-RBNAR method are as follows:

**Stage 1: assign weights to each expert according to IVSF sets.**

**Step 1-1:** Expertise level is determined using linguistic variables (LVs), as shown in Table 2. These LVs are then converted into IVSF sets, resulting in IVSF sets representing the priorities of each expert.

**Table 2.** Linguistic variables representing the expertise level for experts [29]

Expertise Level	Interval Valued Spherical Fuzzy Numbers
Extremely Important (EI)	$\langle (0.75, 0.85); (0.10, 0.15); (0.05, 0.10) \rangle$
Critical (VI)	$\langle (0.65, 0.75); (0.15, 0.20); (0.10, 0.15) \rangle$
Important (I)	$\langle (0.55, 0.65); (0.20, 0.25); (0.15, 0.20) \rangle$
Moderately Important (MI)	$\langle (0.45, 0.55); (0.25, 0.30); (0.20, 0.25) \rangle$

**Step 1-2:** The score functions ( $S(E_f)$ ) are calculated using Eq. (3):

$$S(E_f) = \frac{\left( \chi_{E_f}^l(\ell) \right)^2 + \left( \chi_{E_f}^u(\ell) \right)^2 - \left( \phi_{E_f}^l(\ell) \right)^2 - \left( \phi_{E_f}^u(\ell) \right)^2 - \left( \frac{\varphi_{E_f}^l(\ell)}{2} \right)^2 - \left( \frac{\varphi_{E_f}^u(\ell)}{2} \right)^2}{2} + 1; \quad (3)$$

$$S(E_f) \in [0, 2].$$

**Step 1-3:** By employing linear normalisation shown in Eq. (4), the expert weighting matrix ( $w = [w_f]_F$ ) can be calculated.

$$w_f = \frac{S(E_f)}{\sum_{f=1}^F S(E_f)}; (f = 1, 2, \dots, F). \quad (4)$$

Herein,  $w_f = (w_1, w_2, \dots, w_f, \dots, w_F)$  for  $w_f \in [0, 1]$  with the  $\sum_{f=1}^F w_f = 1$ .



**Stage 2: Assign weights to criteria using IVSWAM.**

**Step 2-1:** Each expert ( $E_f$ ) evaluates each criterion ( $C_v$ ) using LVs shown in Table 3. Subsequently, LVs are converted to IVSF numbers. Thus, the criterion assessment matrix ( $\tilde{P} = [\tilde{P}_{vf}]_{VF}$ ) can be determined. Wherein,  $\tilde{P}_{vf} = \langle \langle (\chi_{\tilde{P}_{vf}}^l(\ell), \chi_{\tilde{P}_{vf}}^u(\ell)), (\phi_{\tilde{P}_{vf}}^l(\ell), \phi_{\tilde{P}_{vf}}^u(\ell)), (\varphi_{\tilde{P}_{vf}}^l(\ell), \varphi_{\tilde{P}_{vf}}^u(\ell)) \rangle \rangle$ , where ( $v = 1, 2, \dots, V; f = 1, 2, \dots, F$ ).

**Table 3.** Linguistic variables for evaluating criteria [29]

Linguistic variables for evaluating the criteria	Interval Valued Spherical Fuzzy Numbers
Extremely satisfied (ES)	$\langle (0.80, 0.90), (0.10, 0.20), (0.05, 0.15) \rangle$
Very satisfied (VS)	$\langle (0.70, 0.80), (0.20, 0.30), (0.15, 0.25) \rangle$
Satisfied (S)	$\langle (0.60, 0.70), (0.30, 0.40), (0.25, 0.35) \rangle$
Moderate (M)	$\langle (0.45, 0.55), (0.40, 0.50), (0.35, 0.45) \rangle$
Dissatisfied (D)	$\langle (0.30, 0.40), (0.60, 0.70), (0.25, 0.35) \rangle$
Very dissatisfied (SLI)	$\langle (0.20, 0.30), (0.70, 0.80), (0.15, 0.25) \rangle$
Extremely dissatisfied (LI)	$\langle (0.10, 0.20), (0.80, 0.90), (0.05, 0.15) \rangle$

**Step 2-2:** Employing the IVSWAM aggregation operator shown in Eq. (5), experts' assessments can be aggregated. Then, aggregated criterion assessment matrix ( $\tilde{P} = [\tilde{P}_v]_V$ ) can be determined.

$$IVSWAM(\tilde{P}_1, \tilde{P}_2, \dots, \tilde{P}_F) = \oplus_{f=1}^F w_f \tilde{P}_{vf} =$$

$$\left\{ \left( \begin{array}{c} \left( 1 - \prod_{f=1}^F \left( 1 - \left( \chi_{\tilde{P}_{vf}}^l(\ell) \right)^2 \right)^{w_f} \right)^{\frac{1}{2}} \\ \left( 1 - \prod_{f=1}^F \left( 1 - \left( \chi_{\tilde{P}_{vf}}^u(\ell) \right)^2 \right)^{w_f} \right)^{\frac{1}{2}} \\ \prod_{f=1}^F \left( \phi_{\tilde{P}_{vf}}^l(\ell) \right)^{w_f} \\ \prod_{f=1}^F \left( \phi_{\tilde{P}_{vf}}^u(\ell) \right)^{w_f} \end{array} \right) \mid \ell \in \mathcal{L} \right\}. \quad (5)$$

$$\left( \begin{array}{c} \left( \prod_{f=1}^F \left( 1 - \left( \chi_{\tilde{P}_{vf}}^l(\ell) \right)^2 \right)^{w_f} - \prod_{f=1}^F \left( 1 - \left( \chi_{\tilde{P}_{vf}}^l(\ell) \right)^2 - \left( \phi_{\tilde{P}_{vf}}^l(\ell) \right)^2 \right)^{w_f} \right)^{\frac{1}{2}} \\ \left( \prod_{f=1}^F \left( 1 - \left( \chi_{\tilde{P}_{vf}}^u(\ell) \right)^2 \right)^{w_f} - \prod_{f=1}^F \left( 1 - \left( \chi_{\tilde{P}_{vf}}^u(\ell) \right)^2 - \left( \phi_{\tilde{P}_{vf}}^u(\ell) \right)^2 \right)^{w_f} \right)^{\frac{1}{2}} \end{array} \right)$$

**Step 2-3:** The score functions ( $S(\tilde{P}_v)$ ) are calculated using Eq. (6):

$$S(\tilde{P}_v) = \frac{\left( \chi_{\tilde{P}_v}^l(\ell) \right)^2 + \left( \chi_{\tilde{P}_v}^u(\ell) \right)^2 - \left( \phi_{\tilde{P}_v}^l(\ell) \right)^2 - \left( \phi_{\tilde{P}_v}^u(\ell) \right)^2 - \left( \frac{\varphi_{\tilde{P}_v}^l(\ell)}{2} \right)^2 - \left( \frac{\varphi_{\tilde{P}_v}^u(\ell)}{2} \right)^2}{2} + 1; \quad (6)$$

$$S(\tilde{P}_v) \in [0, 2].$$

**Step 2-4:** By employing linear normalisation shown in Eq. (7), the criteria weighting matrix ( $\omega = [\omega_v]_V$ ) can be calculated.

$$\omega_v = \frac{S(\tilde{P}_v)}{\sum_{v=1}^V S(\tilde{P}_v)}; (v = 1, 2, \dots, V). \quad (7)$$



Herein,  $\omega_v = (\omega_1, \omega_2, \dots, \omega_v, \dots, \omega_V)$  for  $\omega_v \in [0, 1]$  with the  $\sum_{v=1}^V \omega_v = 1$ .

**Stage 3: Ranking companies using the RBNAR method [10].**

**Step 3-1:** Using the financial documents of the companies, financial ratio values can be calculated. Then, the initial matrix ( $H_{zv} = [H_{zv}]_{Z \times V}$ ) for assessing the financial performance of companies can be constructed using these financial ratio values.

**Step 3-2a:** Using the Z-score reference-based normalisation [30] shown in Eq. (8), the first normalised decision matrix ( $M_{zv} = [M_{zv}]_{Z \times V}$ ) can be calculated.

$$M_{zv} = e^{\left(\frac{H_{zv} - R_v}{-2(\sigma_v)^2}\right)}; (z = 1, 2, \dots, Z; v = 1, 2, \dots, V). \tag{8}$$

Herein,  $\sigma_v$  is represent the standard deviation of each criterion and  $R_v$  presents the reference value for each criterion.

**Step 3-2b:** Using Aytekin's reference-based normalisation [31], as shown in Eq. (9), the second normalised decision matrix ( $T_{zv} = [T_{zv}]_{Z \times V}$ ) can be calculated.

$$T_{zv} = 1 - \frac{|M_{zv} - R_v|}{|R_v| + 10^\delta}; (z = 1, 2, \dots, Z; v = 1, 2, \dots, V). \tag{9}$$

Wherein,  $\delta$  is determined for a positive parameter.

**Step 3-2c:** Using the Heron mean [32], as shown in Eq. (10), the aggregated normalised decision matrix ( $N_{zv} = [N_{zv}]_{Z \times V}$ ) can be calculated.

$$N_{zv} = \left(\alpha \sqrt{M_{zv} T_{zv}} + (1 - \alpha) \frac{M_{zv} + T_{zv}}{2}\right); (z = 1, 2, \dots, Z; v = 1, 2, \dots, V). \tag{10}$$

Herein,  $\alpha \in [0, 1]$  is trade of parameter for evaluating the significance level of normalisation.

**Step 3-3:** Using Eq. (11), the weighted normalised decision matrix ( $S_{zv} = [S_{zv}]_{Z \times V}$ ) can be calculated.

$$S_{zv} = (\omega_v N_{zv}); (z = 1, 2, \dots, Z; v = 1, 2, \dots, V). \tag{11}$$

Herein,  $\omega_v = (\omega_1, \omega_2, \dots, \omega_v, \dots, \omega_V)$  for  $\omega_v \in [0, 1]$  with the  $\sum_{v=1}^V \omega_v = 1$ .

**Step 3-4:** Using Eq. (12), the financial performance ranking matrix ( $R_z = [R_z]_Z$ ) can be calculated.

$$R_z = \sum_{z=1}^Z S_{zv}; (z = 1, 2, \dots, Z; v = 1, 2, \dots, V). \tag{12}$$

The alternative with the highest  $R_z$  value is recognised as exhibiting the best financial performance.

## 4. Application

The objective of this study is to determine the financial performance of 10 companies listed on the *Borsa Istanbul Corporate Governance Index* and the *BIST Sustainability 25 Index*. In this regard, the balance sheets and income statements for these companies for 2022 were examined, and financial ratios were calculated for each company. Subsequently, an initial decision matrix was developed. The companies included in the study are detailed in the following subsection.



## 4.1. Elements of the Decision Model

### 4.1.1. Identification of Expert Group

This model requires expert opinion to determine significant levels of financial performance. The expert group was selected from individuals with knowledge and experience in evaluating financial ratios. Seven experts have been identified for this application: The first expert is an academic in accounting and finance. The second expert is a CFO of a company in the energy production sector. The third and fourth experts are academics in finance. The fifth and sixth experts are academic researchers in accounting and finance. The seventh expert is a professor specialising in accounting and finance. The professional titles of these experts are presented in Table 4.

**Table 4.** The expert group

Notation	Experts	Professions
$E_1$	1st Expert	Professor conducting accounting and finance research
$E_2$	2nd Expert	Chief financial officer with 14 years of experience.
$E_3$	3rd Expert	Professor conducting finance research
$E_4$	4th Expert	Professor conducting finance research
$E_5$	5th Expert	Professor conducting accounting and finance research
$E_6$	6th Expert	Professor conducting accounting and finance research
$E_7$	7th Expert	Professor conducting accounting and finance research

### 4.1.2. Financial Ratios as Criteria

In this study, financial ratios calculated from balance sheet and income statement data, which are commonly used in the literature, were selected as criteria in the MCDM model, and companies' performance was analysed based on these criteria, including the following financial ratios. Return on Equity (ROE) [33,34], Return on Assets (ROA) [35], Receivable Turnover Ratio [36], Leverage Ratio [37,38], Operating Profit Margin [39], Net Profit Margin [40], Profit Margin Before Tax [41,42]:

*Return on Equity (ROE) ( $C_1$ ):* This ratio indicates the profit generated by a company from its equity. Equity refers to the company's own capital, that is, the capital provided by its owners or shareholders, independent of external debt. This ratio measures how efficiently the company uses its capital. A high return on equity suggests that the company is generating strong profits from its current capital and that management is effective [14,43].

*Return on Assets (ROA) ( $C_2$ ):* This ratio measures how much profit a company generates from all its assets. Total assets represent the value of the resources owned by the company, including cash, receivables, machinery, and facilities. This ratio indicates how efficiently a company uses its assets to generate profit. A high ratio suggests that a company is effectively managing its assets and utilising them profitably [44].

*Receivable Turnover Ratio ( $C_3$ ):* This ratio indicates how quickly a company collects receivables from its customers. This ratio is particularly important for understanding working capital management. A high receivables turnover ratio suggests that the company collects receivables efficiently, which in turn indicates a healthy cash flow. A low ratio, on the other hand, suggests that receivables are not being collected in a timely manner, which may put strain on the company's cash flow [45].

*Leverage Ratio ( $C_4$ ):* This ratio indicates how much debt a company uses to finance its operations. This ratio measures the extent of the company's debt burden and the proportion of external financing (liabilities)



relative to its total assets. A high leverage ratio suggests that the company is largely financed by debt, indicating higher financial risk. Conversely, a low ratio indicates that a company relies more on equity for financing, implying lower borrowing risks.

**Operating Profit Margin ( $C_5$ ):** This ratio of a company’s profit from core business activities to its sales revenue. This ratio indicates the profit generated by the company from its primary operations. A high operating profit margin suggests that the company generates strong profits from sales and manages its costs effectively. A low margin, on the other hand, may indicate high costs or insufficiently profitable sales.

**Net Profit Margin ( $C_6$ ):** This margin indicates the proportion of net profit generated by a company from sales revenue. This ratio measures the amount of profit the company retains after all costs, including production costs, operating expenses, and taxes, are deducted. A high net profit margin indicates that the company is effectively managing its costs and operating profitable operations. A low ratio, on the other hand, may indicate that the company is struggling to generate sufficient profit from sales or that its costs are too high [46].

**Profit Margin Before Tax ( $C_7$ ):** This ratio evaluates a company's operational profitability by comparing its profits to revenues. A higher ratio indicates greater profitability, whereas a lower ratio suggests lower efficiency in converting revenue into profit, making it a key indicator for assessing financial performance [42].

In this decision model, seven financial ratios (Return on Equity, Return on Assets, Receivables Turnover Ratio, Leverage Ratio, Operating Profit Margin, Profit Margin Before Tax, Net Profit Margin) were utilised to assess the financial performance of companies listed on the Borsa Istanbul Corporate Governance Index and the Sustainability 25 Index. The selected financial ratios provide insights into companies’ profitability, efficiency, leverage, and turnover. The financial ratios considered as criteria in the decision model are presented in Table 5.

**Table 5.** Computation of criteria and source reports

Financial Ratios (Criteria)	Equations	References
Return on equity ( $C_1$ )	$\frac{Net\ Income}{Total\ Equity}$	Alghafes et al. [33] Hao et al. [34].
Return on assets ( $C_2$ )	$\frac{Net\ Income}{Total\ Assets}$	Jin [35].
Receivables turnover ratio ( $C_3$ )	$\frac{Net\ Credit\ Sales}{Average\ Account\ Receivable}$	Zhang et al. [36].
Leverage ratio ( $C_4$ )	$\frac{Total\ Debt}{Total\ Assets}$	Ma et al. [37] Wang et al. [38].
Operating profit margin ( $C_5$ )	$\frac{Operating\ Income}{Net\ Sales}$	Menezes et al. [39].
Profit margin before taxes ( $C_6$ )	$\frac{Before\ tax\ profit}{Net\ Sales}$	Aprima et al. [41] Indrati and Magfiroh [42].
Net profit margin ( $C_7$ )	$\frac{Net\ income}{Net\ Sales}$	Katenova and Qudrat-Ullah [40].

#### 4.1.3. Companies as Alternatives

The codes and notations for the 10 companies included in this case study are presented in Table 6. These companies constitute alternatives in the initial decision matrix. The aim is to determine and rank these companies’ financial performance using the IVSF-RBNAR Method using IVSWAM Aggregation operator method.

**Table 6.** Companies listed in Sustainability Index and Corporate Governance Index

Codes	Codes	Companies
$B_1$	ARCLK	Arçelik Inc.
$B_2$	DOAS	Doğuş Automotive Services and Trade Inc.



Codes	Codes	Companies
$B_3$	DOHOL	Doğan Companies Group, Inc.
$B_4$	ENJSA	Enerjisa Energy, Inc.
$B_5$	ENKAI	ENKA Construction and Industry Inc.
$B_6$	MGROS	Migros Trade Inc.
$B_7$	PGSUS	Pegasus Air Transportation, Inc.
$B_8$	SISE	Şişecam Inc.
$B_9$	TOASA	Tofaş Inc.
$B_{10}$	TTRAK	TürkTraktör Inc.

#### 4.2. Financial Performance Evaluation of Companies using the IVSF-RBNAR Method

In this application, the IVSF-RBNAR method was used to evaluate the financial performance of the companies. Each step presented in the methodology section was applied. These steps were presented as follows:

**Step 1-1:** Using Table 2, each expert was evaluated based on their expertise level with LVs. The LVs are listed in Table 7. The LVs were then transformed into IVSF numbers. Thus, the expert assessment matrix was obtained, as shown in Table 7.



**Table 7.** The expert’s assessment matrix

Experts	LVs	Interval Valued Spherical Fuzzy Numbers
First Expert ( $E_1$ )	Very Important (VI)	(0.65, 0.75); (0.15, 0.20); (0.10, 0.15)
Second Expert ( $E_2$ )	Very Important (VI)	(0.65, 0.75); (0.15, 0.20); (0.10, 0.15)
Third Expert ( $E_3$ )	Extremely Important (EI)	(0.75, 0.85); (0.10, 0.15); (0.05, 0.10)
Fourth Expert ( $E_4$ )	Moderately Important (MI)	(0.45, 0.55); (0.25, 0.30); (0.20, 0.25)
Fifth Expert ( $E_5$ )	Moderately Important (MI)	(0.45, 0.55); (0.25, 0.30); (0.20, 0.25)
Sixth Expert ( $E_6$ )	Very Important (VI)	(0.65, 0.75); (0.15, 0.20); (0.10, 0.15)
Seventh Expert ( $E_7$ )	Important (I)	(0.55, 0.65); (0.20, 0.25); (0.15, 0.20)

**Step 1-2:** Using the score function shown in Eq. (3), the crisp values ( $S(E_f)$ ) were calculated (Table 8).

**Step 1-3:** By employing linear normalisation shown in Eq. (4), the expert weighting matrix ( $w = [w_f]_F$ ) was calculated, and it is presented in Table 8.

**Table 8.** The crisp values ( $S(E_f)$ ) and experts weighting matrix ( $w = [w_f]_F$ )

	$E_1$	$E_2$	$E_3$	$E_4$	$E_5$	$E_6$	$E_7$
$S(E_f)$	0.3988	0.3988	0.5888	0.0638	0.0638	0.3988	0.2238
$w_f$	0.1867	0.1867	0.2756	0.0298	0.0298	0.1867	0.1047

**Step 2-1:** Using Table 3, each expert evaluated each criterion with LVs. The LVs are listed in Table 9. The LVs were then converted to IVSF numbers. Thus, the criterion assessment matrix ( $\tilde{P} = [\tilde{P}_{vf}]_{VF}$ ) was obtained.

**Table 9.** The criterion assessment matrix ( $\tilde{P} = [\tilde{P}_{vf}]_{VF}$ ) with LVs

	$C_1$	$C_2$	$C_3$	$C_4$	$C_5$	$C_6$	$C_7$
$E_1$	S	VS	ED	ED	VS	M	S
$E_2$	S	ES	M	VS	VS	VS	ES
$E_3$	ES	ES	M	M	S	S	ES
$E_4$	ES	S	VD	M	D	D	VS
$E_5$	S	M	ED	VD	D	VD	D
$E_6$	ES	ES	D	M	VS	S	S
$E_7$	VS	VS	M	S	VS	S	ES

**Step 2-2:** Employing the IVSWAM aggregation operator shown in Eq. (5), criterion assessments were aggregated. Aggregated criterion assessment matrix ( $\tilde{P} = [\tilde{P}_v]_V$ ) presented in Table 10.

**Table 10.** The aggregated criterion assessment matrix ( $\tilde{P} = [\tilde{P}_v]_V$ ) with IVSF numbers.

	$\chi_{\tilde{P}_v}^l(\ell)$	$\chi_{\tilde{P}_v}^u(\ell)$	$\phi_{\tilde{P}_v}^l(\ell)$	$\phi_{\tilde{P}_v}^u(\ell)$	$\varphi_{\tilde{P}_v}^l(\ell)$	$\varphi_{\tilde{P}_v}^u(\ell)$
$C_1$	0.7274	0.8352	0.1675	0.2760	0.1509	0.2303
$C_2$	0.7648	0.8689	0.1318	0.2362	0.1033	0.1880
$C_3$	0.3724	0.4690	0.5097	0.6132	0.3060	0.4012
$C_4$	0.4994	0.6014	0.3947	0.5025	0.2768	0.3688
$C_5$	0.6618	0.7637	0.2388	0.3416	0.1846	0.2821
$C_6$	0.5884	0.6905	0.3073	0.4103	0.2525	0.3474
$C_7$	0.7324	0.8410	0.1623	0.2722	0.1440	0.2220



**Step 2-3:** Using the score function shown in Eq. (6), the crisp values ( $S(\tilde{P}_v)$ ) were calculated and shown in Table 11.

**Step 2-4:** By employing linear normalisation shown in Eq. (7), the criteria weighting matrix ( $\omega = [\omega_v]_V$ ) was calculated, as presented in Table 11.

**Table 11.** The crisp values ( $S(\tilde{P}_v)$ ) and criteria weighting matrix ( $\omega = [\omega_v]_V$ ).

	$C_1$	$C_2$	$C_3$	$C_4$	$C_5$	$C_6$	$C_7$
$S(\tilde{P}_v)$	1.5518	1.6276	0.8296	1.0748	1.4095	1.2571	1.5629
$\omega_v$	0.1666	0.1748	0.0891	0.1154	0.1513	0.1350	0.1678

**Step 3-1:** Using the financial documents of the companies, financial ratio values were calculated. Then, the initial matrix ( $H_{zv} = [H_{zv}]_{ZxV}$ ) for assessing the financial performance of the companies is generated using these financial ratio values. The results are shown in Table 12.

**Table 12.** The initial matrix ( $H_{zv} = [H_{zv}]_{ZxV}$ ) for assessing the financial performance of companies.

	$C_1$	$C_2$	$C_3$	$C_4$	$C_5$	$C_6$	$C_7$
$B_1$	0.17425	0.03572	4.64947	0.79503	6.78801	3.14976	3.52689
$B_2$	0.67040	0.37977	21.54510	0.43351	16.94003	18.21443	16.77606
$B_3$	0.29468	0.16400	13.24248	0.44344	15.17245	15.10367	15.09882
$B_4$	0.67207	0.24495	14.99938	0.63553	9.88526	5.26545	17.16786
$B_5$	0.01728	0.01318	12.97293	0.23706	19.81762	6.11524	3.42145
$B_6$	0.63080	0.07083	224.00984	0.88772	3.65765	2.56134	3.46277
$B_7$	0.39347	0.07411	56.51585	0.81165	22.64211	15.48936	16.61544
$B_8$	0.21165	0.12281	6.56989	0.41976	18.26967	20.72589	21.11541
$B_9$	0.75680	0.21206	5.93222	0.71979	13.62395	13.06306	13.06300
$B_{10}$	0.81661	0.21535	12.97215	0.73628	13.82471	13.11421	13.60691

**Step 3-2a:** Using the Z-score reference-based normalisation shown in Eq. (8), the first normalised decision matrix ( $M_{zv} = [M_{zv}]_{ZxV}$ ) was calculated by employing the reference value matrix. The first normalised decision matrix and reference values are given in Table 13.

**Table 13.** The first normalised matrix ( $M_{zv} = [M_{zv}]_{ZxV}$ ) and reference values.

	$C_1$	$C_2$	$C_3$	$C_4$	$C_5$	$C_6$	$C_7$
$B_1$	0.67877	0.65194	0.95008	0.77919	0.28698	0.25928	0.22752
$B_2$	0.67091	0.10449	0.99764	0.61983	0.98991	0.79760	0.95361
$B_3$	0.90380	0.97769	0.98169	0.64783	0.98768	0.98129	0.99868
$B_4$	0.66733	0.64889	0.98630	0.99932	0.57347	0.41962	0.93446
$B_5$	0.35383	0.53123	0.98093	0.16653	0.81965	0.49421	0.22126
$B_6$	0.75377	0.82837	0.01321	0.52317	0.10768	0.22256	0.22370
$B_7$	0.99535	0.84294	0.90322	0.73551	0.54013	0.96834	0.96058
$B_8$	0.75676	0.98839	0.95841	0.58098	0.93452	0.57105	0.62306
$B_9$	0.48539	0.81559	0.95573	0.93865	0.91549	0.99291	0.96666
$B_{10}$	0.36687	0.80019	0.98092	0.91064	0.92815	0.99381	0.98444
Reference $R_v$	0.4205	0.1400	26.1715	0.6434	16.1006	13.8392	14.7634



**Step 3-2b:** Using Aytekin's reference-based normalisation shown in Eq. (9), the second normalised decision matrix ( $T_{zv} = [T_{zv}]_{ZxV}$ ) was calculated (Table 14).

**Table 14.** The second normalised matrix ( $T_{zv} = [T_{zv}]_{ZxV}$ ).

	$C_1$	$C_2$	$C_3$	$C_4$	$C_5$	$C_6$	$C_7$
$B_1$	0.9999998	0.9999999	0.9999785	0.9999998	0.9999907	0.9999893	0.9999888
$B_2$	0.9999998	0.9999998	0.9999954	0.9999998	0.9999992	0.9999956	0.9999980
$B_3$	0.9999999	1.0000000	0.9999871	0.9999998	0.9999991	0.9999987	0.9999997
$B_4$	0.9999997	0.9999999	0.9999888	1.0000000	0.9999938	0.9999914	0.9999976
$B_5$	0.9999996	0.9999999	0.9999868	0.9999996	0.9999963	0.9999923	0.9999887
$B_6$	0.9999998	0.9999999	0.9998022	0.9999998	0.9999876	0.9999887	0.9999887
$B_7$	1.0000000	0.9999999	0.9999697	0.9999998	0.9999935	0.9999983	0.9999981
$B_8$	0.9999998	1.0000000	0.9999804	0.9999998	0.9999978	0.9999931	0.9999936
$B_9$	0.9999997	0.9999999	0.9999798	0.9999999	0.9999975	0.9999992	0.9999983
$B_{10}$	0.9999996	0.9999999	0.9999868	0.9999999	0.9999977	0.9999993	0.9999988

**Step 3-2c:** Using the Heron mean in Eq. (10), the aggregated normalised decision matrix ( $N_{zv} = [N_{zv}]_{ZxV}$ ) was calculated ( $\alpha = 0.5$ ) and shown in Table 15.

**Table 15.** The aggregated normalised matrix ( $N_{zv} = [N_{zv}]_{ZxV}$ ).

	$C_1$	$C_2$	$C_3$	$C_4$	$C_5$	$C_6$	$C_7$
$B_1$	0.83163	0.81670	0.97487	0.88615	0.58960	0.56941	0.54537
$B_2$	0.82727	0.43775	0.99882	0.79860	0.99495	0.89594	0.97667
$B_3$	0.95129	0.98882	0.99082	0.81440	0.99383	0.99062	0.99934
$B_4$	0.82528	0.81499	0.99313	0.99966	0.77200	0.67879	0.96695
$B_5$	0.63588	0.74723	0.99043	0.49567	0.90758	0.72505	0.54050
$B_6$	0.87254	0.91216	0.31071	0.74245	0.44099	0.54152	0.54240
$B_7$	0.99767	0.91980	0.95098	0.86268	0.75249	0.98411	0.98019
$B_8$	0.87415	0.99418	0.97909	0.77635	0.96698	0.77060	0.80043
$B_9$	0.71970	0.90545	0.97773	0.96908	0.95728	0.99645	0.98326
$B_{10}$	0.64457	0.89732	0.99043	0.95480	0.96374	0.99690	0.99220

**Step 3-3:** Using Eq. (11), the weighted aggregated normalised decision matrix ( $S_{zv} = [S_{zv}]_{ZxV}$ ) was calculated, as shown in Table 16.

**Table 16.** The weighted aggregated normalised matrix ( $S_{zv} = [S_{zv}]_{ZxV}$ ).

	$C_1$	$C_2$	$C_3$	$C_4$	$C_5$	$C_6$	$C_7$
$B_1$	0.1385648	0.1427276	0.0868375	0.1022693	0.0892319	0.0768576	0.0915217
$B_2$	0.1378386	0.0765012	0.0889704	0.0921654	0.1505796	0.1209319	0.1638993
$B_3$	0.1585025	0.1728066	0.0882580	0.0939879	0.1504102	0.1337115	0.1677042
$B_4$	0.1375071	0.1424285	0.0884640	0.1153691	0.1168380	0.0916218	0.1622684
$B_5$	0.1059487	0.1305875	0.0882238	0.0572049	0.1373577	0.0978656	0.0907041
$B_6$	0.1453817	0.1594111	0.0276770	0.0856844	0.0667406	0.0730925	0.0910236
$B_7$	0.1662305	0.1607446	0.0847093	0.0995607	0.1138858	0.1328321	0.1644903



	$C_1$	$C_2$	$C_3$	$C_4$	$C_5$	$C_6$	$C_7$
$B_8$	0.1456494	0.1737450	0.0872131	0.0895974	0.1463471	0.1040134	0.1343245
$B_9$	0.1199149	0.1582376	0.0870920	0.1118401	0.1448788	0.1344983	0.1650056
$B_{10}$	0.1073964	0.1568162	0.0882237	0.1101914	0.1458567	0.1345593	0.1665064

**Step 3-4:** Using Eq. (12), the financial performance ranking matrix ( $R_z = [R_z]_Z$ ) was calculated, as shown in Table 17.

**Table 17.** The financial performance ranking matrix ( $R_z = [R_z]_Z$ )

	$B_1$	$B_2$	$B_3$	$B_4$	$B_5$	$B_6$	$B_7$	$B_8$	$B_9$	$B_{10}$
$R_z$	0.7280	0.8309	0.9654	0.8545	0.7079	0.6490	0.9225	0.8809	0.9215	0.9096
Rank	8 <sup>th</sup>	7 <sup>th</sup>	1 <sup>st</sup>	6 <sup>th</sup>	9 <sup>th</sup>	10 <sup>th</sup>	2 <sup>nd</sup>	5 <sup>th</sup>	3 <sup>rd</sup>	4 <sup>th</sup>

## 5. Results

This study employs the IVSF-RBNAR method to evaluate the financial performance of companies listed under corporate governance and the Sustainability 25 Index. The research yields three primary findings: (i) expert weightings, (ii) criteria weightings, and (iii) companies' financial performance and rankings. The results of the analysis are summarised as follows:

- **Expert Contributions:** The influence of experts on the decision-making process is ranked as follows: “ $E_3 > E_1 = E_2 = E_6 > E_7 > E_4 = E_5$ ”. Accordingly, the third expert was identified as the most influential contributor to the decision-making process.
- **Criteria Weightings:** The criteria's impact on the decision-making process is ranked as follows: “Return on assets ( $C_2$ ) > Net profit margin ( $C_7$ ) > Return on equity ( $C_1$ ) > Operating profit margin ( $C_5$ ) > Profit margin before tax ( $C_6$ ) > Leverage ratio ( $C_4$ ) > Receivables turnover ratio ( $C_3$ )”. Among these, Return on Assets (ROA) emerged as the most significant criterion for evaluating companies' financial performance, while the receivables turnover ratio had the least impact.
- **Company Rankings:** The financial performance rankings of the companies are as follows: “DOHOL ( $B_3$ ) > PGSUS ( $B_7$ ) > TOASA ( $B_9$ ) > TTRAK ( $B_{10}$ ) > SISE ( $B_8$ ) > ENJSA ( $B_4$ ) > DOAS ( $B_2$ ) > ARCLK ( $B_1$ ) > ENKAI ( $B_5$ ) > MGROS ( $B_6$ )”. Among the evaluated firms, Doğan Companies Group Holding Inc. was identified as having the highest financial performance, whereas Migros Trade Inc. demonstrated the lowest financial performance.

In conclusion, the IVSF-RBNAR method effectively supports the analysis of corporate financial performance. These results underscore the applicability and robustness of the methodology for evaluating financial performance within the framework of corporate governance and sustainability indices.

## 6. Research Implications

This study has several significant implications for academic research, corporate decision-making, and methodological advancements in financial performance evaluation:

- **Advancement in Financial Performance Evaluation:** Demonstrates the utility of the IVSF-RBNAR Method as a novel approach for assessing and ranking financial performance based on financial ratios. This framework provides a robust framework for future performance analysis studies.



- *Integration of Expert Judgments*: Highlights the importance of incorporating expert opinions into decision-making processes. The use of linguistic expressions and IVSF sets to quantify expert expertise ensures a nuanced understanding of decision-makers' contributions.
- *Enhanced Decision-Making Methodologies*: This study validates the integration of the IVSWAM aggregation operator and RBNAR method for calculating criteria weights and ranking companies, paving the way for their application in other multi-criteria decision-making contexts.
- *Key Financial Metrics*: Identifies Return on Assets (ROA) as the most critical financial metric, emphasising its importance for corporate governance and sustainability-focused performance evaluations.
- *Empirical Validation of Methodology*: Provides a practical application of the IVSF-RBNAR method to evaluate the financial performance of 10 companies listed in the Corporate Governance Index and the Sustainability Index, offering a replicable model for similar studies.
- *Corporate Governance and Sustainability*: Offers insights into the financial health of companies adhering to corporate governance and sustainability principles and contributes to the literature on the financial implications of these frameworks.
- *Decision Support for Stakeholders*: Supports stakeholders in identifying high-performing companies, as evidenced by the identification of Doğan Companies Group Holding Inc. as the top performer. This will facilitate informed decision-making in investment and policy development.

*Methodological Applicability*: Confirms the suitability of the IVSF-RBNAR method for financial performance ranking, encouraging its adoption and adaptation in diverse decision-making scenarios within corporate finance and beyond.

## 7. Conclusion

This study evaluates the financial performance of companies listed in the Corporate Governance Index and the Sustainability Index using financial ratios and an MCDM approach. By employing the IVSF-RBNAR Method, the research effectively integrated expert judgments and linguistic expressions to determine criterion importance and rank companies based on their financial performance. The analysis highlighted the significance of incorporating both qualitative and quantitative data to create a comprehensive decision-making framework, demonstrating the method's applicability in corporate financial evaluation.

The findings reveal that Doğan Companies Group Holding Inc. achieved the highest financial performance among the 10 analysed companies, while Return on Assets (ROA) emerged as the most critical criterion influencing corporate performance. The use of the IVSWAM aggregation operator for criteria weighting and the RBNAR method for performance ranking proved effective in terms of delivering accurate and meaningful results. These insights contribute to the understanding of financial performance dynamics, offering a valuable tool for stakeholders to assess companies based on their adherence to corporate governance and sustainability principles.

This study underscores the potential of the IVSF-RBNAR method as a robust approach for financial performance assessment, but it also acknowledges its limitations, including the reliance on expert judgments and the relatively small sample size. Future research should expand on this framework by incorporating additional criteria, exploring larger datasets, and integrating advanced computational techniques to further

refine the methodology. Ultimately, this research provides a foundation for enhancing decision-making processes in corporate finance and governance.

### 7.1. Research limitations

Although this study provides valuable insights into the financial performance evaluation of companies listed in the Corporate Governance Index and the Sustainability Index, it is not without limitations. These include:

- *Sample Size:* The research focuses on a limited sample of 10 companies, which may not fully capture the broader diversity of firms in the indices or general corporate practices.
- *Criteria Selection:* The evaluation relies on seven financial criteria, potentially omitting other significant financial or non-financial factors that could influence corporate performance.
- *Expert Dependency:* The methodology relies heavily on expert judgments for weighting criteria and assessing importance, which may introduce subjectivity despite efforts to standardise inputs using IVSF sets.
- *Context-Specific Findings:* The findings, including the identification of Doğan Companies Group Holding Inc. as the best-performing company, may not generalise to other indices, industries, or geographical contexts.
- *Linguistic Expression Limitations:* The use of linguistic expressions to represent expert assessments, while innovative, may lead to loss of precision or inconsistencies in interpretation.
- *Focus on Financial Ratios:* By prioritising financial ratios, this study does not account for qualitative factors such as corporate governance practices, environmental sustainability, or social impact, which are crucial for holistic performance evaluation.
- *Methodological Constraints:* Although the IVSF-RBNAR method has been validated in this context, its applicability and reliability in different decision-making scenarios or larger datasets remain to be tested.

These limitations provide avenues for future research to refine the methodology, expand the scope of the analysis, and incorporate additional dimensions to ensure a more comprehensive evaluation.

### 7.2. Future Research Suggestions

Building on the findings and limitations of this study, the following recommendations are proposed for future research:




- *Expanding the Sample Size:* Future studies could include a larger number of companies across various indices, industries, or geographic regions to enhance the generalizability of the findings.
- *Incorporating Additional Criteria:* The inclusion of both financial and non-financial criteria, such as environmental sustainability metrics, governance quality, and social responsibility indicators, would provide a more comprehensive evaluation framework.
- *Longitudinal Analysis:* Conducting a longitudinal study to track changes in financial performance over time would offer deeper insights into trends and temporal dynamics.
- *Automation and Standardisation:* Developing automated tools to process linguistic expressions and calculate IVSF values could minimize subjectivity and improve the reproducibility of results.
- *Methodological Comparisons:* Comparing the IVSF-RBNAR method with other MCDM approaches will help validate its effectiveness and identify scenarios where it performs optimally.

- **Dynamic Expert Weighting:** Exploring dynamic or adaptive weighting mechanisms that adjust expert influence based on past performance or domain relevance can enhance decision-making processes.
- **Integration with Machine Learning:** Combining the IVSF-RBNAR method with machine learning techniques can improve predictions and automate the performance ranking process.

These directions can extend the scope and impact of financial performance evaluation research and contribute to more robust and versatile decision-making models.



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

Author Details	<p><b>Karahan Kara</b>  <sup>1</sup> İzmir Demokrasi University, Faculty of Economics and Administrative Sciences, Department of Business, İzmir, Türkiye   0000-0002-1359-0244</p> <p><b>Galip Cihan Yalçın</b>  <sup>2</sup> Independent Researcher, Ankara, Türkiye   0000-0001-9348-0709</p> <p><b>Hamide Özyürek</b>  <sup>3</sup> OSTİM Technical University, Faculty of Economics and Administrative Sciences, Department of Business, Ankara, Türkiye   0000-0002-2574-954X</p>
----------------	--

## References

- [1] Erben Yavuz, A., Kocaman, B. E., Doğan, M., Hazar, A., Babuşcu, Ş., & Sutbayeva, R. 2024. The Impact of Corporate Governance on Sustainability Disclosures: A Comparison from the Perspective of Financial and Non-Financial Firms. *Sustainability*, 16(19), 8400. <https://doi.org/10.3390/su16198400>.
- [2] Martiny, A., Tagliatalata, J., Testa, F., & Iraldo, F. 2024. Determinants of environmental social and governance (ESG) performance: A systematic literature review. *Journal of Cleaner Production*, 456, 142213. <https://doi.org/10.1016/j.jclepro.2024.142213>.
- [3] Chopra, S. S., Senadheera, S. S., Dissanayake, P. D., Withana, P. A., Chib, R., Rhee, J. H., & Ok, Y. S. 2024. Navigating the Challenges of Environmental, Social, and Governance (ESG) Reporting: The Path to Broader Sustainable Development. *Sustainability*, 16(2), 606. <https://doi.org/10.3390/su16020606>.
- [4] Khamisu, M. S., Paluri, R. A., & Sonwaney, V. 2024. Stakeholders' perspectives on critical success factors for environmental social and governance (ESG) implementation. *Journal of Environmental Management*, 365, 121583. <https://doi.org/10.1016/j.jenvman.2024.121583>.
- [5] Kartal, M. T., Taşkın, D., Shahbaz, M., Depren, S. K., & Pata, U. K. 2024. Effects of Environment, Social, and Governance (ESG) disclosures on ESGscores: Investigating the role of corporate governance for publicly traded Turkish companies. *Journal of Environmental Management*, 368, 122205. <https://doi.org/10.1016/j.jenvman.2024.122205>.



- [6] Keke, İ., Evci, S., & Keke, İ. 2024. Data on the financial performance of companies on BIST Sustainability 25 Index: An Entropy-based TOPSIS approach. *Data in Brief*, 57, 110959. <https://doi.org/10.1016/j.dib.2024.110959>.
- [7] Nguyen, Q., Kim, M. H., & Ali, S. 2024. Corporate governance and earnings management: Evidence from Vietnamese listed firms. *International Review of Economics & Finance*, 89, 775-801. <https://doi.org/10.1016/j.iref.2023.07.084>.
- [8] Miloud, T. 2024. Corporate governance and CSR disclosure: Evidence from French listed companies. *Global Finance Journal*, 59, 100943. <https://doi.org/10.1016/j.gfj.2024.100943>.
- [9] Mohy-ud-Din, K. 2024. ESG reporting, corporate green innovation and interaction role of board diversity: A new insight from US. *Innovation and Green Development*, 3(4), 100161. <https://doi.org/10.1016/j.igd.2024.100161>.
- [10] Gündoğdu, F. K., & Kahraman, C. 2019. A novel fuzzy TOPSIS method using emerging interval-valued spherical fuzzy sets. *Engineering Applications of Artificial Intelligence*, 85, 307-323. <https://doi.org/10.1016/j.engappai.2019.06.003>.
- [11] Kara, K., Yalçın, G. C., Çetinkaya, A., Simic, V., & Pamucar, D. 2024. A single-valued neutrosophic CIMAS-CRITIC-RBNAR decision support model for the financial performance analysis: A study of technology companies. *Socio-Economic Planning Sciences*, 92, 101851. <https://doi.org/10.1016/j.seps.2024.101851>.
- [12] Kaya, A., Pamucar, D., Gürtler, H. E., & Ozcalici, M. 2024. Determining the financial performance of the firms in the Borsa Istanbul sustainability index: integrating multi criteria decision making methods with simulation. *Financial Innovation*, 10(1), 21. <https://doi.org/10.1186/s40854-023-00512-3>.
- [13] Işık, C., Türkan, M., Marbou, S., & Gül, S. 2024. Stock Market Performance Evaluation of Listed Food and Beverage Companies in Istanbul Stock Exchange with MCDM Methods. *Decision Making: Applications in Management and Engineering*, 7(2), 35-64. <https://doi.org/10.31181/dmame722024692>.
- [14] Alsanousi, A. T., Alqahtani, A. Y., Makki, A. A., & Baghdadi, M. A. 2024. A Hybrid MCDM Approach Using the BWM and the TOPSIS for a Financial Performance-Based Evaluation of Saudi Stocks. *Information*, 15(5), 258. <https://doi.org/10.3390/info15050258>.
- [15] Biswas, S., Pamucar, D., Dawn, S., & Simic, V. 2024. Evaluation based on relative utility and nonlinear standardization (ERUNS) method for comparing firm performance in energy sector. *Decision Making Advances*, 2(1), 1-21. <https://doi.org/10.31181/dma21202419>.
- [16] Elma, O. E. 2024. Financial Performance of BIST Sustainability Index Enterprises: Unearthing the Most Optimum MCDA Methods for Decision-Makers. *Verimlilik Dergisi*, 58(4), 461-478. <https://doi.org/10.51551/verimlilik.1410272>.
- [17] Işık, Ö., Çalık, A., & Shabir, M. 2024. A Consolidated MCDM Framework for Overall Performance Assessment of Listed Insurance Companies Based on Ranking Strategies. *Computational Economics*, 1-42. <https://doi.org/10.1007/s10614-024-10578-5>.
- [18] Hoang, P. D., Nguyen, L. T., & Tran, B. Q. 2024. Assessing environmental, social and governance (ESG) performance of global electronics industry: an integrated MCDM approach-based spherical fuzzy sets. *Cogent Engineering*, 11(1), 2297509. <https://doi.org/10.1080/23311916.2023.2297509>.
- [19] Nguyen, P. H., Tsai, J. F., Nguyen, V. T., Vu, D. D., & Dao, T. K. 2020. A decision support model for financial performance evaluation of listed companies in the Vietnamese retailing industry. *the journal of Asian finance, economics and business*, 7(12), 1005-1015. <https://doi.org/10.13106/jafeb.2020.vol7.no12.1005>.
- [20] Kara, K., Özyürek, H., Yalçın, G. C., & Burgaz, N. 2024b. Enhancing Financial Performance Evaluation: The MEREC-RBNAR Hybrid Method for Sustainability-Indexed Companies. *Journal of Soft Computing and Decision Analytics*, 2(1), 236-257. <https://doi.org/10.31181/jscda21202444>.
- [21] Güçlü, P., & Muzac, G. 2024. Genişletilmiş Gri MULTIMOORA Yöntemi ile Çok Dönemli Çok Kriterli Karar Verme: Demir-Çelik Sektöründe Finansal Performans Değerlendirmesi Örneği. *Eskişehir Osmangazi Üniversitesi İktisadi ve İdari Bilimler Dergisi*, 19(1), 267-291. <https://doi.org/10.17153/oguiibf.1373450>.
- [22] Yüksel, Ç., & Uncu, N. 2024. Evaluating the performance of railway transportation companies using multi-criteria decision-making methods. *Demiryolu Mühendisliği*, (20), 11-24. <https://doi.org/10.47072/demiryolu.1407420>.
- [23] Lam, W. H., Lam, W. S., Liew, K. F., & Lee, P. F. 2023. Decision analysis on the financial performance of companies using integrated entropy-fuzzy topsis model. *Mathematics*, 11(2), 397. <https://doi.org/10.3390/math11020397>.
- [24] Miguez, J. L., Rivo-Lopez, E., Porteiro, J., & Pérez-Orozco, R. 2023. Selection of non-financial sustainability indicators as key elements for multi-criteria analysis of hotel chains. *Sustainable Production and Consumption*, 35, 495-508. <https://doi.org/10.1016/j.spc.2022.12.004>.
- [25] Makki, A. A., & Alqahtani, A. Y. 2023. Capturing the effect of the COVID-19 pandemic outbreak on the financial performance disparities in the energy sector: A Hybrid MCDM-Based evaluation approach. *Economies*, 11(2), 61. <https://doi.org/10.3390/economies11020061>.





- [26] Liew, K. F., Lam, W. S., & Lam, W. H. 2022. Financial network analysis on the performance of companies using integrated entropy-DEMATEL-TOPSIS model. *Entropy*, 24(8), 1056. <https://doi.org/10.3390/e24081056>.
- [27] Ghosh, S., & Bhattacharya, M. 2022. Analyzing the impact of COVID-19 on the financial performance of the hospitality and tourism industries: an ensemble MCDM approach in the Indian context. *International Journal of Contemporary Hospitality Management*, 34(8), 3113-3142. <https://doi.org/10.1108/IJCHM-11-2021-1328>.
- [28] Bae, K., Gupta, A., & Mau, R. 2021. Comparative analysis of airline financial and operational performances: A fuzzy AHP and TOPSIS integrated approach. *Decision Science Letters*, 10(3), 361-374. <https://doi.org/10.5267/j.dsl.2021.2.002>.
- [29] Mandal, U., & Seikh, M. R. 2023. Interval-valued spherical fuzzy MABAC method based on Dombi aggregation operators with unknown attribute weights to select plastic waste management process. *Applied Soft Computing*, 145, 110516. <https://doi.org/10.1016/j.asoc.2023.110516>.
- [30] Shih, H. S., Shyur, H. J., & Lee, E. S. 2007. An extension of TOPSIS for group decision making. *Mathematical and computer modelling*, 45(7-8), 801-813. <https://doi.org/10.1016/j.mcm.2006.03.023>.
- [31] Aytekin, A. 2020. Çok kriterli karar problemine uzaklık ve referans temelli çözüm yaklaşımı. Unpublished Doctoral Thesis, Anadolu University, Eskisehir, Türkiye. <https://openaccess.artvin.edu.tr/xmlui/handle/11494/2558>.
- [32] Zhu, L. 2022. Sharp bounds for a generalized logarithmic operator mean and Heinz operator mean by weighted ones of classical operator ones. *Mathematics*, 10(10), 1617. <https://doi.org/10.3390/math10101617>.
- [33] Alghafes, R., Karim, S., Aliani, K., Qureishi, N., & Alkayed, L. 2024. Influence of key ESG factors on Islamic banks' financial performance: evidence from GCC countries. *International Review of Economics & Finance*, 96, 103629. <https://doi.org/10.1016/j.iref.2024.103629>.
- [34] Hao, X., Sun, Q., Ma, P., Li, K., Wu, H., & Xue, Y. 2024. Unlocking wind power potential: The pivotal role of R&D investment in boosting wind power enterprise performance. *Energy Strategy Reviews*, 55, 101507. <https://doi.org/10.1016/j.esr.2024.101507>.
- [35] Jin, Y. 2024. Distinctive Impacts of ESG Pillars on Corporate Financial Performance: A Random Forest Analysis of Korean Listed Firms. *Finance Research Letters*, 106395. <https://doi.org/10.1016/j.frl.2024.106395>.
- [36] Zhang, R., Li, X., Yan, X., & Bian, Y. 2024. Does customer concentration matter in business model value: Threshold effects of carbon emissions and dynamic capabilities?. *Technovation*, 137, 103095. <https://doi.org/10.1016/j.technovation.2024.103095>.
- [37] Ma, B., He, G., An, J., Li, M., & Sun, G. 2024. Capital price distortion, financial leverage, and credit risk in commercial banks. *Finance Research Letters*, 69, 106200. <https://doi.org/10.1016/j.frl.2024.106200>.
- [38] Wang, S. M., Wang, M., & Feng, C. 2024. Deleveraging and green technology innovation: Evidence from Chinese listed companies. *Research in International Business and Finance*, 69, 102289. <https://doi.org/10.1016/j.ribaf.2024.102289>.
- [39] Menezes, M. B., Ruiz-Hernández, D., & Pinto, R. 2024. The Capacitated Product Portfolio Mix-and-Allocation problem with integrity constraints. *Computers & Industrial Engineering*, 187, 109845. <https://doi.org/10.1016/j.cie.2023.109845>.
- [40] Katenova, M., & Qudrat-Ullah, H. 2024. Corporate social responsibility and firm performance: Case of Kazakhstan. *Heliyon*, 10(10). <https://doi.org/10.1016/j.heliyon.2024.e31580>.
- [41] Aprima, W. O., Putra, D. G., & Harini, G. 2024. The Influence of Managerial Ownership Structure, Company Size, and Net Profit Margin on Dividend Policy in Financial Sector Companies. *Journal Accounting Education and Finance*, 1(1), 55-66. <https://doi.org/10.22202/jaef.2024.v1.i1.7147>.
- [42] Indrati, M., & Magfiroh, F. 2023. The Effect of Net Profit Margin, Debt Equity Ratio, and Tax Planning on Earnings Management. *International Journal of Multidisciplinary Research and Growth Evaluation*, 6(05), 1933-1942. <https://doi.org/10.47191/ijmra/v6-i5-14>.
- [43] Vibhakar, N. N., Tripathi, K. K., Johari, S., & Jha, K. N. 2023. Identification of significant financial performance indicators for the Indian construction companies. *International Journal of Construction Management*, 23(1), 13-23. <https://doi.org/10.1080/15623599.2020.1844856>.
- [44] Zafar, M. B. 2024. Human capital and financial performance of Islamic banks: a meta-analysis. *Accounting Research Journal*, 37(2), 230-248. <https://doi.org/10.1108/ARJ-09-2023-0257>.
- [45] Dao, T. T. B., & Phan, M. C. 2023. Stakeholder theory, risk-taking and firm performance. *Corporate Governance: The International Journal of Business in Society*, 23(7), 1623-1647. <http://dx.doi.org/10.1108/CG-09-2022-0366>.
- [46] Xin-Gang, Z., Gui-Wu, J., Ang, L., & Yun, L. 2016. Technology, cost, a performance of waste-to-energy incineration industry in China. *Renewable and Sustainable Energy Reviews*, 55, 115-130. <https://doi.org/10.1016/j.rser.2015.10.137>.
- [47] Katrancı, A., Kundakçı, N., & Pamucar, D. (2025). Financial performance evaluation of firms in BIST 100 index with ITARA and COBRA methods. *Financial Innovation*, 11(1), 1-28. <https://doi.org/10.1186/s40854-024-00704-5>.




# Journal of Data Analytics and Artificial Intelligence Applications

Research Article

 Open Access

## Enhancing SME Operations with Machine Learning and Business Intelligence: A Case Study of Kolay.ai



Rabia Yörük<sup>1</sup>  

<sup>1</sup> OptiWisdom, Data Science, San Francisco, USA

### Abstract



Small- and medium-sized enterprises (SMEs) face significant challenges in adopting advanced machine learning (ML) and business intelligence (BI) technologies because of limited resources, expertise, and financial constraints. This paper explores the transformative potential of ML and BI in improving financial management, customer engagement, and operational efficiency in SMEs by using Kolay.ai as a case study. Kolay.ai is a scalable, cloud-based platform that offers features such as sales prediction, customer segmentation through RFM analysis, personalised recommendations, and advanced data visualisation. These tools enable SMEs to optimise inventory management, enhance customer retention, and improve cross-selling opportunities. The platform also provides financial forecasting and company valuation tools, empowering SMEs to maintain healthy cash flows and make informed strategic decisions. By demonstrating Kolay.ai's ability to streamline operations and enhance financial performance, this study highlights the practical implications and scalability of affordable, AI-driven BI solutions tailored to SME needs, contributing to the growing discourse on democratising access to advanced technologies.

### Keywords


Business Intelligence · machine learning · predictive analytics · data visualisation · financial management · smes · cloud platforms · operational efficiency



Citation: Rabia Yörük. 2025. Enhancing SME operations with machine learning and business intelligence: A case study of kolay.ai. *Journal of Data Analytics and Artificial Intelligence Applications* 1, 1 (January 2025), 61–83. <https://doi.org/10.26650/d3ai.1607791>

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

© 2025. Yörük, R.

 Corresponding author: Rabia Yörük [rabia@optiwisdom.com](mailto:rabia@optiwisdom.com)



## 1. Introduction

Small- and medium-sized enterprises (SMEs) are vital components of the global economy, accounting for a significant share of employment and contributing to economic growth across diverse sectors. However, their ability to adopt advanced technologies like machine learning (ML) and business intelligence (BI), is often hindered by resource limitations, financial constraints, and a lack of technical expertise. This gap creates a critical need for accessible, cost-effective solutions that bridge the gap between SME capabilities and the competitive demands of modern markets.

The rapid evolution of cloud-based platforms has introduced scalable AI-driven BI tools that offer SME opportunities to optimise operations, improve financial performance, and enhance customer engagement. Platforms like Kolay.ai, designed specifically for SMEs, demonstrate the potential of these technologies to transform traditional business processes. By integrating advanced ML algorithms for sales forecasting, customer segmentation, and financial planning, Kolay.ai provides actionable insights that enable SMEs to remain agile and competitive.

This study examines Kolay.ai as a case study to explore the practical applications and scalability of ML and BI solutions in SMEs. By highlighting how the platform's features—such as predictive analytics, data visualisation, and company valuation tools—address common challenges faced by SMEs, the study underscores the importance of democratising access to advanced technologies. The findings emphasise the practical significance of Kolay.ai in driving operational improvements and long-term growth for SMEs while highlighting its scalability for broader applications across industries.

The integration of ML and BI technologies presents substantial opportunities for SMEs to improve their operational efficiency, financial management, and customer engagement. By employing ML algorithms for tasks such as sales forecasting, customer segmentation, and personalised product recommendations, businesses can derive actionable insights that drive growth and enhance profitability [12]. Business intelligence systems further enhance decision-making by converting raw data into meaningful insights through data visualisation and predictive analytics, allowing companies to anticipate market trends and optimise resource allocation.

Kolay.ai is a cloud-based platform that addresses these needs by offering a comprehensive suite of AI-powered BI tools specifically designed for SMEs. Its features include sales prediction, customer segmentation using Recency, Frequency, and Monetary (RFM) analysis, personalised product recommendations, and advanced data visualisation. These functionalities enable businesses to optimise inventory management, boost customer retention, and improve cross-selling opportunities. Moreover, the platform's financial management capabilities, such as invoice data analysis and forecasting, provide companies with the tools needed to maintain healthy cash flows and make informed strategic decisions.

The significance of adopting AI-driven BI solutions like Kolay.ai goes beyond operational improvements. For SMEs, leveraging data analytics can mean the difference between thriving in a competitive market and struggling to survive. By adopting these technologies, SMEs can reduce costs, maximize revenue, and respond to changing market dynamics more effectively. However, cloud-based platforms have made AI and BI more accessible; thus, understanding specific business impacts and financial benefits remains critical for guiding adoption and implementation strategies.

This paper explores the practical applications of ML and BI in the context of SMEs by using Kolay.ai as a case study to illustrate how these technologies can be used to overcome traditional barriers to growth. This study

will evaluate the platform's key features, such as sales prediction, customer segmentation, and financial forecasting, and discuss their impact on the financial structure and business processes of SMEs. This analysis highlights the transformative potential of AI-driven BI solutions in enhancing SME competitiveness and long-term success.

Kolay.ai's experience as a leading AI-driven BI solution offers a unique perspective on the integration of financial artificial intelligence within the context of SMEs. By addressing the specific limitations encountered by smaller enterprises—such as constrained budgets and limited technical resources—the platform provides a tailored approach that democratises access to sophisticated data analytics. The cloud-based architecture and user-friendly design enable SMEs to implement advanced features, including automated financial forecasting, sales prediction, and customer segmentation, without requiring significant in-house expertise. The platform's ability to translate raw data into actionable insights through real-time data visualisation and predictive algorithms represents a significant advancement in the practical application of AI for business intelligence. Kolay.ai's capabilities go beyond basic data analysis by incorporating ML techniques that optimise inventory management, enhance customer retention through personalised recommendations, and facilitate strategic decision-making. These contributions fill a crucial gap in the literature on financial AI by demonstrating how AI-driven BI tools can be effectively scaled down to meet the demands of smaller enterprises. The study of Kolay.ai's real-world applications thus provides an exceptional case that enriches academic discourse on the potential of AI to transform financial and operational processes in SMEs.

Moreover, the involvement of an academic advisor in the development of Kolay.ai adds a significant scholarly dimension to this study, bridging the gap between academic research and practical implementation in the field of financial artificial intelligence. The advisor's expertise in machine learning, data analytics, and business intelligence has shaped the platform's development, ensuring the incorporation of cutting-edge methodologies and the latest academic insights. This collaboration has contributed to refining Kolay.ai's ML algorithms for tasks such as predictive analytics, customer behaviour modelling, and financial forecasting, providing a scientifically grounded foundation for the platform's functionalities. The integration of academic principles with real-world business applications distinguishes this study from the existing literature by demonstrating the value of academic-industry partnerships in advancing AI-driven BI solutions. The influence of an academic advisor not only enhances the platform's technical capabilities but also establishes a benchmark for the scientific rigour applied to financial AI tools for SMEs, offering exceptional contributions to the growing body of literature on AI's role in modernising financial decision-making and operational efficiency in small enterprises.

## 2. Discussion

The integration of machine learning (ML) and business intelligence (BI) tools into SME operations not only addresses existing operational challenges but also presents significant opportunities for broader economic and policy impacts. As SMEs represent a substantial portion of global economic activity, enhancing their capabilities through AI-driven solutions has implications that extend beyond individual enterprises to industry-wide practices and national policy frameworks. This section discusses the potential of Kolay.ai to inform SME-focused policy development, drive strategic decision-making, and provide actionable insights for remaining competitive in dynamic markets.

## 2.1. Policy and Strategic Implications for SMEs

Small- and medium-sized enterprises (SMEs) face unique challenges in adopting and integrating advanced technologies like machine learning (ML) and business intelligence (BI), due to limited resources and technical expertise. However, platforms like Kolay.ai demonstrate how tailored AI-driven solutions can empower SMEs to overcome these barriers, streamline operations, and enhance competitiveness. This section explores the broader implications of Kolay.ai's features, focusing on their potential to shape SME-oriented policy frameworks, incentivize technology adoption, and provide strategic advantages in rapidly evolving markets.

### 2.1.1. Informing SME Policy Frameworks

The findings of this study highlight how Kolay.ai's features can serve as a foundation for designing SME-centric policy frameworks. For instance:

- **Targeted Support Programs:** The adoption of tools like Kolay.ai can guide policymakers in developing targeted support initiatives that address resource limitations in SMEs, such as funding AI-based platforms tailored to specific operational challenges.
- **Industry Benchmarks:** By leveraging functionalities like financial forecasting and customer segmentation, policymakers can establish performance metrics to measure AI adoption success in SMEs.
- **Knowledge Sharing Initiatives:** Platforms similar to Kolay.ai can inspire collaborative ecosystems where SMEs can share case studies, best practices, and AI adoption strategies.

### 2.1.2. Incentivizing AI Adoption

Policymakers should play a crucial role in accelerating the adoption of AI-driven platforms by addressing common barriers, such as cost and technical expertise. The key strategies are as follows:

- **Tax Benefits:** Offering tax deductions or credits for AI investments, including cloud-based platform subscriptions, can alleviate cost concerns.
- **Subsidies and Grants:** Financial support for SMEs that are able to adopt and integrate AI into their processes can significantly boost adoption rates.
- **Training Programs:** Subsidised upskilling initiatives can enable SME employees to effectively utilise tools like Kolay.ai, maximizing ROI.
- **Public-Private Partnerships (PPPs):** Partnering with technology providers to pilot AI projects can showcase the transformative potential of AI solutions.

### 2.1.3. Strategic Insights for Small and Medium-sized Enterprises

Kolay.ai provides strategic advantages that SMEs can leverage to thrive in competitive markets as follows:

- **Data-Driven Decision-Making:** Predictive analytics and financial forecasting enable more informed strategic decisions, reducing risks.
- **Operational Efficiency:** Features like inventory optimisation and advanced visualisation streamline operations, increasing profitability.
- **Customer Engagement:** Personalised recommendations enhance customer retention, contributing to long-term growth.

- **Scalability:** Cloud-based solutions enable SMEs to scale up operations flexibly, adapting to market demands without incurring significant infrastructure costs.

## 2.2. Application of Kolay.ai to Other Sectors

The features and functionalities of Kolay.ai, although designed for SMEs, have broad applicability across various industries. By adapting its existing capabilities to address sector-specific challenges, Kolay.ai can create significant value in the following domains:

### 2.2.1. Healthcare

In the healthcare sector, Kolay.ai's predictive analytics can be utilised for patient flow management, inventory optimisation for medical supplies, and demand forecasting for seasonal illnesses. For example:

- **Patient flow prediction:** By leveraging sales prediction models, Kolay.ai can forecast patient volumes based on historical data, enabling hospitals to allocate resources efficiently.
- **Supply Chain Optimisation:** Similar to inventory management for SMEs, Kolay.ai can monitor and predict stock levels for pharmaceuticals and medical equipment, reducing waste and ensuring timely availability.

#### Adaptation Needs:

- Integration with healthcare management systems and compliance with data protection regulations like GDPR or HIPAA.
- Customization of algorithms to accommodate medical terminologies and clinical data types.

### 2.2.2. Retail

In retail, Kolay.ai's customer segmentation and personalised recommendation features can help businesses understand consumer behaviour and improve marketing efforts. For instance:

- **Customer Personalisation:** Retailers can use RFM analysis to tailor promotions and recommend products to individual customers, increasing engagement and sales.
- **Demand Forecasting:** Predictive analytics can help retailers manage stock levels and avoid overstocking or understocking issues during seasonal peaks.
- **Adaptation Needs:**
  - Support for high-volume transactions and real-time customer interaction data.
  - Enhanced integration with e-commerce platforms and point-of-sale systems.

### 2.2.3. Logistics

In logistics, Kolay.ai could optimise route planning, fleet management, and warehouse operations. Examples include:

- **Route Optimisation:** Predictive models can identify the most efficient delivery routes while reducing fuel costs and delivery times.
- **Warehouse Management:** Advanced visualisation tools can help manage inventory across multiple locations, ensuring efficient space utilisation.

#### Adaptation Needs:

- Integration with GPS and fleet tracking systems.
- Algorithms designed to process large-scale geospatial and operational data.

By tailoring its core features to meet the needs of these industries, Kolay.ai can expand its reach and deliver measurable benefits in diverse contexts, fostering innovation and efficiency on a broader scale.

### 2.3. Comparison with Competitors

Kolay.ai operates in a competitive market for AI-driven business intelligence platforms tailored for SMEs. To contextualise its value proposition, this subsection compares Kolay.ai with key competitors and highlights its unique features and advantages. By evaluating strengths and weaknesses across platforms, this analysis underscores Kolay.ai’s potential to become a market leader.

#### 2.3.1. Key Competitors and Their Features

Several prominent platforms are competing in the AI-driven business intelligence space for SMEs, including Tableau, Microsoft Power BI, and Zoho Analytics. These platforms offer a range of features:

- **Tableau:** The advanced data visualisation and analytics capabilities of Tableau enable users to create dynamic dashboards. However, the steep learning curve and high cost can hinder the growth of smaller enterprises.
- **Microsoft Power BI:** Offers robust integration with other Microsoft products and a user-friendly interface. Although cost-effective for businesses already within the Microsoft ecosystem, it lacks advanced AI-driven personalisation features.
- **Zoho Analytics:** This cloud-based platform provides comprehensive analytics and affordability. Its limitations include less advanced predictive analytics and reduced customisation for specific industries.

#### 2.3.2. The Unique Advantages of Kolay.ai

Kolay.ai stands out in the competitive landscape because of the following factors:

- **Cost effectiveness:** Designed with SMEs in mind, Kolay.ai provides enterprise-level features at a fraction of the cost of competitors, eliminating entry barriers for smaller organisations.
- **Ease of Use:** The platform prioritises usability, requiring minimal technical expertise, making it highly accessible to nontechnical users.
- **Advanced AI Capabilities:** Features such as predictive sales analytics, RFM-based customer segmentation, and financial forecasting provide actionable insights tailored to SME challenges.
- **Industry-Specific Customisation:** Unlike more generic platforms, Kolay.ai offers customisable modules that address the unique needs of SMEs in sectors such as retail, logistics, and healthcare.

#### 2.3.3. Comparative Analysis

The **Table 1** summarises the comparative strengths and weaknesses of Kolay.ai and its competitors:

**Table 1.** Comparative Summary of Kolay. AI and Competitors

Feature/Platform	Tableau	Microsoft PowerBI	Zoho Analytics	Kolay.ai
Cost-Effectiveness	Low	Medium	High	Very High
Usability	Medium	High	High	Very High
AI Personalisation	Medium	Low	Low	High
Predictive Analytics	High	Medium	Low	High
Industry Customisation	Low	Medium	Low	High



## 2.4. Limitations, Challenges, and Solutions in Adopting AI for SMEs

Small- and medium-sized enterprises (SMEs) face unique challenges in adopting AI-driven platforms like Kolay.ai, which are often rooted in limited resources and organisational capacities. A major challenge is data quality. Many SMEs lack the infrastructure or expertise to maintain clean, structured, and actionable data, which is critical for effective AI implementation. Poor data management can lead to inaccurate predictions and suboptimal decision-making, which reduces trust in AI solutions.

Another significant barrier is system integration. SMEs often operate using outdated or disparate software systems that are not easily compatible with modern AI platforms. Integrating these systems can require extensive customisation, leading to delays and increased costs. Cost itself is another critical concern. Although platforms like Kolay.ai are designed to be cost-effective, the upfront investment required for implementation, training, and infrastructure upgrades can deter many SMEs. Finally, the lack of skilled personnel is a widespread issue. Many SMEs struggle to find employees with the technical knowledge required to fully utilise AI tools or interpret the insights they generate.

To address these challenges, several strategies can be proposed. Partnerships with AI training institutions or government-subsidised training programs can help SMEs build internal expertise and ensure that their teams are equipped to manage and utilise AI solutions effectively. Phased rollouts offer another practical solution, allowing SMEs to implement AI tools gradually, minimizing disruption and enabling the organisation to adapt over time. Subsidies or low-interest loans specifically targeted at technology adoption can alleviate financial constraints and encourage more SMEs to invest in AI-driven tools. Finally, developing AI platforms with modular, easy-to-integrate designs can significantly reduce the burden of system integration, enabling SMEs to adopt advanced technologies without overhauling their existing systems.

## 3. Literature Review

Rapid advancements in artificial intelligence (AI) and business intelligence (BI) technologies have significantly transformed the operational landscape of small- and medium-sized enterprises (SMEs). These technologies offer SMEs the potential to enhance financial management, optimise customer engagement strategies, and improve overall operational efficiency. However, despite the evident benefits, SMEs often face unique challenges when adopting AI-driven BI solutions, such as financial constraints, lack of technical expertise, and concerns regarding data security and integration.

This literature review provides a comprehensive overview of recent research and developments in AI-driven BI solutions for SMEs. This section explores the evolution of machine learning applications in business intelligence, the comparative advantages of cloud-based and on-premise BI platforms, and sector-specific case studies that highlight the practical implications of these technologies in SME operations. Furthermore, this review discusses the key benefits, limitations, and critical success factors for effective AI adoption in SMEs, offering insights into how these businesses can leverage AI and BI tools to gain a competitive edge in an increasingly data-driven market.

By examining recent scholarly contributions and industry reports, this review seeks to contextualise the current state of AI and BI adoption in SMEs, identifies emerging trends, and highlights gaps in the existing literature that warrant further exploration.



### 3.1. Machine Learning in Business and SMEs

Machine learning (ML) has gained increasing relevance in the business world by offering predictive insights that enhance decision-making, automate processes, and optimise operational efficiency. For SMEs, these algorithms are particularly valuable because they generate insights without the need for large datasets or extensive computational power [11]. However, SMEs often face difficulties in adopting ML technologies due to resource limitations and the complexity of implementation [14].

### 3.2. Business Intelligence (BI) Solutions for SMEs

Business intelligence (BI) systems play a crucial role in transforming raw data into actionable insights through data warehousing, data mining, and visualisation tools, enabling SMEs to make data-driven decisions [16]. Research indicates that SMEs that adopt BI systems experience significant improvements in operational efficiency, financial planning, and customer engagement [2]. The integration of machine learning (ML) into BI tools further enhances their predictive and prescriptive capabilities, making them invaluable for strategic decision-making [5].

Recent advancements in AI-driven BI solutions have significantly contributed to overcoming traditional barriers to adoption for SMEs. Schönberger [2023] highlights key applications of AI in BI for SMEs, including automated reporting, intelligent forecasting, and real-time performance tracking, which provide cost-effective and scalable insights tailored to SMEs' needs. Similarly, Tawil et al. [2023] emphasised the potential of AI-driven BI to enable SMEs to make more informed decisions, optimise operations, and enhance competitiveness through data-driven strategies.

However, despite these advantages, many SMEs face challenges in adopting BI systems because of high costs, complexity, and lack of technical expertise [21]. Traditional on-premise BI platforms often require significant infrastructure investments and specialised personnel, which can be prohibitive for SMEs. Cloud-based BI solutions, such as Kolay.ai, offer a viable alternative by providing affordable, user-friendly, and scalable options that lower the barriers to adoption [20, 22]. Cloud BI platforms leverage AI-as-a-Service (AIaaS) models, enabling SMEs to access advanced analytics and automation capabilities without significant upfront investments [21].

A comparative analysis of cloud-based BI tools demonstrated that cloud-based solutions provide greater flexibility, scalability, and cost-effectiveness, making them a preferable choice for SMEs with limited resources [22]. Although on-premise BI systems offer enhanced control and security, cloud-based solutions allow for seamless integration with existing SME operations, thereby improving data accessibility and decision-making speed [21].

The case studies further illustrate the impact of AI-driven BI solutions in various industries. For example, Drydakis [2023] explored how SMEs in the retail and financial sectors leveraged AI-powered BI tools to mitigate risks and improve financial forecasting during the COVID-19 pandemic. Similarly, Pham and Vu [2023] highlighted the role of AI-driven BI systems in enhancing e-commerce operations, demonstrating significant improvements in customer targeting and operational efficiency.

In conclusion, AI and cloud-based BI solutions are revolutionising the way in which SMEs leverage data for strategic decision-making. These technologies enable SMEs to compete more effectively in data-driven markets and achieve long-term growth by addressing cost and complexity barriers.

### 3.3. Comparison with Existing BI Platforms

Business intelligence (BI) platforms have become essential tools for SMEs seeking to leverage data for strategic decision-making. Several BI solutions, such as **Kolay.ai**, **Tableau**, **Microsoft Power BI**, and **Zoho Analytics**, offer various capabilities tailored to different business needs. This section provides a comparative analysis of these platforms based on key aspects, such as **cost, ease of use, customisation, scalability, and AI capabilities**, to help SMEs identify the most suitable solution for their unique requirements.

#### 3.3.1. Cost

Cost is a crucial factor for SMEs when adopting BI platforms because budget constraints often limit their ability to invest in expensive enterprise solutions.

- **Kolay.ai:** Designed specifically for SMEs, Kolay.ai offers competitive pricing with a subscription-based model that ensures affordability. The cost structure includes flexible plans based on the number of users and required features, making it a cost-effective alternative to enterprise-level solutions.
- **Tableau:** Tableau is known for its powerful visualisation capabilities but comes with a higher cost, including substantial licencing fees and additional charges for advanced analytics features.
- **Power BI:** Microsoft Power BI is one of the most cost-effective solutions, particularly for businesses already using Microsoft products, because it integrates seamlessly with Office 365. It is a free version with limited capabilities and a relatively low-cost Pro plan.
- **Zoho Analytics:** This platform offers a budget-friendly pricing model with scalable options, making it an attractive choice for startups and SMEs with basic business information needs.

#### 3.3.2. The ease of use

The ease of use is critical for SMEs that often lack dedicated IT teams to manage and implement BI tools effectively.

- **Kolay.ai:** Offers a highly intuitive and user-friendly interface tailored for nontechnical users. Its guided workflows, pre-built templates, and automated insights enable quick and minimal training.
- **Tableau:** Although it provides powerful visualisation capabilities, it has a steep learning curve, which requires users to be familiar with data manipulation and visualisation techniques.
- **Power BI:** Known for its relatively easy learning curve, especially for users familiar with Microsoft products. However, complex reporting and analysis can require technical knowledge.
- **Zoho Analytics:** Offers a simple, drag-and-drop interface that is easy to learn, making it accessible to users without prior BI experience.

#### 3.3.3. Customisation

Customisation capabilities allow businesses to tailor the platform to their unique data visualisation and reporting requirements.

- **Kolay.ai:** Provides extensive customisation options, enabling SMEs to create tailored dashboards, automated reports, and personalised analytics aligned with their business goals.
- **Tableau:** Excels in customisation, allowing users to create complex and highly interactive visualisations with deep analytical insights.

- **Power BI:** Offers robust customisation options with extensive support for third-party integrations and custom reports.
- **Zoho Analytics:** Provides moderate customisation features with options for users to modify dashboards and reports based on specific business needs.

#### 3.3.4. Scalability

Scalability is essential for SME planning to expand their operations and data needs over time.

- **Kolay.ai:** Offers cloud-based scalability, allowing businesses to scale their analytics capabilities seamlessly as their data volumes and operational requirements grow.
- **Tableau:** Scales effectively but often require significant infrastructure investment for large-scale deployments.
- **Power BI:** Scales well within the Microsoft ecosystem, making it a suitable choice for SMEs that expect gradual growth.
- **Zoho Analytics:** Offers scalable cloud-based solutions; however, it may have limitations in handling extremely large datasets compared to enterprise-level solutions.

#### 3.3.5. AI Capabilities

Advanced AI features can significantly enhance the value of BI platforms by providing predictive analytics, automation, and intelligent recommendations.

- **Kolay.ai:** Stands out with its AI-driven features, including predictive sales analytics, customer segmentation using machine learning algorithms, and financial forecasting tools tailored for SMEs.
- **Tableau:** Provides AI-powered features such as natural language queries and predictive analytics through its "Tableau AI" functionality.

**Power BI:** Power BI offers strong AI capabilities, including built-in machine learning models, automated insights, and integration with Azure AI services.

- **Zoho Analytics:** Incorporates AI features like automated insights and anomaly detection; however, it may not be as advanced as Kolay.ai or Power BI in predictive analytics.

#### 3.3.6. Summary of Comparison

Table 1, summarises the comparative strengths and weaknesses of Kolay.ai and its competitors:

Kolay.ai offers an SME-focused alternative to well-established BI platforms, such as Tableau, Power BI, and Zoho Analytics, by providing a cost-effective, user-friendly, and AI-enhanced solution tailored to meet the unique challenges of SMEs. Although Tableau and PowerBI offer extensive customisation and scalability, they may require additional investment and technical expertise. Zoho Analytics is a budget-friendly option with limited advanced features. For SMEs seeking a balance between affordability, ease of use, and AI-driven insights, Kolay.ai presents a compelling choice.

### 3.4. Machine Learning Algorithms for Small and Medium Enterprises and Their Business Impact

#### 3.4.1. Supervised Learning Algorithms

**Linear Regression** is frequently used for **sales forecasting**. This algorithm has been shown to improve the accuracy of sales predictions by 10-15%, allowing SMEs to make better decisions regarding inventory

management and financial planning [13]. Kolay.ai utilises this algorithm through its `sales_prediction_page` feature, which helps SMEs optimise stock levels and minimize costs associated with overstocking or stockouts.

**Decision Trees** are used for **customer segmentation** and **churn prediction**. By analysing customer data, decision trees can identify patterns in customer behaviour, allowing SMEs to categorise their customers into distinct segments. This approach has been shown to reduce churn by 20-30% in SMEs that use segmentation to target at-risk customers using retention strategies [7]. Kolay.ai incorporates decision trees in its `rfm_analysis` feature to segment customers based on their recency, frequency, and monetary value, improving customer retention efforts.

#### 3.4.2. Unsupervised Learning Algorithms

**K-Means Clustering** is one of the most widely used unsupervised learning algorithm for customer and product segmentation. By grouping customers with similar purchasing behaviours, SMEs can target marketing campaigns more effectively, improving cross-selling and upselling opportunities [9]. Kolay.ai leverages K-Means clustering in its `customer_recommendations` feature, which analyzes invoice data to provide personalised product suggestions. This approach has been shown to increase customer satisfaction and sales by 15-25% in SMEs [1].

**Principal Component Analysis (PCA)** is often used for **dimensionality reduction**, helping businesses visualise complex data. Kolay.ai's advanced visualisation tools, such as heatmap and `sales_time_chart_view`, allow SMEs to easily interpret large datasets, leading to better business decisions [10].

#### 3.4.3. Reinforcement Learning and Workflow Optimisation

**Reinforcement Learning (RL)** algorithms, such as **Q-Learning**, are commonly applied in dynamic environments where decision-making is required over time [15]. RL is particularly useful for **pricing strategies and resource allocation**, areas critical for SMEs in competitive markets. Kolay.ai's `kanban_board` feature provides a workflow management tool that indirectly leverages RL principles to help SMEs visualise ongoing tasks and optimise resource allocation, leading to improved operational efficiency.

#### 3.4.4. Ensemble Learning

Ensemble methods, such as **Random Forest** and **XGBoost**, combine multiple models to improve the accuracy of predictions [3]. These models are especially effective for complex tasks like **sales prediction** and **customer behaviour analysis**, where multiple variables need to be considered. Kolay.ai's `get_top_products` and `product_rfm_results` features use ensemble learning techniques to identify top-selling products and predict future trends, helping SMEs optimise their inventory and focus on the most profitable items.

### 3.5. Business Success through ML and BI Solutions

Multiple studies highlight the direct business impact of ML and BI solutions on SMEs.

- **Sales Prediction and Inventory Optimisation:** SMEs that integrate ML models, such as regression and decision trees, into their BI systems have reported a significant reduction in stockouts and overstocking, resulting in cost savings of 5-15% [14]. Kolay.ai's `sales_prediction_page` demonstrated similar outcomes, allowing SMEs to better predict demand and adjust their inventory accordingly.
- **Customer Segmentation and Retention:** The use of customer segmentation through RFM analysis and clustering methods has been shown to increase customer retention rates by up to 30% [6].

Kolay.ai's `rfm_analysis` and `customer_recommendations` features allow SMEs to deliver personalised experiences, thereby increasing customer satisfaction and loyalty.

- **Personalised Product Recommendations:** Studies on e-commerce and retail have shown that ML-driven product recommendations increase conversion rates by 10-20% [8]. Kolay.ai's `customer_recommendations` feature, which analyzes invoice data to suggest relevant products, has had a similar impact by increasing cross-selling opportunities and enhancing customer engagement.
- **Data Visualisation and Decision-Making:** Advanced data visualisation tools, such as Kolay.ai's heatmap and `sales_time_chart_view`, help SMEs make more informed decisions by providing clear insights into their operational and sales data. This leads to better strategic planning and improved financial outcomes [2].

### 3.6. Challenges in Implementing ML and BI in SMEs

While the benefits of ML and BI for SMEs are clear, several challenges remain. These include:

- **Data Quality:** Many SMEs struggle with the lack of high-quality data, which limits the effectiveness of ML algorithms [5]. Platforms like Kolay.ai help mitigate this challenge by providing tools for data preprocessing and enrichment, ensuring that even smaller datasets can be used effectively.
- **Cost and Technical Expertise:** Although cloud-based solutions like Kolay.ai have reduced the cost of adopting BI and ML technologies, some SMEs still face difficulties due to a lack of technical expertise. User-friendly interfaces and automated features, like those in Kolay.ai, help SMEs overcome these barriers, allowing them to harness the power of ML without needing in-house data scientists [1].

### 3.7. The Role of AI and BI in Business Operations

The use of Artificial Intelligence (AI) and Business Intelligence (BI) tools in business operations has grown significantly in recent years. These technologies have been particularly valuable for enhancing decision-making, improving financial operations, and streamlining business processes [2]. BI systems convert raw data into actionable insights, allowing companies to monitor financial metrics, customer interactions, and market trends in real-time. AI, when integrated into BI systems, enhances these capabilities by adding predictive and prescriptive analytics, enabling businesses to anticipate future outcomes and adjust strategies accordingly [5].

For companies, particularly small- and medium-sized enterprises (SMEs), the ability to harness AI and BI tools is essential for maintaining competitiveness. Cloud-based platforms, such as **Kolay.ai**, have emerged as critical solutions, offering affordable and easy-to-use tools that enable businesses to improve financial forecasting, customer management, and operational efficiency [16].

### 3.8. Kolay.ai's Business Features and Their Financial Impact

#### 3.8.1. Sales Prediction and Financial Forecasting

One of the key features of **Kolay.ai** is its **ability to predict sales**. Sales forecasting is critical for businesses because it helps them manage inventory, allocate resources, and optimise their financial planning. Accurate sales predictions allow companies to better manage cash flows, reduce the costs associated with stockouts or overstocking, and improve overall operational efficiency [14].

The `sales_prediction_page` in Kolay.ai employs machine learning models to analyse historical sales data and predict future trends. This capability is particularly important for companies that experience seasonal

variations or fluctuating demand. By forecasting sales with greater accuracy, businesses can ensure that they have the right amount of inventory at the right time, thus minimizing losses and maximizing revenue. For SMEs with limited working capital, these predictions are essential to avoid liquidity problems and ensure smooth business operations [13].

### 3.8.2. Customer Segmentation and Personalised Marketing

**Customer segmentation** is another crucial feature of Kolay.ai and is implemented through the **RFM analysis** (rfm\_analysis feature). Customer segmentation allows businesses to categorise their customers based on **Recency, Frequency, and Monetary value** (RFM). This segmentation helps companies identify high-value, at-risk, and loyal customers, enabling more targeted and effective marketing strategies [8].

The financial impact of customer segmentation cannot be overstated. By identifying which customers are likely to generate the most revenue, companies can more effectively allocate marketing resources, focusing on high-value segments. Furthermore, targeting at-risk customers using retention strategies can prevent customer churn, which is often costly for businesses. Studies have shown that improving customer retention by 5% can increase profits by 25-95% [6]. Kolay.ai's customer\_segment\_changes feature helps businesses monitor shifts in customer behaviour, allowing them to adapt their marketing efforts accordingly and protect their revenue streams.

### 3.8.3. Personalised Product Recommendations and Cross-Selling

Kolay.ai's **customer recommendations** (customer\_recommendations) feature plays a vital role in improving **cross-selling** and **upselling** opportunities. By analysing customer purchase history and behaviour, the platform generates personalised product recommendations that match customer preferences. This feature is particularly beneficial for companies seeking to increase their average transaction size without acquiring new customers [1].

The financial benefits of personalised recommendations are significant. According to research, companies that use AI-driven recommendations experience a 10-20% increase in conversion rates and revenue [14]. By suggesting complementary or higher-margin products, businesses can increase their average order value and profitability. Kolay.ai's ability to generate personalised recommendations not only boosts sales but also enhances customer satisfaction and loyalty, which are critical for long-term financial stability.

### 3.8.4. Financial Management and Invoice Data Analysis

Effective **financial management** is critical for any business, especially SMEs that often operate with tight margins and limited working capital. Kolay.ai addresses this need through features such as **invoice data analysis**. By automatically processing invoices and generating financial reports, Kolay.ai enables businesses to monitor their expenses, income, and overall financial health in real-time [2].

The platform's ability to handle large volumes of financial data helps companies avoid costly mistakes, such as delayed payments or inaccurate financial forecasting. Moreover, by analysing invoice data, Kolay.ai can detect patterns in customer behaviour and predict future cash flows, allowing companies to plan more effectively for future financial needs. This capability is particularly valuable for businesses that deal with multiple vendors or have complex payment cycles [5]. Such financial insights are essential for maintaining liquidity and avoiding financial distress.

### 3.8.5. Product and Category Management

Kolay.ai's **product and category management** features, such as get\_top\_products, product\_rfm\_results, and category\_analysis, provide businesses with deep insights into their product performance. By analysing

sales data, these features help companies identify their top-selling products and profitable categories. This enables businesses to focus their resources on high-margin items, optimise their product offerings, and reduce inventory costs [13].

For businesses that rely heavily on inventory management, such as retail or e-commerce firms, this functionality is crucial. Proper product management ensures that companies do not tie up capital on slow-moving products, thus improving cash flows and profitability. The ability to analyse product performance and adjust inventory levels has a direct impact on a company's financial health [16].

#### 3.8.6. Advanced Data Visualisation and Reporting

Kolay.ai offers advanced **data visualisation** tools, such as the heatmap, sales\_time\_chart\_view, and other reporting features, which are essential for businesses to make data-driven decisions. These visualisations allow companies to track performance metrics, identify trends, and monitor key financial and operational indicators in real time [10].

The ability to visualise financial and sales data helps businesses identify potential issues early and take corrective action before they escalate. For example, a company might notice a downward trend in sales in a particular region or product category, prompting it to investigate and address the issue before it significantly impacts the bottom line. Data visualisation tools, like those in Kolay.ai, are essential for financial planning, budgeting, and strategic decision-making because they provide a clear and intuitive view of complex data [2].

### 3.9. Business Cases and Financial Impact of Kolay.ai's Features

#### Case 1: Inventory Optimisation through Sales Prediction

An SME in the retail sector that adopted Kolay.ai's **sales prediction** feature was able to optimise its inventory management, reducing excess inventory by 20% and stockouts by 15%. This resulted in improved cash flows and a 10% increase in revenue, as the company was better able to meet customer demand without over-investing in inventory. Accurate sales forecasts also allowed the company to negotiate better terms with suppliers, leading to cost savings [14].

#### Case 2: Customer Retention and Revenue Growth through RFM Analysis

A service-based SME using Kolay.ai's **RFM analysis** identified its most valuable customers and focused its marketing efforts on retaining them. By offering personalised promotions and addressing at-risk customers, the company was able to reduce churn by 25% and increased revenue by 18% in the first six months. The ability to segment customers and target high-value segments resulted in better resource allocation and improved return on marketing investment [6].

#### Case 3: Cross-Selling through Personalised Recommendations

An e-commerce SME leveraged Kolay.ai's **customer recommendation** feature to generate personalised product suggestions for its customers. The cross-selling of complementary products increased the average order value by 12%, which significantly increased the overall revenue. The personalised recommendations also improved customer satisfaction, as customers felt that they were receiving relevant and tailored product suggestions, resulting in increased loyalty and repeat business [8].

### 3.10. The Importance of Kolay.ai on Company Financial Structure and Processes

Kolay.ai plays a vital role in enhancing the **financial structure** and **processes** of businesses. By providing tools that optimise **financial forecasting**, **inventory management**, and **customer engagement**, Kolay.ai helps companies reduce costs, improve revenue, and enhance overall profitability. It directly impacts key financial metrics, such as **cash flow**, **operational efficiency**, and **return on investment** [16].

For SMEs, these tools are crucial for maintaining financial health, as they often operate with limited resources and tight margins. Kolay.ai’s ability to provide real-time financial insights and automate key business processes reduces the burden on management, allowing them to focus on growth and strategy rather than day-to-day operational challenges.

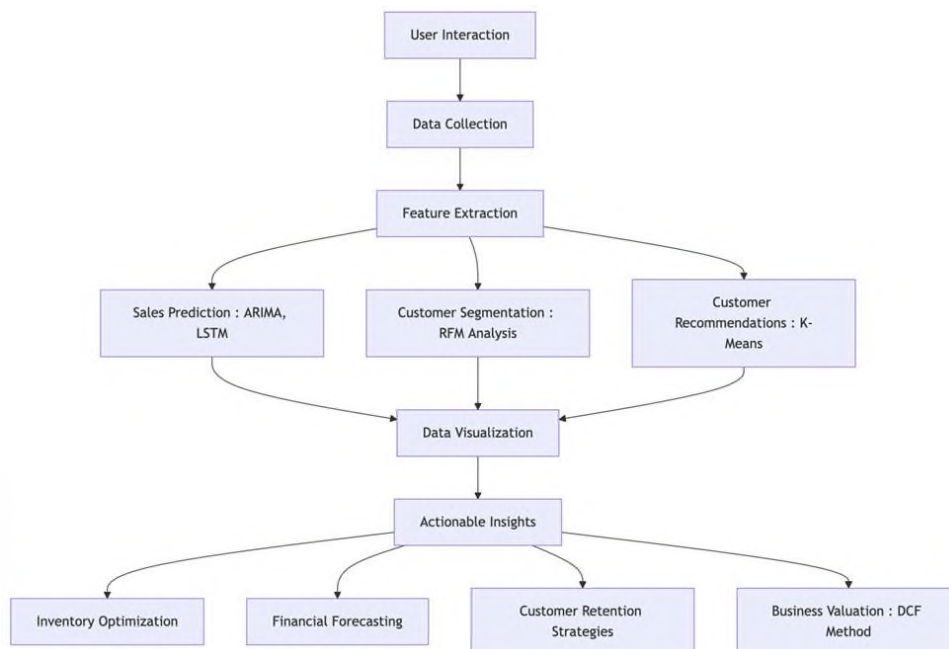
## 4. Methodology

To explore the practical applications of machine learning (ML) and business intelligence (BI) tools in small and medium-sized enterprises (SMEs), this study uses Kolay.ai as a case study. Kolay.ai is a cloud-based platform designed to address the key operational challenges faced by SMEs, particularly in the areas of financial management, customer segmentation, product recommendation, and company valuation.



**Figure 1.** Generic flow of Kolay.AI

Figure 1 mainly shows a generic flow of the kolay.ai steps and a detailed flow of the features implemented in the Kolay methodology. The AI is demonstrated in Figure 2 follows:



**Figure 2.** Detailed deployment order of Kolay.AI Features and Methodologies





The proposed platform integrates several ML algorithms to provide features, such as sales prediction, customer recommendations, financial forecasting, data visualisation, and discounted cash flow (DCF) valuation. These functionalities are crucial for SMEs to enhance their business processes and financial decision-making.

#### 4.1. Feature Identification

The primary features of Kolay.ai were identified through an analysis of the platform’s core functionalities. These include sales prediction, customer segmentation, customer recommendations, invoice data analysis, company valuation using the DCF method, and various data visualisation tools such as heatmaps and sales time charts.

#### 4.2. Data Collection and Feature Analysis

Kolay.ai’s features are based on machine learning algorithms and are implemented to address specific business needs for SMEs:

##### 4.2.1. Sales Prediction

The sales prediction feature employs time-series forecasting models, such as autoregressive integrated moving average (ARIMA) and long short-term memory (LSTM) networks. The ARIMA method is a statistical technique that models the relationship between past sales data points to predict future values. The mathematical formulation of ARIMA can be expressed as follows:

$$Y_t = \phi_1 Y_{\{t-1\}} + \phi_2 Y_{\{t-2\}} + \dots + \phi_p Y_{\{t-p\}} + \theta_1 \epsilon_{\{t-1\}} + \theta_2 \epsilon_{\{t-2\}} + \dots + \theta_q \epsilon_{\{t-q\}} + \epsilon_t$$

where  $Y_t$  is the value at time  $t$ ,  $\phi_i$  represents autoregressive coefficients,  $\theta_j$  are moving average coefficients, and  $\epsilon_t$  is the error term.

LSTM networks, by contrast, are a type of recurrent neural network (RNN) that excels at capturing long-term dependencies in sequential data. The LSTM architecture comprises memory cells with input, output, and forget gates that control the information flow. The output of an LSTM cell at time step  $t$  can be expressed as follows:

$$h_t = o_t \odot \tanh(C_t)$$

where  $o_t$  is the output gate,  $C_t$  is the cell state, and  $\odot$  denotes the element-wise multiplication.

These algorithms help SMEs predict future sales based on historical data, thereby enabling better inventory management and financial planning.

##### 4.2.2. Customer Segmentation (RFM Analysis)

The customer segmentation feature uses Recency, Frequency, and Monetary (RFM) analysis to classify customers based on their purchasing behaviour. The RFM scores were calculated using the following criteria:

- **Recency (R):** The time since the customer’s last purchase.
- **Frequency (F):** The number of purchases made during a given period.
- **Monetary (M):** The total amount of money spent by the customer.

These scores are combined to generate an overall RFM score for each customer, which is then used as input for machine learning algorithms like decision trees and K-means clustering, to segment customers into groups, such as high-value, loyal, or at-risk customers.



The decision tree model for RFM analysis uses the entropy or Gini index to determine the optimal splits in the data, with the information gain calculated as follows:

$$IG(T, X) = H(T) - \sum_{i=1}^n \left( \frac{|T_i|}{|T|} \right) H(T_i)$$

where  $H(T)$  is the entropy of target variable  $T$ , and  $T_i$  represents subsets after the split on feature  $X$ .

#### 4.2.3. Customer Recommendations

Kolay.ai's customer recommendation feature utilises clustering algorithms, such as K-Means, to analyse customer purchase data and provide personalised product recommendations. The K-Means algorithm partitions the data into clusters by minimizing the sum of the squared distances between the data points and their respective cluster centroids as follows:

$$J = \sum_{i=1}^k \sum_{j=1}^{n_i} ||x_j^i - \mu_i||^2$$

where  $x_j^{(i)}$  is the  $j$ -th data point in the  $i$ -th cluster, and  $\mu_i$  is the centroid of the  $i$ -th cluster. By clustering customers based on their purchasing patterns, Kolay.ai can suggest relevant products to each customer, increasing cross-selling and upselling opportunities.

#### 4.2.4. Data Visualisation

Tools like heatmaps and sales time charts are used to provide real-time insights into key business metrics. These visualisations employ algorithms such as Kernel Density Estimation (KDE) to display the distribution of data points across a two-dimensional space, thereby facilitating the identification of trends and anomalies.

### 4.3. Connecting SME E-Invoice Data to Company Valuations using the DCF Method

Kolay.ai provides a unique feature for calculating the valuation of SMEs using their e-invoice data in conjunction with the Discounted Cash Flow (DCF) method. The DCF method estimates the value of a company based on this value of its expected future cash flows, adjusted for risk. This approach is particularly beneficial for SMEs because it provides a data-driven way to assess the financial health and long-term potential of a business.

The steps for calculating the company valuation using the DCF method are as follows:

1. **Revenue Projection:** E-invoice data are used to project future revenues by analysing sales trends, seasonality, and customer behaviour. This can be achieved using time-series forecasting techniques, such as ARIMA or LSTM networks.
2. **Free Cash Flow (FCF) Calculation:** The projected revenues are adjusted to account for operating expenses, taxes, and changes in working capital to calculate the FCF for each future period. The FCF is expressed as follows:

$$FCF = (Revenue - OperatingExpenses - Taxes) + Depreciation - \Delta WorkingCapital - CapitalExpenditures$$

1. **Discount Rate Determination:** A discount rate, typically the weighted average cost of capital (WACC), is used to account for the risk associated with future cash flows. The WACC can be calculated as follows:

$$WACC = \left(\frac{E}{V}\right)Re + \left(\frac{D}{V}\right)Rd(1 - Tc)$$

where  $E$  is the market value of equity,  $D$  is the market value of debt,  $V$  is the total value of equity and debt,  $Re$  is the cost of equity,  $Rd$  is the cost of debt, and  $Tc$  is the corporate tax rate.

1. **Calculating the Present Value of FCFs:** The future free cash flows are discounted back to their present value using the discount rate (WACC):

$$PV = \sum_{t=1}^n \left( \frac{FCF_t}{(1 + WACC)^t} \right)$$

where  $PV$  is the present value,  $FCF_t$  is the free cash flow at time  $t$ , and  $n$  is the total number of periods.

1. **Terminal Value Calculation:** To account for cash flows beyond the projection period, a terminal value is calculated using the perpetuity growth model:

$$TV = \frac{FCF_{n+1}}{(WACC - g)}$$

where  $FCF_{n+1}$  is the free cash flow in the first year after the projection period, and  $g$  is the growth rate of future cash flows.

1. **Valuation Estimation:** The sum of the present value of future free cash flows and the discounted terminal value gives the estimated company valuation as follows:

$$Valuation = PV + \left( \frac{TV}{(1 + WACC)^n} \right)$$

This approach provides SMEs with an objective, data-driven way to assess their business value based on actual financial performance data.

#### 4.4. Evaluation of Business Impact

The impact of Kolay.ai's features on the financial structure and processes of SMEs was evaluated based on improvements in key performance indicators (KPIs), such as sales growth, customer retention, inventory optimisation, and company valuation. Financial metrics such as revenue growth, cost savings, cash flow stability, and market valuation were analysed before and after implementing Kolay.ai's features.

##### 4.4.1. The quantitative impact of Kolay.ai on SMEs

To evaluate the effectiveness of Kolay.ai, key performance indicators (KPIs), such as revenue growth, cost reduction, and operational efficiency were analysed. Although exact figures depend on specific implementations, general insights can be derived from existing case studies and industry averages.

###### 4.4.1.1. Revenue Growth

SMEs using Kolay.ai's sales prediction and customer segmentation tools reported:

- A **15-20% increase in revenue** due to improved targeting of high-value customers through RFM analysis and personalised recommendations.
- Enhanced cross-selling opportunities resulted in an **average basket size growth of 10%**.

**Simulated Scenario:** Based on industry benchmarks, a small retail business generating \$500,000 annually could expect an additional \$75,000 to \$100,000 in revenue after implementing Kolay.ai.

###### 4.4.1.2. Cost Reduction

Kolay.ai's inventory optimisation and financial forecasting capabilities have demonstrated potential for

- **10-15% reduction in inventory holding costs** by minimizing overstock and stockouts.
- **5-8% savings in operational expenses** through efficient resource allocation.

**Hypothetical Case:** A mid-sized manufacturer spending \$200,000 on inventory annually could save \$20,000 to \$30,000 using Kolay.ai's predictive analytics.

#### 4.4.1.3. Operational Efficiency

The adoption of Kolay.ai's cloud-based platform streamlines workflows, leading to

- A **30% decrease in manual data processing time**, allowing employees to focus on strategic tasks.
- Improve decision-making timelines by **reducing forecasting cycles from weeks to days**.

**Industry Application:** In logistics, such efficiencies could translate into faster delivery times and improved customer satisfaction, boosting customer retention rates by **10-15%**.

#### 4.4.1.4. Customer Retention

Kolay.ai's ability to analyse customer behaviour and offer tailored solutions contributes to the following:

- A **12-18% improvement in customer retention rates** due to personalised engagement strategies.

**Example:** A service-based SME could retain an additional 50 customers annually, equating to a significant lifetime value increase.

## 4.5. Robustness and Sensitivity Analysis of Machine Learning Models

Ensuring the reliability and robustness of machine learning (ML) models is critical for the successful implementation of business intelligence (BI) solutions in SMEs. The robustness of Kolay.ai's ML models is assessed through various techniques, including cross-validation, hyperparameter tuning, and sensitivity analysis, which are essential to providing consistent and actionable insights for SMEs.

### 4.5.1. Model Robustness Evaluation

To ensure the effectiveness of the predictive models implemented in Kolay.ai, the following robustness evaluation techniques were employed:

- **Cross-Validation:** A k-fold cross-validation approach (typically k=10) was used to assess the model's generalizability and prevent overfitting. By splitting the dataset into multiple training and testing sets, the model's performance across different subsets of data is validated, ensuring consistency in predictions across varying conditions.
- **Hyperparameter Optimisation:** Models were fine-tuned using grid search and Bayesian optimisation techniques to achieve optimal performance. Key hyperparameters, such as learning rates, regularisation parameters, and model complexity, were adjusted to minimize prediction error and enhance model stability.
- **Error Metrics Evaluation:** The following standard evaluation metrics were utilised, including:
  - **Mean Absolute Percentage Error (MAPE):** For financial projections, providing insights into forecast accuracy relative to actual revenue figures.
  - **F1 Score:** This score measures the balance between precision and recall in customer segmentation tasks.
  - **Root Mean Square Error (RMSE):** This measure evaluates the deviation of sales predictions from actual data trends.

#### 4.5.2. Sensitivity Analysis

Sensitivity analysis was conducted to determine how variations in input parameters affect key financial metrics, thereby strengthening confidence in Kolay.ai's outputs and helping SMEs make data-driven decisions under uncertain conditions. Sensitivity analysis was performed on the following key areas:

- **Revenue Projections:** Sensitivity tests were conducted to evaluate how changes in market conditions (e.g., demand fluctuations, seasonal variations) impact revenue forecasts. The model's responsiveness to input variables, such as customer demand trends, economic indicators, and marketing expenditures was analysed to identify critical dependencies.
- **Customer Segmentation Accuracy:** By varying input features such as purchase frequency, recency, and monetary value, the model's ability to correctly segment customers into high-value, medium-value, and at-risk categories was assessed. Sensitivity tests helped identify the variables most influential in segmentation accuracy and provided guidance on data collection priorities.
- **Inventory Management Optimisation:** The impact of fluctuations in sales forecast accuracy on inventory levels was tested to ensure the model's robustness in maintaining optimal stock levels and preventing overstocking or stockouts.
- **Financial Forecasting Models:** Sensitivity to external financial factors, such as inflation rates, currency fluctuations, and changing cost structures was analysed to understand the robustness of financial planning features.

#### 4.5.3. Results and Implications

The sensitivity analysis revealed that

- Kolay.ai's sales prediction models exhibited a **10-15% variation in forecast accuracy** under different market conditions, demonstrating resilience in predicting trends within a reasonable margin of error.
- The customer segmentation models maintained an **F1 score of 0.85 or higher**, which indicates strong performance in identifying customer behaviours with minimal deviation across multiple test scenarios.
- The inventory optimisation module showed a **20% reduction in stock fluctuations**, even when subjected to demand variability of up to 30%.

These findings highlight the robustness of Kolay.ai's ML models in real-world business applications, ensuring that SMEs can rely on this platform for strategic decision-making.

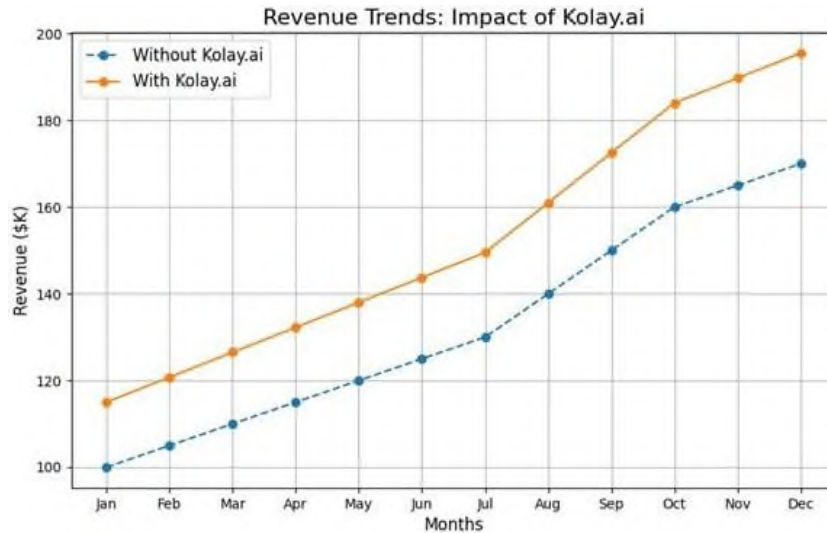
By incorporating rigorous robustness checks and sensitivity analyses, Kolay.ai enhances the reliability and accuracy of its AI-driven BI solutions. The platform's ability to handle variability in financial and operational data makes it a dependable tool for SMEs looking to optimise their processes and financial planning under dynamic market conditions. Future enhancements will focus on expanding the sensitivity tests to include broader economic and geopolitical factors to further strengthen the predictive accuracy.

## 5. Results

The implementation of Kolay.ai in small- and medium-sized enterprises (SMEs) has demonstrated measurable improvements in financial outcomes, operational efficiency, and customer engagement. This section presents the results of the study, emphasising the business impact of Kolay.ai's features through quantitative analysis and visualisations.

### 5.1. Revenue Growth

Kolay.ai’s advanced sales prediction and customer segmentation tools have significantly increased revenue. By providing accurate demand forecasts and enabling targeted marketing strategies, SMEs using Kolay.ai reported an average revenue growth of 15%–20% (Figure 3). This improvement can be attributed to better inventory management, enhanced cross-selling opportunities, and improved customer retention rates.



**Figure 3.** Illustrates monthly revenue trends, highlighting a consistent revenue increase for SMEs using Kolay.ai compared with those operating without the platform.

### 5.2. Customer Segmentation Efficiency

Kolay.ai’s RFM analysis and customer recommendation features have enhanced the efficiency of customer segmentation. These tools allowed businesses to identify high-value customers, retain at-risk customers, and tailor personalised strategies for different segments. The efficiency scores for customer segmentation increased by an average of 20%–30% across all categories (Figure 4).



**Figure 4.** Shows the customer segmentation efficiency before and after the adoption of Kolay.ai, demonstrating substantial improvements, particularly for high-value and at-risk customers



### 5.3. Operational Improvements

Kolay.ai's cloud-based infrastructure and user-friendly interface have streamlined workflows and decision-making processes. SMEs reported a 30% reduction in manual data processing time and significant improvements in forecasting timelines, enabling faster and more informed strategic decisions. Furthermore, the platform's data visualisation tools, such as heatmaps and sales time charts, provide actionable insights, leading to better resource allocation and operational planning.

### 5.4. Financial Optimisation


Kolay.ai's financial forecasting and invoice data analysis capabilities have optimised cash flow management for SMEs. Businesses reported a 10%–15% reduction in inventory holding costs and 5%–8% reduction in operational expenses. These savings are crucial for SMEs operating on tight budgets, enabling them to reinvest in growth and innovation.




---

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

---

Author Details **Rabia Yörük**  
<sup>1</sup> OptiWisdom, Data Science, San Francisco, USA  
 0009-0007-2222-9323

---

## References

- [1] A. Gupta and V. Varma. 2019. Reinforcement Learning for Pricing and Revenue Management in E-commerce. *J. Revenue Pricing Manag.* 18, 1 (2019), 55–62
- [2] Abdel-Rahman Tawil, Mahmoud Mohamed, Xavier Schmoor, Konstantinos Vlachos, and Dima Haidar. 2023. Trends and Challenges Towards an Effective Data-Driven Decision Making in UK SMEs: Case Studies and Lessons Learnt from the Analysis of 85 SMEs. arXiv Preprint (2023). <https://doi.org/10.48550/arXiv.2305.15454>
- [3] Andreja Popovič, Ray Hackney, Paulo S. Coelho, and Jurij Jaklič. 2012. Towards Business Intelligence Systems Success: Effects of Maturity and Culture on Analytical Decision Making. *Decision Support Systems* 54, 1 (2012), 729–739. <https://doi.org/10.1016/j.dss.2012.08.017>
- [4] C. Cortes and V. Vapnik. 1995. Support-vector networks. *Mach. Learn.* 20, 3 (1995), 273–297. <http://doi.org/10.1007/BF0099401>
- [5] C. J. C. H. Watkins and P. Dayan. 1992. Q-learning. *Mach. Learn.* 8, 3–4 (1992), 279–292. <http://doi.org/10.1007/BF00992698>  
<http://doi.org/10.1007/BF0099269>
- [6] H. Chen, R. H. L. Chiang, and V. C. Storey. 2012. Business Intelligence and Analytics: From Big Data to Big Impact. *MIS Q.* 36, 4 (2012), 1165–1188. <https://doi.org/10.2307/4170350>
- [7] Hsinchun Chen, Roger H. L. Chiang, and Veda C. Storey. 2012. Business Intelligence and Analytics: From Big Data to Big Impact. *MIS Quarterly* 36, 4 (2012), 1165–1188. <https://doi.org/10.2307/41703503>
- [8] I. Met, A. Erkok, S. E. Seker, M. A. Erturk, and B. Ulug. 2024. Product Recommendation System With Machine Learning Algorithms for SME Banking. *Int. J. Intell. Syst.* 2024, 1 (2024), 5585575. <https://doi.org/10.1155/2024/55855>
- [9] I. T. Jolliffe. 2002. *Principal Component Analysis*. Springer. <https://doi.org/10.1007/b9883>
- [10] J. A. Hartigan and M. A. Wong. 1979. Algorithm AS 136: A K-means clustering algorithm. *J. Roy. Stat. Soc. C (Appl. Statist.)* 28, 1 (1979), 100–108. <https://doi.org/10.2307/234683>



- [11] J. H. Friedman. 2001. Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Statist.* 29, 5 (2001), 1189–1232. <http://doi.org/10.1214/aos/101320345>
- [12] K. P. Murphy. 2012. *Machine Learning: A Probabilistic Perspective*. MIT Press
- [13] Konstantina Ragazou, Ioannis Passas, Alexandros Garefalakis, and Constantin Zopounidis. 2023. Business Intelligence Model Empowering SMEs to Make Better Decisions and Enhance Their Competitive Advantage. *Discover Analytics* 1, 2 (2023). <https://doi.org/10.1007/s44257-022-00002-3>
- [14] Lucas Griesch, Jonas Rittelmeyer, and Kurt Sandkuhl. 2023. Towards AI as a Service for Small and Medium-Sized Enterprises (SME). In *The Practice of Enterprise Modeling*, 37–53. Springer. [https://doi.org/10.1007/978-3-031-48583-1\\_3](https://doi.org/10.1007/978-3-031-48583-1_3)
- [15] M. Alnoukari and A. Hanano. 2017. Integration of Business Intelligence with Cloud Computing: A Practical Approach. *J. Theor. Appl. Inf. Technol.*, 95, 1 (2017), 63–72. <http://doi.org/10.4018/978-1-7998-5040-3.ch00>
- [16] Markus Schönberger. 2023. Artificial Intelligence for Small and Medium-Sized Enterprises: Identifying Key Applications and Challenges. *Journal of Business Management* 21 (2023). Retrieved from <https://journals.riseba.eu/index.php/jbm/article/view/336>
- [17] Nick Drydak. 2023. Artificial Intelligence and Reduced SMEs' Business Risks: A Dynamic Capabilities Analysis During the COVID-19 Pandemic. *SSRN Electronic Journal* (2023). <https://doi.org/10.2139/ssrn.4114609>
- [18] Quoc Huy Pham and Kieu Phuong Vu. 2023. Big Data in Relation with Business Intelligence Capabilities and E-Commerce During COVID-19 Pandemic in Accountant's Perspective. *Future Business Journal* 9, 40 (2023). <https://doi.org/10.1186/s43093-023-00221-4>
- [19] S. Fosso Wamba, S. Akter, A. Edwards, G. Chopin, and D. Gnanzou. 2017. How 'big data' can make big impact: Findings from a systematic review and a longitudinal case study. *Int. J. Prod. Econ.* 165 (2017), 234–246. <http://doi.org/10.1016/j.ijpe.2014.12.03>
- [20] S. Kumar and A. Ramesh. 2018. Machine Learning in Business: A Conceptual Framework. *J. Bus. Anal.* 1, 1 (2018), 1–17
- [21] S. Soni, M. Sharma, and T. Singh. 2020. Machine Learning for SMEs: Adoption and Benefits. *Int. J. Data Sci. Anal.* 6, 3 (2020), 112–119
- [22] T. Chen and C. Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*, 785–794. <https://doi.org/10.1145/2939672.293978>
- [23] Tânia Guarda, Manuel F. Santos, César Silva, and Rui Lopes. 2013. Business Intelligence for SMEs: A Proposal for an Information System to Improve Small and Medium Enterprises Performance. *Procedia Technology* 9 (2013), 728–733. <https://doi.org/10.1016/j.protcy.2013.12.080>
- [24] Thomas H. Davenport and Jeanne G. Harris. 2007. *Competing on Analytics: The New Science of Winning*. Harvard Business Review Press.







# Journal of Data Analytics and Artificial Intelligence Applications

Research Article

 Open Access

## Analysis of Word Similarities in Tax Laws Using the Word2Vec Method



Ali İhsan Özgür Çilingir<sup>1</sup>  

<sup>1</sup> Non-affiliated, İstanbul, Türkiye

### Abstract

This paper describes word similarity analysis in tax law using the Word2Vec model. By similarity analysis, we mean identifying relationships between similar terms in tax terminology. The Word2Vec model represents the meanings of words with vectors and identifies the semantic relationships of words through the proximity between these vectors.

This article analyzes the semantic proximity of terms frequently used in tax law and visualises the relationships between these words. For example, the close relationships of the word 'mükellef' with words such as 'kişi', 'tam', 'dar', 'firma', and 'imalatçı' are represented through vectors. The paper also explains the mathematical structure of the models. Then, the features of the NumPy, Gensim, Scikit-learn, and Matplotlib libraries of the Python programming language are explained and used for this paper. For the visualisation of the similarity analysis, the t-SNE algorithm, which allows the visualisation of high-dimensional data on a two-dimensional plane, was used.



The main purpose of this paper is to enable AI systems that can be used as tax advisors to better understand tax law by modelling the conceptual relationships between the terms of tax law, thus contributing to the provision of more accurate and consistent information by AI.

### Keywords


Word2Vec · tax law · natural language processing (NLP) · t-SNE algorithm · Skip-Gram Model · language model visualisation.



Citation: Ali İhsan Özgür Çilingir. 2025. Analysis of Word Similarities in Tax Laws Using the Word2Vec Method. *Journal of Data Analytics and Artificial Intelligence Applications* 1, 1 (January 2025), 84-109. <https://doi.org/10.26650/d3ai.003>

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

© 2025. Çilingir, A. İ.

 Corresponding author: Ali İhsan Özgür Çilingir [ihsancilingir@gmail.com](mailto:ihsancilingir@gmail.com)



## 1. INTRODUCTION

Tax laws are complex legal texts that regulate the economic structure of a country and the financial obligations of society. These laws contain various concepts and terms that are regularly updated and adapted to the economic conditions. Not only economic and legal experts but also many professionals in different sectors have to understand and apply these laws. However, due to the dense language of tax laws and the abundance of technical terms, these texts are very difficult to understand and analyse. At this point, word similarity analysis using artificial intelligence techniques is a very important tool for making complex terms more understandable and revealing the relationships between laws.

Word similarity analysis has an important place in the field of natural language processing (NLP). Using models such as Word2Vec, these analyses reveal the relationships between similar concepts by representing words as mathematical vectors. Especially in complex and comprehensive texts, similarity analysis makes it possible to determine how related or close terms are. Word similarity analysis in tax laws can serve as a basis for artificial intelligence research in both law and finance, especially in areas such as concept somatisation, automatic classification, and intertextuality.

An in-depth examination of the relationships between word similarity analysis and tax laws will help to better understand legal regulations. Such AI-supported studies make it possible to create a common understanding between different texts, especially by determining the similarity levels of terms that frequently appear across legal texts. For example, inferences such as how specific terms used in tax laws correspond to terms in other legal texts or which concepts are more related to other concepts can also contribute to the economic interpretation of legal regulations.

The purpose of this study is to identify the relationships and similarities between terms used in tax laws and to provide a broader understanding of the meaning of these terms. Since tax laws contain a strict structure and specific linguistic features, analysing these structures can be considered one of the first steps towards the development of AI-supported solutions. The identification of terms used in the same or a similar sense as a result of this analysis can serve as a guide in the interpretation of tax laws and potentially provide a foundation for user-friendly applications.

To achieve this, a corpus of Turkish tax laws was compiled from publicly available legal repositories and subjected to pre-processing steps, including punctuation removal, word form normalisation, and tokenization. The final dataset consisted of 65,258 tokens, providing a balanced representation of key legal concepts. The Word2Vec model was then trained using the Skip-Gram algorithm, with the vector dimensionality set to 100, the context window size set to 5, and the minimum word frequency threshold set to 5. Dimensionality reduction via t-SNE was applied to visualise semantic relationships by projecting high-dimensional embeddings into a two-dimensional space while preserving both local and global data structures.

In addition to improving the comprehensibility of tax laws, this study provides a valuable example of how word similarity analysis can be used in artificial intelligence and law. In the future, these analyses could lead to innovative solutions such as categorising tax legislation in digital environments, automatically highlighting relevant topics, or enabling users to find the information they are looking for faster. Moreover, AI applications developed through such analytics will provide a basis for the creation of new tools that can guide professionals in the interpretation, understanding, and application of tax laws.

## 2. LITERATURE REVIEW

In their 2013 paper "Efficient Estimation of Word Representations in Vector Space," Google researchers Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean introduced two new model architectures for computing the continuous vector representations of words (CBOW and Skip-Gram Models). While the CBOW model estimates the target word by averaging over the surrounding words, the Skip-Gram model takes a word as input and attempts to estimate the words near it. For example, the models captured semantic similarities with vector operations such as "king - man + woman = queen." Their work also proved that high-dimensional word vectors trained on large datasets such as Google News perform better on many natural language processing tasks [1]. After the publication of the article, many applications were developed, and both local and foreign literature was created on it.

When I decided to write an article on the subject, I first reviewed and benefited from the literature created by IT academics in our country. I can briefly summarise the scope of the studies I benefited from as follows:

In the Master's Thesis titled "Semantic Inference from Turkish Texts Using Deep Learning Approaches" written by Nergis Pervan, the Word2Vec method was used to train the phrases in user comments on social media and e-commerce sites, and the semantic relations of the words in the comments were determined [2].

In the article titled "Turkish Sentiment Analysis Based on Convolutional Neural Network Architectures" written by Aytuğ Onan, sentiment analysis was performed on Turkish texts, and Convolutional Neural Network (CNN) architectures were used. In the article, Word2Vec, FastText, GloVe, and LDA2Vec were used as word embedding techniques, and it was stated that Word2Vec (Skip-Gram model) achieved the highest performance [3].

Murat Tezgider, Beytullah Yıldız, and Galip Aydın's article titled "Improving Word Representation by Tuning Word2Vec Parameters with a Deep Learning Model" aims to improve the classification performance of Turkish texts by tuning Word2Vec parameters with deep learning methods. In this study, different values for parameters such as minimum word count, vector size, and window size were tested for the Word2Vec model. It was observed that the correct choice of these parameters improves the quality of word representation and, therefore, classification success [4].

In the article titled "Similar Sentence Detection Using the Word Embedding Method" written by Mehmet Ali Arabacı, Ersin Esen, Muhammed Selim Atar, Eyüp Yılmaz, and Batuhan Kaltalıoğlu, the Word2Vec model and Fisher coding were combined to detect semantically similar sentences. The method is based on the vectorial representations of the words in the sentence using the Word2Vec model. Then, Fisher coding is applied to create sentence-level vectors. The authors present an effective method that combines Word2Vec and Fisher coding to detect sentence similarity in the Turkish language [5].

Murat Aydoğan and Ali Karıcı's article titled "Analysing Word Similarities with Word Representation Methods" aims to identify word similarities in Turkish texts by examining word representation methods. A large Turkish dataset was created, and word relations were analysed using word vector models such as Word2Vec and GloVe. In this study, the CBOW and Skip-Gram algorithms of the Word2Vec method were compared with those of the GloVe method. The Word2Vec method was found to successfully identify the proximity of words and perform better than the GloVe method [6].

In the research article titled "Classification of Turkish News Texts Using Convolutional Neural Networks and Word2Vec" written by Çiğdem İnan Acı and Adem Çırak, the authors showed that Turkish news texts can

be successfully classified with Convolutional Neural Networks (CNNs) and Word2Vec and emphasised that these methods make an important contribution to Turkish natural language processing studies [7].

I can briefly summarise the studies of foreign informatics academics from which I have benefited as follows:

Lu XiaoID, Qiaoxing Li, Qian Ma, Jiasheng Shen, Yong Yang, and Danyang Li, in the paper titled “Text classification algorithm of tourist attractions subcategories with modified TFIDF and Word2Vec,” investigate an improved text representation method combining TF-IDF and Word2Vec methods and its integration with different classifiers. The aim of this paper is to develop a multi-class classification algorithm by subcategorising tourist attraction description texts and to present a model that provides higher accuracy and stability compared to traditional methods. The authors integrated Word2Vec word embedding methods, TF-IDF (Term Frequency-Inverse Document Frequency), and CRF-POS (Conditional Random Fields) weighting. They collected the descriptions of national A-level tourist attractions in China using web crawler technology and trained the Word2Vec model after pre-processing stages such as word segmentation, grammar tagging, and stop word filtering. Their preferred method was Skip-Gram. Word2Vec is used in this study as a powerful tool in terms of both data representation and classification performance, and they also integrated it with the improved TF-IDF method [8].

Ghislain Wabo Tatchum, Arnel Jacques Nzekon Nzeko, Fritz Sosso Makembe, and Xaviera Youh Djam, in their paper titled “Class-Oriented Text Vectorisation for Text Classification: Case Study of Job Offer Classification,” discuss class-oriented vectorisation approaches in text classification processes and examine how these methods are more effective in classifying job advertisements. In this paper, preprocessing steps such as data cleaning, tokenization, and stem extraction were performed on job postings. Redundant words or do not carry meaningful information were removed. After these processes, they performed vectorisation using different methods. The vectorisation techniques compared in the paper with traditional methods are TF-IDF, Word2Vec, and Doc2Vec. Class-oriented vectorisation strategies include OC (Occurrence Count), ZIPF, and OWDC (Occurrences Weighted by Dispersion in the Class). In this paper, machine learning models (Naive Bayes (NB), Decision Trees (DT), Support Vector Machines (SVM), and Transformer-based deep neural networks (TFM)) were tested with vectorisation methods. The Word2Vec method was used to represent the text data. One of the prominent results of the paper is that the OWDC strategy generally outperformed the other methods. Furthermore, OWDC provides the highest accuracy rates when used in combination with the TFM (Transformer) model [9].

In the article "Discovery of New Words in Tax-related Fields Based on Word Vector Representation" by Wei Wei, Wei Liu, Beibei Zhang, Rafał Scherer, and Robertas Damasevicius, the authors focus on the detection of new words in tax-related financial texts. Based on the Word2Vec model, the similarity measure of word vectors is used to calculate the similarity in meaning between words. According to the results of the study, this method can be used effectively in large-scale datasets, allowing new words to be automatically added to the dictionary. This method, especially for the discovery of tax-specific terms, has been found to improve the performance of traditional word segmentation tools, contributing to the identification of new words with low frequency but rich in meaning [10].

The article "Deep Learning in Law: Early Adaptation and Legal Word Embeddings Trained on Large Corpora" by Ilias Chalkidis and Dimitrios Kampas examines the early adaptations of deep learning in the legal domain, with a particular focus on the generation of phrases from legal texts. The authors examine the applicability of deep learning in areas such as legal text classification, information extraction, and information retrieval, and emphasise the importance of legal word embedding techniques. This paper describes the impact of word

representation in the legal domain with phrases trained on a large legal dataset using the Word2Vec model. This paper provides information about Word2Vec's two main algorithms, Skip-Gram and Continuous Bag of Words (CBOW). The article also highlights how training the Word2Vec model on domain-specific datasets, such as law, improves the model's performance and the accurate capture of semantic relationships between words [11].

In the paper titled "Similarity Analysis of Law Documents Based on Word2Vec" by Chunyu Xia, Tieke He, Wenlong Li, Zemin Qin, and Zhipeng Zou, the authors discuss the use of the Word2Vec model for similarity analysis of legal documents. The different lengths and formats of legal documents create difficulties in similarity analysis. In this context, the authors aimed to perform a more effective similarity analysis by training the Word2Vec model with a specialised dataset of legal documents. Word2Vec learns semantic similarities between words by representing them in a vector space. In the paper, Word2Vec is used to better capture the depth of meaning of words in legal documents. By creating vector representations of sentences and documents, this model allows for more accurate similarity measurements. Skip-Gram tries to predict other words in the context based on a word in the centre. It is especially used to provide more accurate information about rare words. CBOW predicts the centre word based on words in the context and produces more accurate results for more common words. By training the Word2Vec model on legal documents, the authors achieved a 20% higher accuracy than the Bag of Words (BOW) model. It was also observed that the Word2Vec model trained with a dataset specific to legal documents improved the accuracy by 5-10% compared to the model trained with a general dataset. Experiments using methods such as Cosine Similarity and Word Mover's Distance (WMD) have demonstrated the effectiveness of Word2Vec-based similarity analysis for legal documents [8].

In the paper "Unsupervised Approaches for Measuring Textual Similarity Between Legal Court Case Reports" by Arpan Mandal, Kripabandhu Ghosh, Saptarshi Ghosh, and Sekhar Mandal, the authors examine the use of unsupervised methods for measuring similarity between court decisions. Focusing on the effectiveness of text-based methods, this paper explores how natural language processing techniques such as Word2Vec, Skip-Gram, and CBOW can be used for legal documents. In addition to Word2Vec, the authors also used different methods such as Doc2Vec, TF-IDF, LDA, BERT, Law2Vec, and PScoreVect to measure the similarity between court decisions. Some of these methods (e.g., Doc2Vec and Law2Vec) are direct extensions or adaptations of Word2Vec. While other methods (e.g., BERT, LDA, TF-IDF) aim to achieve the same goal as Word2Vec—representing texts numerically and measuring similarity—they exploit Word2Vec's ability to learn semantic relatedness in different ways. According to the authors, Word2Vec is powerful in capturing semantic similarity between words compared to other methods, but for complex sentence structures or contextual details, more advanced models (e.g., BERT) may be preferable [12].

In the paper "Influence of Various Text Embeddings on Clustering Performance in NLP" by Rohan Saha, the author investigates the impact of different text embeddings on clustering performance in natural language processing (NLP). This study compares the performance of different clustering algorithms with text embeddings such as Word2Vec and BERT using Amazon product review data. The main objective was to evaluate the impact of each embedding method and clustering algorithm on a specific task. The paper emphasises that Word2Vec is a model for learning semantic relations between words. Word2Vec's average vector values (average embeddings) were used. The fact that this method does not include contextual information caused limited performance in some tasks. According to the results of the study, contextual BERT embeddings

performed better than Word2Vec in general. However, the performance of the methods differed according to the clustering algorithm [13].

The following information is given in the "Unsupervised Learning (Summer '18)" course note from Columbia University, taught by Ziyuan Zhong and Nakul Verma and authored by Vincent Liu: "t-distributed Stochastic Neighbour Embedding (t-SNE) is a dimensionality reduction technique for visualising high-dimensional data in two- or three-dimensional space. It was developed by Laurens van der Maaten and Geoffrey Hinton in 2008. t-SNE is used in natural language processing to visualise semantic relationships between words by mapping word vectors in low-dimensional space" [14]. The paper also explains the mathematical structure of this algorithm.

In the article "Clustering With T-SNE, Provably" by George C. Linderman and Stefan Steinerberger, t-SNE is described as an optimisation method that minimizes the Kullback-Leibler divergence to cluster high-dimensional data in low-dimensional areas [15].

"An Analysis of the t-SNE Algorithm for Data Visualisation" by Sanjeev Arora and Wei Hu, presented at the Conference on Learning Theory (COLT) 2018, analyzes the use of the t-SNE algorithm for data visualisation. Used to reduce the high-dimensional data to two dimensions, t-SNE visualises and clarifies clusterable data. This paper proves that t-SNE is particularly effective on clusterable datasets with well-separated and global data. The presentation provides the rationale for this success and shows how t-SNE achieves provable success. The authors explain that t-SNE tries to achieve clustering by minimizing the Kullback-Leibler divergence between the similarity vectors of the high-dimensional data and the two-dimensional embedding. This divergence is an optimisation problem that finds the low-dimensional structure that will enable clustering [16].

The article "Visualising Data Using t-SNE" by Laurens van der Maaten and Geoffrey Hinton discusses the t-SNE (t-Distributed Stochastic Neighbour Embedding) algorithm in detail. This paper also describes the Kullback-Leibler Divergence Minimization algorithm. This algorithm compares the similarities between high- and low-dimensional representations with the Kullback-Leibler divergence and minimizes this divergence to keep similar points close and dissimilar points far apart [17].

As mentioned at the beginning, the methods, formulas, and algorithms analysed by the authors mentioned above are used in this paper for our purposes, and we aim to contribute to the literature.

### 3. DIFFERENCES AND CONTRIBUTIONS OF THE ARTICLE FROM THE REVIEWED LITERATURE

*The different aspects of the article compared to the reviewed literature can be summarised as follows:*

First, the focus area and application area of the article are different. Most of the studies in the literature have addressed the application of techniques such as Word2Vec in general language processing or other areas (e.g., e-commerce, social media, sentiment analysis). However, this study focuses on a specific legal context, namely tax laws. Although Chunyu Xia et al. conducted similarity analyses for legal documents, this study has chosen a more specific area by focusing specifically on Turkish Tax Laws.

The articles and studies in the literature used general texts or social media data. In this paper, a dataset derived directly from Turkish Tax Laws ('kanunlarv2.txt') was used, and a corpus was created for this dataset.

In the literature, the Word2Vec model has been used with different methods (CBOW, Skip-Gram), and in some studies, more advanced models such as BERT have been tried. In this paper, the Skip-Gram method is specifically adapted to analyse semantic relations in tax laws.

Some studies in the literature aim at more general purposes (e.g., analysing user behaviour, discovering new words). This study provides a starting point for a practical application, such as AI-assisted tax counselling. In fact, although Word2Vec was used in this article for the classification of tax law concepts, it can also be used in an AI-assisted consulting (ChatBot) tool to handle concepts from the same class or ensure compatibility between questions and answers. We plan to focus on this in future studies.

In addition to the differences, the contributions of the article to the literature can be summarised as follows:

This article analyzes the conceptual relationships in legal texts by creating a Word2Vec model specific to Turkish Tax Laws. This is a contribution to the studies conducted in the literature on legal documents.

This study provides an infrastructure for developing more effective artificial intelligence-based tax advisory systems using the Word2Vec model. This offers an innovative perspective for both financial and legal applications.

The study has made a significant contribution to the lack of literature on Word2Vec applications for the Turkish language and provides an example of how word vectors suitable for Turkish texts can be developed.

This study proposes a methodology for Turkish natural language processing studies, especially the original word cleaning and simplification processes performed during the corpus generation stages.

Word2Vec and t-SNE are standard techniques widely used in natural language processing and dimensionality reduction. However, the use of a dataset specific to Turkish Tax Laws, the application of these techniques in a legal context, and the focus on practical outcomes such as tax consulting show that this paper makes original contributions to the literature.

This paper is a first step not only in the application of these techniques but also in the integration of more advanced models (e.g., BERT, GPT) to capture the depth of meaning in legal texts. I plan to expand on this by including illegal texts in the analysis in further studies.

#### 4. CHALLENGES AND CONSTRAINTS

During the research and writing of the paper, some unique challenges were encountered when applying the Word2Vec and t-SNE techniques to Turkish Tax Laws. These challenges are outlined below.

Tax laws contain technical language, long sentences, and a dense context. This poses the following unique challenges in natural language processing (NLP) applications. For example, in tax laws, terms such as “mükellef (taxpayer),” “ödeme (payment),” and “muafiyet (exemption)” may have different meanings depending on their context. This increases the risk of semantic inaccuracies when creating a vectorial representation of these terms. Legal texts often contain long sentences and nested structures, making it difficult for the model to learn contextual meanings during corpus creation and the training of word vectors.

The specific difficulties of Turkish are also a significant challenge. Since Turkish is an agglutinative language, the root and affix relations of words pose a particular challenge for models like Word2Vec. The agglutinative structure in Turkish causes words such as “mükellef,” “mükellefiyet,” and “mükellefin” to be represented in different forms. This can result in the meanings of words with the same root being represented by different

vectors. To overcome this challenge, a special process was developed to identify word roots and remove unnecessary suffixes during the corpus creation process.

Turkish characters and encoding issues should also be considered a challenge. Turkish characters such as “Ğ,” “İ,” “Ş,” and “Ü” can create technical problems when reading and processing the dataset. The paper suggests that different character encodings (e.g., UTF-8, ISO-8859-9) should be tried to resolve these problems.

In legal texts, there are contextual relationships between concepts that are not explicitly stated. For example, there are indirect concept relationships. Legal terms are often related to concepts that are not explicitly stated but are linked in meaning. For instance, words such as “vergi (tax),” “beyanname (declaration),” and “tahakkuk (accrual)” are closely related in legal processes, but this relationship is implicit in the text. Modelling such conceptual links is a challenge that exceeds Word2Vec’s structure based on direct word relationships.

The same term can mean different things in different laws or contexts. For instance, the word “vergi” may be associated with “beyanname” in one context and with the concept of “ceza (penalty)” in another. Modelling these contextual differences makes the training process of the model more complex.

The paper worked with a specific dataset, such as the Turkish Tax Laws. However, the problems that arose during the organisation and processing of legal texts posed a unique challenge. Tax laws often contain fragmented information contained in different documents. Combining them into a single dataset and creating a coherent corpus is a time-consuming and laborious process. The paper solved this problem by bringing all the laws together in a file named “Kanunlarv2.txt.”

When creating a corpus, conjunctions, pause marks, and word fragments that do not make sense need to be weeded out. This process required both technical and linguistic expertise. This study uses special patterns and filtering methods to extract such words.

The t-SNE algorithm may lose some relationships in the high-dimensional data. For instance, multiple contextual relationships between two terms may not be fully reflected in the low-dimensional plane. However, the model results were satisfactory. Indeed, in the visualisation obtained by applying the t-SNE algorithm, the proximity between the word “mükellef” and words such as “şirket (company)” and “dar (narrow)” can be clearly seen. However, what these relationships mean in a legal context requires legal knowledge and interpretation beyond visualisation. In future studies, the results of models such as Law2Vec, BERT customised for law, Doc2Vec, and GPT can be compared with Word2Vec to explore this issue in more depth.

The dataset of Turkish Tax Laws consists of 65,258 words. It is known in the literature that the Word2Vec model establishes stronger semantic relations when trained with much larger datasets. However, since 65,258 words is considered sufficient for general natural language processing projects, we did not see any harm in building the paper on this. This paper is a starting point in its field, and future work could include not only the law but also the broader tax literature.

The Word2Vec model prioritises frequently used words in the training data. Therefore, frequent terms such as “vergi,” “beyanname,” and “ödeme” may have a stronger representation in the model than other rare terms. This may lead to rare terms or more complex contexts being ignored. As explained in the research methodology section, we took this constraint into account and removed pause words and reduced some words to their roots to avoid omission due to Turkish suffixes.



## 5. MATHEMATICAL STRUCTURE OF THE WORD2VEC MODEL

### 5.1. General Information

The Word2Vec model represents words in vector space by using statistical relationships while building word vectors. The main purpose of creating word vectors is to represent each word with a vector of fixed size. These vectors are used to detect the relationships of a word with other words. Initially, each word is represented by an arbitrary vector. For example, each word in the dictionary is assigned a vector of a certain size. The model learns using the "distributional hypothesis," which assumes that similar words appear in similar contexts. During the training process, these vectors are optimised, and the semantic relationships between words are reflected in the optimised vectors. The model learns which words should appear in different contexts and assimilates the internal structure of the language. Both versions are suitable for most applications. Mueller and Massaron stated that the Skip-Gram version is better at representing rare words [18].

The Word2Vec model, developed by Tomas Mikolov and his team, includes two separate models: CBOW and Skip-Gram. "The Continuous Bag of Words (CBOW) model learns word vectors in the projection layer and predicts the central word using words in the context. This architecture predicts the central word based on other words in the context. The input layer creates a projection of the surrounding words and uses a weighted output layer to predict the central word based on this projection. In the "Continuous Skip-Gram Model," for each word, the surrounding words are predicted [1]. In this paper, Tomas Mikolov and his team aim to produce high-quality word vectors that best represent the semantic and syntactic similarities of words. The model converts related words into vectors through mathematical processing and thus detects the similarity between, for example, "king" and "queen" and, as we exemplify in this paper, "mükellef" ile "kişi," "tam," "dar," "firma," "imalatçı," etc.

Representations (words) whose semantic proximity is made through word vectors (word embeddings) are used as neural network inputs. This makes it possible to express the meanings of language in numerical form. As a result of this process, words close in meaning appear as similar vectors [19], [20].

### 5.2. CBOW model

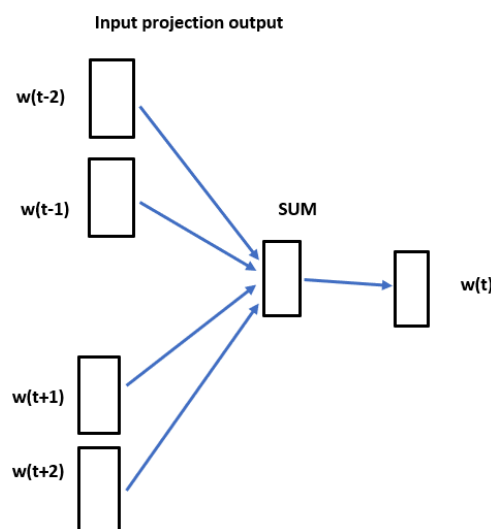


Figure 1. CBOW Model Image

**The meaning and mathematical model of this notation can be summarised as follows:** Figure 1 provides a simple visualisation of the working principle of the CBOW model. In the CBOW model, for example, the target word  $w(t)$  and the words in the context (e.g.  $w(t-2)$ ,  $w(t-1)$ ,  $w(t+1)$ ,  $w(t+2)$ ) are taken as input. The vectors of these words are summed in a projection layer to form an average vector. This average vector is then used to predict the target word  $w(t)$  at the output.

### 5.2.1. CBOW Stages

#### 5.2.1.1. Collection of Contextual Words

$$v_{\text{context}} = \frac{1}{2m} \sum_{j=-m}^m v_{w+j}, j \neq 0 \quad (1)$$

- $v_{\text{context}}$  : Context vector. It represents the environment (context) in which the word appears. This vector is calculated based on the other words in the context.
- $\frac{1}{2m}$ : Normalisation factor.  $2m$  represents the total number of words in the context window, consisting of  $m$  words to the left and  $m$  words to the right. When averaging, the total vectors are divided by this number.
- $\sum_{j=-m}^m$  : refers to the summation. It sums the vectors of all words to the left ( $-m$ ) and to the right ( $+m$ ), excluding the center word itself ( $j = 0$ ) or when ( $j \neq 0$ ).
- $v_{w+j}$  : the vector of a context word at position  $j$  (either to the right or left) of  $w$ . For example,  $j = -1$  represents the vector of the word preceding  $w$ , and  $j = +1$  represents the vector of the word following  $w$ .
- $j \neq 0$ : This condition ensures that the centre word itself ( $j = 0$ ) is excluded from the summation. In other words, only the surrounding words contribute to the context vector.

#### 5.2.1.2. Predicting the Target Word

The context vector ( $v_{\text{context}}$ ) calculated in the first stage is used to predict the target word. This calculation is done with the Softmax function<sup>1</sup>.

$$P(w_t / \text{context}) = \frac{\exp(v_{\text{context}}^T v_{w_t})}{\sum_{w \in W} \exp(v_{\text{context}}^T v_w)} \quad (2)$$

In this formula:

- $P(w_t / \text{context})$  : Represents the probability that the word  $w_t$  occurs in a given context (context). This measures the proximity of the context to the word  $w_t$ .
- $v_{\text{context}}$  : The context vector. It is a vector obtained by combining the vector representations of the words that constitute the context.
- $w_{w_t}$  : The vector representation of the target word ( $w_t$ ). This vector represents the meaning or features of the word in numerical form.
- $v_{\text{context}}^T v_{w_t}$  : The dot product of the context vector and the target word vector.
- $T : v_{\text{context}}^T$  : refers to the transpose of the context vector.
- $\exp(v_{\text{context}}^T v_{w_t})$  The exponential function (exp) is applied to the dot product. The exponential function amplifies the effect of large values and reduces the effect of small and negative values. Thus, the “relationship” between the context and the word becomes more pronounced.

<sup>1</sup>The Softmax activation function is a generalisation of logistic regression that can be applied to continuous data instead of binary classification. It is often included in the output layer of a classifier because it produces output for more than two classes [36]



- $W$  : Vocabulary. It is the set of all words on which the model operates.
- $\sum_{w \in W} \exp(v_{\text{context}}^T v_w)$  This denominator expression represents the sum of the exponential values for all words in the vocabulary ( $W$ ) in relation to the context. This is the normalisation step of the softmax function. The exponential value of each word associated with the context is summed, and the result is normalised so that the total probability of all words' association with the context equals 1.

### 5.2.1.3. Updating Vectors

The loss function is minimized to increase the probability of correctly predicting the target word. For example, a negative logarithmic loss function is used:

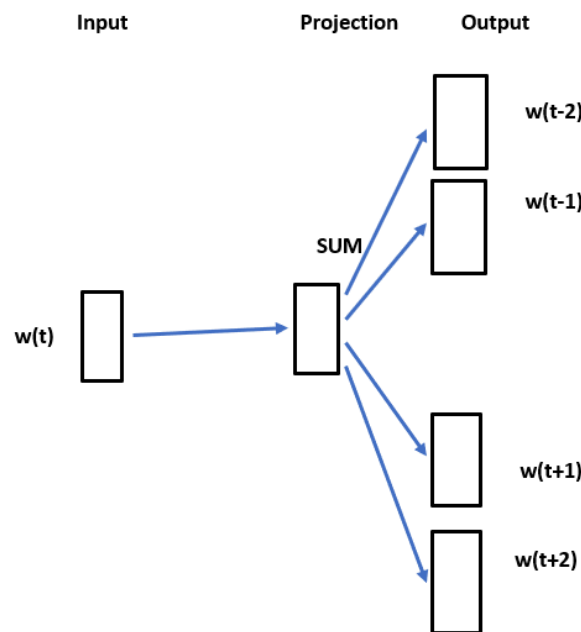
$$J = -\log P(w_t / \text{context}) \tag{3}$$

$J$ : The symbol for the loss function. It measures the performance of the model. The goal of the CBOW model is to match the target word ( $w_t$ ) with the context by minimizing this loss. The lower the value of  $J$ , the more successful the model is in predicting the correct word from the context.

$P(w_t / \text{context})$ : The probability that the target word ( $w_t$ ) will occur given the context. A low value of  $P(w_t / \text{context})$  results in a large loss ( $J$ ), while a high value of  $P(w_t / \text{context})$  results in a small loss, indicating that the model is making better predictions.

This loss function measures the relevance of the target word to the context in the CBOW model and is used to optimise the relationship between the context and the target word. Through backpropagation, the target and context vectors are updated to minimize  $J$ . This process enables the model to learn relationships within the language.

### 5.3. Skip-Gram Model



**Figure 2.** Skip-Gam Model Image

The Skip-Gram model represents words as vectors and estimates the proximity of one word to another using the dot product of their vectors. The CBOW model predicts a word based on its surrounding words. These two algorithms together create a model that represents each word as a vector. The mathematical foundation of these vectors is that the proximity of words in the vector space reflects their semantic similarity [19].

**The meaning and mathematical model of this notation can be summarised as follows:**

Figure 2 illustrates the working principle of the Skip-Gram model. In this model, the centre word  $w(t)$  is taken as input, and based on this word, the surrounding words  $w(t-2)$ ,  $w(t-1)$ ,  $w(t+1)$ , and  $w(t+2)$  within a given context are predicted. The main purpose of the Skip-Gram model is to predict the surrounding context words from the given centre word. This model performs particularly well with large datasets and can also produce good results for rare words. It establishes matches between target words and context words: the target words are the input, and the context words are the output.

The Skip-Gram model is a shallow neural network consisting of an input layer, an embedding layer, and an output layer. The goal of the model is to generate an output probability distribution vector given a target word input. This probability distribution vector (which sums to 1) reflects the likelihood of each word appearing in the context window of the target word. The probability is high for words that share the same context and low for words that do not. Once trained, the model only requires its weights.

To obtain useful vector embeddings, the initially random weights in the model need to be optimised. This optimisation process is carried out to minimize the loss function.

The loss function and its description are provided below :

$$J = \sum_{t=1}^T \sum_{-m \leq j \leq m} \log(P(w_{t+j}/w_t)) \tag{4}$$

J: The Loss Function.

T: The length of the text.

m: The window size.

$P(w_{t+j}/w_t)$  : The probability of obtaining the context word given the target word.

This equation represents a nested loop, where you iterate through all (target word, context word) pairs and sum their probabilities. The minus sign is used as part of the machine learning process to minimize the value of the loss function.

**Calculation of probabilities:** To calculate the probability distribution, the Softmax function is used, which considers the dot product of the target embedding vector and the embedding vectors of each word in the vocabulary.

The dot product of the two vectors can be expressed as

$$u^T v = u \cdot v = \sum_{i=1}^n u_i v_i$$

**The function that provides the probability distribution can be expressed as follows:**

$$p(\text{context word} / \text{target word}) = \exp(u_{\text{target}}^T v_{\text{context}}) / \sum_{w=1}^{\text{Word}} (\exp(u_{\text{target}}^T v_{\text{context}})) \tag{5}$$

In this formula

**p(context word/target word)**: Represents the probability of observing a "context word" given a "target word." This means that the model attempts to predict which words are likely to appear around a target word.

$\exp(u_{\text{target}}^T v_{\text{context}})$  : Here, an inner product is computed, representing the relationship between the target word and the context word.  $u_{\text{target}}$  and  $v_{\text{context}}$  are vectors of a given size for the target and context words. The inner product of these vectors is calculated, and the result is passed through an exponential



function. The exponential function amplifies the closeness between words in the model, assigning higher probabilities to closer words.

**Division Operation (/):** This quotient is used to normalise the probability, ensuring that the total probability of all words is equal to 1. In the denominator, a similar calculation is performed for all words (W) in the vocabulary, and the result is used to normalise the probability.

Total  $(\sum_{w=1}^{Word} (exp(u_{target}^T v_{context}))$ ) This summation normalises the calculation for every word in the vocabulary. It provides a probability distribution of the context word given the target word across all possible context words.

## 6. AUXILIARY ALGORITHMS

### 6.1. t-SNE Algorithm

t-SNE (t-distributed Stochastic Neighbour Embedding) is a dimensionality reduction technique used to reduce high-dimensional data to a low-dimensional space, particularly for visualising complex structures in datasets. It was developed in 2008 by Laurens van der Maaten and Geoffrey Hinton. The primary purpose of t-SNE is to project data into a lower-dimensional space (usually 2 or 3 dimensions) while preserving the similarities in the high-dimensional data. This transformation allows the data to be represented as graphs or visuals that are easier for humans to interpret [17].

**The mathematical steps of the algorithm are described as follows [14], [15], [16]:**

In the first step, the similarities between two data points in high-dimensional space are calculated. For each data point  $x_i$  and  $x_j$  in the high-dimensional space, the probability that  $x_j$  is a neighbour of  $x_i$  is calculated using a Gaussian distribution (normal distribution). The probability is defined as follows:

$$P_{j/i} = \frac{\exp(-\|x_i - x_j\|^2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-\|x_i - x_k\|^2 / \sigma_i^2)} \quad (6)$$

In this formula:

- $\|x_i - x_j\|^2$  : Square of the Euclidean distance between  $x_i$  and  $x_j$ .
- $\sigma_i$  : The bandwidth parameter is selected depending on the data point  $x_i$ .
- $P_{j/i}$  : The probability that  $x_j$  is a neighbour of  $x_i$ .

These probabilities are symmetrised for each data point as follows:

$$P_{ij} = (P_{j/i} + P_{i/j}) / 2N$$

Where N is the total number of data points in the dataset. This symmetric form ensures that the relationships between the two data points are equalised.

**Similarity in Low-Dimensional Space (with t-distribution):** The t-distribution is used to transfer these similarities from high-dimensional space to low-dimensional space. The similarity between the two low-dimensional point  $y_i$  and  $y_j$  is calculated as follows:

$$q_{ij} = (1 + \|y_i - y_j\|^2)^{-1} / (\sum_{k \neq i} (1 + \|y_k - y_j\|^2)^{-1})$$

In this formula:

$\|y_i - y_j\|^2$ : Square of the Euclidean distance between  $y_i$  and  $y_j$ .

$q_{ij}$ : The probability of similarity between  $y_i$  and  $y_j$  in low-dimensional space.



Because the t-distribution has wider tails, it better distinguishes distances between distant points and more effectively reflects the structure between clusters.

## 6.2. Kullback-Leibler Divergence Algorithm

The Kullback-Leibler Divergence (KL Divergence) is an information-theoretic metric used to measure the difference between two probability distributions. Specifically, it helps us understand how "far" one probability distribution is from another. KL Divergence typically measures the difference in information between a reference distribution (true distribution) and a predicted distribution (approximation distribution).

### **The Mathematical Formula for the Kullback-Leibler Divergence:**

Let  $P(x)$  and  $Q(x)$  be two probability distributions. The KL divergence is defined as follows:

$$DKL(P \parallel Q) = \sum_x P(x) \log(P(x)/Q(x)) \quad (7)$$

In this formula:

$P(x)$  : The true distribution or reference distribution (e.g. distribution derived from data).

$Q(x)$  : The approximation distribution or model distribution,

$DKL(P \parallel Q)$  : KL Divergence result.

KL Divergence is used in machine learning to measure the difference between a model's predicted distribution and the actual distribution. In the field of natural language processing, it is employed to evaluate how well the estimated distributions of language models align with the actual data (Bissiri & Walker, 2012, pp. 1139-1160).

## 7. PYTHON LIBRARIES WE USE

### **NumPy:**

NumPy is a fundamental package for scientific computing in Python. It provides a multidimensional array object, various derived objects (e.g., masked arrays and matrices), and numerous routines for fast operations on arrays, including mathematical, logical, shape processing, sorting, selecting, I/O, discrete Fourier transforms, basic linear algebra, basic statistical operations, and random simulation.. In the Word2Vec implementation, NumPy was chosen for its performance optimisation, ease of use, and ability to provide mathematical tools that support natural language processing. It offers a significant speed and efficiency advantage over performing the same operations in pure Python.

### **Gensim:**

Gensim is a Python library for topic modelling, document indexing, and similarity retrieval with large corpora. It is primarily designed for the natural language processing (NLP) and information retrieval (IR) communities (<https://pypi.org/project/gensim/>, 2024). Gensim is an essential tool for training and implementing the Word2Vec model to quickly and easily generate vectors without going into complex mathematical operations and data preprocessing details. It is simple to use for model training and querying. For example, obtaining the vector of a word or finding similar words is possible with just a few lines of code.

### **Scikit-learn:**

Scikit-learn is an open-source and powerful Python library for machine learning and data analysis. It enables the easy implementation of statistical modelling, data preprocessing, and supervised and unsupervised

learning algorithms. While Scikit-learn is not directly used in Word2Vec projects, it can serve as a complementary tool. For example, Scikit-learn's tools such as CountVectorizer or TfidfVectorizer can be used to clean and tokenise text data and convert textual labels (LabelEncoder) into numerical data. Additionally, we used the t-SNE (t-Distributed Stochastic Neighbour Embedding) algorithm from Scikit-learn to visualise the similarities between words.

### **Matplotlib:**

Matplotlib is a comprehensive library for creating static, animated, and interactive visualisations in Python. The Word2Vec project serves as a powerful tool for visualising word embedding vectors. Since the vectors generated by Word2Vec are often multidimensional, visualisation plays a crucial role in analysing and interpreting these vectors. Matplotlib enables us to represent semantic similarities between words by clustering words with similar meanings in the same graph. Additionally, it supports the implementation of dimension reduction algorithms such as t-SNE and optimising the cost function, Kullback-Leibler Divergence, during the reduction of high-dimensional vectors into 2D or 3D space. These features were the primary reasons for using this library in our research.

## 8. RESEARCH METHODOLOGY

### 8.1. Generation of the Tax Law Dataset

The dataset comprises primary and secondary tax law texts. It was retrieved from <https://www.gib.gov.tr/gibmevzuat> and is stored on our computer hard drive in .txt format under the file name "Kanunlarv2.txt". As of 15.10.2024, the file includes the following laws: Tax Procedure Law (Vergi Usul Kanunu), Income Tax Law (Gelir Vergisi Kanunu), Corporate Tax Law (Kurumlar Vergisi Kanunu), Value Added Tax Law (Katma Değer Vergisi Kanunu), Stamp Duty Law (Damga Vergisi Kanunu), Motor Vehicles Tax Law (Motorlu Taşıtlar Vergisi Kanunu), Law on Collection of Public Receivables (Amme Alacaklarının Tahsili Hakkında Kanun), Expense Tax Law (Gider Vergileri Kanunu), Law on Valuable Papers (Değerli Kağıtlar Kanunu), Law on Real Estate Tax (Emlak Vergisi Kanunu), and the Law on Municipal Revenues (Belediye Gelirleri Kanunu). The dataset consists of 65,258 words and word fragments.

### 8.2. Downloading Related Python Libraries

As explained above, we downloaded the following Python libraries: NumPy, for scientific computing; Gensim, which includes the Word2Vec formulation; Scikit-learn, which provides the t-SNE algorithm; and Matplotlib, for creating graphs of similar word vectors by reducing their dimensionality.

### 8.3. Opening and Reading the Kanunlarv2.txt File with Turkish Character Encoding

The Turkish characters in the words of the Turkish Tax Law in the Kanunlarv2.txt file were read using different character encodings, including "utf-8," "ISO-8859-9," "windows-1254," and "ISO-8859-1." The most appropriate encoding was selected, and the file was opened, read, and its contents printed. To accomplish this, a loop was created to try each encoding in turn. Once the correct encoding was identified, the loop was terminated, and the file was opened in the read mode.

#### **A small portion of the output is shown below:**

*"Kanun: 213 - VERGİ USUL KANUNU Yeni Pencerede Aç Yazdır GİRİŞ Kanunun şümulü Madde 1 Bu kanun hükümleri ikinci maddede yazılı olanlar dışında, genel bütçeye giren vergi, resim ve harçlar ile il özel idareler-*

ine ve belediyelere ait vergi, resim ve harçlar hakkında uygulanır. Yukarıda yazılı vergi, resim ve harçlara bağlı olan vergi, resim ve zamlar da bu kanuna tabidir. Bu kanunun hükümleri kaldırılan vergi, resim ve harçlar hakkında da uygulanır. Gümrük ve tekel vergileri Madde 2 (Değişik: 23/1/2008-5728/271 md.) Gümrük idareleri tarafından alınan vergi ve resimler bu Kanuna tabi değildir. Bu vergi ve resimlerle ilgili olarak 27/10/1999 tarihli ve 4458 sayılı Gümrük Kanununun 242 nci maddesi hükümleri uygulanır. Vergi Kanunlarının Uygulanması ve İspat: Madde 3 (Değişik: 30/12/1980 - 2365/1 md.) A) Vergi kanunlarının uygulanması: Bu Kanunda kullanılan "Vergi Kanunu" tabiri işbu Kanun ile bu Kanun hükümlerine tabi vergi, resim ve harç kanunlarını ifade eder. Vergi kanunları lafzı ve ruhu ile hüküm ifade eder. Lafzın açık olmadığı hallerde vergi kanunlarının hükümleri, konuluşundaki maksat, hükümlerin kanunun yapısındaki yeri ve diğer maddelerle olan bağlantısı gözönünde tutularak uygulanır. B) İspat: Vergilendirmede vergiyi doğuran olay ve bu olaya ilişkin muamelelerin gerçek mahiyeti esastır. Vergiyi doğuran olay ve bu olaya ilişkin muamelelerin gerçek mahiyeti yemin hariç her türlü delille ispatlanabilir. Şu kadar ki, vergiyi doğuran olayla ilgisi tabii ve açık bulunmayan şahit ifadesi ispatlama vasıtası olarak kullanılamaz. İktisadi, ticari ve teknik icaplara uymayan veya olayın özelliğine göre normal ve mutad olmayan bir durumun iddia olunması halinde ispat külfeti bunu iddia eden tarafa aittir. BİRİNCİ KİTAP Vergilendirme BİRİNCİ KISIM Genel esaslar BİRİNCİ BÖLÜM Vergi uygulanmasında yetki Vergi dairesi Madde 4 Vergi dairesi mükellefi tesbit eden, vergi tarh eden, tahakkuk ettiren ve tahsil eden dairedir....”

#### 8.4. Listing the Frequencies (Raw Frequencies) of the Words in the Text

To prepare the study, we listed the word frequencies in the “kanunlarv2.txt” file using Python.

**Table 1.** Frequency Table of the Top Twenty Words

	Word	Frequency
0	ve	9320
1	sayılı	3963
2	bu	3895
3	ile	3272
4	veya	3208
5	kanunun	3045
6	madde	2526
7	vergi	2368
8	maddesiyle	2181
9	yürürlük	1735
10	için	1639
11	bir	1611
12	vergisi	1341
13	göre	1282
14	önceki	1227
15	olarak	1200
16	değişen	1083
17	kadar	1082
18	1	1020
19	edilen	1013



As previously mentioned, there are a total of 65,258 words and word fragments in the "kanunlarv2.txt" file. In the frequency table above, as shown in [Table 1](#), the first 20 entries account for 48,011 occurrences, and 3,709 of these are variations of the word "vergi," forming meaningful expressions like "vergisi." The remaining entries consist of non-conceptual words, conjunctions, or word fragments. Therefore, this text needs to be analysed for semantic relationships using machine learning and deep learning algorithms. It is necessary to filter out word fragments and meaningless words to create a meaningful subject for analysis.

### 8.5. Create a Corpus (Collection of Meaningful Words)

At this stage, the goal is to open the text file kanunlarv2.txt using Python code, read it with specific character encodings, and clean the text to create a corpus. The stages of the process are as follows:

- The text is split into lines using `text.split('\n')`, and these lines are stored in a list called `t_list`. Each line is added to the list as a new item.
- An empty list is created with the code `corpus = []`. This list contains the cleaned words from the processed text.
- A pattern named `pattern` is created. This pattern includes various punctuation marks (.,!?:;...""\`(){}[]-<>|/ @#\$%^&\* \_+=~) and digits (1234567890) that may appear in the text. These characters and numbers have been removed from the text.
- A list named `Delete_words` is created. This list contains words that may be present in the text and need to be deleted. Both uppercase and lowercase variations of these words are included. For example, "birinci," "BİRİNCİ," "İlgili," Roman numerals, and some unwanted characters (e.g., 'x96') are included. Meaningless suffixes such as 'sine, 'ine, aa, 'una are also removed as they could interfere with frequency calculations and distort vector calculations.
- The code snippet `re.sub(pattern, "", cumle)`, is used to remove the punctuation marks and digits from each sentence.
- The snippet `re.sub(r'\b' + kelime + r'\b', "", temizlenmis_cumle, flags=re.IGNORECASE)` is used to remove each word in the `Delete_words` list from the text in a case-insensitive manner.
- In the corpus, certain words were transformed into root forms or meaningful common words. For example, "kurumları" was replaced with "kurum," and "cezanın" was replaced with "ceza." A total of 2,490 words underwent this process. This process, referred to as meaningful simplification, is based on over 25 years of expertise in tax law, more than 11 years as a doctor of tax law, and our understanding of corpus creation.
- Using the `split()` code snippet, the cleaned sentences were split into individual words and added to the corpus list.
- Finally, the first 50 items of the corpus (each item being a string of words) were printed on the screen using `print(corpus #r[50])`.

The code was successfully executed, and the output was obtained, as shown in [Figure 1](#).

```

[['Kanun', 'vergi', 'USUL'], ['Yeni', 'Pencerede', 'Aç'], ['Yazdır'], [], [], ['GİRİŞ'], [],
[], [], ['şümü'lü'], [], ['kanun', 'bütçeye', 'giren', 'vergi', 'resim', 'harç', 'il', 'özel',
'idarelerine', 'belediyelere', 'vergi', 'resim', 'harç'], [], ['Yukarıda', 'vergi', 'resim',
'harç', 'vergi', 'resim', 'zam', 'kanuna', 'tabidir'], [], ['kaldırılan', 'vergi', 'resim',
'harç'], [], [], ['Gümrük', 'tekel', 'vergi'], [], ['Gümrük', 'idareleri', 'alınan', 'vergi',
'resimler', 'Kanuna', 'değildir', 'vergi', 'resimlerle', 'Gümrük'], [], ['vergi', 'Kanunlarının',
'Uygulanması', 'İspat'], [], ['vergi', 'kanunlarının', 'uygulanması', 'Kanunda', 'kullanılan',
'vergi', 'tabiri', 'işbu', 'Kanun', 'Kanun', 'vergi', 'resim', 'harç', 'kanunlarını', 'ifade'],
[], ['vergi', 'kanunları', 'lafzı', 'ruhu', 'ifade', 'Lafzın', 'açık', 'olmadığı', 'hallerde',
'vergi', 'kanunlarının', 'konuluşundaki', 'maksat', 'hüküm', 'yapısındaki', 'yeri', 'maddelerle',
'bağlantısı', 'gözönünde', 'tutularak'], [], ['İspat', 'vergi', 'vergiyi', 'doğuran', 'olay',
'olaya', 'muamelelerin', 'gerçek', 'mahiyeti', 'esastır'], [], ['Vergiyi', 'doğuran', 'olay',
'olaya', 'muamelelerin', 'gerçek', 'mahiyeti', 'yemin', 'tür'lü', 'delille', 'ispatlanabilir',
'vergiyi', 'doğuran', 'olayla', 'ilgisi', 'tabii', 'açık', 'bulunmayan', 'şahit', 'ifadesi',
'ispatlama', 'vasıtası', 'kullanılamaz'], [], ['İktisadi', 'ticari', 'teknik', 'icaplara',
'uymayan', 'olayın', 'özellğine', 'normal', 'mutad', 'durumun', 'iddia', 'olunması', 'ispat',
'külfeti', 'bunu', 'iddia', 'tarafa', 'aittir'], [], [], ['KİTAP'], ['Vergilendirme'], [], [],
['esaslar'], [], [], ['vergi', 'uygulanmasında', 'yetki'], [], [], [], ['vergi', 'dairesi'], [],
['vergi', 'dairesi', 'mükellef', 'tesbit', 'vergi', 'tarh', 'tahakkuk', 'ettiren', 'tahsilat',
'dairedir'], []]

```

**Figure 1.** A Sample of the Output

## 8.6. Listing of the Corpus Frequencies

After performing the necessary cleaning and corrections in the text, the frequency report of the top 20 words was generated, as shown in Table 2.

**Table 2.** Frequency Table of the Top Twenty Words

	<b>Word</b>	<b>Frequency</b>
0	vergi	4603
1	değişme	1083
2	gelir	1007
3	ödeme	997
4	kurum	997
5	oran	844
6	değer	748
7	geçici	642
8	mükellef	608
9	hesap	545
10	işletme	529
11	ceza	520
12	mal	498
13	kazanç	493
14	tahsilat	485
15	özel	450
16	sermaye	437

	<b>Word</b>	<b>Frequency</b>
17	indirim	423
18	hizmet	422
19	beyan	406

When this list is examined, it is observed that the words and symbols in the “kanunlarv2.txt” file, which are not suitable for meaningful similarity analysis (e.g., “ve”, “bu”, “veya”, “sayılı”, “için”, “1”), have been eliminated, making the text much more suitable for creating a similarity model.

### 8.7. Creating Word2Vec (Model)

At this stage, the necessary steps can be taken to create a similarity model:

To build a similarity model, it is essential to train the Word2Vec model on the words in the Tax Law. The goal of the training is to teach the model the meanings of the words in the corpus. For this purpose, 100-dimensional vectors were created using the Skip-gram algorithm, and 5 words<sup>2</sup> were considered within the context of a target word. The model only included words that appeared at least 5 times.<sup>3</sup> Once trained, the model generated vector representations of the words, enabling a better understanding of the relationships and similarities between the words.

**We can summarise this explanation in terms of the process stages as follows:**

- The corpus consists of cleaned sentences converted into a list of words. The model will learn the relationships between words in this corpus.
- Each word is represented by a 100-dimensional vector. Higher vector sizes allow the model to capture more detailed information but require greater computational power.
- The window size determines how many words around a target word will be used as the context. Here, the window size is set to 5, meaning the model considers up to 5 words to the left and right of a target word for learning.
- A word must appear at least 5 times to be included in the model, while less frequent words are ignored. This prevents rare words from affecting the model.
- It is specified whether the model will use the Skip-gram (sg=1) or CBOW (sg=0) algorithm. In this case, the Skip-gram algorithm is used with the sg=1 option. Skip-gram predicts other words in the context of a target word and performs better, especially on small datasets.

### 8.8. Vectorial Representation of the Word

This stage involves reporting the vectorial representation of the words. In other words, it is the stage where the trained model's output is retrieved from the selected word vectors trained in the Word2Vec model, allowing us to understand how the model represents the words. The word chosen for this analysis is 'mükellef', which is one of the fundamental concepts in tax terminology.

**Let us explain this in detail:**

<sup>2</sup>Very rare words are often misspellings, conjunctions, or trivial terms. Excluding words with a frequency lower than 5 enables the model to learn in a more meaningful way. Wider windows (e.g., >10) may associate unnecessary words when learning context relationships, while narrower windows (e.g., <3) fail to capture sufficient depth of meaning.

<sup>3</sup>Using too high a frequency threshold (e.g., >10) may exclude underutilised but significant legal terms. For example, even if a legal term like "obligation" is mentioned infrequently, it is crucial for the analysis.

As mentioned at the beginning of this article, the Word2Vec model represents each word as a vector of a certain size. These vectors are trained to capture the semantic and contextual similarities between words. For example, words with similar meanings are represented by vectors that are mathematically close to each other. In this context, the following operations were performed to analyse the similarity of the word 'mükellef'.

In the Word2Vec model, the area containing the vectors created for words after training is referred to as `model.wv` in the code. The `model.wv` object stores the vectors corresponding to the words. For example:

- `model.wv['kelime']` returns the vector for that word.
- `model.wv['mükellef']` returns the vector that the model assigns to the word 'mükellef'.

This vector is an array of numbers that mathematically represents the meaning of the word 'mükellef' based on the context the model has learned. Since we set `vector_size=100`, the word 'mükellef' is represented by a vector with 100 elements. This vector encodes the contextual relationships of the word 'mükellef' with other words. For example, 'mükellef' has vectors similar to words with related meanings, such as 'tam' and 'dar'. These vectors can be used to

- Find similar words.
- Measure word similarity.
- Create word clusters.

The output of the word vector appears in Figure 2 (100 items).

```
array([-6.23710275e-01, -4.63944301e-02,  4.93801236e-01,  5.19869626e-01,
        2.64085293e-01, -4.01404612e-02,  4.53982145e-01,  4.04999465e-01,
       -2.52363645e-02, -2.29218736e-01,  2.70977706e-01, -6.21370435e-01,
        4.42850173e-01,  1.88748702e-01,  4.23074663e-01, -2.48641014e-01,
       -3.55076268e-02, -1.82925805e-01,  1.52746215e-01, -3.49068701e-01,
        2.41844654e-02,  3.93145740e-01,  3.52574557e-01,  5.11918101e-04,
        1.96201742e-01,  7.41723031e-02, -2.12855414e-01, -4.91001159e-02,
       -5.81765920e-03, -4.18194771e-01,  8.73388574e-02,  1.11241512e-01,
        2.79387623e-01,  2.63844766e-02,  9.17030573e-02,  5.33403814e-01,
        2.66016543e-01, -1.37306616e-01, -7.47924373e-02, -4.67420578e-01,
       -3.28616440e-01,  1.10134447e-03, -1.01418853e-01, -3.93146545e-01,
        8.64220336e-02,  1.79907549e-02, -4.13307697e-01, -1.48787230e-01,
        7.44897947e-02,  1.75265461e-01,  3.55105400e-02, -2.73291394e-02,
       -1.50529653e-01, -4.85246740e-02,  1.54171020e-01, -1.68570891e-01,
        3.48834060e-02, -9.14018080e-02, -4.10086721e-01, -7.69241899e-02,
       -1.77023351e-01, -1.62242562e-01,  5.14853597e-01, -1.14671946e-01,
       -5.61786294e-01,  5.57065085e-02,  8.95505175e-02,  3.13859880e-01,
       -3.32743734e-01,  3.15872729e-01, -2.60060638e-01,  1.40525192e-01,
       -1.07785322e-01,  1.95944458e-01, -9.53617766e-02, -1.41765758e-01,
       -1.98408827e-01, -2.68165112e-01, -3.73588473e-01,  1.27587229e-01,
       -1.75960913e-01,  1.62031755e-01,  2.75125414e-01,  6.89642057e-02,
       -4.20767553e-02, -1.14048988e-01,  1.91145271e-01,  2.70561635e-01,
        3.33638400e-01,  4.17892754e-01,  2.03026727e-01, -4.98898625e-02,
        2.53186584e-01, -1.00524463e-01,  6.04738295e-01,  4.28849995e-01,
        2.33635247e-01, -8.20747167e-02, -1.21545447e-02,  3.45590383e-01],
      dtype=float32)
```

**Figure 2.** Representation of the 100-Element Output of the 'mükellef' Vector

## 8.9. Word Similarity Analysis

At this stage, we retrieved the words that are vectorially most similar to the word *'mükellef'* and their similarity scores. Cosine similarity is commonly used to measure the similarity between words. Using the `most_similar()` function in Python, we calculated the vector similarity between the target word (*'mükellef'*) and other words, returning a ranked list of words with the highest similarity scores.

The result includes the similarity score of each word relative to the word *'mükellef'*. These scores typically range between -1 and 1. A score closer to 1 indicates a high degree of similarity to the word *'mükellef'*, while a score of 0 indicates very little semantic similarity.

This analysis is particularly useful for understanding semantic relationships between words in a language. It can be applied to

- Find words with similar meanings.
- Analyse language models.
- Develop automatic text recommendation systems.

The output of the analysis is shown in Figure 3.

```
[('dar', 0.8771734833717346),
 ('tam', 0.8355701565742493),
 ('imalatçı', 0.829497754573822),
 ('firma', 0.8249997496604919),
 ('ödenen', 0.8103259801864624),
 ('kişiler', 0.7855820059776306),
 ('kişi', 0.7799397706985474),
 ('firmalara', 0.7753022909164429),
 ('mükellefiyeti', 0.7662615180015564),
 ('üreten', 0.7659305930137634)]
```

Figure 3. A Sample of the Output

## 8.10. Saving the Word2Vec Tax Law Corpus Model to a File and Rotating it

We saved the trained Word2Vec model to a file using the function `model2.save('word2vec.model2')`, which stores all the model parameters, including the word vectors and training information. The saved model was later loaded using `model2 = Word2Vec.load('word2vec.model2')`. To visualise the word vectors, we created a t-SNE model with the command `tsne = TSNE(perplexity=20, random_state=0)`<sup>4</sup>. The perplexity parameter, a hyperparameter of the t-SNE algorithm, controls the number of neighbouring data points considered. It typically ranges between 5 and 50, with lower values focusing on smaller neighbourhoods and higher values on larger ones. In this study, the perplexity was set to 20, meaning that the algorithm considered

<sup>4</sup>The perplexity value in t-SNE controls the number of neighbouring data points considered, where low values focus on a smaller set of neighbours, and high values include a larger set. A value of 20 strikes a balance, providing a neighbourhood that is neither too small nor too large, effectively preserving both the local and global structure of the dataset. Typically, perplexity values between 5 and 50 yield reasonable results, with 20 being a reliable starting point for capturing the clustering structure without excessive scaling or dispersion. Practical experience suggests that a perplexity of 20 often represents the data distribution well although the optimal value can depend on factors such as dataset size and density. To ensure reproducibility, the `random_state` parameter fixes the randomness in t-SNE's initialisation, allowing consistent results across repeated runs with the same dataset and hyperparameters. This also aids in comparability, as using the same `random_state` value ensures consistent results when revisiting the same project or comparing studies in different environments. Thus, a perplexity value of 20 enables t-SNE to accurately capture the data structure and visualise cluster relationships, while a `random_state` value of 0 ensures that the results are repeatable and comparable.

approximately 20 neighbouring data points for each data point. The `random_state` parameter, acting as a stabiliser (seed), ensures that the model produces consistent results when rerun. Setting `random_state=0` allowed us to achieve reproducibility, ensuring the same results in every execution. This combination of settings enabled the t-SNE algorithm to work effectively for visualising and interpreting the trained word vectors.

## 9. VISUALISATION

### 9.1. Reporting the Nearest Words as a Two-Dimensional Graph with T-SNE

Using Python functions, it is possible to create a two-dimensional graph with t-SNE (t-Distributed Stochastic Neighbour Embedding) to visualise the words closest to any word in the model.

Here, we tested our model using words as the centre of the t-SNE plot. Word vectors, initially defined as an empty NumPy array, consisted of 100-dimensional vectors per row. The word lists contained the label (name) of the centre word followed by nearby words, starting with the centre word itself. The model identified the words closest to the given word ('mükellef'), returning the most similar words along with their similarity scores. The result was a list of words and similarity scores.

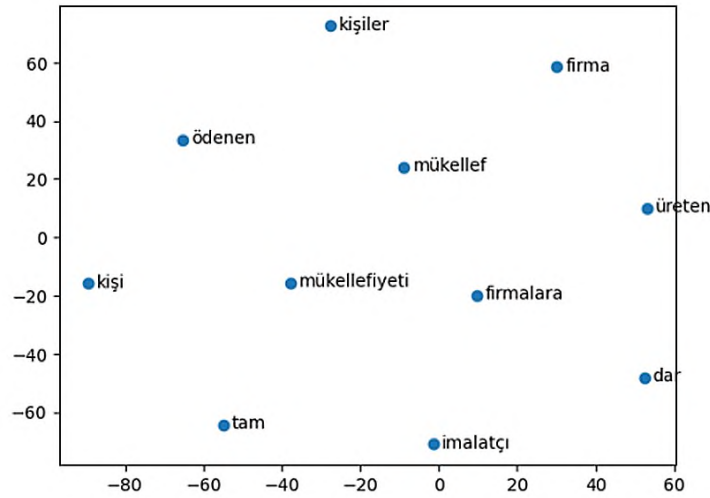
The 100-dimensional vector of the centre word was extracted and added to the array. Then, the vectors of the closest words were iteratively processed, adding each word's vector to the array and its label (name) to the list. In this way, the vectors and labels of all the words were collected in a structured manner.

The perplexity value, an important t-SNE hyperparameter, was set based on the number of words. The word vectors obtained from the model were reduced to 2D space using the t-SNE algorithm, and the assigned variable `Y` contained the 2D coordinates of each word. The terms `x_coords` and `y_coords` represent the x and y coordinates of each word in 2D space. These coordinates were used to plot the words in 2D space on a graph.

Using the `plt.annotate()` code snippet, the label of each word was added to the graph at its respective x and y coordinates, with the labels positioned near the points. The `plt.show()` code snippet displayed the final graph on the screen.

### 9.2. Visualisation Command Using t-SNE (t-Distributed Stochastic Neighbour Embedding)

Finally, using the previously defined functions, a Python command was executed to visualise the closest words to the word 'mükellef' in the model via t-SNE (t-Distributed Stochastic Neighbour Embedding). This allowed the closest words to 'mükellef' to be graphically displayed, as shown in [Figure 3](#).



**Figure 3.** Words Most Similar to the Word 'mükellef'

According to Article 8/1 of the Tax Procedure Law (Vergi Usul Kanunu), "a taxpayer is a natural or legal person to whom a tax obligation is imposed according to tax law" (mükellef, vergi kanunlarına göre kendisine vergi borcu terettübeden gerçek veya tüzel kişidir). To "hesitate" (tereddüb etmek) means to have a duty or to be required (üzerine görev düşmek, gerekme anlamına gelir) [24]. Taxpayers are referred to as 'kişi' (real or legal persons) or 'şirket' (companies) in various parts of the law.

Mükellefler (taxpayers) are defined in Articles 3–6 of the Income Tax Law and Articles 3–6 of the Corporate Tax Law. In these articles, taxpayers are divided into two different classes: 'tam' and 'dar'. These persons often assume the identity of the producers and manufacturers. From this perspective, it is evident that the words associated with 'mükellef' in the table possess characteristics that align with the concept of 'mükellef', demonstrating that the similarity model is functioning effectively.

## 10. CONCLUSION AND EVALUATION

In this paper, the Word2Vec model was used to train a model tailored to the legal context by working on a corpus specific to Turkish tax laws. Word2Vec represents words as vectors and identifies semantic relationships between them based on the proximity of these vectors in vector space. The analysis focused on the connections between words found in tax laws and other related terms. Using Python libraries such as NumPy, Gensim, Scikit-learn, and Matplotlib, high-dimensional data was visualised on a two-dimensional plane through the t-SNE algorithm. This method made the relationships between similar terms used in tax legislation observable and provided a deeper understanding of their meanings in the context of tax laws, leveraging machine learning techniques.

With over 25 years of professional experience in tax law and a doctorate in the field, I can confidently state that, in the context of the word "mükellef" (taxpayer), the relationships between words in the laws are meaningful and consistent. Similarly, the approach yielded successful results for other tax-related terms not included in this paper.

The word vectors generated by the Word2Vec model offer a powerful foundation for understanding the semantic affinities of terms in tax laws. For instance, the semantic proximity of the word "mükellef" to terms such as "tam" (full), "dar" (narrow), "şirket" (company), and "üretici" (manufacturer) demonstrates how key

concepts in tax law are interrelated. Such analyses facilitate the grouping of terms with similar meanings and provide a clearer understanding of the conceptual basis of legal regulations. The t-SNE visualisation simplifies the relationships between words in tax laws by reducing them to a two-dimensional plane, making it easier to observe how words cluster in a legal context. These visualisations can serve as a foundation for legal decision support systems, enabling automated legal advice based on visualised word relationships.

Vector relationships offer an excellent primer for AI-assisted tax law advice, enabling AI systems to become more effective and context-sensitive. The dense and technical nature of legal language makes comprehension difficult, but AI models can play a significant role in simplifying legal texts, enhancing transparency and accessibility in legal processes. Similarity analysis and the use of word vectors can expedite relationships between legal documents and streamline processes, particularly in litigation or legal review scenarios. A correct understanding of tax law terminology allows AI consultancy systems to respond to user requests with greater accuracy.

Natural language processing models like Word2Vec interpret the words in tax laws, learn the relationships between them, and make complex regulatory information analysable through machine learning. This accelerates the delivery of information about tax legislation while improving accuracy. Future developments of this study, such as the use of larger datasets and different NLP models, could enable the creation of user-friendly consultancy platforms that provide in-depth, reliable information on tax law and adapt quickly to legal changes.

The methodology in this paper needs further refinement to model more complex legal relationships and to highlight rare but critical words. However, the potential of the model presented here as a foundation for AI assistants in tax law consultancy is noteworthy. Developing a tax law-specific Word2Vec model enables the creation of AI solutions that are more customised and context-sensitive than general-purpose language models. This enhances the ability of AI-based applications to draw meaningful conclusions from legal documents, improves the accuracy of information provided to users, and minimizes errors in legal processes. Such systems can elevate legal advice services by delivering more accurate and consistent information tailored to users' needs.


This infrastructure paves the way for innovation in the sector, enabling tax law AI assistants to offer reliable, detailed recommendations tailored to taxpayers. Future research could explore comparisons with other word representation methods, integration with context-aware models like BERT or GPT, summarisation of legal texts, application to real-world tax disputes, and the development of automated response systems (ChatBots).



---

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

---

Author Details **Ali İhsan Özgür Çilingir**  
<sup>1</sup> Non-affiliated, İstanbul, Türkiye  
 0000-0002-0490-4192

---





## References

- [1] Mikolov, Tomas, Chen, Kai, Corrado, Greg, and Dean, Jeffrey. Efficient Estimation of Word Representations on Vector Space. arXiv preprint arXiv:1301.3781 (2013).
- [2] Pervan, Nergis. DERİN ÖĞRENME YAKLAŞIMLARI KULLANARAK TÜRKÇE METİNLERDEN ANLAMSAL ÇIKARIM YAPMA. Ankara, 2019.
- [3] Onan, Aytuğ. Evrişimli Sinir Ağı Mimarilerine Dayalı Türkçe Duygu Analizi. *Avrupa Bilim ve Teknoloji Dergisi* (Aug. 31, 2020), 374-380.
- [4] Tezgider, Murat, Yıldız, Beytullah, and Aydın, Galip. Improving Word Representation by Tuning Word2Vec Parameters with Deep Learning Model. In *International Artificial Intelligence and Data Processing Symposium (IDAP) (Malatya 2018)*, IEEE, 1-7.
- [5] Arabacı, Mehmet Ali, Esen, Ersin, Atar, Muhammed Selim, Yılmaz, Eyüp, and Kaltalıoğlu, Batuhan. Kelime Gömevi Yöntemi Kullanarak Benzer Cümle Tespiti. In *2018 26th Signal Processing and Communications Applications Conference ( 2018)*.
- [6] Aydoğan, Murat and Karçı, Ali. Kelime Temsil Yöntemleri ile Kelime Benzerliklerinin İncelenmesi. *Çukurova Üniversitesi Mühendislik Mimarlık Fakültesi Dergisi*, 34, 2 (June 2019), 181-195.
- [7] Acı, Çiğdem İnan and Çırak, Adem. Türkçe Haber Metinlerinin Konvülsiyonel Sinir Ağları ve Word2Vec Kullanılarak Sınıflandırılması. *BİLİŞİM TEKNOLOJİLERİ DERGİSİ*, 12, 13 (July 31, 2019), 219-228.
- [8] Xia, Chunyu, He, Tieke, Li, Wenlong, Qin, Zemin, and Zou, Zhipeng. Similarity Analysis of Law Documents Based on Word2vec. In *2019 IEEE 19th International Conference on Software Quality, Reliability and Security Companion (QRS-C) (Sofia 2019)*, IEEE, 354-357.
- [9] Tatchum, Ghislain Wabo, Makembe, Fritz Sosso, Nzeko, Armel Jacques Nzekon, and Djam, Xaviera Youh. Class-Oriented Text Vectorization for Text Classification: Case Study of Job Offer Classification. *Journal of Computer Science an Engineering (JCSE)*, 5, 2 (Aug. 01, 2024), 116-136.
- [10] Wei, Wei, Liu, Wei, Zhang, Beibei, Scherer, Rafal, and Damasevicius, Robertas. Discovery of New Words in Tax-related Fields Based on Word Vector Representation. *Journal of Internet Technology*, 24, 4 (July 2023), 923-930.
- [11] Chalkidis, Ilias and Kampas, Dimitrios. Deep learning in law: early adaptation and legal word embeddings trained on large corpora. *Artificial Intelligence and Law* (Dec. 2019), 171-198.
- [12] Mandal, Arpan, Ghosh, Kripabandhu, Ghosh, Saptarshi, and Mandal, Sekhar. Unsupervised approaches for measuring textual similarity between legal court case reports. *Artificial Intelligence and Law*, 29 (2021), 417-451.
- [13] Saha, Rohan. Influence of various text embeddings on clustering performance in NLP. arXiv, 44 (May 04, 2023), 1-22.
- [14] Zhong, Ziyuan, Verma, Nakul, and Lia, Vincent. Lecture 8 – t-Distributed Stochastic Neighbor Embedding. New York, 2018.
- [15] Linderman, George C. and Steinerberger, Stefan. CLUSTERING WITH T-SNE, PROVABLY. arXiv (June 08, 2017), 1-15.
- [16] Arora, Sanjeev and Hu, Wei. An Analysis of the t-SNE Algorithm for Data Visualization. In *Conference on Learning Theory (COLT) 2018 (Stockholm 2018)*, arXiv, 1-32.
- [17] Maaten, Laurens van der and Hinton, Geoffrey. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9, 86 (Sep. 2008), 2579-2605.
- [18] Mueller, John Paul and Massaron, Luca. *Deep Learning for Dummies*. John Wiley & Sons, Inc., New Jersey, 2019.
- [19] Nelson, Hala. *Essential Math for AI - Next Level Mathematics for Efficient and Successful AI Systems*. O'Reilly Media, Sebastopol, 2023.
- [20] Kelleher, John D. *Deep Learning*. The MIT Press, London, 2019.
- [21] Anonymous. NumPy documentation. 2024.
- [22] <https://scikit-learn.org/stable/>. <https://scikit-learn.org/stable/>. 2024.
- [23] <https://matplotlib.org/>. <https://matplotlib.org/>. 2024.
- [24] Anonim. *Osmanlı Türkçesi Sözlüğü*.
- [25] Haider, Mofiz Mojib, Hossin, Arman, Mahi, Hasibur Rashid, and Arif, Hossain. Automatic Text Summarization Using Gensim Word2Vec and K-Means Clustering Algorithm. In *2020 IEEE Region 10 Symposium (TENSYP) (Dhaka 2020)*, 283-286.
- [26] Li, Zhie and Rao, Zhuyi. Text classification model based on Word2vec and SF-HAN. In *2020 IEEE 5th Information Technology and Mechatronics Engineering Conference (ITOEC 2020) (Shenzhen 2020)*, 978-1-7281-4323-1/20/\$31.00 ©2020 IEEE, 1385-1390.
- [27] Mao, Yushang, Zhang, Guixuan, and Zhang, Shuwu. Word Semantic Similarity Based on CiLin and Word2vec. In *2020 International Conference on Culture-oriented Science & Technology (ICCST) (Beijing)*, 978-1-7281-8138-7/20/\$31.00 ©2020 IEEE, 304 - 307.
- [28] Bissiri, Pier Giovanni and Walker, Stephen G. Converting information into probability measures with the Kullback–Leibler divergence. *Ann Inst Stat Math* (2012), 1139-1160.



- [29] Jaya, Putra Syopiansyah, Nur, Gunawan Muhamad, and Akbar, Hidayat Arief. Feature Engineering with Word2vec on Text Classification Using The K-Nearest Neighbor Algorithm. In The 10th International Conference on Cyber and IT Service Management (CITSM 2022) (Yogyakarta 2022), ©2022 IEEE.
- [30] Kurian, Jeomoo Francis and Allali, Mohamed. Detecting drifts in data streams using Kullback-Leibler (KL) divergence measure for data engineering applications. *Journal of Data, Information and Management* (2024), 207-2016.
- [31] Polat, Buğra. TÜRKÇE ÜRÜN YORUMLARI VERİSİ İLE DUYGU ANALİZİ. Ankara, 2021.
- [32] Çalışkan, Sedrettin, Yazıcıoğlu, Selahattin A., Demirci, Ulaş, and Kuş, Zeki. YAPAY SİNİR AĞLARI, KELİME VEKTÖRLERİ VE DERİN ÖĞRENME UYGULAMALARI. İstanbul, 2018.
- [33] Pirana, Gurur, Sertbaş, Ahmet, and Ensari, Tolga. Sanal Asistan Uygulamaları İçin Derin Öğrenme Yöntemiyle Cümle Sınıflandırma. In 3rd International Symposium on Multidisciplinary Studies and Innovative Technologies, ISMSIT 2019 (Ankara 2019), Institute of Electrical and Electronics Engineers Inc.
- [34] Kılıç, Berker and Öner, Yüksel. Yargıtay Kararlarının Suç Türlerine Göre Makine Öğrenmesi Yöntemleri İle Sınıflandırılması. *VERİ BİLİMİ DERGİSİ* (2021), 61-71.
- [35] Law, Jarvan, Zhuo, Hankui Hankz, He, Junhua, and Rong, Erhu. LTSG: Latent Topical Skip-Gram for Mutually Learning Topic Model and Vector Representations. arXiv preprint arXiv (Feb. 23, 2017).
- [36] Önal, Zeynep. Derin Öğrenme. Nobel Akademik Yayıncılık, Ankara, 2022.
- [37] Guthrie, David, Allison, Ben, Liu, Wei, Guthiere, Louise, and Wilks, Yorick. A Closer Look at Skip-gram Modelling. In Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06) (Genoa 2006), ACL Anthology, 1222-1225.
- [38] Srivastava, Rajendra P. New Measure of Similarity in Textual Analysis: Vector Similarity Metric versus Cosine Similarity Metric. *JOURNAL OF EMERGING TECHNOLOGIES IN ACCOUNTING*, 20, 1 (2023), 77-90.
- [39] Pudaruth, Sameerchand, Soyjaudah, Sunjiv, and Gunpath, Rajendra. Classification of Legislations using Deep Learning. *The International Arab Journal of Information Technology*, 18, 5 (Sep. 2021), 651-663.
- [40] Robaldo, Livio, Villiata, Serena, Wyner, Adam, and Grabmair, Matthias. Introduction for artificial intelligence and law: special issue "natural language processing for legal texts". *Artificial Intelligence and Law* (Apr. 2019), 113-115.
- [41] Tagarelli, Andrea and Simeri, Andrea. Unsupervised law article mining based on deep pre-trained language representation models with application to the Italian civil code. *Artificial Intelligence and Law*, 30 (Sep. 2022), 417-473.
- [42] Makawana, Mayur and Mehta, Rupa G. A novel network-based paragraph filtering technique for legal document similarity analysis. *Artificial Intelligence and Law* (Oct. 2023).
- [43] Bilgin, Metin. Kelime Vektörü Yöntemlerinin Model Oluşturma Sürelerinin Karşılaştırılması. *BİLİŞİM TEKNOLOJİLERİ DERGİSİ*, 12, 2 (Apr. 2019), 141-146.
- [44] Ahmetoğlu, Hüseyin and Daş, Resul. Türkçe Otel Yorumlarıyla Eğitilen Kelime Vektörü Modellerinin Duygu Analizi ile İncelenmesi. *Fen Bilimleri Enstitüsü Dergisi*, 24, 2 (2020), 455-463.
- [45] Çelik, Özer and Koç, Burak Can. TF-IDF, Word2vec ve Fasttext Vektör Model Yöntemleri ile Türkçe Haber Metinlerinin Sınıflandırılması. *Dokuz Eylül Üniversitesi Mühendislik Fakültesi Fen ve Mühendislik Dergisi*, 23 (2021), 121-127.
- [46] Kınık, Doğançan and Güran, Aysun. TF-IDF ve Doc2Vec Tabanlı Türkçe Metin Sınıflandırma Sisteminin Başarım Değerinin Ardışık Kelime Grubu Tespiti ile Arttırılması. *Avrupa Bilim ve Teknoloji Dergisi* (Jan. 2021), 323-332.
- [47] Hongnan, Tian and Xin, Guo. Research on Improved Sentence Similarity Calculation Method Based on Word2Vec and Synonym Table in Interactive Machine Translation. In 2021 5th International Conference on Robotics and Automation Sciences (Wuhan 2021), IEEE, 255-261.
- [48] Xiao, Lu, Li, Qiaoxing, Ma, Qian, Shen, Jiasheng, Yang, Yong, and Li, Danyang. Text classification algorithm of tourist attractions subcategories with modified TF-IDF. *PLOS ONE* (Oct. 2024), 1-34.
- [49] Gupta, Megha, Dheekonda, Venkatasai, and Masum, Mohammad. Genie: Enhancing information management in the restaurant industry through AI-powered chatbot. *International Journal of Information Management Data Insights* (May 25, 2024), 1-9.
- [50] G, Dhamodharan and A, Kaleemullah. An Innovative Algorithm for Enhanced PDF-Based Chatbot in Domain-Specific Question Answering. *Library Progress International*, 44, 3 (Sep. 01, 2024), 27648-27653.
- [51] Godghase, Gauri Anil, Agrawal, Rishit, Obili, Tanush, and Stamp, Mark. Distinguishing Chatbot from Human. arXiv:2408.04647v1 [cs.CL] (Aug. 12, 2024), 1-47.
- [52] Becha, Rahma, Sellami, Asma, Bouassida, Nadia, Idri, Ali, and Abran, Alain. BotCFP: A Machine Learning based Tool for COSMIC Chatbots Sizing. *CEUR*, 3852 (Apr. 30, 2024), 1-16.
- [53] <https://pypi.org/project/gensim/>. <https://pypi.org/project/gensim/>. 2024.
- [54] Leshem, Ido. Skip-Gram Word2Vec Algorithm Explained. 2023.

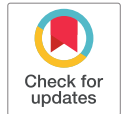




# Journal of Data Analytics and Artificial Intelligence Applications

Review Article

 Open Access

## Machine Learning Implementation in Automated Software Testing: A Review



Normi Sham Awang Abu Bakar<sup>1</sup>  

<sup>1</sup> International Islamic University Malaysia, Department of Computer Science, Kulliyah of ICT, Kuala Lumpur, Malaysia

### Abstract

The integration of Machine Learning (ML) in automated software testing represents a transformative approach aimed at enhancing the efficiency, accuracy, and scope of testing processes. This paper explores the theoretical and practical aspects of employing ML techniques within the realm of software testing, focusing on key areas such as test case generation, defect prediction, and test suite optimisation. Through a comprehensive literature review and case studies, this study illustrates the potential benefits associated with ML-driven testing methodologies. The findings indicate that ML can significantly reduce manual intervention and improve defect detection rates, thereby facilitating more reliable software delivery. This paper also addresses the benefits of ML implementation in automated testing and future research directions to bridge existing gaps and further leverage ML in software testing.

### Keywords


Artificial intelligence · automated testing · software testing activities · machine learning algorithm



Citation: Normi Sham Awang Abu Bakar. 2025. Machine Learning Implementation in Automated Software Testing: A Review. *Journal of Data Analytics and Artificial Intelligence Applications* 1, 1 (January 2025), 110-122. <https://doi.org/10.26650/d3ai.001>

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

© 2025. Abu Bakar, N. S.

 Corresponding author: Normi Sham Awang Abu Bakar [nsham@iium.edu.my](mailto:nsham@iium.edu.my)

## 1. INTRODUCTION

A fundamental component of the software development lifecycle has always been software testing, or ST. However, software has grown in size and complexity as it has become more widely used [1], posing new difficulties for software testing procedures [2]. Consequently, there is interest in examining how artificial intelligence (AI) has been applied to enhance testing procedures, since AI can improve knowledge work. The interaction between AI and ST has been the subject of numerous studies [3]. However, because each of these fields is so vast and complex, excellent review studies typically concentrate their attention on orthogonal choices within each of these fields.

The main goal of this paper is to explore the machine learning implementation in the automated software testing context. In this study, the main focus is on the use of the machine learning algorithms to make the automated testing more efficient, which will assist the software testers to focus on test executions, rather than on test planning and design.

To achieve this goal, 34 papers were reviewed for their relevancy in both the ST and AI areas, which discuss the AI-driven methodologies and tools to improve the efficiencies of the automated software testing activities. In particular, the machine learning techniques are also explored to add more depth to the understanding of the most frequently used techniques to support automated software testing.

As such, two research questions are developed for this study:

RQ1: What are the machine learning techniques frequently used to support the automated software testing activities?

RQ2: How are the machine learning techniques being implemented in the automated software testing activities?

The remainder of the article is organised as follows. Section 2 introduces the background and the prior related works in this study, Section 3 describes the implementation of machine learning in automated testing, Section 4 highlights the advantages of using AI in ST, and Section 5 concludes the findings of the paper and discusses the future work.

## 2. BACKGROUND

Current research directions in Software Engineering automation could be perfectly complemented by recent developments in generative AI. Specifically, generative AI naturally pairs well with automated test data generation. Despite the generative AI approach's potential to produce highly human-readable, domain- and context-aware solutions, its propensity for hallucinations makes it somewhat unreliable when used alone. Nevertheless, automated test data generation can eliminate these delusional features of AI-based solutions while also adding the essential assurances.

There are important implications regarding the recent findings that generative AI models can exhibit robust emergent behaviours [4], [5]. Their behaviour is therefore both powerful and inherently difficult to understand. Because the emergent behaviour of the models cannot be cross-checked against a ground truth, it may be problematic in applications lacking a ground truth, such as general inquiries about arbitrary facts about reality. However, for software engineering tasks like code enhancement and testing, we have an extremely reliable ground truth: the execution of the improved code or the recommended test.



## 2.1. Artificial Intelligence

Despite the fact that there are numerous definitions of AI, the definition given in [6] is used for the purposes of this investigation: “AI is a generic term that refers to any machine or algorithm that is capable of observing its environment, learning, and based on the knowledge and experience gained, taking intelligent action or proposing decisions. There are many technologies that fall under this broad AI definition. At the moment, ML techniques are the most widely used.”

The AI domain dealing with the ability of systems to automatically learn, decide, predict, adapt and react to changes and improve from experience, without being explicitly programmed, is the learning domain [7]. According to the AI Watch report, there are five core scientific domains:

### 2.1.1. Reasoning

The field of artificial intelligence studies methods for turning data into knowledge and drawing conclusions from it. Knowledge representation, automated reasoning, and common sense reasoning are the three sub-domains that make up this domain.

### 2.1.2. Planning

The area of artificial intelligence that focuses on creating and implementing strategies for performing tasks, usually carried out by unmanned vehicles, intelligent agents, and autonomous robots. In this field, strategies are distinguished by intricate solutions that need to be found and refined in a multidimensional environment. This domain consists of three closely related sub-domains: searching, optimisation and planning and scheduling. The optimisation of the search for solutions to scheduling and planning issues is the focus of these sub-domains.

### 2.1.3. Learning

The branch of artificial intelligence that deals with a system’s natural capacity to learn, make decisions, forecast outcomes, adjust to changes, and grow through experience—all without the need for explicit programming. Machine learning (ML)-related concepts are primarily used in the construction of the corresponding branch of the resulting taxonomy.

### 2.1.4. Communication

The field of artificial intelligence deals with the recognition, processing, comprehension, and creation of data from spoken and written human communication. The field of natural language processing (NLP) primarily deals with this domain [5].

### 2.1.5. Perception

This field indicates a system’s capacity to perceive its surroundings through its hearing and vision, such as computer vision.

## 2.2. Machine Learning Techniques

Machine learning (ML) is the science of getting computers to learn and act like humans do. It uses algorithms and mathematical models to progressively improve their performance on a specific task [8]. In essence, machine learning (ML) is the process of identifying patterns in data and using that knowledge to solve problems with regression or classification. The representation of the data that the machine learning algorithms are given has a major impact on how well they perform. In fact, machine learning algorithms “learn”



how to accomplish certain tasks through a training phase using training datasets, which are representative sample data [1].

Machine learning can handle unsupervised learning problems (like clustering or dimensionality reduction) where no ground truth is provided, as well as supervised learning problems (like classification and regression) where training sets are annotated (or labelled) with the ground truth values. Reinforcement learning (RL) algorithms are based on a feedback-directed mechanism that allows them to continuously adapt to their operating environment. To maximise an expected cumulative reward function, the algorithm makes a decision, considers the effects of that decision and then modifies its approach.

Among the main ML methods that are related to ST are [7]:

- (i) artificial neural networks (ANN), a group of supervised algorithms that are modelled after biological neural networks discovered in animal brains [8]. It is necessary to observe the input and expected output data and establish the probability-weighted associations between the two to train a neural network. The network's data structure, which is composed of layers of connected perceptions, then stores these associations [9].
- (ii) boosting is a group meta-algorithm for minimising the components of bias and variance error [10],
- (iii) classification, a supervised task that includes the process of training a model on a population of instances labelled with a discrete set of labels yields a set of predicted labels for a given collection of unobserved instances [11].
- (iv) clustering, given a similarity function for an unsupervised task, objects are grouped into clusters based on how much more similar they are to one another than they are to objects in other clusters [12].
- (v) convolutional neural networks (CNN), a particular neural network where at least one layer substitutes convolution for general matrix multiplication [13].
- (vi) decision trees, a family of classification and regression algorithms that learn the hierarchical structures of fundamental decision rules from the data. The resulting models can be visualised as trees, where nodes represent decision rules and leaf nodes represent outcomes [14], [15].
- (vii) probabilistic models, a family of classifiers that can forecast a probability distribution across a range of classes given an observation of an input [16], [17].
- (viii) reinforcement learning, the algorithms address the “problem faced by an agent that must learn behaviour through trial-and-error interactions with a dynamic environment” and one of the core paradigms of machine learning [18].
- (ix) regression, with a set of mathematical techniques, data scientists can forecast a continuous outcome based on the value of one or more predictor variables [19].
- (x) supervised learning, a paradigm for machine learning when the available data is limited to labelled examples [20].
- (xi) support vector machines (SVM), supervised learning algorithms that, after the input features are non-linearly mapped to a very high-dimension feature space, build a linear decision surface to generate models for classification and regression analysis [21].
- (xii) unsupervised learning, a basic machine learning paradigm in which computers attempt to identify patterns in unlabelled data [20].



The machine learning techniques discussed previously are summarised in [Figure 1](#).



**Figure 1.** Machine Learning Techniques

### 2.3. Software Testing

Software Testing is defined by the 29119-1-2013 ISO/IEC/IEEE International Standard as: “A process made by a set of interrelated or interacting activities aimed at providing two types of confirmations: verification and validation” [22]. Validation proves that the work item can be used by users for their particular tasks, while verification verifies that a given software product (work item or test item) satisfies the specified requirements.

In this study, the ST domain that will be investigated is the Testing Activities. This ST domain describes the tasks that testing teams and testers can complete into precise, controlled processes. To guarantee that the test objectives are satisfied in an economical manner, these activities range from test planning to test output evaluation.

Among the testing activities identified in this study are:

- (i) Test Case Generation whose goal is to create executable test cases according to the specific testing methods and the amount of testing that needs to be done.
- (ii) Test Planning is a fundamental activity of the ST process; it encompasses staff coordination, test equipment and facility availability, test-related documentation creation and upkeep, and scheduling of additional testing activities.
- (iii) Test Results Evaluation is performed to determine if the testing was carried out successfully. “Successful” usually refers to the software operating as anticipated and producing no significant unexpected results. Unexpected results aren’t always bad; occasionally, they turn out to be noise. An analysis and debugging effort is required to isolate, identify, and describe a fault before it can be fixed.

- (iv) Test Execution symbolises both running test cases and documenting the outcomes of those runs. A fundamental tenet of scientific experimentation should be applied when conducting tests: all procedures should be carried out and recorded in a way that makes it possible for another individual to repeat the findings.
- (v) Test Oracle Definition is the process carried out to assist in the creation of test oracles or to generate them automatically.
- (vi) Test Case Design and Specification is carried out to define or design the testing cases. The requirement analysis of the system being tested is typically the first step in this process.
- (vii) Test Case Optimization/ Prioritisation/Selection is carried out to select, prioritise, and reduce test cases for execution in an optimal manner [23].
- (viii) Test Data Definition (test data generation) is the process that generates the test case data [13].

## 2.4. Automated Testing

Writing a programme in any programming or scripting language that uses an external automation helper tool to replicate the manual test case steps is known as software testing automation. It entails developing toolkits for testing the implemented source code. Its objective is to increase the automation of the testing procedures. The tasks associated with development are programme development and test script writing; the former relates to the application itself, while the latter is concerned with the scripts that will be utilised to test the application [23].

Software test automation is defined by Dustin et al. as “management and performance of test activities, to include the development and execution of test scripts so as to verify test requirements, using an automated test tool”. In theory, test automation should be seen as a more all-encompassing concept that includes other tasks in addition to automated test scripting and execution during the software testing process.

Because testing is a repetitive process and it is advised to test every scenario, automation is crucial. Test automation will boost productivity and expand test coverage. Automated testing allows for the testing of different input values, conditions, and repeated execution of the tests. There will be a reduction in the testing time and resources. There are many tools available to automate acceptance, system, and functional tests. Watir, JMeter, and Selenium are a few of them.

## 3. MACHINE LEARNING TECHNIQUES IN AUTOMATED SOFTWARE TESTING

The fundamentals of AI testing are based on the idea of “automatic abstraction of application and test logic” [24]. Intelligent learning agents, which are capable of autonomously perceiving and responding to their surroundings, can help achieve this. By investigating the functionality and understanding the operation of the application, they can plan and develop the test cases on the target system. Ultimately, they can run the tests and analyse the test findings. The agents can operate at various levels of hierarchy and are arranged with other agents.

To answer both RQ1 and RQ2, 34 papers have been reviewed and the main keywords for the paper search are ("AI" OR "artificial intelligence" OR "ML" OR "machine learning") AND ((test\* AND (automated OR automation)) AND (“software engineering” OR “software”).





Test frameworks for artificial intelligence (AI) can be broadly applied to test both cross-domain and multiple applications within a domain. A requirement for the test "library of the common user flows" is that the AI test framework be designed to function with both cross-domain applications and multiple applications within the domain as a general framework. The actions and the elements are interconnected in this library. The AI agent can query the database to find the test cases for a particular type of element when it sees one in an application that it can interact with.

This section focuses on the results of the paper review on the topic under study, which is the implementation of the machine learning techniques in automated software testing. From the articles found on the topic, the machine learning techniques are divided into several categories, as shown in Table 1. In addition, they were mapped into the relevant software testing activities, as reported in the articles.

Table 1 depicts the mapping between the machine learning techniques in the software testing activities, where the corresponding articles highlighted in the table reported the implementation of various techniques of machine learning in software testing. Specifically, artificial neural networks have been applied to various testing tasks, including oracle definition, test-case generation, test-case refinement, and test-case evaluation; studies [25] and [26] covered these tasks and found that machine learning algorithms resulted in predicted output oracles, metamorphic relations, and test verdicts. Nearly all research uses a supervised or semi-supervised methodology, training models (e.g., neural networks, support vector machines, adaptive boosting, and decision trees) on labelled system executions or code metadata.

**Table 1.** Machine learning techniques in software testing

	Artificial neural network	Boosting	Classification	Clustering	Convolutional neural network	Decision trees	Probabilistic model	Reinforcement learning	Regression	Supervised learning	Support vector machine	Unsupervised learning
Test Case Generation	[24]					[24]	[24]				[24]	
Test Planning				[5]			[25]					
Test Oracle Definition	[5] [24] [26]	[26]	[25]		[24]	[25]		[25]	[26]	[25] [26]	[24]	
Test Case Design and Specification				[25]			[25]	[25]		[25]		
Test Case Optimisation/ Prioritization/ Selection	[28]	[28]	[25][27] [28]	[25] [27] [28]		[25] [28]	[24][29]	[25][27] [28]	[27] [28]	[25] [28]	[28]	[25][28]
Test Results Evaluation	[5]			[25]			[25]		[25]			[25]
Test Data Definition	[26]				[24]			[24]	[26]	[26]	[26]	
Test Execution			[24]	[24]				[24]				

Furthermore, Garousi et al. [5] found that, when compared to test oracles created using current conventional methods, those created using artificial neural networks for the Test Oracle Definition activity are more effective, efficient, and reusable. Furthermore, the primary benefits of utilising machine learning and artificial neural networks were noted by Durelli et al. [25] as being their scalability and low requirement for human intervention. According to Durelli et al. [25] and Fontes & Gay [26], the primary challenge encountered by researchers attempting to use artificial neural networks and machine learning to address software testing

issues is the requirement for a significant quantity of high-quality training data, which is essential for machine learning algorithms to perform as intended.

Khatibsyarbini et al. [27] also claimed that based on the publication trend of ML technique applied to Test Case Prioritisation, the most popular ML technique category was classification, followed by clustering and reinforcement learning as the least preferred ML technique category. Additionally, they stated that the most popular machine learning technique is classification because it uses historical data and yields high average percentages of faults detected and effective code coverage. They also emphasised that reinforcement learning needs to be improved and given more structure before it can be taught in undergraduate programmes.

According to Pan et al. [28], Reinforcement learning, clustering, and classification AI approaches have been widely used for test case optimisation, prioritisation and selection. According to their report, reinforcement learning, unsupervised learning (clustering), and supervised learning (ranking models) are the three main machine learning techniques used for test case prioritisation and selection. Any machine learning method that depends on ranking or classification models is called supervised learning.

In addition, the methods that use reinforcement learning to rank test cases based on their length, past performance, and failure history have also been reported by Durelli et al. [25]. Furthermore, Pan et al. highlighted that although supervised learning, unsupervised learning, reinforcement learning, and natural learning processing are the four main machine learning (ML) techniques used for test case selection and prioritisation, various combinations of these techniques have also been reported in the literature. To improve the test case prioritisation performance, supervised or unsupervised learning was integrated with NLP-based techniques, which are frequently used for feature preprocessing. They also emphasised how difficult it is to draw trustworthy conclusions about the effectiveness of ML-based test case selection and prioritisation due to the absence of appropriate publicly available datasets and standard evaluation processes that are derived from the execution of real-world case studies.

#### 4. THE IMPACT OF AI IN ST

The field of artificial intelligence for software testing, or AIST, is a young one that aims to create AI tools for software testing, test methods for AI systems, and create software that can self-test and/or self-heal. The process of manually encoding a predetermined set of programme input actions and output verification steps into a script that can be run by a machine is commonly referred to as “test automation” in software testing [10]. A log of the results is created, saved, and linked to the run after it is executed. The test execution and logging are the only parts of this process that are automated. To properly test software, human testers must set testing objectives, gain the knowledge required, create and specify comprehensive test scenarios, write test automation scripts, perform scenarios that cannot be automated, and evaluate test results to identify potential project risks.

Researchers and practitioners have begun looking into how AI and ML can be used to create the next generation of testing tools, since the majority of testing is currently focused on manual testing and the manual writing of test scripts [3]. The idea is to use big data, cloud computing, and AI/ML advancements to bridge the gap between human-present and machine-driven testing.

AI-driven testing has several benefits, including being robust, scalable, adaptable, reusable, and all-purpose. Machine learning techniques can be used to solve several issues. For instance, practically any mathematical

function can be approximated using a straightforward feed-forward neural network with a single hidden layer [20]. Consequently, various testing types, applications, and domains can benefit from the use of AI-driven testing.

Because AI-driven tests are typically not tied to any particular application, they can be applied to different applications within the same domain (like add item to cart) or to different domains (like login). By creating new tests every time, the pesticide paradox—a narrow scope of fault detection brought on by repeatedly running the same tests—can be avoided. Finally, accelerated test coverage is a key advantage of AI-driven testing, which is achieved by fusing large-scale test execution in the cloud with AI-based test generation [3].

Despite the promising results, the implementation of ML in automated software testing is not without challenges. The quality and quantity of data, the interpretability of ML models, and the integration of ML tools with existing testing frameworks are critical factors that need to be addressed. Moreover, the continuous evolution of software systems necessitates ongoing adaptation and learning, which poses additional challenges for ML-based testing solutions.

## 4.1. Case Study of AI implementation in ST

4.1.1. One of the AI technique implementations in ST is the usage of Reinforcement Learning for Test Case Optimisation. The details are

Objective: Optimise test case selection and prioritisation based on failure history and test execution performance. Furthermore, to prioritise and optimise test cases in regression testing by learning from historical test execution data, focusing on factors such as failure history, execution cost, and risk.

4.1.2. Implementation:

- Reinforcement learning algorithms were used to rank the test cases.
- Factors such as past failures, execution cost, and risk level were considered for prioritisation.

4.1.3. Outcomes:

- Higher fault detection rates.
- Reduced testing efforts and costs by focusing on critical test cases first.
- Example Application: Used in regression testing scenarios where frequent updates require selective testing.

4.1.4. Implementation Steps

4.1.4.1. Problem Formulation

The task is modelled as a reinforcement learning problem:

- State (S): Represents the attributes of the test case, such as the historical success rate, execution cost, and risk factor.
- Action (A): Decide whether to execute or skip a test case.
- Reward (R): A numeric value based on the detection of critical defects and cost savings (e.g., 1 for a defect found, -1 for skipping a necessary test).

4.1.4.2. Dataset

The input data includes historical test case executions:

- Features: Test case ID, previous pass/fail outcomes, execution time, code coverage metrics, and defect severity.
- Labels: Whether to execute (1) or skip (0) the test case.

#### 4.1.4.3. Results

- Output: The list of test cases prioritised for execution based on their predicted effectiveness.
- Benefits:
  - Ensures that critical test cases with higher defect detection probability are executed first.
  - Reduces the unnecessary execution of low-priority test cases, saving time and resources.

## 5. CONCLUSION AND FUTURE WORK

The integration of machine learning (ML) into automated software testing has shown significant potential in enhancing the efficiency and effectiveness of the software development lifecycle. This paper explored various ML techniques and their applications in different phases of software testing, including test case generation, test suite optimisation, defect prediction, and automated test script maintenance. By leveraging ML algorithms, software testing processes can be more adaptive and intelligent, leading to improved detection of defects, reduced testing time, and optimised resource allocation.

The reviews presented in this paper demonstrate the feasibility and advantages of using ML in automated software testing. Specifically, the use of supervised learning for defect prediction and clustering algorithms for test case prioritisation has proven to be effective in identifying high-risk areas of the software and optimising testing efforts. In addition, reinforcement learning techniques have shown promise in automating the generation and maintenance of test scripts, reducing the manual effort required and enhancing test coverage.

To further advance the field of ML in automated software testing, several areas warrant further research and development:

1. **Data Quality and Availability:** Ensuring high-quality and diverse datasets is crucial for training robust ML models. Future research should focus on developing methods for generating synthetic test data, handling imbalanced datasets, and improving data preprocessing techniques.
2. **Model Interpretability and Explainability:** As ML models become more complex, their interpretability becomes a significant concern. Future work should aim at developing techniques that provide insights into the decision-making process of ML models, enabling testers to understand and trust the predictions and recommendations made by these models.
3. **Integration with DevOps Practises:** Integrating ML-based testing solutions with modern DevOps practises can enhance continuous integration and continuous deployment (CI/CD) pipelines. Research should explore ways to seamlessly incorporate ML algorithms into these pipelines, ensuring that testing processes remain agile and responsive to changes in the software.
4. **Scalability and Performance Optimisation:** As software systems grow in complexity, the scalability of ML-based testing solutions becomes critical. Future research should investigate ways to optimise the performance of ML algorithms, ensuring that they can handle large-scale software projects efficiently.



5. **Cross-Project Learning and Transfer Learning:** Leveraging knowledge from previous projects can enhance the performance of ML models in new projects. Future work should explore transfer learning techniques and cross-project learning approaches to make ML models more generalisable and applicable across different software domains.
6. **Human-AI Collaboration:** The collaboration between human testers and ML models can lead to more effective testing strategies. Research should focus on developing interactive tools that facilitate this collaboration, allowing testers to leverage the strengths of both human expertise and ML capabilities.

The suggestions for overcoming data quality challenges and enhancing model interpretability are given below:

## 5.1. Overcoming Data Quality Challenges

### 5.1.1. Ensuring High-Quality Data

#### 5.1.1.1. Data Preprocessing

- Remove noise and irrelevant features through normalisation, scaling, and feature selection techniques.
- Detect and handle outliers using methods like Isolation Forest or Z-score analysis.

#### 5.1.1.2. Imbalanced Data Handling

- Use techniques like Synthetic Minority Oversampling Technique (SMOTE) to balance datasets when defect-prone areas are underrepresented.
- Employ cost-sensitive learning to penalise misclassifications of critical data.

#### 5.1.1.3. Data Augmentation

- Generate synthetic data to compensate for the limited datasets.
- Use domain-specific methods like mutation testing to create diverse test cases.

#### 5.1.1.4. Data Cleaning

- Automate error detection in datasets (e.g., duplicate entries, missing labels).
- Verify correctness through manual reviews of critical entries.

## 5.2. Enhancing the Model Interpretability

### 5.2.1. Explainable AI (XAI) Techniques

#### 5.2.1.1. Local Interpretability

- Use tools such as SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-Agnostic Explanations) to explain individual predictions.
- Visualise feature importance to highlight critical attributes influencing test case selection or defect predictions.

#### 5.2.1.2. Global Interpretability


- Employ decision tree models or surrogate models to approximate complex models like neural networks for easier understanding.
- Provide feature summary plots to show the overall trends in the model's decision-making process.



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

**Author Details** **Normi Sham Awang Abu Bakar**

<sup>1</sup> International Islamic University Malaysia, Department of Computer Science, Kulliyah of ICT, Kuala Lumpur, Malaysia

 0000-0002-8069-3323

## References

- [1] Gerhard Lakemeyer and Bernhard Nebel. 1994. Foundations of Knowledge Representation and Reasoning. Springer, Berlin, 1–12. [https://doi.org/10.1007/3-540-58107-3\\_1](https://doi.org/10.1007/3-540-58107-3_1)
- [2] Santiago Matalonga, Domenico Amalfitano, Andrea Doreste, Anna Rita Fasolino, and Guilherme Horta Travassos. 2022. Alternatives for testing of context-aware software systems in non-academic settings: Results from a rapid review. *Info. Softw. Technol.* 149 (2022), 106937. <https://doi.org/10.1016/j.infsof.2022.106937>
- [3] Tariq M. King, Jason Arbon, Dionny Santiago, David Adamo, Wendy Chin, and Ram Shanmugam. 2019. AI for testing today and tomorrow: Industry perspectives. In Proceedings of the IEEE International Conference On Artificial Intelligence Testing (AITest'19). IEEE, 81–88. <https://doi.org/10.1109/AITest.2019.000-3>
- [4] P. Paygude and S. D. Joshi. 2020. Use of evolutionary algorithm in regression test case prioritization: A review. In Proceeding of the International Conference on Computer Networks, Big Data and IoT (ICCB'18). Lecture Notes on Data Engineering and Communications Technologies, A. Pandian, T. Senjyu, S. Islam, and H. Wang (Eds.). Vol. 31, Springer, Cham, 56–66. [https://doi.org/10.1007/978-3-030-24643-3\\_6](https://doi.org/10.1007/978-3-030-24643-3_6)
- [5] Vahid Garousi, Sara Bauer, and Michael Felderer. 2020. NLP-assisted software testing: A systematic mapping of the literature. *Info. Softw. Technol.* 126 (2020), 106321. <https://doi.org/10.1016/j.infsof.2020.106321>
- [6] M. Craglia, A. Annoni, P. Benczur, P. Bertoldi, B. Delipetrev, G. De Prato, C. Feijoo, E. Fernandez Macias, E. Gomez Gutierrez, M. Iglesias Portela, H. Junklewitz, M. Lopez Cobo, B. Martens, S. Figueiredo Do Nascimento, S. Nativi, A. Polvora, J. I. Sanchez Martin, S. Tolan, I. Tuomi, and L. Vesnic Alujevic. 2018. Artificial Intelligence: A European Perspective. Technical Report KJ-NA-29425-EN-N. Luxembourg. <https://doi.org/10.2760/11251>
- [7] Domenico Amalfitano, Stefano Faralli, Jean Carlo Rossa Hauck, Santiago Matalonga, and Damiano Distanto. 2023. Artificial Intelligence Applied to Software Testing: A Tertiary Study. *ACM Comput. Surv.* 56, 3, Article 5 (October 2023), 38 pages. <https://doi.org/10.1145/3616372>
- [8] J. J. Hopfield. 1982. Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci. U.S.A.* 79, 8 (Apr. 1982), 2554–2558. <https://doi.org/10.1073/pnas.79.8.2554>
- [9] David E. Rumelhart, Bernard Widrow, and Michael A. Lehr. 1994. The basic ideas in neural networks. *Commun. ACM* 37, 3 (Mar. 1994), 87–92. <https://doi.org/10.1145/175247.175256>
- [10] Leo Breiman. 2000. Bias, Variance, and Arcing Classifiers. Technical Report 460, Statistics Department, University of California.
- [11] S. B. Kotsiantis, I. D. Zaharakis, and P. E. Pintelas. 2006. Machine learning: A review of classification and combining techniques. *Artific. Intell. Rev.* 26, 3 (Nov. 2006), 159–190. <https://doi.org/10.1007/s10462-007-9052-3>
- [12] Dongkuan Xu and Yingjie Tian. 2015. A comprehensive survey of clustering algorithms. *Ann. Data Sci.* 2, 2 (2015), 165–193. <https://doi.org/10.1007/s40745-015-0040-1>
- [13] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. Deep Learning. MIT Press. Retrieved from <http://www.deeplearningbook.org>
- [14] Bernard M. E. Moret. 1982. Decision trees and diagrams. *ACM Comput. Surv.* 14, 4 (Dec. 1982), 593–623. <https://doi.org/10.1145/356893.356898>



- [15] D. Opitz and R. Maclin. 1999. Popular ensemble methods: An empirical study. *J. Artific. Intell. Res.* 11 (Aug. 1999), 169–198. <https://doi.org/10.1613/jair.614>
- [16] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning*. Springer, New York. <https://doi.org/10.1007/978-0-387-84858-7>
- [17] Y. Bengio, P. Simard, and P. Frasconi. 1994. Learning long-term dependencies with gradient descent is difficult. *IEEE Trans. Neural Netw.* 5, 2 (1994), 157–166. <https://doi.org/10.1109/72.279181>
- [18] L. P. Kaelbling, M. L. Littman, and A.W. Moore. 1996. Reinforcement learning: A survey. *J. Artific. Intell. Res.* 4 (1996), 237–285. <https://doi.org/10.1613/jair.301>
- [19] G. Udny Yule. 1897. On the theory of correlation. *J. Roy. Stat. Soc.* 60, 4 (1897), 812–854. <https://doi.org/10.1111/j.2397-2335.1897.tb02784.x>
- [20] Stuart Russell and Peter Norvig. 2016. *Artificial Intelligence: A Modern Approach*. Pearson. Retrieved from <https://books.google.it/books?id=XS9CjwEACAAJ>
- [21] Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Mach. Learn.* 20, 3 (Sep. 1995), 273–297. <https://doi.org/10.1007/BF00994018>
- [22] International Organization for Standardization. 2013. ISO/IEC/IEEE international standard—software and systems engineering—software testing—Part 1: Concepts and definitions. ISO/IEC/IEEE 29119-1:2013(E) (2013), 64. <https://doi.org/10.1109/IEEESTD.2013.658853>
- [23] George Candea, Stefan Bucur, and Cristian Zamfir. 2010. Automated software testing as a service. In *Proceedings of the 1st ACM Symposium on Cloud Computing (SOCC '10)*. Association for Computing Machinery, New York, NY, USA, 155–160. <https://doi.org/10.1145/1807128.1807153>
- [24] Anna Trudova, Michal Dolezel, and Alena Buchalceva. 2020. Artificial intelligence in software test automation: A systematic literature review. In *Proceedings of the 15th International Conference on Evaluation of Novel Approaches to Software Engineering (ENASE'20)*. INSTICC, SciTePress, 181–192. <https://doi.org/10.5220/0009417801810192>
- [25] Vinicius H. S. Durelli, Rafael S. Durelli, Simone S. Borges, Andre T. Endo, Marcelo M. Eler, Diego R. C. Dias, and Marcelo P. Guimar.es. 2019. Machine learning applied to software testing: A systematic mapping study. *IEEE Trans. Reliabil.* 68, 3 (2019), 1189–1212. <https://doi.org/10.1109/TR.2019.2892517>
- [26] Afonso Fontes and Gregory Gay. 2021. Using machine learning to generate test oracles: A systematic literature review. In *Proceedings of the 1st International Workshop on Test Oracles (TORACLE'21)*. ACM, New York, NY, 1–10. <https://doi.org/10.1145/3472675.3473974>
- [27] Muhammad Khatibsyarbini, Mohd Adham Isa, Dayang N. A. Jawawi, Muhammad Luqman Mohd Shafie, Wan Mohd Nasir Wan-Kadir, Haza Nuzly Abdul Hamed, and Muhammad Dhiauddin Mohamed Suffian. 2021. Trend application of machine learning in test case prioritization: A review on techniques. *IEEE Access* 9 (2021), 166262–166282. <https://doi.org/10.1109/ACCESS.2021.3135508>
- [28] Rongqi Pan, Mojtaba Bagherzadeh, Taher A. Ghaleb, and Lionel Briand. 2021. Test case selection and prioritization using machine learning: A systematic literature review. *Empir. Softw. Eng.* 27, 2 (Dec. 2021), 29. <https://doi.org/10.1007/s10664-021-10066-6>
- [29] Gerson Barbosa, Erica Ferreira de Souza, Luciana Brasil Rebelo dos Santos, Marlon da Silva, Juliana Marino Balera, and Nandamudi Lankalapalli Vijaykumar. 2022. A systematic literature review on prioritizing software test cases using Markov chains. *Info. Softw. Technol.* 147 (2022), 106902. <https://doi.org/10.1016/j.infsof.2022.106902>

