Lecture Notes in Networks and Systems 1217

Marcin Hernes
Jaroslaw Wątróbski
Artur Rot *Editors*

Emerging Challenges in Intelligent Management Information Systems

Proceedings of 27th European Conference on Artificial Intelligence ECAI 2024—IMIS Workshop. Volume 1

Springer

Lecture Notes in Networks and Systems 1217

The series "Lecture Notes in Networks and Systems" publishes the latest developments in Networks and Systems—quickly, informally and with high quality. Original research reported in proceedings and post-proceedings represents the core of LNNS.

Volumes published in LNNS embrace all aspects and subfields of, as well as new challenges in, Networks and Systems.

The series contains proceedings and edited volumes in systems and networks, spanning the areas of Cyber-Physical Systems, Autonomous Systems, Sensor Networks, Control Systems, Energy Systems, Automotive Systems, Biological Systems, Vehicular Networking and Connected Vehicles, Aerospace Systems, Automation, Manufacturing, Smart Grids, Nonlinear Systems, Power Systems, Robotics, Social Systems, Economic Systems and other. Of particular value to both the contributors and the readership are the short publication timeframe and the worldwide distribution and exposure which enable both a wide and rapid dissemination of research output.

The series covers the theory, applications, and perspectives on the state of the art and future developments relevant to systems and networks, decision making, control, complex processes and related areas, as embedded in the fields of interdisciplinary and applied sciences, engineering, computer science, physics, economics, social, and life sciences, as well as the paradigms and methodologies behind them.

Indexed by SCOPUS, EI Compendex, INSPEC, WTI Frankfurt eG, zbMATH, SCImago.

All books published in the series are submitted for consideration in Web of Science.

For proposals from Asia please contact Aninda Bose (aninda.bose@springer.com).

Marcin Hernes · Jaroslaw Wątróbski · Artur Rot
Editors

Emerging Challenges in Intelligent Management Information Systems

Proceedings of 27th European Conference on Artificial Intelligence ECAI 2024—IMIS Workshop. Volume 1

Springer

Editors
Marcin Hernes
Center for Intelligent Management Systems
Department of Process Management
Wroclaw University of Economics
and Business
Wroclaw, Poland

Jaroslaw Wątróbski
Institute of Management
University of Szczecin
Szczecin, Poland

Artur Rot
Department of Information Systems
Wroclaw University of Economics
and Business
Wrocław, Poland

ISSN 2367-3370 ISSN 2367-3389 (electronic)
Lecture Notes in Networks and Systems
ISBN 978-3-031-78464-4 ISBN 978-3-031-78465-1 (eBook)
https://doi.org/10.1007/978-3-031-78465-1

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

This volume contains the first part of proceedings of the ECAI 2024 Workshop on Intelligent Management Information Systems (IMIS 2024). IMIS 2024 was part of the 27th European Conference on Artificial Intelligence ECAI 2024, held in Santiago de Compostela from October 19, 2024, to October 24, 2023.

Following the topics of IMIS 2024, this book discusses emerging challenges related to implementing artificial intelligence in Management Information Systems. The book is devoted to models, methods, and approaches addressing the development of artificial intelligence solution for improving the functionality of information systems that support management. The main focus is put on machine learning, including deep learning to support business processes, artificial intelligence for financial systems and cryptocurrencies, intelligent human-computer interfaces, knowledge management in business organizations, hybrid artificial intelligence, and multiple criteria decision analysis methods. The book has an interdisciplinary character; therefore, it is intended for a broad scope of readers, including researchers, students, managers, and employees of business organizations, software developers, IT and management specialists.

The volume is divided into three major parts covering the main issues related to the topic.

Part I deals with the application of artificial intelligence in information systems.

Huyen Trang Phan and Ngoc Thanh Nguyen ("A Dual LSTM-Based Multimodal Method for Fake News Detection") propose a method based on three elements: a BiLSTM-based fake headline extractor to capture and represent features of context and semantics on the headlines, a BiLSTM-based fake body extractor to capture and represent features of context and semantics on the news bodies, and a dual BiLSTM-based multimodal fake news detector to combine two fake headline and body features based on early fusion mechanism. The proposed model was tested on ISOT and FAKES datasets and achieved better results than previous methods regarding the same content-based multimodal approach.

Anna Sołtysik-Piorunkiewicz and Witold Chmielarz ("Conception of Intelligent Systems for the Creation Sustainable Use of Human Resources in the IT Sector") present the concept of an intelligent system for creating sustainable development of labor resources in the employment market in Poland. The original achievement of this research is the association of these three main elements of the constructed system, which, after systematic and consistent supplementation of its elements, can be used by both academics and business activists for research on the sustainable development of the labor market in the IT sector in Poland.

Anna Borawska and Adrianna Mateja ("Eye Tracking Insights: Analyzing Cognitive Load Across Media Types") explore the relationship between eye tracking and cognitive load across different media types., including images and videos. By leveraging eye-tracking technology, the research provides precise insights into visual attention

and cognitive effort, ultimately improving the effectiveness of digital communication strategies.

Maciej Krzan, Helena Dudycz, and Maciej Suchomski ("Artificial Intelligence in the Automation Process of Customs Declaration Preparation: A Case Study of an International Logistics Company") present a prototype for modifying an existing system for the preparation of customs declarations by using Optical Character Recognition technology and an artificial intelligence system. The contribution of the research is the proposal to automate the customs declaration process by using OCR and artificial intelligence, which significantly reduces the time required to prepare the finished document.

Krystian Wojtkiewicz, Rafał Palak, Zbigniew Telec, and Filip Litwinienko ("Modelling Cognitive Load of Computer Game Users—Case Study") evaluated the effectiveness of various methods for measuring cognitive load in arcade racing game players and their usefulness in modelling cognitive load. Significant changes in physiological measures were observed corresponding to different cognitive load levels, with the most accurate classification results achieved by combining physiological and performance data.

Ewa Ziemba, Dariusz Grabara, Katarzyna Renik, and Ewa Maruszewska ("Behavioral Intention and Use of ChatGPT Among Accounting and Finance Students: A Two-Year Comparative Study") assessed the behavioral intention and use of ChatGPT among accounting and finance higher education students, highlighting differences over the years 2023 and 2024. The results confirmed that both the levels of behavioral intention and use of ChatGPT increased significantly over the examined years.

Andrzej Greńczuk, Iwona Chomiak-Orsa, Katarzyna Tryczyńska, Sahinguvu William, Hilaire Nkunzimana, and Piotr Ogonowski ("Artificial Intelligence as a Tool to Support the Translator's Work. Bibliometric Analysis") present the first stage of the author's research dedicated to identifying and assessing the quality of translation processes supported by artificial intelligence tools. The bibliometric analysis carried out showed that the recent years, including the development of new tools, have made it possible to quickly translate texts (in the form of files, loose passages, and real-time statements) at a good level.

Adrianna Mateja, Dawid Subocz, Marta Stępień-Słodkowska, and Małgorzata Nermend ("Cognitive Load and Online Customer Decisions: The Role of Visual Perception, Memory, and Brain Activity") developed an innovative approach to measuring the impact of cognitive load on the purchasing decision-making process in online stores. The authors focus on psychological aspects, including reducing the influence of external factors, to obtain more reliable results.

Part II is devoted to advanced machine learning and data processing methods to support business processes.

Mateusz Piwowarski, Paweł Ziemba, and Jacek Cypryjański ("Machine Learning-Based Exploration of Eye-Tracking Data to Predict Offer Selection") present the research results on developing a classification model to predict decisions on selecting sales offers, based on specific eye-tracking metrics calculated for designated areas of these offers. The research results show that the use of appropriate classification algorithms, methods to reduce the impact of data imbalance on results, or attribute filtering methods provide effective methodological approaches to develop such a classifier.

Dorota Jelonek, Magdalena Graczyk-Kucharska, Robert Olszewski, Maciej Szafrański, and Magdalena Rzemieniak ("Competencies of the Future as a Criterion for Segmentation of Generation Z Candidates: Machine Learning and the CART Model") explore the possibility of modeling the segmentation of candidates with selected competencies of the future. Data on young generation Z's characteristics, competencies possessed, candidates' job expectations, or location were used for analysis. The key finding is that machine learning and other artificial intelligence methods are widely applicable in segmenting candidates, and that the competencies of the future can be linked to their expectations of where and how they work, including such things as work atmosphere, company innovation, and the ability to work remotely.

Marek Karwanski, Boleslaw Borkowski, Wieslaw Szczesny, and Artur Wilinski ("The Effectiveness of AI Methods Compared to Statistical Analysis and Data Modeling In the Construction of Scoring Models"), a comparative analysis of the effectiveness of two analytical cultures, statistical analysis, and data modeling (DMC) and algorithmic AI analysis (AMC) using machine learning methods. They proposed variable selection methods in the DMC and AMC models by analyzing the importance of variables.

Grzegorz Wojarnik ("The Impact of Current Temperatures on Forecasting Natural Gas Futures Prices in the USA Using Machine Learning Algorithms") analyzes the impact of current temperatures in major US cities on the accuracy of forecasts for natural gas futures prices using various machine learning algorithms. The conclusions of the study suggest that incorporating temperatures into predictive models may be beneficial for "end-of-day trading" investment strategies, leading to more accurate investment decisions.

Anna Borawska, Mariusz Borawski, and Małgorzata Łatuszyńska ("The Role of Data Normalization Algorithm in Measuring of EU Countries Digitalization Degree") analyze the impact of selected normalization algorithms on the Digital Economy and Society Index ranking and to identify the algorithm that is least sensitive to unusual indicator values. The findings indicate that the linear normalization sum-based algorithm is the most effective, providing the most stable ranking where changes in the Digital Economy and Society Index of some countries are minimally affected by changes in the indicator values of other countries in the ranking.

Grzegorz Szyjewski ("Enhancing Examination Question Quality through Data-Driven Analysis and Systematic Maintenance") developed an approach for maintaining a high-quality question test database, implemented in a real-world case study. The study concludes that this tool is valuable for systematic improvements, maintaining the integrity and reliability of the assessment process.

Ryszard Zygała, Wiesława Gryncewicz, and Agnieszka Rosa ("Analyzing Data Science Labor Market Trends in Poland Using NLP Techniques") analyze trends within the Polish data science job market. The study points to hypotheses and theories about the growing demand for data science occupations, the impact of artificial intelligence on job creation and skill requirements, and the influence of social processes on the observed changes.

Mateusz Anders, Jozef Zurada, and Paweł Weichbroth ("An Empirical Study of a Dynamic Stop Loss Strategy with Deep Reinforcement Learning on the NASDAQ Stock Market") empirically investigate the efficacy of using Deep Reinforcement Learning

to maximize investment returns by incorporating expected optimal closing prices of long positions into a daily strategy. Authors found a positive impact of using DRL compared to other tested strategies when encountering entirely new data, suggesting positive serial market correlations. The results suggest that appropriate closing rules and active management of stop levels can increase investment returns without necessarily reducing portfolio return volatility.

Ewa Walaszczyk, Michał Nadolny, Marcin Hernes, Agata Kozina, and Bogdan Franczyk ("Machine Learning for Prediction of the Importance of Factors Influencing Prosumer Attitudes") develop a method for classifying prosumer survey respondents depending on their method of assessing each decision-making factor separately using selected machine learning and deep learning methods. The results research show that there is a possibility of predicting the importance of decision-making factors based on respondents' characteristics. It may be especially important for smaller companies that do not have much money or time to take large market surveys.

Part III is devoted to intelligent multiple-criteria decision analysis methods.

Bartłomiej Kizielewicz, Arkadiusz Marchewka, and Wojciech Sałabun ("RRF-EDAS—An Extended Approach Free from the Rank Reversal Paradox") developed Rank Reversal Free EDAS (RRF-EDAS) approach to address rank reversal paradox issue. Results show that RRF-EDAS effectively eliminates rank reversal, providing more consistent and reliable rankings than the classical EDAS method.

Paula Bajdor ("MCDM Approach to Quality Assessment of Functioning of e-Commerce Platforms Operating in Poland") presents the possibility of using a multi-criteria approach to evaluate the quality of functioning of e-commerce platforms. The obtained results indicated the completeness of criteria set identified, as all of them are relevant to assess the quality of functioning of e-commerce platforms.

Aleksandra Bączkiewicz, Jarosław Wątróbski, Anna Shkurina, and Wojciech Książek ("Multi-Criteria Evaluation of Floating Photovoltaics Considering a Strong Sustainability Paradigm Regarding Sustainable Solutions") developed the Strong Sustainability Paradigm based Technique for Order Preference by Similarity to Ideal Solution (SSP-TOPSIS) method for modeling the degree of compensation reduction of criteria according to the strong sustainability paradigm. The results confirmed the suitability of the proposed multi-criteria evaluation approach in intelligent decision supporting requiring consideration of compensation reduction and different preferences.

Andrzej Greńczuk, Kamila Łuczak, Iwona Chomiak-Orsa, Estera Piwoni-Krzeszowska, Sudha Prathyusha Jakkaladiki, Martina Janečková, and Dominik Palla ("Transforming the Tutor and Tutee Matching Process Using ICT: The Problems of Multi-Criteria Decision Analysis") developed a methodology for transforming the tutor and tutee matching process with a multi-criteria analysis using ICT. The study contributes to the existing body of knowledge on the integration of ICT in educational processes, particularly in improving the accuracy of decision-making through MCDA.

Aleksandra Bączkiewicz, Jarosław Wątróbski, Robert Król, and Rafał Pawlak ("Sustainable Development Strategies Assessment Using the New Multi-Target TOPSIS Method ") propose a novel MT-TOPSIS method that allows one to take into account the ideal reference points assigned individually for each evaluated alternative, taking into account its current objectives and the possibility of achieving them. The results

demonstrate the complexity of evaluating sustainable development involving multiple targets and how to take them into account.

We want to express our gratitude to all the authors for their interesting, novel, and inspiring contributions. Peer reviewers also deserve deep appreciation because their insightful and constructive remarks and suggestions have considerably improved many contributions. Last but not least, we wish to thank Prof. Janusz Kacprzyk for his help in implementing and finishing this large publication project on time and maintaining the highest publication standards.

September 2024

Marcin Hernes
Jaroslaw Wątróbski
Artur Rot

Organization

International Programme Committee

Wolfgang Dorner	Technische Hochschule Deggendorf, Germany
Marcin Fojcik	Western Norway University of Applied Sciences, Norway
Mykola Dyvak	Ternopil National Economic University, Ukraine
Bogdan Franczyk	University of Leipzig, Germany
Jan Stępniewski	Université Paris 13, France
Philippe Krajsic	University of Leipzig, Germany
Artur Rot	Wrocław University of Economics, Poland
Ewa Walaszczyk	Wrocław University of Economics, Poland
Wiesława Gryncewicz	Wrocław University of Economics, Poland
Monika Eisenbardt	Wrocław University of Economics, Poland
Dorota Jelonek	Czestochowa University of Technology, Poland
Paweł Weichbroth, Gdańsk	University of Technology, Poland
Dragan Pamučar	University of Belgrade, Serbia
Muhammet Deveci	University College London, United Kingdom
Witold Chmielarz	University of Warsaw, Poland
Mykola Dyvak	West Ukrainian National University, Ukraine
Adrianna Kozierkiewicz	Wrocław University of Science and Technology, Poland
Kesra Nermend	University of Szczecin, Poland
Boleslaw Borkowski	Warsaw University of Life Sciences, Poland
Andriy Melnyk	West Ukrainian National University, Ukraine
Paweł Ziemba	University of Szczecin, Poland
Artur Karczmarczyk	West Pomeranian University of Technology, Poland
Zbigniew Pastuszak	Maria Curie-Sklodowska University, Poland
Joanna Kołodziejczyk	National Institute of Telecommunications, Poland
Richard Chbeir	University of Pau and the Adour Region, France
Joanna Paliszkiewicz	Warsaw University of Life Sciences, Poland
Wojciech Sałabun	West Pomeranian University of Technology, Poland
Józef Żurada	University of Louisville, USA

Contents

Application of Artificial Intelligence in Information Systems

A Dual LSTM-Based Multimodal Method For Fake News Detection

Huyen Trang Phan[1] and Ngoc Thanh Nguyen[2(✉)]

[1] Faculty of Information Technology, HCMC University of Technology and Education, Ho Chi Minh, Vietnam
trangpth@hcmute.edu.vn

[2] Department of Applied Informatics, Wroclaw University of Science and Technology, Wroclaw, Poland
Ngoc-Thanh.Nguyen@pwr.edu.pl

Abstract. Along with the appearance and development of many social networking platforms in many different fields, there is an explosion of freely shared information. The benefits of these sources of information depend on whether the information is real or fake. The problem today is that we cannot distinguish factual information from fiction when receiving information on social networking platforms. To solve this problem, many different approaches to detect fake news have been proposed with promising performance, most notably approaches based on shared textual content. However, previous methods based on text content focus too on extracting features within the news's leading content and consider the news title only as a relevant feature. This is not true for the habits of some users, who only read the news title without paying attention to the content. To overcome this limitation, we propose a dual BiLSTM-based multimodal method to detect fake news. The proposed method includes three main elements: (i) a BiLSTM-based fake headline extractor to capture and represent features of context and semantics on the headlines, (ii) a BiLSTM-based fake body extractor to capture and represent features of context and semantics on the news bodies, and (iii) a dual BiLSTM-based multimodal fake news detector to combine two fake headline and body features based on early fusion mechanism. The difference in the proposed method is that we do not consider the headline simply as one of the features; we view it as an independent text containing important features of context and semantics. The proposed model was tested on ISOT and FA-KES datasets and achieved better results than previous methods regarding the same content-based multimodal approach.

Keywords: Fake news detection · Dual-BiLSTM · Fake headline extractor · Fake body extractor · Multimodal fake news detector

1 Introduction

Along with technology development comes the birth and development of social networking platforms, leading to an explosion of shared news. These news are almost free, and whether they bring benefits or harm depends on whether the information is real or

M. Hernes et al. (Eds.): ECAI 2024, LNNS 1217, pp. 3–14, 2025.
https://doi.org/10.1007/978-3-031-78465-1_1

fake. Fake news is often information that is not true. They are created by fabricating or distorting real news to spread information contrary to real news [1]. Fake news often contains metadata including source, headline (title), body content (text), image/video, or links. The purpose of fake news is to propagate false things, aiming to change them into real news so basically, they violate ethics and law, so stopping their spread is necessary [2].

Many different fake news detection methods have been proposed and are usually classified into four methods: context-based, propagation-based, multilabel learning-based, content-based, including knowledge-based and style-based, and hybrid-based [3]. Among them, the content-based category is one of the dominant and exciting groups. The ideas in this category are proposed based on fake news's authenticity and news characteristics. Specifically, knowledge-based methods aim to employ external sources to fact-check news text, and the objective of style-based techniques is to capture the distinct writing style of fake news. Many approaches to detect fake news have been proposed, such as using deep learning, machine learning, knowledge graphs, and graph neural networks, which provide promising performance, especially transformer models.

Why are transformer models used so much in developing fake news detection methods? First, they are highly parallelizable, meaning they can simultaneously process multiple sequence parts, significantly speeding up the training and inference process. Additionally, they can capture long-term textual dependencies, allowing them to understand context better and produce more coherent text. They are also more flexible and scalable, making adapting them to different domains and datasets easier. For example, Ma et al. [4] modeled the sequential relationship between news using recurrent neural networks. Khattar et al. [5] proposed a multimodal variational autoencoder to extract hidden multimodal representations of multimedia news. Nasir et al. [6] used a charged neural network to establish high-level relationships between news. Transformer based fake news detection includes single-modal and multimodal methods.

Transformer-based multimodal fake news detection methods on news content have been quite successful, such as [7–12]. Some basic information of the mentioned methods is listed in Table 1.

Table 1. Some transformer-based multimodal fake news detection methods

Method	News part	Combination of Models
Hybrid CNN-RNN [7]	Body	CNN, RNN
LR, SVM, NB [8]	Body, Headline	LR, SVM, NB
DSS [9]	Body	propagation tree, stance network
UMLARD [10]	Body	view-wise attention network, capsule attention network
Das et al. [11]	Body	statistical feature fusion network novel heuristic algorithm
FND-NS [12]	Body	transformer encoder and decoder

Table 1 shows some of the most recent multimodals for detecting fake news on text data type. We can see that most previous methods only focus on representing and analyzing the news body; only one method [8] considers the role of the headline, but it only uses the headline as an additional feature. This is not close to the real situations of many users, who only read the news title without paying attention to the news body. We propose a Dual BiLSTM-based multimodal method (DuBiM) to overcome this limitation, and DuBiM is experimented only on text datasets. The DuBiM model includes the following steps: (i) Build two separate models to detect the fake body and fake headline based on BiLSTM, (ii) Build a multimodal fake news detection method by combining a fake body detection model and a fake headline detection model using the early fusion mechanism, (iii) Train the model on benchmark datasets to compare and evaluate the achieved performance. Our contributions are:

- We construct two separate feature extractor of fake news detection model called BiLSTM-based fake body feature extractor and BiLSTM-based fake headline feature extractor.
- We propose a novel dual BiLSTM-based multimodal method (DuBiM) by fusion of BiLSTM-based fake body feature extractor and BiLSTM-based fake headline feature extractor.
- We conduct extensive experiments on benchmark datasets to evaluate effectiveness of the DuBiM model and variants.

We organize the remainder of this paper as follows. Section 2 presents an overview of related research. Section 3 describes the research problem of the proposal. Section 4 offers mathematical models for solving the research problem step by step. The data acquisition, setting up, experimental results, and evaluation of the proposal is provided in Sect. 5. Conclusions and research directions for the future are presented in Sect. 6.

2 Related Work

In this section, we focus on surveying the most recent multimodal methods to detect fake news.

Ajao et al. [7] proposed a framework to detect and classify fake news in Twitter posts using a combination of CNNs and LST-RNN models. This approach can intuitively capture relevant features related to fake news without requiring domain knowledge. Agarwalla et al. [8] combined three conventional machine learning models and the Lidstone smoothing technique to the body and headline to find a reliable and accurate model for fake news detection. Although this method has considered the role of the headline, the authors only use the headline as a relevant feature without building a separate extractor to extract the context and semantics hidden in that headline entirely. Davoudi et al. [9] proposed a DSS model that includes three main components: RNN-based dynamic analyzer to encode over time the evolutionary patterns of the propagation tree and the stance network; fully connected network-based static analyzer to represent all features extracted by RNN-based dynamic analyzer accurately; the node2vec algorithm to encode the structure of RNN-based dynamic analyzer during the structural analysis. This model can detect early fake news effectively. Das et al. [11] proposed a robust framework for

identifying fake news on sensitive topics by combining pre-trained language models based on an ensemble mechanism with soft voting. The system with a statistical feature association network and a new heuristic-based post-processing algorithm is also applied to integrate statistical features, e.g., source, username, URL domain, and author, into the news content features. Nasir et al. [6] proposed a hybrid neural network model combining CNN and an RNN. The Conv1D layer extracted local features from the input text vector. This vector is then fed into the RNN layer of the following LSTM units/cells to learn local features extracted from the previous layer and extract their long-term dependencies.

From the above surveys, we have some observations as follows:

- All methods have achieved expected results, but no method has achieved absolute results, meaning improvements are still needed to improve performance.
- All five methods surveyed are multimodal based on deep learning and machine learning and tested on text data. Multimodals are built by combining different models; no method combines dual BiLSTM.
- Only one of the five methods considers headline content as input text. Still, headline content is extracted as one of the relevant features that must be added to enrich the news body representation without being considered an independent text.

3 Research Problem

The main goal of this research is to propose a new multimodal fake news detection method by combining fake body and fake headline detection models to improve performance in terms of accuracy and loss. The highlight of the proposed method is the ability to take advantage of the content of both news body and headline. The research problem of this article is presented as follows:

Given a set of n news $A = \{a_i | i \in [1, n]\}$ and $a_i = \{t_i, h_i, l_i\}$, where t_i, h_i, l_i are news body, headline, and label, respectively. This article considers this task a binary classification problem, meaning that each sentence can be assigned one of two labels: fake and real. In other words, assuming we have a news t_i and a h_i, we use $l_i \in R^{m \times 1}$ to denote the new labels. $l_i = 0$ if t_i and h_i contain characteristics of fake news. $l_i = 1$ if t_i and h_i contain features of real news. With the above symbols, the article's research problem is as follows: *Given news body a, news headline h and vector labeled l. This research aims to propose a new multi-model fake news detection method by combining fake body and fake headline detection models to predict the* $\hat{l}$ *vector of news that is not labeled effectively.*

4 Proposed Method

The input of the DuBiM model is a set of news. For each news, the DuBiM includes the main steps are: (i) a BiLSTM-based fake headline extractor to capture and represent features of context and semantics on the headlines, (ii) a BiLSTM-based fake body extractor to capture and represent features of context and semantics on the news bodies, and (iii) a dual BiLSTM-based multimodal fake news detector to combine two fake headline and body features based on early fusion mechanism. Output is a set of news that are assigned labels. The workflow of DuBiM is illustrated in Fig. 1.

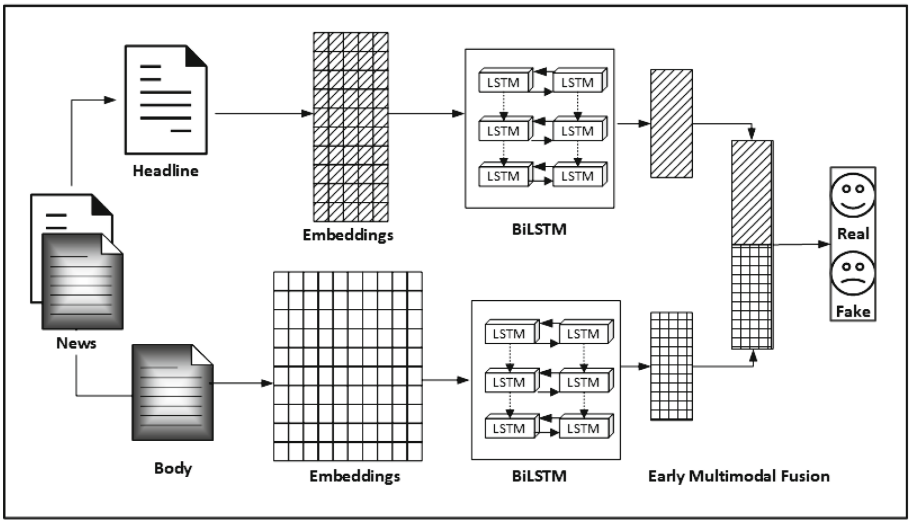


Fig. 1. Workflow of DuBiM model.

4.1 BiLSTM-Based Fake Headline Feature Extractor

Input layer: The input of this extractor is headline $h = \{w_j^h | j \in [1, m_1]\}$ where w_j^h is the j-th word in headline h and m_1 is the number of words in the headline.

Embeddings layer: This layer is to convert the headline into a matrix of word vectors. For $w_j^h \in h, j = [1, m_1]$, each word w_j^h is mapped to a vector, denoted by $v_j^h \in R^d$, where d is dimension of vector v_j^h. The value of vector v_j^h is identified as follows:

$$v_j^h = GloVe(w_j^h) \tag{1}$$

where $GloVe(w_j^h)$ is GloVe embeddings [13].

The output of this layer is an embeddings matrix $v^h \in R^{m_1 \times d}$.

BiLSTM layer: This layer extracts the most significant context and semantics in the headline following both directions: left to right using a forward LSTM and right to left using a backward LSTM [14]. That means each word vector v_j^h has a pair of hidden vectors $\overrightarrow{h_j}$ and $\overleftarrow{h_j}$. This process is determined as follows:

$$\overrightarrow{h_j} = \overrightarrow{lstm}(W_{v^h \overrightarrow{h}} v_j^h + W_{\overrightarrow{h}\overrightarrow{h}} \overrightarrow{h}_{j-1} + b_{\overrightarrow{h}}) \in R^{d_h}, j = [1, m_1] \tag{2}$$

$$\overleftarrow{h_j} = \overleftarrow{lstm}(W_{v^h \overleftarrow{h}} v_j^h + W_{\overleftarrow{h}\overleftarrow{h}} \overleftarrow{h}_{j+1} + b_{\overleftarrow{h}}) \in R^{d_h}, j = [1, m_1] \tag{3}$$

$$h_j = W_{\overrightarrow{h}\,h} \overrightarrow{h_j} \oplus W_{\overleftarrow{h}\,h} \overleftarrow{h_j} + b_h \tag{4}$$

where W is the weight matrix; for example, $W_{v^h \overleftarrow{h}}$ is the weight matrix between the input vector and forward hidden vectors. h_j is the contextualized representation. h_{j+1} is

the next hidden vector of h_j and $h_{m_1+1} = 0$. h_{j-1} is the previous hidden vector of h_j and $h_0 = 0$.

The output of this layer is a significant context and semantics features extracted from the headline $h = \{h_j | j \in [1, m_1]\} \in R^{m_1 \times d}$. The vector of significant context and semantics features extracted from the headline is formed by linearizing this matrix as $h^a = linear(h)$.

4.2 BiLSTM-Based Fake Body Feature Extractor

Input layer: The input of this extractor is news body $t = \{w_j^t | j \in [1, m_2]\}$ where w_j^t is the j-th word in news body t and m_2 is the number of words in the news body.

Embeddings layer: This layer is also to convert the news body into a matrix of word vectors. For $w_j^t \in t, j = [1, m_2]$, each word w_j^t is mapped to a vector, denoted by $v_j^t \in R^d$, where d is dimension of vector v_j^t. The value of vector v_j^t is identified as follows:

$$v_j^t = GloVe(w_j^t) \tag{5}$$

where $GloVe(w_j^t)$ is GloVe embeddings [13].

The output of this layer is an embeddings matrix $v^t \in R^{m_2 \times d}$.

BiLSTM layer: This layer extracts the most significant context and semantics in the news body following both directions: left to right using a forward LSTM and right to left using a backward LSTM [14]. That means each word vector v_j^t has a pair of hidden vectors $\overrightarrow{t_j}$ and $\overleftarrow{t_j}$. This process is determined as follows:

$$\overrightarrow{t_j} = \overrightarrow{lstm}(W_{v^t \overrightarrow{t}} v_j^t + W_{\overrightarrow{h} \overrightarrow{t}} \overrightarrow{t}_{j-1} + b_{\overrightarrow{t}}) \in R^{d_t}, j = [1, m_2] \tag{6}$$

$$\overleftarrow{t_j} = \overleftarrow{lstm}(W_{v^t \overleftarrow{t}} v_j^t + W_{\overleftarrow{t} \overleftarrow{t}} \overleftarrow{t}_{j+1} + b_{\overleftarrow{t}}) \in R^{d_t}, j = [1, m_2] \tag{7}$$

$$t_j = W_{\overrightarrow{t} t} \overrightarrow{t_j} \oplus W_{\overleftarrow{t} t} \overleftarrow{t_j} + b_t \tag{8}$$

where W is the weight matrix; for example, $W_{v^t \overleftarrow{t}}$ is the weight matrix between the input vector and forward hidden vectors. t_j is the contextualized representation. t_{j+1} is the next hidden vector of t_j and $t_{m_2+1} = 0$. t_{j-1} is the previous hidden vector of t_j and $t_0 = 0$.

The output of this layer is a matrix $t = \{t_j | j \in [1, m_2]\} \in R^{m_2 \times d}$. The vector of significant context and semantics features extracted from the news body is formed by linearizing this matrix as $t^a = linear(t)$.

4.3 Dual BiLSTM-Based Multimodal Fake News Detector

This detector is designed to combine and learn the features of context and semantics information from two extracted feature vectors to capture higher-level features to improve the performance of fake news detection. This detector includes the following steps.

Input layer: The input of this layer is two vectors t^a and h^a.

Early fusion layer: Early fusion [15] is more potent since it merges data sources at the beginning of the processing. In this study, early fusion combines two vectors of

significant context and semantics extracted from the news body and headline to extract highlight-level multimodal features as follows:

$$g = fusion(t^a \oplus h^a) \tag{9}$$

where *fusion* is the linear layer-based early fusion, $\oplus$ is a combination operation, g is the vector of highlight-level multimodal features.

Multimodal fake news classifier: The multimodal features vector is fed into the fully connected layer, and the softmax function calculates the output of the fully connected layer to identify the distribution value of news label as follows:

$$p = softmax(W \cdot g + b) \tag{10}$$

where W and b are a weight matrix and a bias of the softmax function.

Model training: The proposed model is trained by minimizing the cross-entropy error of the predicted and true label distributions as the following equation:

$$\hat{l} = -\sum[ylogp + (1-y)log(1-p)] \tag{11}$$

where y indicates the real distribution value of the news label, and p is the predicted distribution value of the news label.

5 Experiment Results

5.1 Datasets

The ISOT[1] dataset includes 44,919 news items distributed almost evenly between real and fake categories. Genuine articles were collected from the Reuters website, and fake articles from various sources were flagged as fake sources by Wikipedia and Politifact. The datasets include the full content of each news, title, date, subject, and labels. Each news is assigned two labels: 0 (indicating "fake") or 1 (indicating "real"). The FA-KES [16] dataset includes 804 news about the war in Syria. Each news contains content, title, date, location, news source, and label. Each news is also assigned one of two labels: 0 (indicating fake) or 1 (indicating real). However, to comply with the proposal, only three components of each news are used in this study: news content, title, and label.

The detailed dataset is listed in Table 2.

Table 2. Experimental dataset

Dataset	Total	Fake	Real
ISOT	44,919	23,502	21,417
FA-KES	804	378	426

Table 3. Hyperparameters of the proposed method

Key	Value
Training set	$n\times$ 80%
Testing set	$n\times$ 20%
Way to split dataset	Random
Epochs	20
Hidden size	50
Optimization	Adam
Learning rate	0.001
Criterion	Cross Entropy Loss

n is the number of news

5.2 Experimental Setting

The hyperparameters of the proposed method are listed in Table 3.

Baseline methods: In this study, we use the methods analyzed and evaluated in Sect. 2 as baselines.

Ablation methods: To confirm the necessity of all four types of extracted features to build a set of rules used to determine the sentiment of conditional sentence types, we also proceed to develop ablation methods of the proposed method as follows:

- Ablation#1: BiLSTM-based fake headline detection indicates the ablation method by considering only the content of the headline.
- Ablation#2: BiLSTM-based fake body detection indicates the ablation method by considering only the content of the body.

Evaluation metrics: We used the accuracy as the metric to evaluate and compare the performance of our proposed on the mentioned dataset. Accuracy is calculated by comparing the actual and predicted test set values.

5.3 Results and Discussions

General Results: The results of the proposed method is shown in Figs. 2 and 3, Tables 4 and 5.

Table 4. Training time

Dataset	CPU	GPU
ISOT	162.0s	51.3s
FA-KES	64.0s	14.6s

[1] https://www.uvic.ca/engineering/ece/isot/datasets/fake-news/index.php.

Looking at Table 4, we see that the training time of the proposed method is relatively fast when implemented on both CPU and GPU. Especially the ISOT dataset on GPU; although the dataset contains a relatively large number of samples, the training time is still only in seconds. The training time proves that the time complexity of the proposed method is quite reasonable.

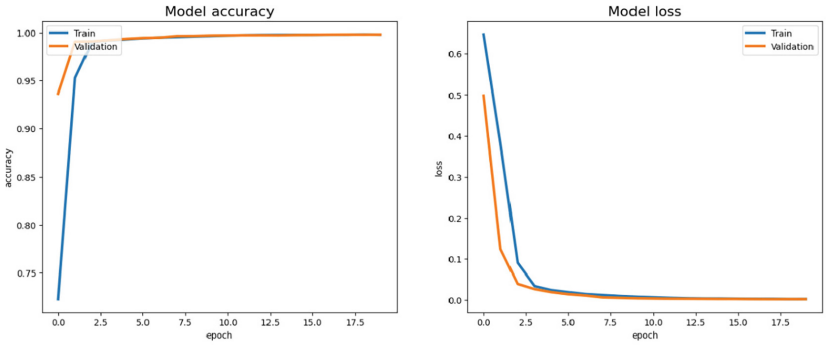


Fig. 2. Training performance of the proposed method on ISOT

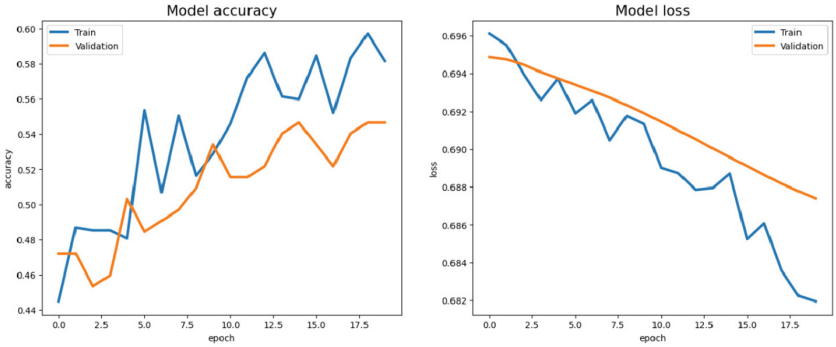


Fig. 3. Training performance of the proposed method on FA-KES

Figures 2 and 3 show that when training, the accuracy and loss values the proposed method achieves on the ISOT dataset are better and more stable than on the FA-KES dataset. At the same time, the number of samples on the ISOT dataset is five times larger and contains fewer noise components than the FA-KES data. This proves that the proposed method suits big and low-noise datasets.

Table 5 presents the testing performance of the proposed method on both baseline datasets. Like the training performance, we see that the testing performance achieved on the ISOT dataset is much better than the FA-KES dataset. Although both datasets are balanced in terms of the number of samples labeled “fake” and “real,” with the ISOT dataset, the proposed method achieves stable performance for both “fake” and “real.” In contrast, the FA-KES dataset performs better with “real” labeled sample groups. This confirms that the proposed method suits large and less noisy datasets.

Table 5. Testing performance of the proposed method

Metric	ISOT		FA-KES	
	fake	real	fake	real
precision	1.00	1.00	0.48	0.56
recall	1.00	1.00	0.15	0.87
f1 score	1.00	1.00	0.23	0.68
accuracy	1.00		0.55	

Ablation Results: Table 6 presents the performance achieved by the proposed method compared to its pruning versions.

Table 6. Testing performance of ablation methods

Metric	ISOT		FA-KES	
	loss	accuracy	loss	accuracy
Ablation#1	0.0086	0.9977	0.6898	0.5155
Ablation#2	0.0077	0.9973	0.6927	0.5280
Proposed method	0.0013	0.9982	0.6874	0.5466

Looking at the data, we can see that similar to the proposed method; the ablations method also achieves better and more stable performance on the ISOT dataset. The reason is that the ISOT dataset is more extensive and less noisy than FA-KES. Although the performance difference between the methods on the ISOT dataset is not high, this difference is shown quite clearly on the FA-KES dataset, which is enough to demonstrate the necessity of analyzing both body and headline content in improving fake news detection performance.

Comparison Results: Table 7 presents the testing accuracy of the proposed method compared to the baseline methods on two datasets. Note that we acknowledge the results reported by the authors who proposed the baseline methods without performing a reproduction.

Looking at the data in the table, we can see that on the ISOT dataset, the proposed method achieves better accuracy than the most recent modern multimodal-based methods. This once again confirms that (i) building a separate model to extract and learn contextual and semantic features from news headlines can potentially improve the performance of fake news detection methods; (ii) BiLSTM can well extract contextual and semantic features in text data. However, with the FA-KES dataset, the proposed method achieves poorer results than the Hybrid CNN-RNN method [6]. This again confirms the proposed method's limitation, which is poor performance on small and noisy datasets.

Table 7. Testing accuracy of methods

Method	ISOT	FA-KES
Hybrid CNN-RNN [7]	0.83	–
LR, SVM, NB [8]	0.83	–
DSS [9]	0.95	–
Das et al. [11]	0.94	–
Hybrid CNN-RNN [6]	0.99 ± 0.02	0.60 ± 0.007
Proposed method	1.00	0.55

6 Conclusion

This paper has presented a multi-model fake news detection method based on text data, called the Dual-BiLSTM-based multimodal method, to simultaneously take advantage of contextual and semantic knowledge learned shown in news bodies and headlines to improve the performance of previous methods. The proposed method includes three main parts: a fake headline feature extractor to extract fake information expressed in headlines, a fake body feature extractor to extract fake information expressed in bodies, and a multi-model fake news detector by fusion two extractors of fake headline feature and fake body feature. The proposed method is tested on two benchmark datasets. Overall, the proposed method achieves promising results on the ISOT dataset and comparable results on the FA-KES dataset. After evaluating the results, the paper also presents the advantages and disadvantages of the proposed method. In terms of advantages, the proposed method achieves high and stable performance on large datasets in terms of accuracy, loss, and training time. Regarding disadvantages, the proposed method is quite sensitive to noisy data and unsuitable for small datasets.

References

1. Phan, H.T., Nguyen, N.T., Hwang, D.: Fake news detection: a survey of graph neural network methods. Appl. Soft Comput., 110235 (2023)
2. Phan, H.T., Nguyen, N.T., Hwang, D.: Content-context-based graph convolutional network for fake news detection. In: International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, pp. 571–582. Springer (2022)
3. Zhou, X., Zafarani, R.: A survey of fake news: fundamental theories, detection methods, and opportunities. ACM Comput. Surv. (CSUR) **53**(5), 1–40 (2020)
4. Ma, J., et al.: Detecting rumors from microblogs with recurrent neural networks (2016)
5. Khattar, D., Goud, J.S., Gupta, M., Varma, V.: MVAE: multimodal variational autoencoder for fake news detection. In: The World Wide Web Conference, pp. 2915–2921 (2019)
6. Nasir, J.A., Khan, O.S., Varlamis, I.: Fake news detection: a hybrid CNN-RNN based deep learning approach. Int. J. Inf. Manag. Data Insights **1**(1), 100007 (2021)
7. Ajao, O., Bhowmik, D., Zargari, S.: Fake news identification on twitter with hybrid CNN and RNN models. In: Proceedings of the 9th International Conference on Social Media and Society, pp. 226–230 (2018)

8. Agarwalla, K., Nandan, S., Nair, V.A., Hema, D.D.: Fake news detection using machine learning and natural language processing. Int. J. Recent. Technol. Eng. (IJRTE) **7**(6), 844–847 (2019)
9. Davoudi, M., Moosavi, M.R., Sadreddini, M.H.: DSS: a hybrid deep model for fake news detection using propagation tree and stance network. Expert Syst. Appl. **198**, 116635 (2022)
10. Chen, X., Zhou, F., Trajcevski, G., Bonsangue, M.: Multi-view learning with distinguishable feature fusion for rumor detection. Knowl.-Based Syst. **240**, 108085 (2022)
11. Das, S.D., Basak, A., Dutta, S.: A heuristic-driven uncertainty based ensemble framework for fake news detection in tweets and news articles. Neurocomputing **491**, 607–620 (2022)
12. Raza, S., Ding, C.: Fake news detection based on news content and social contexts: a transformer-based approach. Int. J. Data Sci. Anal. **13**(4), 335–362 (2022)
13. Pennington, J., Socher, R., Manning, C.D.: Glove: global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543 (2014)
14. Zhou, P., et al.: Attention-based bidirectional long short-term memory networks for relation classification. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (vol. 2: short papers), pp. 207–212 (2016)
15. Gadzicki, K., Khamsehashari, R., Zetzsche, C.: Early vs late fusion in multimodal convolutional neural networks. In: 2020 IEEE 23rd International Conference on Information Fusion (FUSION). IEEE, pp. 1–6 (2020)
16. Salem, F.K.A., Al Feel, R., Elbassuoni, S., Jaber, M., Farah, M.: FA-KES: a fake news dataset around the Syrian war. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 13, pp. 573–582 (2019)

Conception of Intelligent Systems for the Creation Sustainable Use of Human Resources in the IT Sector

Anna Sołtysik-Piorunkiewicz[1,2(✉)] and Witold Chmielarz[1,2]

[1] University of Economics in Katowice, 1 Maja 50, 40-287 Katowice, Poland
anna.soltysik-piorunkiewicz@uekat.pl, witek@wz.uw.edu.pl
[2] University of Warsaw, Krakowskie Przedmieście 26/28, 00-927 Warsaw, Poland

Abstract. The main goal of this article is to present the concept of an intelligent system for creating sustainable development of labor resources in the employment market in Poland. The study is of a pilot nature and covers three basic elements of such a system: the general concept of the system, the framework of the basic model of shaping the employment market, and the conceptualization of a two-factor mechanism for balancing this market. To achieve the aim of the article, conceptual modeling was used based on the authors' practical experience, identification of employment model factors based on previous research on the career of an IT specialist, and their own experience using the production function to support sustainable development. The original achievement of this article is the association of these three main elements of the constructed system, which, after systematic and consistent supplementation of its elements, can be used by both academics and business activists for research on the sustainable development of the labor market in the IT sector in Poland.

Keywords: Intelligent System · Sustainable Development · IT Employment Model

1 Introduction

The concept of sustainable development of the labor market in the IT industry in Poland is one of the most important problems in the current development of our country. The economic crisis that has lasted for many years, deepened by the COVID-19 pandemic, has resulted in high, although decreasing, inflation and imbalance in the labor market. This particularly applies to the market related to modern technologies, and especially to people working in the IT sector. Without the development of this labor market, investment in both "clean" technology, and, above all, in the staff who will implement and maintain this development at all times, it is impossible to quickly lift Poland out of the existing crisis. At the same time, to global trends and European Union policy, it should be a sustainable, conscious technological development of an intensive nature ensuring the use of IT in at least three domains: the natural environment (preventing activities destructive to the natural environment, such as excessive deforestation, development of

M. Hernes et al. (Eds.): ECAI 2024, LNNS 1217, pp. 15–26, 2025.
https://doi.org/10.1007/978-3-031-78465-1_2

coal-fired power plants), related to including meeting social needs, including ensuring work in professions that do not pose a threat to the environment (e.g. IT) and guarantee constant future progress and transformation towards sustainable development, as well as satisfying the organizational framework of such processes (efficient administration, capable of managing changes).

There is an employment gap of 150,000 in the Polish IT sector, and if it is not eliminated on time, it will increase in the future [9]. The proposals to avert this danger [15] do not seem up-to-date or comprehensive in the current situation because they do not include, for example, graduates of all fields of study, and in particular, they pay too little and incompetent attention to women as potential employees of the IT sector [7].

For this reason, the implementation of an intelligent system of a sustainable strategy, taking into account the most important factors affecting employment in the IT sector, should result in ensuring an increase in employment in this industry related to the latest technologies, which will influence the direction of future investments that are relatively harmless to the environment and provide jobs for the current and future generations. It also seems that in Poland's current situation, this system can be implemented with relatively low expenditure, compared to investments that destroy the environment.

That is why the main goal of the article is to build a framework for a management IT system in which a mechanism for intelligent balancing of factors related to human resources and investments in the latest information and communication technologies will be proposed. One of the elements of such a system should be a general model of the impact of factors on employment in the IT sector and a proposal of a method to solve the problem of employment balance in the IT sector in the context of balance between personal and technical-organizational factors.

Since this article is part of the research on the career of an IT specialist, the structure of the model and the most important factors influencing employment in IT will refer to the research results [3].

The intelligent mechanism for balancing factors influencing employment in the IT sector is based on a two-factor VES production function. Paul Samuelson, defining the production function (…) The production function is the technical relationship between physical inputs and mental outputs, ceteris paribus, pointed out that it describes (…) the maximum feasible output of a firm or an economy, for every possible combination of inputs given existing knowledge and technology (…) [19]. Based on this definition and remembering that the production function determines the optimal use of personal, organizational, and technological factors, a form of an original function was proposed to solve this problem, which could serve as a reference point for the results obtained and forecasts of substitute development in this field.

To achieve the above-defined goal, the following structure of the article was adopted. After introducing the topic and specifying the purpose of the work, the second section deals with a literature review on the factors, models, and the role of AI that influence decision-making and shape the labor market in the IT sector. The next section presents a method for solving the problem, which consists of presenting a decision-making model framework. Based on previous research, the fourth section presents a structural model for identifying factors influencing employment in the IT sector in

Poland and a solution enabling achieving optimal use of factors characterizing personal and technical-organizational resources in terms of broadly defined sustainability.

The research procedure adopted to solve the problem was as follows:

1. Literature analysis of the issue using conventional methods and searching the Internet by keywords.
2. Building a framework for an intelligent decision support system.
3. Creating a procedure for the operation of the system.
4. Construction of a structural model of the IT market based on previously conducted research.
5. Creating a procedure for solving the problem of balancing factors affecting the employment market in Poland.
6. Summary, conclusions, and further directions of work on the development of the system.

The implementation of these subsequent stages will allow for the creation of solid foundations for building a comprehensive intelligent system supporting a sustainable market in the IT sector in Poland.

2 Literature Background

The management information system is treated here as an integrated computer-based network, along with surrounding applications, which allows the collection and analysis of data to help managers make management decisions. That is a system that is designed to provide past, present, and future information appropriate for planning, organizing, and controlling the operations of functional areas in an organization [5]. In practice, MIS consolidates raw data from multiple sources, turn it into useful information, and distribute customized reports to final users [12]. Elements of intelligent decision support appeared in expert models, and their greatest development so far has been in the structures of business intelligence systems.

Therefore, a system equipped with elements of artificial business intelligence is referred to here as a system that extracts (…) valuable data from various enterprise systems, facilitating storage, analysis, and internal data management (…) BI systems consist of three main components: data collection, organized presentation, and efficient delivery to those in need (…) [10]. It consists of three main elements connected by relationships: data warehouse, BI software, and users with appropriate analytical. (…) Business intelligence is the process by which enterprises use strategies and technologies for analyzing current and historical data, with the objective of improving strategic decision-making and providing a competitive advantage (…) [11]. In our case, it is planned to create appropriate business analytics that will determine options for an optimal, sustainable IT employment strategy in Poland.

The definitions of business intelligence systems and their applications depend on the needs of the company distributing this type of system, the industry, the size of the project, and individual, specific areas of application, which is why the definitions of these systems and their analyst differ significantly [8, 12]. Nevertheless, the basic elements and their relationships in all these systems perform similar functions defined

by convergence. However, business intelligence systems for shaping employment policy in the IT industry are rarely found in source materials.

The second important aspect discussed in the article is the issue of sustainable development. The classic approach to a sustainable economy seeks solutions that integrate its three basic dimensions: social, environmental, and economic, which will result in an improvement in the living conditions of society, without violating the quality of the environment [14]. These three dimensions translate into specific goals: preservation of water and air biodiversity; minimizing atmospheric pollution and waste production; increasing human and intellectual capital; improving social capital - flow between professions, environments, and regions; dynamization of economic development, among other modern technologies; development of national and international institutions that promote democracy, transparency and support sustainable development [18]. The development of the IT sector is, in a way, inherently part of the sustainable development of the entire country: it is pro-ecological (low pollution, limited resource consumption in remote work, rapid development of the modern economy, ensuring lasting, continuous, and well-paid work in the long term, etc. However, the literature so far did not see significant connections between the development of the IT sector and sustainable development, focusing on agriculture, ecology, and transforming the functioning of administration [24].

The third aspect is modeling the relationship between the increase in the number and quality of employment in the IT sector and the determinants of this development. The model is always a certain reflection of reality, omitting some less important details and focusing on the most important factors and their mutual relations. There are numerous examples in the literature of the impact of various factors on the development or potential development of employment in the IT sector. However, no comprehensive model is determining their impact on employment in the entire sector.

The article proposes a combination of the structural model and the balanced model. Referring to P. Samuelson's statement quoted at the beginning, the production function model may become such a model. Balance in this case is achieved by determining the optimal proportion between environmental, organizational, and technological factors in the development of the IT industry. Variants of this development are possible due to the assumed substitutability between these two components.

Assuming this assumption for the construction of the mechanism, we can choose from three basic Cobb-Douglas production functions, CES (Constant Elasticity of Substitution) and VES (Variable Elasticity of Substitutions) [23].

Cobb-Douglas function [22]:

$$P_1 = A * \left(K^{\alpha} * L^{\beta}\right) \tag{1}$$

where: P - total product, K - capital input, L - labor input, α - product elasticity concerning capital, β - product elasticity concerning labor, A - constant indicating technological progress - scale parameter, efficiency of the production process; α and β – scale effects. In the classical Cobb-Douglas function, $\alpha + \beta = 1$, which results in no scale effects. Production substitution = 1.

This is a two-factor function, where instead of labor input and capital input, we can substitute other categories. CES function, generalization of the Cobb-Douglas function

[23]:

$$P_2 = A * \left(\alpha K^{-\sigma} + \beta L^{-\sigma}\right)^{-v/\sigma} \quad (2)$$

where: $\beta = 1 - \alpha$; $\delta = 1/1 - \sigma$; σ-elasticity of substitution, v- parameter of scale effects, generally = 1,

VES function, generalization of the CES function [17, 19]:

$$P_3 = A\left[(1+\beta) * K * L^{\beta} + \alpha L^{1+\beta}\right]^{1/1+\beta} \quad (3)$$

where: α and β generalized parameters of the shares of production factors K and L. Currently, there is no standardized VES function; its above form was presented in a two-factor form by Beckman and Sato [1]. Alternative parametric forms of the VES production function have been derived, analyzed, and tested on various datasets, including Revankar [19]. To ensure a balanced set of factors influencing the IT labor market, the authors used their form of the VES function.

The use of various forms of production functions mainly concerns research on industrial sectors and short-term forecasts of both the entire economy and individual companies. There have been attempts to use them in the sphere of finance [4] and in the economy [2]. However, there is no application of the production function to the analysis of the employment market, especially due to its sectoral balance. In the situation of the economic crisis in Poland, the most generalized form of the VES production function seems to be the most suitable for undertaken research, providing a flexible approach to the dynamically changing economic situation [6].

The above review shows that the available literature did not take into account the creation of a system supporting optimal strategic decisions in the area of employment in the IT sector supported by an intelligent sustainable development mechanism. Therefore, the following research questions arise:

RQ1 – Is it possible to build an intelligent system supporting decision-making in the area of shaping employment in the IT sector?
RQ2 – What factors influence the employment model in the IT sector?
RQ3 – Is it possible – within the model – to create an intelligent mechanism supporting decisions in the field of sustainable development based on the VES-type production function?

The following sections will try to answer these questions.

3 Methods of Solving the Problem: Building a Structural Employment Model

3.1 Research Sample

The data on which the construction of the structural model was obtained from a survey questionnaire in the IT career study [4].

The characteristics of the research sample obtained at four Polish universities: A, B, C, and D are presented below. The study was limited to organizational units related to

the field of economics and management and economic informatics. In total, over 800 students were surveyed in the first half of 2023, in randomly selected years of study. At the university level, the study was conveniently conducted by limiting it to faculties related to management, economics and economic informatics. At the level of individual years of study and student groups, it was conducted randomly, by submitting proposals to complete the survey in student groups available at the time of the study.

Demographic data was collected by: gender, age, education and its field, place of origin, and financial status. Generally, based on the collected data it can be concluded that:

- the faculties of distinguished universities related to management are dominated by women - on average they constitute over 60% of the population (men - almost 40%), which is typical for studies of this type, only D differs from this pattern due to the IT-related specializations existing at this university (63% men, 37% women),
- the largest difference calculated by Euclidean distance (almost 30%) occurs between C, and slightly smaller (29%) between A and D.
- most of the respondents (almost 90%) are aged 19–24, which is also typical for this population, only over 10% are aged 25–34 and above, the largest (3.68%) variation calculated by the Euclidean distance occurs between A and C,
- according to declarations about studies in a specific year of studies, almost 56% of respondents have secondary education, basic vocational education (mostly C - 78%), and over 34% have a bachelor's degree (mostly A - almost 53% - due to conducting surveys during studies second level). For this reason, the greatest difference (30%), calculated by the Euclidean distance, occurred between C and D.
- as expected, the dominant field of education was social sciences - almost 54% and exact and technical sciences - on average almost 36% (mostly for D, almost -74%, due to the survey of students of econometrics and economic informatics),
- on average, most people (34%) come from cities with over 200,000 inhabitants. Inhabitants (mainly C - 60% - due to the central location) and from rural areas - on average over 34% (mainly due to A - 58%). In these two categories, the greatest difference measured by the Euclidean distance occurs between A and C,
- financial status is rated as good on average (53%), although in second place almost 21% of respondents described it as bad or average. Here, in the very good and good categories, C dominates (70% in total) due to the greatest employment opportunities for students. Students B assess their financial situation the worst. The greatest number of respondents (almost 28%) assessing their financial situation as very good are in D. The greatest variation (2.2% Euclidean distance) occurs here between A and D.

3.2 Elements Influencing Employment in the IT Sector

This section proposes a model of solutions that, thanks to the appropriate organizational redirection of funds and social preferences, can fill the employment gap and put the IT industry on the path of intensive development. This will also allow us to understand the importance of personal development in the IT industry for sustainable development focusing on increasing social well-being while protecting the natural environment. We dedicate this publication, especially to people who have problems understanding how the

development of modern technologies in one country, projected on a regional or global scale, can contribute to the creation of a sustainable society.

Actions are recommended that can balance the growing demand for people working in the IT profession with the possibilities of meeting it in connection with the requirements of the rapid development of modern societies. The basis for achieving this goal is to understand the development of employment in IT by identifying the areas in which the greatest problems with employing broadly understood IT workers have so far occurred; educational opportunities for people who could be employed in these areas; changes in education methods corresponding to the demand for IT employees; attracting social groups that have not been involved in this industry so far and identifying areas where, for social reasons, there is the greatest demand for this type of employees. In other words, demonstrating that balance in the IT labor market serves to achieve sustainable development in the environmental, social, and economic areas. Identification of these attributes will constitute a basis for proposing framework solutions that can systemically solve not only the problem of the employment gap in the IT sector, but also ensure constant, long-term, sustainable development of the entire country and region.

The most important determinants of employment in IT are, of course, personal reasons. Their subsequent employment depends on the quantity, quality, and preferences of potential staff. On the other hand, employment in the IT labor market is strongly dependent on organizational and technological conditions.

There is substitutability between the groups of factors determining personal and intellectual resources and technological and organizational resources constituting the infrastructure of possible employment in IT. An optimal balance between these two groups may result not only in sustainable employment in the IT sector, but also through synergy - in other areas of the economy, especially those that are the newest and most technologically developed, and therefore environmentally friendly.

3.3 Identification of Structural Model Variables

Based on the research [4], we can build a model of the formation and growth of the labor market in the IT sector.

The final results of the research were influenced by the fact that it was conducted with a view to a possible career in the IT sector for students and graduates of economic universities. All four analyzed universities or their departments were universities/faculties of economics/management. Therefore, the average respondent was a student (70%), aged 19–24 (85%), with completed secondary education (64%), mainly a student working without a contract (44%) in the second year of studies (41%).), majoring in social sciences (45%), coming from a city with over 200,000 inhabitants. People who consider their financial status to be good (53%). Nevertheless, the opinion of the remaining 30% also significantly influenced the results.

The set of factors influencing the employment of this group will consist of two basic groups: personal factors and organizational and technological (infrastructural) factors. Each of these basic groups can be divided into specific factors (percentage of acceptance by respondents in brackets).

Personal factors (55.2% importance according to respondents):

- experience - seniority, and experience (4.3% among personnel; 2.4% in total),
- security - certainty of finding a job now and in the future (4.6%; 2.5%),
- education – the possibility of continuous development, acquiring new skills (2.5%; 1.4%),
- convenience - ability to work remotely, irregular working hours (6.8%; 3.8%),
- existing competencies - use of office applications, analysis of mass data, knowledge of application programming (8.2%; 4.5%),
- desired competencies - knowledge of Business Intelligence (in the field of e.g. reporting and analysis of business data, CRM, etc.), programming skills, knowledge of modifying and developing applications (programming - 8.4%; 4.7%),
- expected benefits (high wages with a rapid upward trend (6.2%; 3.4%),
- emotions – the desire to start a career as an IT specialist: answer - yes, sometimes, and sometimes (13.9%; 7.7%),
- conditions facilitating becoming an IT specialist - a position with current competencies (consultant/manager in the field of business contacts, project team leader, consultant tester – a specialist in detecting and eliminating substantive and technical errors (8.9%; 4.9%)) and after training (analyst business data (Business Intelligence Analyst), IT security analyst (IT Security Analyst) 6.4%;
- substantive and psychophysical predispositions according to respondents - quick acquisition of new information and IT knowledge, ability to work in a project team, openness to non-standard methods, mental resistance to real or apparent shortcomings (13.9%; 7.7%),
- substantive and psychophysical predispositions according to employers - technical skills, ability to think logically and solve problems independently (7.5%; 4.2%),
- relationship of work effort in a specific position to expected salary - programmer, technical support, back office - EUR 3,000, project team leader, a team of 5–15 people - EUR 5,000, business analyst, system analyst - EUR 4,000 (8.2%; 4.5%),
- organizational and technological (infrastructural) factors - 44.8% of importance according to respondents:
- progressive organizational and social changes caused by IT - IT specialists will be divided into many distant professions with IT, technical, economic, sociological, and psychological background, continuous education process, knowledge is improved throughout life, only remote work (15.4%; 6.9%),
- knowledge of current technology - office software, integrated system subsystems, use of Business Intelligence (e.g. reporting and analysis of business data, - 14.6%; 6.5%),
- gaps in knowledge of current technology – ability to program applications, ability to design IT systems, general theoretical knowledge of IT systems and their applications (11.2%; 5.0%).
- the need to supplement knowledge about technology - methods of human-computer interaction, in the field of Business Intelligence Systems - the value of information and its interpretation in business, The use of mobile applications in society (m-commerce, m-marketing, m-location, etc. - (10.1%; 4.5%),
- technological and social trends – creation and use of intelligent medical, transport, and media equipment; constantly increasing speed of information flow - Big Data, progressive automation of all areas of life, readiness to accept the use of AI (20.5%; 9.2%),

– resistance to the introduction of technology - high level of stress during the implementation of urgent and non-traditional projects, lack of permanent working hours, congestion during the implementation of urgent projects, inconvenience of contact with the client (10.1%; 4.5%),
– trends in organizational and social changes - key skills in information analysis, knowledge optimization, concluding, ability to quickly adapt; constant changeability of technology and organization; information and knowledge are the key resources of the organization (18.0%; 8.1%).

4 Construction of an Intelligent System Supporting Employment Decisions in the IT Sector

The proposed IT system framework for making decisions regarding the optimal proportions of increasing intellectual human resources versus implementing the most modern technological and organizational resources can be supported by an AI system that coordinates and improves the appropriate selection of factors regulating employment in the IT labor market. The main elements of the logical architecture of such a system placed on the central administration platform would be reduced to at least the following constructs:

– Coordinator - equipped with a management cockpit - an interactive intranet service that provides such basic functionalities as a list of available attributes describing human resources; a list of possible available technological and organizational resources; a list of decision-making situations used so far (procedures and algorithms); communication with other organizational partners influencing employment in the IT industry; categorization of participants based on specified attributes and recommendations of cooperation partners; access to patterns of conduct implemented and collected in the knowledge base and the possibility of modification to a limited extent; the possibility of creating acceptable scenarios for the development of their distribution; preview of the results obtained in various cross-sections and their graphic illustration, i.e. visualization in any cross-section. We will equip you with a query and suggestion system as well as access to communication media,
– Accommodator – an AI auxiliary module used to compare the planned development variant with those proposed by other partners, with an algorithm allowing to development of a compromise solution; after conducting an estimate, sending proposals to the coordinator module. The moderator of a given project may take these suggestions into account or use his judgment,
– Identifier - tools for describing and quantifying the problem: attribute, functional, and process - creator of goal(s), development theses and hypotheses and their opportunities based on statistical analyses, with access to patterns of conduct contained in the knowledge base and the ability to create reference lists (ChatGPT),
– Generator and translator of forms for new solutions - a tool based on a knowledge base, allowing the selection of the entire template and/or its components (based on existing patterns from the past), regarding the declared direction of decision-making, equipped with (or with access to the existing) in language translator. The generator should also be equipped with access to demographic standards and statistical data, as well as a transformer and converter to other international standards,

- Modeler - a module for creating and modifying development models of the optimal structure of the labor market and IT technological and organizational resources and their adaptation. Additionally equipped with models for the adoption and assimilation of the latest ICT technologies to the needs of a future employee of the IT sector, in conditions ranging from standardized to high risk, equipped with a Selector tool for categorizing data and an Evaluator to assess the significance of models and their possible changes,
- Knowledge base - a set of models used and the results of their operation - patterns of behavior in decision-making situations,
- Solver - a computational module - equipped with a package of statistical and numerical methods for assessing data obtained from survey questionnaires.
- Associator – tools for automatic association of data and models based on the user's knowledge and mechanisms of referring to patterns contained in the knowledge base,
- Concluder - a subsystem redirecting to calculations on the solver or adopting a solution indicated by a pattern from the knowledge base.
- Databases of interactive models – supplemented with relationships created in the modeler between model elements, appropriate for the types of research.

It is obvious that human cooperation with decision support systems, especially AI modules, can significantly improve the decision-making process. However, this cannot be an automatic process (e.g. AI is always right) for many reasons. Firstly, the IT system(s) used for this cooperation should be actually involved in the important problems of project management versus automatic copying of activities performed traditionally. Secondly, the processes taking place there should be continuously improved, integrated, and implemented as quickly as possible. Thirdly, be flexible - towards dynamic changes in the rules for achieving success in a rapidly changing reality. Fourthly, be resistant to attempts to manipulate individual partners for one-sided goals, and therefore rather be based not only on the optimization of goals but also on Pareto-optimal equilibrium.

The authors proposed future research based on this framework to create a social adaptation model for the IT market in Poland, using or supplementing one of the theories of acceptance models (i.e. TAM, UTAUT, UTAUT2) or creating a new model to contain any premises for pursuing a sustainable employment policy. The model should be implemented in the interactive model database, and equipped with a balancing mechanism in the knowledge base, with AI backend. In its current form, it does not contain the mechanism for ensuring substitution equilibrium, which will be presented for the implication of the research for science and practice more deeply.

5 Conclusions

The considerations presented above on the construction of an IT system that will support strategic decision-making regarding employment in the IT sector constitute the first stage of research on the sustainable development of this industry. The above article formulates a general concept of such a system containing elements of artificial intelligence and establishes a procedure for using it in the decision-making process, which answers the first research question (RQ1). Moreover, based on previous research, factors influencing

the employment of an economic IT specialist in Poland were identified in a broad economic and social context. This provides the basis for formulating a structural or adaptive model. This answered the second question (RQ2). To formulate the mechanism of sustainable development of this market, a new form of the VES class production function was used, along with a general interpretation of its coefficients, which should be treated as a preparatory stage for further research on methods for assessing the proportion of labor market development in the IT sector. The substitution discovered - during individual expert assessments - and confirmed by their subsequent opinions between the factors influencing the quantitative results of this market, seems so important that it requires additional simulation research on the employment market in the IT sector (RQ3).

This study is a pilot study extending previous research on career opportunities for an IT specialist in Poland. This is where the limitations of this study come from. The first is the fact that it covers only three key areas: creating a logical structure of an intelligent IT system, identifying factors of the IT employment market development model, and the concept of a two-factor function that can constitute the basis for creating simulation variants of sustainable employment market development strategies. However, there are no relations and elements connecting these three main areas between them, which should constitute the basis for further research in this direction.

References

1. Beckmann, M.J., Sato, R.: Aggregate production functions and types of technical change: a statistical analysis. Am. Econ. Rev., 88–101 (1969)
2. Chmielarz, W., Stachurski, A.: A class of VES production function: properties and estimation results. Control Cybern. **3–4**, 367–381 (1986)
3. Chmielarz, W., Zborowski, M., Jelonek, D., Sołtysik-Piorunkiewicz, A., Wiechetek, Ł: Świadomość możliwości kształtowania kariery informatyka. Analiza porównawcza na wybranych uczelniach w Polsce, Problemy Organizacji **3**, 185–196 (2023)
4. Chmielarz, W.: VES function (variable elasticity of substitution) in the evaluation of factors substitution shaping a website's usability. In: Korczak, J., Dudycz, H., Dyczkowski, M. (eds.) Advanced Information Technologies for Management –AITM 2008. Research Papers Wrocław University of Economics, no. 35. Publishing House of Wrocław University of Economics, Wrocław, pp. 47–57 (2008)
5. Girdhar, J.: Management Information Systems, p. 328. Oxford University Press, New Delhi (2013)
6. Grassetti, F., Mammana, Ch., Michetti, E.: Sustainability between production factors and growth. An analysis using VES production functions. Chaos Solitons Fractals Nonlinear Sci. Nonequilib. Complex Phenom. **113**, 53–62 (2018)
7. http://www.dziewczynynapolitechniki.pl/pdfy/raport-kobiety-na-politechnikach-2022.pdf. Accessed 1 Aug 2023
8. https://nofluffjobs.com/pl/log/praca-w-it/business-intelligence-czy-to-dobra-sciezka-kariery/. Accessed 1 Aug
9. https://pie.net.pl/wp-content/uploads/2022/11/PIE_Raport_Ilu-specjalistow-IT-brakuje-w-Polsce.pdf. Accessed 1 Aug 2023
10. https://www.finereport.com/en/bi-tools/business-intelligence-system.html. Accessed 1 Aug 2023
11. https://www.heavy.ai/technical-glossary/business-intelligence. Accessed 1 Aug 2023

12. https://www.sap.com/poland/products/technology-platform/cloud-analytics/what-is-business-intelligence.html. Accessed 1 Aug 2023
13. https://www.shopify.com/blog/what-are-management-information-systems. Accessed 1 Aug 2023
14. Keiner, M.: Re-emphazing sustainable development – the concept of evolutionability. Environ. Dev. Sustain. **6**, 381 (2004)
15. Kemme D.M.: Discriminating among alternative production functions in Polish industry. Empir. Econ. **9**, 59–66 (1984)
16. Laudon, K.C., Laudon, J.P.: Management Information Systems: Managing the Digital Firm, 11 edn., p. 164. Prentice Hall/CourseSmart (2009)
17. Lovell, C.A.: Knox, CES and VES production functions in a cross-section context; source. J. Polit. Econ. **81**(3), 705–720 (1973). The University of Chicago Press A Class of Variable Elasticity of Substitution Production Functions
18. Quental, N., Lourenco, J.M.: References, authors, journals and scientific disciplines underlying the sustainable development literature: a citation analysis. Scientometrics **90**, 362 (2012)
19. Revankar, N.: Econometrica Econom. Soc. **39**(1), 61–71 (1971)
20. Samuelson, P.: Economics. An Introductory Analysis. 13 edn., Chapter 9, p. 372 (1989)
21. Samuelson, P.: Nordhaus, W.: Ekonomia, Wydawnictwo Naukowe PWN, Warszawa (2007)
22. Solow, R.M.: A contribution to the theory of economic growth. Q. J. Econ. **70** (1956)
23. Voosholz, F.: The influence of different production functions on modeling resource extraction and economic growth CAWM, Discussion Paper no. 7 (2014)
24. Turban, E., Pollard, C.: Gregory Wood Information Technology for Management: Driving Digital Transformation to Increase Local and Global Performance, Growth and Sustainability, 12th edn. Wiley, Hoboken (2021)

Eye Tracking Insights: Analyzing Cognitive Load Across Media Types

Anna Borawska and Adrianna Mateja(✉)

University of Szczecin, Cukrowa 8, 71-004 Szczecin, Poland
adrianna.mateja@usz.edu.pl

Abstract. In the realm of digital communication, understanding how individuals process information across various media types is crucial. This paper explores cognitive load across different media types, including images and videos. Drawing on the Limited Capacity Model of Motivated Mediated Message Processing and Cognitive Load Theory, the study employs eye-tracking technology alongside traditional methods to measure cognitive load. Our study aims to explore the relationship between eye tracking and cognitive load across different media types. The study's novelty lies in its use of eye-tracking technology combined with traditional methods to offer a comprehensive understanding of visual attention and cognitive processing in social campaigns. The authors were motivated to explore this area to enhance the effectiveness of digital communication strategies by assessing how images and videos impact cognitive load. By leveraging eye-tracking technology, the study provides precise insights into visual attention and cognitive effort, ultimately improving the effectiveness of digital communication strategies.

Keywords: cognitive overload · eye tracking study · media types · visual attention

1 Introduction

In the era of digital communication, which is constantly evolving and bringing a variety of content and messages, understanding how people process information becomes a key element [1]. The Limited Capacity Model of Motivated Mediated Message Processing (LC4MP) suggests that the human brain has limited resources for processing information [2], which can lead to cognitive overload [3], especially with an excessive influx of messages [4]. On the other hand, the Cognitive Load Theory [5] indicates that there is a certain point beyond which an individual's ability to process information becomes limited, leading to difficulties in understanding and retaining the message.

Social campaigns are an important tool for promoting positive societal changes [6]. They build and increase awareness of many important and significant issues, influencing the perceptions and attitudes of the audience [1]. In today's world, where communication primarily occurs through various media channels such as television, radio, the internet, and print, social campaigns strive to reach audiences through different communication

M. Hernes et al. (Eds.): ECAI 2024, LNNS 1217, pp. 27–38, 2025.
https://doi.org/10.1007/978-3-031-78465-1_3

channels [7], so studying individuals' reactions to these messages becomes extremely important in this context. There are several factors that can influence the cognitive load induced by social advertisements. These factors include the way messages are edited [8], the presence and nature of emotional content [9], the quantity and strength of arguments [10], the format of the message [11], and its structure [12]. Considering the LC4MP model and the cognitive load theory, it becomes essential to examine media messages in terms of the cognitive load they induce and their attractiveness to the audience.

Previous research on cognitive overload has mainly focused on phenomena such as fatigue resulting from information overload on social media [13], the influence of users on advertising effectiveness [14], and information overload [15]. Conducted experiments confirm that high problematic involvement improves message processing, and thus may result in increased or decreased acceptance of persuasive message content [16]. Based on numerous survey, questionnaire, and focus group studies [4, 15, 17] cognitive overload has been identified as a significant barrier to evaluating the effectiveness and assimilation of media messages [18]. Despite the fact that the scale developed by Paas [19] is the most popular method for measuring cognitive load, in recent years, increasing attention has been paid to the use of new, more objective research methods, such as eye tracking [20, 21]. This technique can capture the dynamics of social campaign content as it provides continuous data unlike traditional methods [22]. Eye tracking enables precise monitoring of participants' eye movements, allowing for the recording of which content elements they focus on, how long they concentrate on them, and how quickly they shift their gaze between different areas of the content [21, 23]. In many previous experiments examining cognitive overload and advertising effectiveness (engagement with the advertisement as well as the intensity and arousal of emotions elicited by the advertisement), this tool has been used to measure consumers' unconscious reactions [24].

The aim of the article is to investigate the relationship between eye tracking and cognitive load by examining various types of media. The key research questions were formulated as follows:

1. Which type of multimedia (images or videos) exerts the greatest cognitive load?
2. Is there consistency between subjective and objective measures?
3. Is there a correlation between cognitive load level and subjective attractiveness of stimuli?

To address these research questions, an eye tracking study was conducted among students, utilizing various forms of social campaign advertisements. This paper begins with an introduction to the importance of understanding cognitive load in digital communication. In the next section details of an experimental case study using Polish-language social advertisements are described. The results section presents findings from both subjective and objective measures, highlighting correlations and differences in cognitive load and attractiveness between images and videos. The conclusion summarizes key insights, discusses the implications for media consumption and social campaigns, acknowledges study limitations, and suggests directions for future research.

2 Experimental Case Study

2.1 Stimuli

In our study we used publicly available social advertisements promoting safe driving, which were created as part of previous campaigns in Poland. This is because all participants in the experiment were Polish, and we wanted to make sure that they clearly understood the content of the ads in their native language. We deliberately avoided using foreign languages to prevent additional cognitive load that could impact the experiment results. The experimental materials consisted of five images and four videos aimed at investigating the impact of different modalities on cognitive load. All stimuli aimed to reduce the number of accidents caused by using a phone while driving. Each stimulus was presented for a similar duration, approximately 30 s, to facilitate better comparison. The materials were divided into two categories (images and videos). Both stimuli within each category and the categories themselves were presented randomly to avoid biased assessments.

2.2 Participants

As with many traditional laboratory studies, we recruited students as a pool of important experiment participants [25] as they often represent the age group targeted by social campaigns [26–29]. Additionally, statistical analysis regarding road accidents [30, 31] and mobile phone usage while driving suggests that young drivers, who often include students, are a significant target group concerning these issues.

The study involved 33 participants, comprising 25 males and 8 females aged between 22 and 24 years. This sample size aligns with the standard for research utilizing cognitive neuroscience methods [21, 32]. The average age was 23 years, with a standard deviation of 0.72. The participant pool encompassed both licensed drivers (94%) and non-drivers, with 64% of the participants being car owners. On average, participants had possessed a driver's license for 4 years, with a standard deviation of 1.43. One in four participants drove a car daily, while over half drove at least once a week or more frequently.

2.3 Procedure

The research took place at the Cognitive Neuroscience Laboratory and received approval from the Bioethics Committee of the Regional Medical Chamber. Each participant spent an average of 25 min undergoing the research procedure. Before starting, participants were informed about the research goals, procedures, and measurement tools, and then gave their informed consent to participate. To ensure precise measurement, the study initiated with eye tracker calibration.

As part of the experiment, participants were tasked with familiarizing themselves with materials from social campaigns regarding safe driving and subsequently assessing the level of mental effort required to understand the presented content as well as the attractiveness of the presented stimuli. The initial phase involved presenting stimuli while concurrently gathering real-time psychophysiological data. At the beginning, a

relaxation slide was presented to calm the participants. Following this, stimuli from various categories were presented randomly. Post each stimulus presentation, participants completed a questionnaire evaluating their cognitive load and rating the visual appeal of the displayed content. Intervals of relaxation were interspersed between stimulus groups to mitigate cognitive overload before the next set. The subsequent phase involved completing a survey to collect demographic information. Participants disclosed their reasons for using phones while driving and subjectively assessed different forms of material presentation in social campaigns.

2.4 Data Collection and Processing

Eye movements were analyzed using a Tobii Pro X3-120 commercial eye tracker, operating at a sampling rate of 120 Hz [33]. The eye tracker was affixed to the lower portion of a 15-inch monitor with a resolution of 1920 × 1080 pixels (Full HD), which presented stimuli and questionnaires. Data visualizations such as heatmaps and areas of interest were promptly generated using the iMotions software [34].

2.5 Measures

The study used traditional and psychophysiological methods. The cognitive load assessment utilized a 9-point Paas scale [19], ranging from minimal (1 - Very, very low mental effort) to maximal mental effort (9 - Very, very high mental effort). The specific question used to evaluate cognitive load was: "Invest your understanding into the presented material". Subjective attractiveness ratings of the stimulus were made on a 10-point scale, with value 1 indicating poor and 10 indicating very good attractiveness. This assessment approach is commonly utilized in research [35]. The analysis of data collected using an eye tracker included various measures such as saccades (frequency, amplitude, duration, velocity), fixations (duration, number, frequency), pupils (pupil diameter) and blinks (frequency, duration, number). These selected measures are frequently used in the literature for studying cognitive load [36–39]. They are categorized into two groups: the first group consists of metrics where higher values indicate increased cognitive load, and the second group comprises metrics where lower values are interpreted as indicators of greater cognitive load (Table 1).

Table 1. Groups of eye-tracking metrics in the context of cognitive load assessment

Group Description	Metrics Included
Higher Values = Increased Cognitive Load	Saccade frequency [40], saccade amplitude [41], saccade duration [42], saccade velocity [42], fixation duration [43], fixation number [44], fixation frequency [45], pupil diameter [46], blink duration [42]
Lower Values = Increased Cognitive Load	Blink frequency [42], blink number [42]

3 Results

The analysis of the questionnaire revealed that videos are deemed the most effective medium in social campaigns by 85% of respondents. Approximately half of the participants see images as helpful in conveying information.

To assess the internal consistency of the questionnaire data, Cronbach alpha's [47] value was calculated. For the ratings in the Paas scale the obtained result was 0.83 and for the attractiveness ratings – 0.82. This indicates excellent internal consistency since both values are greater than 0.8. To assess the significance of the assessment difference between the various groups of stimuli (images and videos) Mann-Whitney test was applied, since the normality Shapiro-Wilk test has indicated that collected data in majority (for both Paas and attractiveness ratings) do not have normal distribution.

The Mann-Whitney test showed that there is no significant difference between the median assessment of perceived cognitive load induced by different types of stimuli using the Paas scale ($U = 357.5$, $p = 0.0842$). On the other hand, considering the attractiveness ratings, the differences between medians for images (median = 6) and videos (median = 7) are statistically significant ($U = 318.0$, $p = 0.0225$), indicating that videos were assessed as more appealing than images.

We examined the correlations between attractiveness and Paas assessments to determine if more cognitively demanding stimuli were seen as more attractive. Except for image-1, correlation coefficients were positive but low, indicating a weak to moderate correlation (Fig. 1).

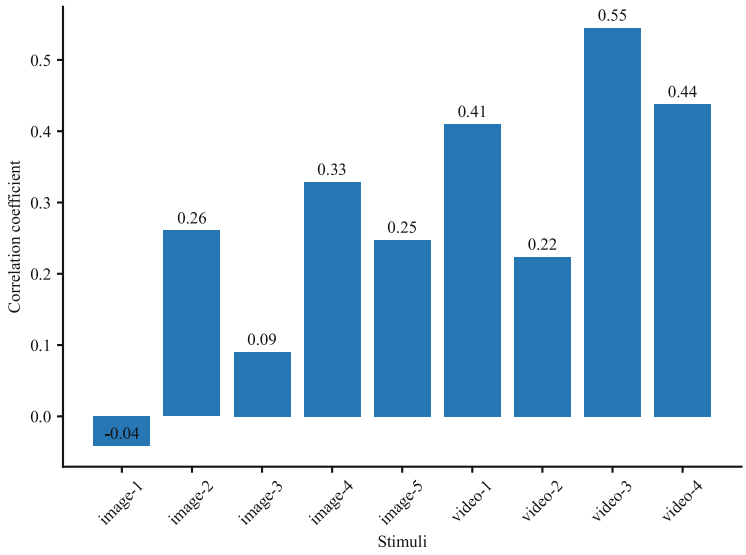


Fig. 1. Correlation between attractiveness and Paas ratings.

Supporting analyses of eye tracking data were conducted to compare their consistency with the results obtained by subjective measures. Before performing any statistical

test, eye tracker data were normalized using Z-score based on the baseline activity (relaxation) to allow for comparison of cognitive load between stimulus groups at the level of the entire group of participants.

Eye tracking data were also tested in terms of the normality of their distribution with the use of Shapiro-Wilk test. Out of 99 series of data only 46 have normal distribution (46.5%). Due to this fact the rest of statistical analyses concerning the eye tracking were conducted using non-parametric tests.

To compare the results of subjective and objective measures of cognitive load and to check whether they are consistent with each other, Spearman's correlation was used. The results in the form of the heatmap are shown in Fig. 2. For most of the eye tracking measures there is no correlation with Paas assessment, or it is weak. In general, these results suggest that objective measures yield different outcomes compared to subjective ones, and there appears to be a weak relationship between the variables.

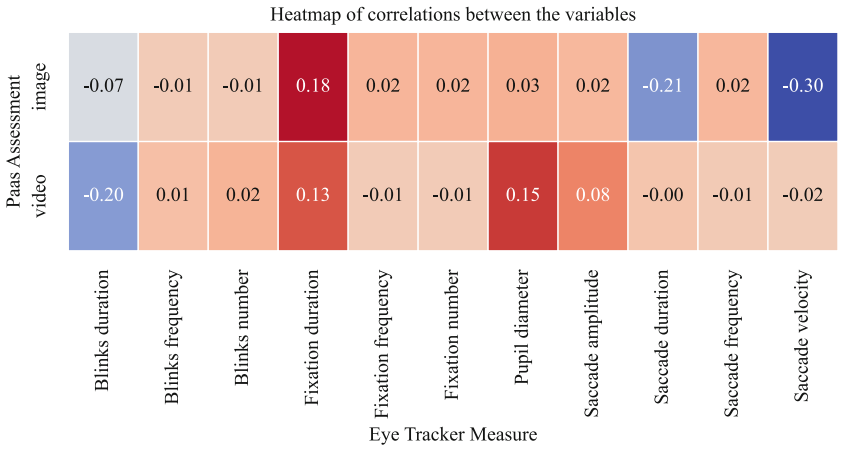


Fig. 2. Spearman's correlation for Paas assessment and eye tracking measures.

Table 2 presents the overall cognitive load for all examined stimuli. The stimuli are categorized into two types: "v" for video and "i" for image. The positions in parentheses indicate their ranking. Number 1 signifies that the stimulus is the most cognitively demanding in terms of the eye tracking measure. To analyze which measures indicate higher cognitive load based on rankings, we have used the median of the ranks occupied by images and videos. A summary based on the median rank will allow us to determine which stimuli are systematically ranked higher, suggesting greater cognitive load. This analysis reveals that certain eye-tracking metrics suggest higher cognitive load when participants view images compared to films. Metrics indicating a higher cognitive load for images include fixation frequency, blink frequency, saccade frequency, blink number, and blink duration, with median rankings predominantly positioned at 3.0. These metrics highlight the repetitive visual scanning and increased blink activity associated with processing static images, which may require more effort to interpret detail and context compared to dynamic stimuli.

Conversely, metrics such as saccade amplitude, fixation duration, saccade duration, fixation number, saccade velocity, and pupil diameter suggest a higher cognitive load for films, with median ranks varying from 2.5 to 3.5. These findings could indicate that dynamic visual stimuli, like films, demand more extensive visual tracking and sustained attention, leading to larger and more frequent eye movements and possibly greater pupil dilation as viewers try to capture and process moving images and changing scenes.

Table 2. Mean values of eye-tracking measures for each stimulus

Metric	Stimulus								
	v-1	v-2	v-3	v-4	i-1	i-2	i-3	i-4	i-5
Saccade amplitude	−0.135 (2)	−0.219 (3)	−0.249 (4)	**0.019 (1)**	−1.042 (9)	−0.765 (6)	−0.877 (8)	−0.437 (5)	−0.868 (7)
Fixation duration	−0.345 (4)	**−0.189 (1)**	−0.412 (7)	−0.281 (2)	−0.339 (3)	−0.369 (6)	−0.576 (9)	−0.436 (8)	−0.368 (5)
Saccade duration	0.123 (3)	0.124 (2)	−0.021 (7)	0.103 (4)	−0.245 (9)	−0.103 (8)	0.069 (5)	**0.217 (1)**	−0.020 (6)
Fixation frequency	0.415 (8)	0.376 (9)	0.521 (3)	0.466 (5)	0.516 (4)	0.455 (6)	**0.770 (1)**	0.453 (7)	0.576 (2)
Blink frequency	0.742 (7)	0.663 (6)	0.999 (9)	0.984 (8)	0.209 (3)	0.585 (4)	0.088 (2)	0.608 (5)	**−0.172 (1)**
Saccade frequency	0.399 (8)	0.361 (9)	0.502 (4)	0.448 (7)	0.507 (3)	0.451 (5)	**0.756 (1)**	0.448 (6)	0.558 (2)
Fixation number	**5.403 (1)**	4.656 (2)	0.532 (5)	0.477 (7)	0.516 (6)	0.455 (8)	0.770 (3)	0.453 (9)	0.576 (4)
Blink number	6.245 (9)	5.308 (8)	1.011 (7)	0.997 (6)	0.209 (3)	0.585 (4)	0.088 (2)	0.608 (5)	−0.172 (1)
Saccade velocity	−0.102 (3)	−0.179 (4)	−0.247 (5)	**−0.007 (1)**	−0.562 (9)	−0.398 (7)	−0.354 (6)	−0.060 (2)	−0.453 (8)
Blink duration	−0.600 (8)	−0.445 (6)	−0.407 (5)	−0.756 (9)	0.416 (2)	0.248 (3)	0.056 (4)	**0.480 (1)**	−0.479 (7)
Pupil diameter	**−0.373 (1)**	−0.600 (7)	−0.458 (4)	−0.442 (3)	−0.515 (5)	−0.392 (2)	−0.755 (9)	−0.564 (6)	−0.724 (8)

v – video.
i - image.

Table 3 provides additional insights into the frequency of occurrence of stimuli at specific ranks in cognitive load measures. Stimuli v-1, v-4, i-3, i-4, and i-5 more often occupy the top positions in the ranking, suggesting their dominant role in eliciting high cognitive load. Images have a slightly higher average number of top-ranked occurrences (positions 1–3) on cognitive load measures than videos (3.8 vs. 3.5). This suggests that images may be slightly more cognitively taxing than videos.

To statistically determine which type of stimuli induces more cognitive load the Mann-Whitney tests were conducted to determine whether the difference between the medians of images and video groups are statistically significant. The tests were performed for each of the eye tracking measures separately. The difference between the medians was significant in case of three different metrics: saccade amplitude, fixation duration and blink number (see Table 4). The first two measures show that videos are more cognitively demanding. The third indicates contrary conclusion. Since the results are not unambiguous, further investigation in that matter would be advisable.

Table 3. The number of occurrences of stimuli in specific places in the cognitive load measures ranking

Rank	Stimulus								
	v-1	v-2	v-3	v-4	i-1	i-2	i-3	i-4	i-5
1	2	1	0	2	0	0	2	2	2
2	1	2	0	1	1	1	2	1	2
3	2	1	1	1	4	1	1	0	0
4	1	1	3	1	1	2	1	0	1
5	0	0	3	1	1	1	1	3	1
6	0	2	0	1	1	3	1	2	1
7	1	1	3	2	0	1	0	1	2
8	3	1	0	1	0	2	1	1	2
9	1	2	1	1	3	0	2	1	0

v – video.
i - image.

Table 4. Results of Mann-Whitney tests for groups of stimuli (images and videos) in various measures of cognitive load calculated from eye-tracker data

Metric	U-stat	P-value	Median - images	Median - videos
Saccade amplitude	0.0	**0.016**	−0.951	−0.344
Saccade duration	2.0	0.063	−0.515	−0.286
Fixation duration	1.0	**0.032**	−0.594	−0.445
Blink number	1.0	**0.032**	−0.019	2.417
Blink frequency	3.0	0.111	−0.019	0.369
Saccade frequency	12.0	0.730	0.702	0.486
Fixation number	6.0	0.413	0.721	2.943
Blink duration	18.0	0.063	0.258	−0.552
Pupil diameter	8.0	0.730	−0.620	−0.505
Fixation frequency	12.0	0.730	0.721	0.496
Saccade velocity	2.0	0.063	−0.700	−0.402

4 Conclusion

The article examines how individuals process information in the digital age, with particular emphasis on social campaigns. According to the LC4MP model and cognitive load theory, the importance of understanding cognitive load induced by media messages and their impact on audience reception is emphasized.

The literature review and study analysis reveal that stimuli vary in cognitive load, crucial for creating multimedia content. No simple rule determines the highest cognitive load, but certain tendencies exist. For example, longer saccade and fixation durations, higher fixation counts, and lower blink frequency may indicate greater cognitive effort.

Our study indicates that images may induce greater cognitive load than videos, as supported by subjective ratings and eye-tracking measures. Analysis of saccade duration, fixation duration, and fixation count suggests images are more cognitively demanding. Although no significant difference in perceived cognitive load was observed between images and videos on the Paas scale, our preliminary results suggest that videos may be more effective in capturing attention and engaging viewers. To confirm this hypothesis, further studies are needed that also consider stimulus retention.

By employing non-parametric tests, it was observed that there is a weak correlation between the results of subjective measures (such as Paas ratings) and objective measures (such as eye tracking metrics). This means that participants' subjective ratings of cognitive load may not always align with objective indicators, such as eye movements. These results suggest that participants' perception does not always reflect the actual cognitive load measured using more objective methods.

The analysis of the correlation between attractiveness and cognitive load ratings (Paas scale) showed a weak to moderate positive correlation for most stimuli, except image-1. However, low correlation coefficients suggest that stimuli with higher cognitive

load are not consistently rated as more attractive, indicating that attractiveness does not necessarily correspond with cognitive load.

The study methodology, encompassing both traditional psychophysiological methods and eye tracking, provided a comprehensive insight into how individuals process media messages. However, it is important to note the study's limitations, such as the non-normal distribution of some data and the weak correlation between subjective and objective measures. Future research can overcome these limitations by increasing the sample size and incorporating additional, more diverse measures of cognitive load to achieve more robust and generalizable results.

Overall, the study highlights the complexity of assessing cognitive load induced by social campaign materials and underscores the need for a detailed understanding of cognitive processes involved in media consumption, especially in the context of promoting social change through digital communication channels. The results of our study may have practical applications in creating more effective social campaigns that better capture attention and engage audiences.

Acknowledgments. Co-financed by the Minister of Science under the "Regional Excellence Initiative" Program.

References

1. Lee, N., Kotler, P.: Social Marketing: Influencing Behaviors for Good. SAGE Publications, Thousand Oaks (2011)
2. Pittman, M., Haley, E.: Cognitive load and social media advertising. J. Interact. Advert. **23**, 33–54 (2023)
3. Gunawardena, N., et al.: Assessing surgeons' skill level in laparoscopic cholecystectomy using eye metrics. In: Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications, Denver Colorado, pp. 1–8. ACM (2019)
4. Park, C.S.: Does too much news on social media discourage news seeking? Mediating role of news efficacy between perceived news overload and news avoidance on social media. Soc. Media Soc. **5**, 205630511987295 (2019)
5. Sweller, J.: Cognitive load during problem solving: effects on learning. Cogn. Sci. **12**, 257–285 (1988)
6. Borawska, A., Oleksy, T., Maison, D.: Do negative emotions in social advertising really work? Confrontation of classic vs. EEG reaction toward advertising that promotes safe driving. PLoS ONE **15**, e0233036 (2020)
7. Borawska, A., Borawski, M., Łatuszyńska, M.: The concept of virtual reality system to study the media message effectiveness of social campaigns. Procedia Comput. Sci. **126**, 1616–1626 (2018)
8. Southwell, B.G.: Information overload? advertisement editing and memory hindrance. Atl. J. Commun. **13**, 26–40 (2005)
9. Chung, S., Sparks, J.V.: Motivated processing of peripheral advertising information in video games. Commun. Res. **43**, 518–541 (2016)
10. Huskey, R., Mangus, J.M., Turner, B.O., Weber, R.: The persuasion network is modulated by drug-use risk and predicts anti-drug message effectiveness. Soc. Cogn. Affect. Neurosci. **12**, 1902–1915 (2017)

11. Dunlop, S.M., Wakefield, M., Kashima, Y.: Pathways to persuasion: cognitive and experiential responses to health-promoting mass media messages. Commun. Res. **37**, 133–164 (2010)
12. Jensen, J.D., Ratcliff, C.L., Yale, R.N., Krakow, M., Scherr, C.L., Yeo, S.K.: Persuasive impact of loss and gain frames on intentions to exercise: a test of six moderators. Commun. Monogr. **85**, 245–262 (2018)
13. Hattingh, M., Dhir, A., Ractham, P., Ferraris, A., Yahiaoui, D.: Factors mediating social media-induced fear of missing out (FoMO) and social media fatigue: a comparative study among Instagram and Snapchat users. Technol. Forecast. Soc. Change **185**, 122099 (2022)
14. Pleyers, G., Vermeulen, N.: How does interactivity of online media hamper ad effectiveness. Int. J. Mark. Res. **63**, 335–352 (2021)
15. Cheng, P., Ouyang, Z., Liu, Y.: The effect of information overload on the intention of consumers to adopt electric vehicles. Transportation **47**, 2067–2086 (2020)
16. Petty, R.E., Cacioppo, J.T.: Issue involvement can increase or decrease persuasion by enhancing message-relevant cognitive responses. J. Pers. Soc. Psychol. **37**, 1915–1926 (1979)
17. Kaufhold, M.-A., Rupp, N., Reuter, C., Habdank, M.: Mitigating information overload in social media during conflicts and crises: design and evaluation of a cross-platform alerting system. Behav. Inf. Technol. **39**, 319–342 (2020)
18. Otondo, R.F., Van Scotter, J.R., Allen, D.G., Palvia, P.: The complexity of richness: media, message, and communication outcomes. Inf. Manag. **45**, 21–30 (2008)
19. Krieglstein, F., Beege, M., Rey, G.D., Sanchez-Stockhammer, C., Schneider, S.: Development and validation of a theory-based questionnaire to measure different types of cognitive load. Educ. Psychol. Rev. **35**, 9 (2023)
20. Ryu, K., Myung, R.: Evaluation of mental workload with a combined measure based on physiological indices during a dual task of tracking and mental arithmetic. Int. J. Ind. Ergon. **35**, 991–1009 (2005)
21. Borawska, A., Mateja, A.: The use of cognitive neuroscience tools for evaluating the cognitive overload caused by social advertising. In: AMCIS 2023 Proceedings (2023)
22. Guixeres, J., et al.: Consumer neuroscience-based metrics predict recall, liking and viewing rates in online advertising. Front. Psychol. **8**, 1808 (2017)
23. Zagermann, J., Pfeil, U., Reiterer, H.: Studying eye movements as a basis for measuring cognitive load. In: Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems, Montreal, QC, Canada, pp. 1–6. ACM (2018)
24. Ausin-Azofra, J.M., Bigne, E., Ruiz, C., Marín-Morales, J., Guixeres, J., Alcañiz, M.: Do you see what I see? Effectiveness of 360-degree vs. 2D video ads using a neuroscience approach. Front. Psychol. **12**, 612717 (2021)
25. Druckman, J.N., Kam, C.D.: Students as experimental participants: a defense of the "narrow data base. In: Druckman, J.N., Green, D.P., Kuklinski, J.H., Lupia, A. (eds.) Hand Book of Experimental Political Science, pp. 41–57. Cambridge University Press, New York (2011)
26. Perkins, H.W., Linkenbach, J.W., Lewis, M.A., Neighbors, C.: Effectiveness of social norms media marketing in reducing drinking and driving: a statewide campaign. Addict. Behav. **35**, 866–874 (2010)
27. Philbrook, J.K., Franke-Wilson, N.A.: The effectiveness of a peer lead smart driving campaign on high school students' driving habits. J. Trauma Acute Care Surg. **67**, S67 (2009)
28. Cismaru, M., Lavack, A.M., Markewich, E.: Social marketing campaigns aimed at preventing drunk driving: a review and recommendations. Int. Mark. Rev. **26**, 292–311 (2009)
29. Barrie, L.R., Jones, S.C., Wiese, E.: "At least I'm not drink-driving": formative research for a social marketing campaign to reduce drug-driving among young drivers. Australas. Mark. J. **19**, 71–75 (2011)
30. Caamaño-Isorna, F., Moure-Rodriguez, L., Corral Varela, M., Cadaveira, F.: Traffic accidents and heavy episodic drinking among university students. Traffic Inj. Prev. **18**, 1–2 (2017)

31. Indawati, R., Bagus Qomaruddin, M.: The probability of the traffic accidents on students. J. Int. Dent. Med. Res. **11**, 348–351 (2018)
32. Ramsøy, T.Z.: Building a foundation for neuromarketing and consumer neuroscience research: how researchers can apply academic rigor to the neuroscientific study of advertising effects. J. Advert. Res. **59**, 281–294 (2019)
33. Dalrymple, K.A., Manner, M.D., Harmelink, K.A., Teska, E.P., Elison, J.T.: An examination of recording accuracy and precision from eye tracking data from toddlerhood to adulthood. Front. Psychol. **9** (2018)
34. Razavi, M., Janfaza, V., Yamauchi, T., Leontyev, A., Longmire-Monford, S., Orr, J.: OpenSync: an open-source platform for synchronizing multiple measures in neuroscience experiments. J. Neurosci. Methods **369**, 109458 (2022)
35. Vecchiato, G., et al.: Neurophysiological tools to investigate consumer's gender differences during the observation of TV commercials. Comput. Math. Methods Med. **2014**, e912981 (2014)
36. Johannessen, E., et al.: Psychophysiologic measures of cognitive load in physician team leaders during trauma resuscitation. Comput. Hum. Behav. **111**, 106393 (2020)
37. Cho, Y.: Rethinking eye-blink: assessing task difficulty through physiological representation of spontaneous blinking. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, Yokohama, Japan, pp. 1–12. ACM (2021)
38. Hebbar, P.A., Bhattacharya, K., Prabhakar, G., Pashilkar, A.A., Biswas, P.: Correlation between physiological and performance-based metrics to estimate pilots' cognitive workload. Front. Psychol. **12**, 555446 (2021)
39. Hu, X., Lodewijks, G.: Exploration of the effects of task-related fatigue on eye-motion features and its value in improving driver fatigue-related technology. Transp. Res. Part F Traffic Psychol. Behav. **80**, 150–171 (2021)
40. Appel, T., et al.: Predicting cognitive load in an emergency simulation based on behavioral and physiological measures. In: 2019 International Conference on Multimodal Interaction, Suzhou, China, pp. 154–163. ACM (2019)
41. Armougum, A., Gaston-Bellegarde, A., Joie-La Marle, C., Piolino, P.: Physiological investigation of cognitive load in real-life train travelers during information processing. Appl. Ergon. **89**, 103180 (2020)
42. Behroozi, M., Lui, A., Moore, I., Ford, D., Parnin, C.: Dazed: measuring the cognitive load of solving technical interview problems at the whiteboard. In: Proceedings of the 40th International Conference on Software Engineering: New Ideas and Emerging Results, Gothenburg, Sweden, pp. 93–96. ACM (2018)
43. Jacob, R.J.K., Karn, K.S.: Eye tracking in human-computer interaction and usability research. In: The Mind's Eye, pp. 573–605. Elsevier (2003)
44. Han, Y., Yin, Z., Zhang, J., Jin, R., Yang, T.: Eye-tracking experimental study investigating the influence factors of construction safety hazard recognition. J. Constr. Eng. Manag. **146**, 04020091 (2020)
45. Memar, A.H., Esfahani, E.T.: Physiological measures for human performance analysis in human-robot teamwork: case of tele-exploration. IEEE Access **6**, 3694–3705 (2018)
46. Bernhardt, K.A., et al.: The effects of dynamic workload and experience on commercially available EEG cognitive state metrics in a high-fidelity air traffic control environment. Appl. Ergon. **77**, 83–91 (2019)
47. Cronbach, L.J.: Coefficient alpha and the internal structure of tests. Psychometrika **16**, 297–334 (1951)

Artificial Intelligence in the Automation Process of Customs Declaration Preparation: A Case Study of an International Logistics Company

Maciej Krzan[1,2(✉)], Helena Dudycz[1], and Maciej Suchomski[2]

[1] Wroclaw University of Economics and Business, Wrocław, Poland
{maciej.krzan,helena.dudycz}@ue.wroc.pl
[2] XTRAS forward thinking GmbH & Co. KG, Bergiusstraße 1, 28816 Stuhr, Germany
suchomski@xtras-log.de

Abstract. International logistics companies operating in the market also provide customs declaration preparation services to their clients. These documents are filled out by employees based on received commercial invoices. This process is time-consuming, and errors often occur during data entry. The article aims to present a prototype for modifying an existing system for the preparation of customs declarations by using Optical Character Recognition (OCR) technology and an artificial intelligence (AI) system. The article discusses the process of preparing a customs declaration in the examined international logistics company, along with a proposal to use OCR and AI to automate filling the document with the necessary data. The verification of this concept using a prototype is also presented, along with calculations based on real data regarding the profitability of implementing automation into the customs declaration preparation system. The contribution of the article is the proposal to automate the customs declaration process by using OCR and artificial intelligence, which significantly reduces the time required to prepare the finished document.

Keywords: customs declaration · data entry automation · logistics efficiency optimization

1 Introduction

In today's global logistics, the topic of importing and exporting goods from a country is dealt with by local customs offices, which control and monitor the country's border crossings, ports, airports and railway stations. One of the basic documents for declaring the movement of goods is the customs declaration submitted to the office. Currently in the European Union, the rules on what should be filed, as well as how and where it should be filed, are the same for all members [1, 2]. Nowadays, offices are already using specially developed IT systems to handle the country's document flow digitally [3]. The transfer of customs management from the desk to the computer has increased the efficiency of declaration, and logistics companies are looking for newer and newer

M. Hernes et al. (Eds.): ECAI 2024, LNNS 1217, pp. 39–51, 2025.
https://doi.org/10.1007/978-3-031-78465-1_4

technological possibilities to control and speed up this process [4]. Some of the technologies worth implementing in customs declaration preparation systems are automation and artificial intelligence [5]. This demand has raised the following question: how can artificial intelligence technologies be applied to improve the customs declaration preparation process in international logistics companies? The purpose of this article is to present a prototype for modifying an existing automated system for the preparation of customs declarations by adding an artificial intelligence (AI) system responsible for a new form of data preparation needed for declarations in an international logistics company.

The structure of the article is as follows. The next section discusses the explanation of the modern approach to customs declarations and the rationale for using new technologies in this field. This is followed by a presentation of the methodology used in the study in order to discuss its stages in the following sections, i.e., the process of preparing the customs declaration in the international logistics company under study, the proposal to use AI in the process and the verification of this solution with a prototype. The next section presents calculations based on data on the cost-effectiveness of implementing automation into the customs declaration preparation system. The article concludes with a summary.

2 Approaches to Automating the Process of Processing Customs Declarations

One of the key stages of the international transport of goods is the preparation of a customs declaration, which is based, for example, on a commercial invoice received from the customer [6]. In the preparation of this document, a logistics company plays an important role as an intermediary between contractors [7]. The essence of the correct preparation of a customs declaration is the stage of accurately collecting information from many sources, such as the description of the goods, matching the HS Code (Harmonized Commodity Description and Coding System) to them and calculating their value. One of the proposals to improve the preparation of customs declarations is to use IT support to automatically classify goods according to HS codes [8]. The difficulty of achieving automation in this area is due to the use of different types and HS codes of goods and different types of sets by stakeholders of this process [4].

The conditions related to the implementation of a customs declaration are multifaceted and require taking into account various factors. Understanding these elements is key to making changes and automating the process of preparing the document [9]. The digitization of tax control processes, including the implementation of electronic communication with customs offices, and allowing the sending of customs declaration documents, contributed to minimizing time and material costs while increasing trust in employees of regulatory authorities [10, 11]. The implementation of Robotics Process Automation in import and export customs clearance should contribute to maximizing work efficiency and improving the speed of import and export handling [12].

The literature attaches more and more importance to the need to automate the preparation of customs declarations using IT solutions [8]. The implementation of new technologies and automation of customs procedures are aimed at increasing the efficiency of the entire process of creating and circulating customs declarations [4]. The benefits

obtained from solutions of this type include shortening the time of necessary customs clearance procedures, reducing the number of physical and documentation checks and improving the efficiency of customer service and their safety [13]. One of the proposals is to use IT support for automatic classification of goods' HS codes [8]. Another proposal is to use blockchain technology when creating smart contracts that will automatically aggregate relevant information and allow for its cross-verification [6].

The literature also indicates the possibility of using artificial intelligence solutions in the process of preparing and implementing customs declarations. Artificial intelligence-based technologies can support smarter customs operations and efficiency, contributing to streamlining processes, enhancing security and transaction accuracy, as well as providing more effective service for importers and exporters [14]. In literature, one of the proposed solutions is to use technologies such as Optical Character Recognition (OCR), Natural Language Processing methods and deep learning to automate both the preparation of such documents and their verification during cargo clearance at the border [5, 15]. S.R. Landge and others point out the possibility of using artificial intelligence to verify transaction records (process automation while detecting inconsistencies) and to store and search documents [16]. Another proposal concerns the use of artificial intelligence elements to decide on the exemption of goods from customs duty when sending documents via electronic communication [17]. The use of artificial intelligence solutions in the customs declaration process should contribute to reducing the costs associated with the business [18].

To sum up, automating the process of preparing and processing a customs declaration together with the use of artificial intelligence can improve the process, contributing to increased efficiency while reducing costs.

3 Research Methodology

The research was conducted in a real, international freight forwarder company XYZ, headquartered in the Netherlands. The company has been in existence for 25 years. Since its founding, it has been mainly involved in sea and air freight worldwide, along with related customs formalities. Additionally, it offers services such as road transport, warehousing and project logistics. Customs formalities are a significant part of its operations, impacting other areas of its business in the global market. The company's main areas of transport are the exports and imports of goods, both within the European Union and beyond. The following research procedure was applied:

1. Identification of issues related to the preparation of customs declarations in company XYZ.
2. Development of a concept for modifying the information system for preparing customs declarations in XYZ.
3. Modification of the information system for the preparation of customs declarations in XYZ.
4. Testing of the customs declaration prototype in XYZ.
5. Conclusions drawn from the conducted modification of the information system in XYZ.

The next section describes the customs declaration process in XYZ.

4 The Process of Preparing a Customs Declaration Document in XYZ

XYZ uses an ERP system (Microsoft Dynamics NAV Business Central) which includes dedicated additional modules. Among them is the customs declaration module, responsible for direct electronic communication with the customs office. It operates based on strict technical documentation received from the customs office concerning message types, data format, field validation, communication authorization, etc. It includes an expert system for identifying errors in already prepared documents. The main functionalities of this module are as follows: creating various types of customs declarations (e.g., import declaration, export declaration), sending the prepared customs declaration electronically to the customs office, receiving feedback information regarding a given customs declaration (e.g., document status, customs duty information), sending auxiliary or explanatory documents. The customs declaration module allows for the creation of various types of necessary documents, but these are filled out by employees of XYZ based on received commercial invoices from customers. The customs declaration preparation process was carried out in XYZ as follows (Fig. 1):

1. Preparation of customer-specific master data for automation. This stage involves a one-time definition of patterns for a given client (taking into account the format and layout of the data sent as a spreadsheet) and the mapping of item identification numbers from the client's system to the item tables in the ERP system in XYZ. The preparation of this source data then allows for automatic operations including grouping multiple lines of goods into one HS code, summing the number of goods and determining their value and weight.
2. Receipt of commercial invoice. The document, created in a spreadsheet format, contains a list of different goods that must be declared together. It contains information such as the identification code of the goods (coming from the customer's system), description of the goods, their quantity and sometimes additional information such as country of origin, destination country, container number, account number and currency.
3. Preparation of data for customs declaration. This stage involves both the automatic import of data from the commercial invoice received from the customer and the automatic preparation of data necessary to complete the customs declaration (e.g. matching HS codes to each item of goods on the invoice, grouping goods according to the same codes, calculating the weight of the items, assigning value to a given item, etc.).
4. Creating a data entry form in the ERP system. An empty customs declaration document is generated.
5. Filling in the form in the ERP system. The generated blank customs declaration document is automatically filled with the prepared data. If there is a need to complete the prepared customs declaration with additional data, it is done in the ERP system by an employee.

6. Checking of the form by the expert system. The prepared customs declaration document is checked by a dedicated module of the expert system. If errors or omissions are identified, the document is forwarded to an employee for correction. If the expert system does not identify any errors, the customs declaration is forwarded to the customs office.
7. Sending the data to the customs office. The prepared customs declaration document is sent to the customs office via electronic communication.
8. Cyclical response server check. Periodic checking of the customs office server by the ERP system to download incoming messages from the customs office.
9. Processing of responses. Registration of the response received from the customs office in the ERP system. Two situations may occur. The first case is the acceptance of the customs declaration document (incl. customs tax notice). This means that the preparation of the document for the client of company XYZ has been completed (and by fiscal representation the customs tax is automatically posted on GL accounts and paid). The other case is the rejection of the sent customs declaration, with the customs office indicating the reason for such a situation. Such a declaration can be corrected and resent.

The presented process of automating the preparation of customs declarations in XYZ concerns regular customers for whom patterns of sent commercial invoices have been defined in the ERP system and the identification numbers of goods from customer systems have been mapped. This is a good solution, provided that the commercial invoice sent by customers is in a spreadsheet file. Additionally, such automation is used for clients outsourcing many logistics services to XYZ and having patterns mapped in the ERP system. In the case of XYZ, approximately 70–80% of the customs declaration forms prepared are documents for regular customers in a spreadsheet. For other customers, customs declarations are created manually by an employee. These are customers who place orders on a one-off basis or several times a year and who send a commercial invoice as a PDF file. In XYZ, a need has been identified for automatic reading of commercial invoices without the necessity of creating templates for client documents in the ERP system and sending them in various file formats.

5 A Proposal to Use Artificial Intelligence Solutions in the Process of Preparing Customs Declarations in XYZ

Current research efforts have discovered multiple ways to include AI into the customs declaration preparation process. Even though these approaches focus on various phases of the preparation process, they all aim to enhance it—whether it be by decreasing errors or accelerating the creation of documents. In this context, the scope of AI will be broadened in order to increase customer reach. The focus of the use of AI lies in the provision of data for the preparation of declarations. Currently, the system only works with Excel files after configuration. As part of the proposed solution, support for PDF files is envisaged thereby extending XYZ's service portfolio to clients using this document format. The planned support for PDF files involves AI-driven extraction of unstructured data from documents, followed by the application of trained machine learning algorithms to match it with the information required to prepare customs declarations.

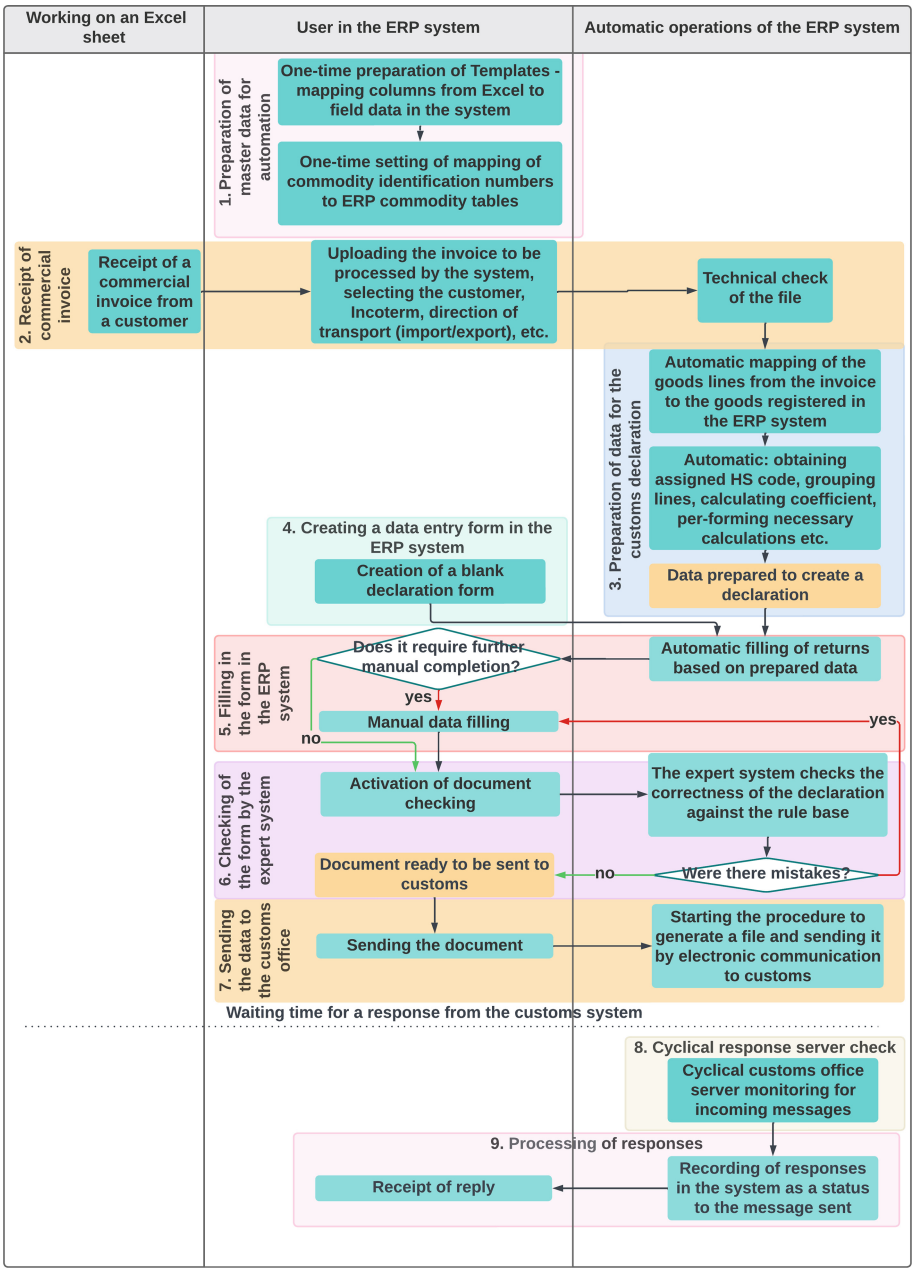


Fig. 1. The process of preparing a customs declaration at company XYZ with automatic data preparation

The process of applying AI in the preparation of a customs declaration was carried out at XYZ as follows (Fig. 2):

1. Receipt of a commercial invoice and upload of the file to the system for processing in the ERP system.
2. Transfer of the PDF file to the AI system via API communication.
3. Work in the AI system in the following way: receipt of PDF and OCR files, splitting files into documents, classification – assignment of the document to one of the skills, extraction of information from the document for configured labels in the skill, preparation of a JSON with data mapped to information, setting a JSON file ready for download.
4. Download and receipt of the JSON from the AI system to the ERP system.
5. Reading the file and importing data into buffer tables in the ERP system.
6. Transfer of data to the customs module where it is used to prepare the declaration.

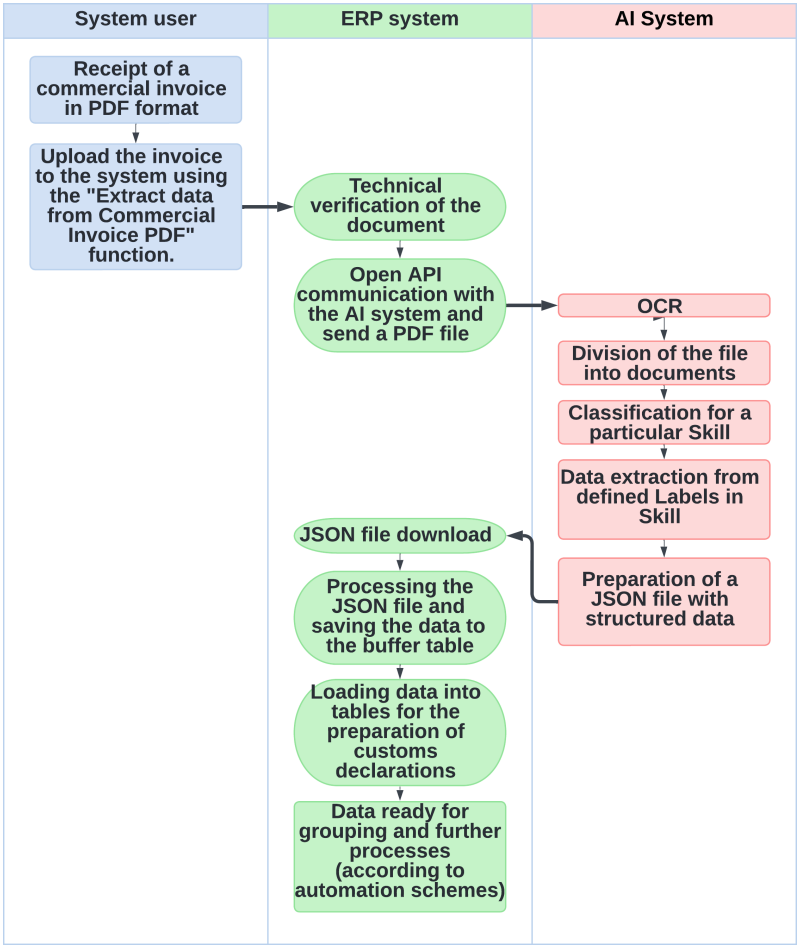


Fig. 2. Working prototype of an AI-enhanced system for extracting data from PDF files

The process of preparing the declaration continues. The described part relates to the improvement of part 2 and, to a lesser extent, 1 and 3 of Fig. 1. The next steps are not changed. During the development and testing of the prototype, two aspects in particular had to be checked. Firstly, it must be certain that the data would not come in late compared to the standard process And secondly – that the data would not be altered or incorrectly created by the AI system. Only with due correctness and speed it will be possible to determine whether the prototype built on the basis of this proposal is in fact functional.

6 A Prototype for Automating the Customs Declaration Preparation Process

A prepared proposal for the use of artificial intelligence in the customs declaration preparation process was accepted by XYZ as a proof of concept to test whether such an extension would realistically work. Thanks to access to ERP technology and the selection of a partner with a configurable AI system [19], a prototype could be developed and tested on a commercial invoice document with test data for the purposes of the study.

The document was uploaded by the new function to the ERP system, which in the background automatically transmitted it via electronic communication to the AI partner's software for further processing. There, the document was OCR-processed and classified into the previously rehearsed Skill responsible for Commercial Invoice. According to the prepared Skill, the required data was extracted from the document with the help of machine learning algorithms, which were then used to create a JSON with the necessary information (Fig. 3).


```
{
    "Version": "3.0",
    "Fields": {
        "Consignee": {
            "Address": "7894\n10115 Berlin\nGermany",
            "Name": "Himmelstraße"
        },
        "Invoice Date": "2024-10-05",
        "Invoice Number": "INV474125",
        "Line Items": [
            {
                "Description": "steel rails",
                "Quantity": "850",
                "Total Price": "1700",
                "Item": "45216"
            },
            {
                "Description": "hairdressing supplies",
                "Quantity": "10",
                "Total Price": "25000",
                "Item": "66528"
            },
```


Fig. 3. Part of the JSON file obtained as a result of the employment of the AI system

The JSON file is loaded into the ERP system as a response to the PDF file sent previously. The system then reads and saves the data from the file to the system tables,

first buffered from the API communication and then already target-related to the customs declaration data (Fig. 4).

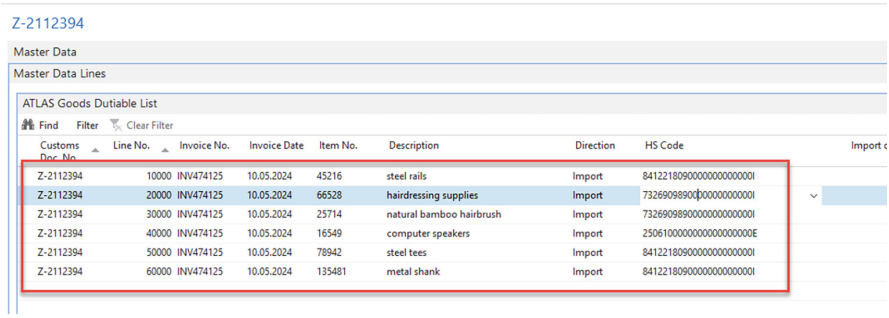


Fig. 4. A window from the ERP system with uploaded data ready for further work

At this point, the data is ready for further work. The whole process, from the user sending the file to receiving the finished data in the system table, is fully automatic (see the tracks in Fig. 2) and takes, including API communication, about 1 min. The data prepared for the declaration was correct and delivered promptly. The prototype can be considered to have worked correctly and is a solid foundation on which to build a fully functioning solution.

7 The Benefits of Implementing Automation of the Customs Declaration Preparation Process in XYZ

This part compares the time needed to prepare a customs declaration before and after the automation of this process in XYZ. In January 2024, XYZ prepared 695 customs declarations. The documents consisted of a varying number of lines describing the goods. The declarations covered one product, several, or about 100 items.

The time needed to prepare a customs declaration before implementing automation of this process was calculated according to the following formula:

$$y = m * x + b \tag{1}$$

where:

- y is the time in minutes,
- x is the number of items,
- m is the slope coefficient (i.e. the rate of time increase depending on the number of items),
- b is the intercept (time needed to create a declaration with one item).

By measuring the time needed to prepare a customs declaration by an employee at XYZ before automating this process, the following results were obtained:

- in the case of 1 line of goods in the customs declaration, its preparation took 20 min;

- in the case of 10 lines of goods in the customs declaration, its preparation took 60 min;
- in the case of 100 lines of goods in a customs declaration, its preparation took 480 min.

These data was used to calculate m and b as follows:

$$20 = m * 1 + b \tag{2}$$

$$60 = m * 10 + b \tag{3}$$

$$480 = m * 100 + b \tag{4}$$

By solving Eqs. (2–4), the following results were obtained:

$$m = (480 - 20)/(100 - 1) = 460/99 \approx 4.6465 \tag{5}$$

$$b = 20 - 4.6465 * 1 = \mathbf{15.3535} \tag{6}$$

The equation to calculate the average time to prepare a customs declaration for any number of items x of goods before implementing automation is as follows:

$$\boldsymbol{y = 4.65 * x + 15.35} \tag{7}$$

After implementing automation of the customs declaration preparation process, an employee of company XYZ needs 10 min to prepare this document, regardless of the number of lines of goods. Table 1 contains a fragment of the summary containing the preparation times of all customs declarations in January 2024 before and after the implementation of automation. As a result of the automation implementation, one person now prepares as many customs declarations per month as over four individuals did before automation.

8 Conclusions

The use of artificial intelligence in the customs declaration preparation process proved to be an effective way to increase the operational efficiency of company XYZ. The development of the concept and subsequent preparation of an AI-based prototype proved its feasibility and validity, demonstrating the progress the company has made in handling customs declarations. The prototype uses OCR and machine learning technologies to extract unstructured data and interpret it into specific information. The change works quickly, streamlines the process and reduces human error, achieving all this in one minute. The main benefit of implementing automation is the reduction in the time taken to prepare declarations. The productivity increase calculated is a very good result. Importantly, the scalability of the declaration preparation time to the number of goods has been removed. The current process is not only fast but also repeatable. This provides opportunities for specific customers who expect very large declarations to be drawn up, and XYZ can now offer this without loss of productivity.

Table 1. Summary of customs declaration preparation times before and after the implementation of automation

Customs declaration number	Number of product lines	Without automation	After implementing automation
Declaration no. 1	11	66.5	10
Declaration no. 2	1	20	10
Declaration no. 3	24	126.95	10
Declaration no. 4	1	20	10
Declaration no. 5	4	33.95	10
…..	…..	…..	…..
Declaration no. 694	1	20	10
Declaration no. 695	8	52.55	10
Total minutes		**31,086.4**	**6,950**
Total of hours		**518**	**116**
Automation efficiency factor		**4.47**	**1**

The limitations of the research are also worth noting. First of all, it should be noted that a European company was analyzed in the case study, which is obliged to adhere to the regulations, laws and processes strictly defined by the European Union. It would also be worthwhile for future research to be carried out in other countries such as America or Asia which have different regulations and procedures. Another limiting aspect is the type of business documents used. The study focused on the most popular, i.e. commercial invoice. In practice, others are also used, such as packing lists, Bill of Lading or a mix of information from several document types. The next stage of the study will be the complete automation of the customs declaration preparation process, which will eliminate tasks performed manually by the employee. Artificial intelligence algorithms could be used to classify HS codes based on information extracted from documents.

In conclusion, the successful implementation of artificial intelligence in the customs declaration preparation process at XYZ has set a new standard for efficiency and accuracy in the logistics industry. The prototype developed in this study can be used as a template for other companies showing how to implement automation and artificial intelligence tools in their own processes. All the indications are that it is this type of innovation that will have a real impact on competitive advantages and sustaining service quality. The example of company XYZ shows the benefits of transformation to modern logistics.

References

1. Lyons, T.: EC Customs Law, 2nd edn. Oxford University Press, Oxford (2018)
2. Gwardzińska, E., Gwardzińska-Chowaniec, Ż., Chowaniec, J.: Principle of choice of customs procedure in EU customs traffic. Front. Law **2**, 110–118 (2023). https://doi.org/10.6000/2817-2302.2023.02.13
3. Yakovleva, M.: Electronic customs as a currency control authority. In: Proceedings of the 2nd International Scientific and Practical Conference on Digital Economy (ISCDE 2020), Advances in Economics, Business and Management Research, vol. 156, pp. 246–249 (2020)
4. Liutkevičius, M., Pappel, K.I., Butt, S.A., Pappel, I.: Automatization of cross-border customs declaration: potential and challenges. In: Viale Pereira, G., et al. (eds.) EGOV 2020. LNCS, vol. 12219, pp. 96–109. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-57599-1_8
5. Han, C., Wang, B.: Research on intelligent customs declaration generation in Guangdong-Hong Kong cross-border road cargo clearance. In: Proceedings of the SPIE, vol. 12918 (2023). https://doi.org/10.1117/12.3009242
6. Segers, L., Ubacht, J., Tan, Y.-H., Rukanova, B. D.: The use of a blockchain-based smart import declaration to reduce the need for manual cross-validation by customs authorities. In: Proceedings of the 20th Annual International Conference on Digital Government Research, pp. 196–203 (2019). https://doi.org/10.1145/3325112.3325264
7. Lux, M.: Import and the responsible person(s): what would change under the EU commission's reform proposal? Glob. Trade Cust. J. **19**(5), 348–359 (2024)
8. Lux, M., Matt, Ch.: Classification of goods: what are the hurdles and pitfalls in the use of automation or IT support? Glob. Trade Cust. J. **16**(6), 237–247 (2021). https://doi.org/10.54648/gtcj2021027
9. Omar, N.H., Selamat, M.N., Aziz, S.F.A., Mohd, R.H., Ismail, R.: Integrity risk in the customs automation system. Int. J. Acad. Res. Account. Finance Manag. Sci. **11**(3), 206–222 (2021)
10. Pavlova, K.S., Smolina, E.S.: Digitalization of tax and customs control of foreign trade operations. In: Ashmarina, S.I., Horák, J., Vrbka, J., Šuleř, P. (eds.) IES 2020. LNNS, vol. 160, pp. 684–691. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-60929-0_88
11. Stupak, Y., Syzdykova, E., Jazykbayeva, B., Shirshova, L., Issayeva, A., Niyazbekova, S.: Unification of technical solutions used in customs control as a factor in the development of economic systems. In: Web of Conferences, vol. 371 (2023). https://doi.org/10.1051/e3sconf/202337105053
12. Lee, J.: Robotics process automation (RPA) and the import/export customs declaration process. Glob. Trade Cust. J. **18**(10), 384–390 (2023)
13. Popa, I.B., Mihaela, G., Paraschiv, D.M., Marinoiu, A.M.: Best practices in customs procedures. Amfiteatru Econ. **17**(40), 1095–1107 (2015)
14. Popov, G., Popova, A.: Structural model of an innovative information system for prevention and detection of financial customs violations. In: 3rd International Conference on High Technology for Sustainable Development (HiTech 2020), Sofia, Bulgaria, pp. 1–4 (2020). https://doi.org/10.1109/HiTech51434.2020.9364001
15. Han, C., Wang, B., Lai, X.: Research on the construction of intelligent customs clearance information system for cross-border road cargo between Guangdong and Hong Kong. In: Zhao, F., Miao, D. (eds.) AIGC 2023. CCIS, vol. 1946, pp. 181–190. Springer, Singapore (2024). https://doi.org/10.1007/978-981-99-7587-7_15
16. Landge, S.R., Gunturu, S.R., Josyula, H.P., Thangavel, K., Fatima, E.: Exploring the use of artificial intelligence for automated compliant transaction processing. In: 2nd International Conference on Disruptive Technologies (2024). https://doi.org/10.1109/ICDT61202.2024.10489553

17. Boikova, M.V., Vorona, A.A., Gubarev, D.V., Maksimov, Y.A., Timchenko, T.N.: Transformation of customs administration and the impact of automation on decision-making by customs authorities. In: Popkova, E.G. (eds.) ISC 2020. LNNS, vol. 368, pp. 12–20. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-93244-2_2
18. Kugonza, J., Mugalula, C.: Effectiveness and efficiency of artificial intelligence in boosting customs performance: a case study of RECTS at Uganda customs administration. World Cust. J. **14**(2), 177–192 (2020)
19. Krzan, M.: Metoda wielokryterialnego wspomagania decyzji SAW w wyborze dostawcy systemów informatycznych. In: Dudycz, H., Hernes, M., Pondel, M., Rot, A. (eds.) Informatyka w zarządzaniu. Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu, Wrocław, pp. 148–161 (2023). https://doi.org/10.15611/2023.51.0.08

Modelling Cognitive Load of Computer Game Users - Case Study

Krystian Wojtkiewicz(✉), Rafał Palak, Zbigniew Telec, and Filip Litwinienko

Faculty of Computer Science and Telecommunication Technologies, Wrocław University of Science and Technology, Wybrzeże Stanisława Wyspiańskiego 27, 50-370 Wrocław, Poland
krystian.wojtkiewicz@pwr.edu.pl

Abstract. This study evaluated the effectiveness of various methods for measuring cognitive load (CL) in arcade racing game players and their usefulness in modelling CL. The techniques used were EEG, GSR, eye-tracking, pupilometry, and subjective surveys. The "Trackmania Nations Forever" game had three difficulty levels: easy, medium, and hard. Forty-three participants were involved, and data were collected on player performance, survey responses, and physiological signals. The analysis included retries per minute (RPM), advancement measures, NASA-TLX, and SEQ scores. Significant changes in physiological measures were observed corresponding to different cognitive load levels, with the most accurate classification results achieved by combining physiological and performance data.

Keywords: Bio-metrics · User experience · Cognitive load

1 Introduction

With the rise of video games not only as a form of entertainment but also as a medium for learning and skill development, understanding the cognitive load (CL) imposed by different gaming tasks has become increasingly important. Cognitive load theory (CLT), introduced by John Sweller in 1988 [15], suggests that minimizing mental strain when presenting information and tasks enhances information retention and long-term memory access. This theory is particularly relevant in gaming, where players often encounter complex and dynamic tasks that can significantly impact their cognitive resources. This study uses various biosensors and subjective measures to model players' cognitive load in an arcade racing game. The objective is to explore the correlation between cognitive load and the game's difficulty level, ultimately proposing a model to automatically assess game task difficulty. Such a model could streamline the design and evaluation of game levels, ensuring they are appropriately challenging for players of different skill levels.

This paper's key contributions include exploring the relationship between cognitive load and game difficulty levels, providing insights into how varying challenges affect player cognition, and a detailed analysis of the most effective measures for classifying cognitive load levels, specifically in arcade racing games.

M. Hernes et al. (Eds.): ECAI 2024, LNNS 1217, pp. 52–63, 2025.
https://doi.org/10.1007/978-3-031-78465-1_5

This study aims to contribute to the broader understanding of cognitive load in gaming and its potential applications in game design and player experience optimization by investigating the effectiveness of different measurement methods and their integration into a cognitive load model. The structure of this paper includes a review of relevant literature on cognitive load theory, its application in gaming, and methods for measuring it with biosensors. It outlines the research problem, describes the experimental setup, presents the results, and concludes with implications and future directions.

2 Literature review

2.1 Cognitive Load Theory

John Sweller introduced Cognitive Load Theory (CLT) in [15]. He proposed that information and tasks should be presented to minimize mental strain, affect information retention, and facilitate positive access to long-term memory. Sweller theorized that memory and cognitive capabilities are inherently limited at any given moment. Thus, an essential aspect of the learning and problem-solving processes is to maximize the use of cognitive resources on the task at hand. He categorized cognitive load into three types, each differently influenced by the task structure. The diagram in Figure 1 illustrates Sweller's classification.

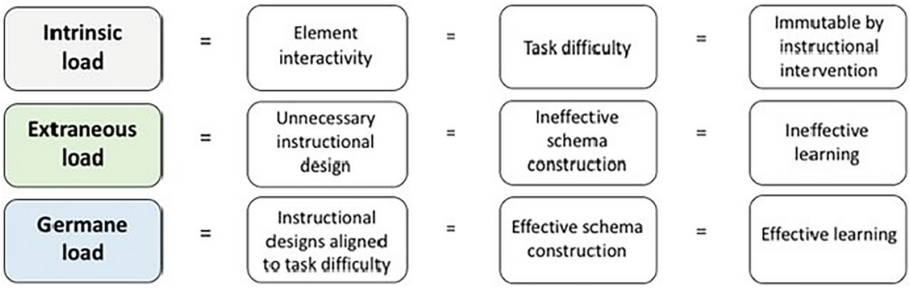


Fig. 1. Presentation of types of CL and their nature (source [11])

Sweller emphasized that while intrinsic and extraneous loads are invariably present during learning or task performance, germane load arises only when sufficient cognitive resources are available. The interactions among these loads are depicted in Figure 2.

Cognitive Load Theory finds broad application in various fields such as education [6] or user interface design [8].

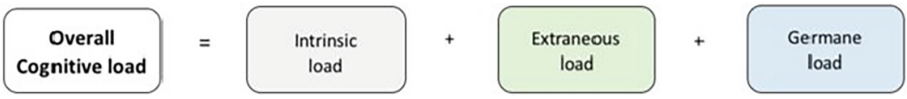


Fig. 2. Relation between different loads and how they influence CL (source [11])

2.2 CLT in Video Games

Numerous studies have sought to measure cognitive load in video game players. For instance, Fortin-Cote et al. [5] attempted to predict the enjoyment derived from video games using physiological, behavioral, and psychological data.

Pusey et al. [13] introduced a framework for assessing cognitive load in puzzle games. It utilizes performance-based metrics that eliminate the need for bio-sensors, which could be stressful, particularly for those unfamiliar with such technology. This framework was validated across various games.

Lecoutre et al. [7] evaluated the usefulness of eye-tracking technology in assessing workload or cognitive load. Their study, which also utilized the NASA-TLX subjective survey, demonstrated that eye-tracking can differentiate between perceptual and cognitive loads, primarily through variations in fixation duration for cognitive tasks.

2.3 Methods for Measuring CL

The literature suggests two primary approaches to measuring the potential cognitive load of a task, namely analytical and empirical. The analytical method involves mathematical models and expert opinions to estimate the mental workload of a task. Due to the challenges of applying this method across various types of tasks, researchers in the 1990s shifted focus towards refining empirical methods [12, 14], which provides the following distinction of CL measures:

- subjective:
  - questionnaires
- objective:
  - task performance
  - behavioral
  - physiological:
    - intrusive
    - non-intrusive

A systematic review in [2] indicated that the NASA-TLX questionnaire was used in nearly half of the analyzed studies to assess CL, suggesting it has become a standard for subjective measures.

Regarding objective measures, task performance is particularly challenging to frame due to its reliance on the task's completion status. However, Pusey [13] suggested metrics that could assess task performance using the number of incorrect solutions entered, time spent on individual puzzles, and the number of actions taken relative to the minimum required.

Physiological measures are categorized into intrusive and non-intrusive methods, with non-intrusive methods being preferable as they do not require physical contact with the participant, thus minimizing stress.

Our study will focus on non-intrusive methods such as pupilometry, eye-tracking, emotion recognition, and intrusive methods like EEG and GSR. While we won't take into consideration behavioral metrics derived from user-computer interactions due to their specificity to the task nature.

Despite the significant advancements in cognitive load theory and its application to video games, the existing literature exhibits several gaps that our study seeks to address. First, while numerous studies have focused on assessing cognitive load using standardized tools such as the NASA-TLX questionnaire, there remains a lack of consensus on the optimal methodologies for integrating physiological and performance data. Moreover, much of the current research predominantly relies on intrusive physiological measures, which can potentially alter the cognitive load by introducing stress or discomfort to the participants.

Additionally, the current frameworks for measuring cognitive load in video game environments often neglect the dynamic interplay between different types of cognitive loads (intrinsic, extraneous, and germane) and their cumulative effect on a player's performance and experience. Many studies also fail to adequately explore the implications of different game genres and difficulty levels on cognitive load, providing a somewhat limited understanding of how these factors might interact in more complex gaming scenarios.

The necessity for our study arises from these shortcomings. Our research aims to develop a more holistic and nuanced model of cognitive load in arcade racing games by employing physiological and performance-based metrics in a controlled experimental setup across varying game difficulty levels. Our approach seeks to validate the integration of diverse measurement techniques and contribute to a more standardized methodology for assessing cognitive load in gaming contexts. Through this, we intend to fill the critical gaps identified in the existing literature and pave the way for future research that can benefit from our findings and methodological innovations.

3 Problem definition

The main focus of this paper is to determine if physiological and subjective measures can accurately model a player's cognitive load in arcade racing video games. To explore this, we have formulated the following research questions that shaped the experimental design of the study:

RQ1 Are the changes in physiological measures statistically significant across different levels of cognitive load?

RQ2 How relevant are the responses from biosignals for classifying cognitive load levels in arcade racing video games?

RQ3 Which measures are most effective for classifying changes in cognitive load within this gaming context?

4 Experiment Setup

The data was collected by masters students pursuing their degrees [1, 3, 16] and involved 43 participants: 37 males and six females, of Polish, Indian, Congolese, Nigerian, Kenyan, and Ukrainian origin. 23 participants identified themselves as beginner-level racing game players, while the remaining were intermediate-level players. There were no participants who considered themselves experts.

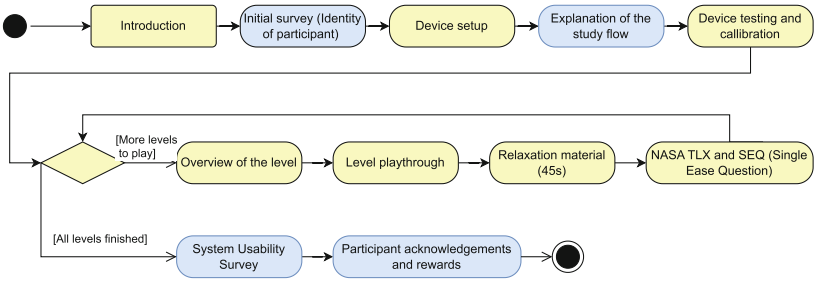


Fig. 3. UML activity diagram for the study flow presentation

Figure 3 depicts the flow of the experiment. Below, we comment on the most important aspects of the data collection.

We used the suite of tools paired with iMotions software for data collection. It included a NIC EEG helmet, a Shimmer S3 for galvanic skin response (GSR), a Gazepoint G3 eye tracker, and a webcam for emotion recognition and pupilometry. The summary of equipment and Software used in experiments:

Eye tracker Gazepoint GP3, 150Hz
EEG helmet NIC EEG, 64 channels, 500 Hz sampling rate, filtration: notch noise 50HZ artifact threshold 102 μV, band pass 0.5-100 Hz
GSR Shimmer S3, calibrated before each session
Software

- iMotions, version 8.0
- TrackMania 2008 version

Trackmania Nations Forever was our game of choice due to its multilingual support, freeware licensing, and the ease with which races can be designed. Its straightforward rules allow newcomers to quickly understand and engage with the game.

Three levels of difficulty were designed: easy, medium, and hard. They differ in terms of the maximum time to complete the map, the map length, the number of checkpoints, the complexity level of the obstacles, and the number of dead ends. The maximum time needed to finish each map, also known as the maximum driving time, was determined based on the completion times recorded during test drives conducted by the authors.

The difficulty levels were as follows:

Easy shortest route, five checkpoints, no dead ends, consistent surface, low number of obstacles, and places where the player may fall off the route

Medium medium-length route, four checkpoints, several dead ends, the same surface on the entire route, but part of the road was without borders, a significant number of obstacles and places where the player may fall off the route

Hard longest route, 3 checkpoints, many dead ends, different surfaces on the route: normal road, dirt road, a road without border and very narrow road, a very large number of obstacles and places where the player may fall off the route, very complicated obstacles that required skills and many attempts to overcome

The surveys included NASA-TLX, Single Ease Question (SEQ), and Subjective User Survey (SUS). For the NASA-TLX and SEQ, the rating scale ranged from 1 ("Extremely low") to 7 ("Extremely high"). The SUS was scored according to its standard, ranging from 1 to 5.

5 Experimental results

5.1 Data analysis methodology

A summary of participants' performance based on the time required to complete tasks is presented in Table 1.

Given the variability in completion times across different tracks, two new standardized measures were prepared to normalize the data: RPM – Retries Per Minute (Eq. 1) and Advancement (Eq. 2). The results for these measures were presented in Table 2.

$$RPM = \frac{retries}{T} \tag{1}$$

where *retries* refers to the total number of retries made by a participant, and T refers to the time needed by the participant to complete the task.

$$advancement = 1 - \frac{playthroughTime}{maxTrackTime} \tag{2}$$

where *playthroughTime* is the time the participant spends on a level, and *maxTrackTime* is the expected completion time.

The analysis of subjective data was straightforward. The formula for interpreting NASA-TLX scores is defined in Equation 3.

$$tlxInterprated = \frac{\sum questionRating}{maxScale * numberOfQuestions} \tag{3}$$

where *questionRating* is the score for a given question, *maxScale* is the maximum possible score (7), and *numberOfQuestions* was 6 (Table 3).

Table 1. Time-based performance of participants per difficulty level.

Track	Min time	Max time	Not finished
Easy	02:40:00	05:57:00	4
Medium	04:46:14	10:17:21	21
Hard	13:27:00	14:20:00	35

During experiments, we explored different measures to quantify cognitive load, such as Percentage Change in Pupil Size (PCPS) [9], Mean Pupil Diameter Change (MPDC) [10], and Low-High Index of Pupillary Activity (LHIPA) [4]. Besides the baseline correction, z-scoring standardization was also used. The results of our experiments are

Table 2. Avarage RPM and advancement per difficulty level.

Track	Avg. RPM	Avg. advancement
Easy	0,62	0,452
Medium	1,41	0,255
Hard	1,42	0,047

Table 3. Avarage normalized NASA-TLX score and SEQ score per difficulty level

Track	Avg. NASA-TLX	Avg. SEQ
Easy	0,48	6,94
Medium	0,66	5,17
Hard	0,72	3,43

shown in Table 4, where we pointed out the statistical method used for each result. The table shows measures and statistical differences between easy and hard (E-H p-value), easy and medium (E-M p-value), and medium and hard (M-H p-value) difficulty levels. Since some value sets did not come from a normal distribution, using the *t*-test wasn't always possible. Therefore, we used the Wilcoxon Signed-Rank test in such cases.

Among the 15 measures evaluated, 8 displayed statistically significant differences across all difficulty levels (with p-value $= 0.05$), indicating their reliability as indicators of difficulty and cognitive load changes. Furthermore, it was determined that z-standardization does not negate these differences, allowing the process to be safely applied without compromising the ability to distinguish between difficulty levels. In one specific case (standard deviation), the standardized value revealed significant differences, while the raw value did not. The following measures consistently showed differences across all difficulty levels:

- Mean pupil diameter and mean z-scored pupil diameter
- Median pupil diameter and median z-scored pupil diameter
- Peak dilation
- PCPS variance
- LHIPA and z-scored LHIPA
- Standard deviation of z-scored pupil diameter

5.2 Results presentation

Various datasets were tested separately and in combination to evaluate the impact of different data types on cognitive load (CL) modeling. These included physiological data, performance data, and the combination of all datasets. The following classifiers were employed to ascertain the most effective model:

- XGBoost,

Table 4. Cognitive load measures presentation

Measure	E-H p-value	E-M p-value	M-H p-value	Statistic
Mean pupil	3,88E-07	2,35E-03	1,64E-06	T-test
Median pupil	0,000001	0,004556	0,000007	T-test
Peak dilation	0,000392	0,006617	0,022277	Wilcoxon
Pupil variance	0,005255	0,151802	0,000001	Wilcoxon
Mean PCPS	0,142182	0,160501	0,004958	T-test
Median PCPS	0,237795	0,160357	0,010168	T-test
PCPS variance	0,009887	0,040195	0,000003	Wilcoxon
MPDC	0,068541	0,218849	0,003718	T-test
LHIPA	1,65E-07	1,94E-04	5,48E-07	Wilcoxon
LHIPA z-scored	1,52E-07	2,46E-04	4,20E-06	Wilcoxon
Z-scored mean pupil diameter	1,64E-08	3,09E-03	1,85E-07	T-test
Z-scored median pupil diameter	0,000002	0,008288	0,000008	Wilcoxon
Z-scored std pupil diameter	2,34E-03	4,35E-02	1,03E-08	T-test
Z-scored variance pupil diameter	0,004565	0,085462	0,000001	Wilcoxon
Pupil std	0,003425	0,151802	0,000001	Wilcoxon

- Random Forest(RF),
- AdaBoost,
- Support Vector Machine(SVM),
- K-Nearest Neighbours(KNN),
- Gradient Bossting Machine(GBM),
- Naive Bayes(NB),
- Bagging,
- Quadratic Discriminant Analysis(QDA),
- Multi-Layer Percepton(MLP).

The classification aimed to distinguish between three levels of game difficulty: easy, medium, and hard. The number of features utilized in the classification varied to optimize performance. To quantify classifier performance, we used the F1-score defined in equation 4

$$F1 = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \tag{4}$$

where TP represents true positives, FP false positives, and FN false negatives.

The training process utilized models and transformations from the sci-kit-learn library. The optimal combination of K features and hyperparameters for a given classifier was sought at the beginning of the process. Models were tested for the following features: 2, 3, 4, 5, 6, 7, 8, 10, 12, 15, 17, 20, and all available features. Feature selection

was conducted using the SelectKBest method, which selects features with the highest k values, with $f_classif$, based on the F value from ANOVA, as the ranking function. The features were scaled to a range from 0 to 1 using MinMaxScaler.

Hyperparameter search was conducted using the BayesSearchCV function from the sci-kit-optimize library. Bayesian search is an iterative algorithm where each subsequent step is based on the results of the previous ones, optimizing the selection of the next parameters to test. To avoid overfitting, 5-fold stratified cross-validation was used.

In the validation phase, the classifiers had hyperparameters set according to the result obtained in the first phase. Using a loop with 100 repetitions, the performance of the classifiers was calculated, with data being randomly split into training (70% of total data) and test sets. After completing the loop, the average accuracy and F1 score were calculated as the final performance measures of the classifier. The loop was necessary due to the high variance of results between different training and test set splits. In extreme cases, the accuracy was significantly lower or higher (up to 0.4 less than the average). In table 5, we present the results for all classification algorithms.

In tests utilizing only physiological data, the highest *F1 score* score was 0.776, with the MLP classifier based primarily on pupilometry and eye-tracking data. When using only the eye-tracker data, the *F1 score* score dropped to 0.517 with the XGBoost classifier; other approaches produced even lower results. When we utilized physiological and performance data, the highest F1 score was 0.877, achieved by the SVM and KNN classifiers, with the MLP classifier also obtaining a high score of 0.867. Combining these two data types significantly improved the modeling results, though some classifiers, such as Naive Bayes and QDA, had limited ability to effectively utilize this integration.

Integrating all datasets did not enhance the modeling results. In some cases, the performance was equivalent to the two-dataset model, but notably, the KNN classifier experienced a slight decrease in performance, dropping by 0.11 in its F1 score. This suggests that the extensive range of information might obscure optimal classifier choices, especially considering the potentially skewed nature of some datasets, such as the subjective data influenced by the participant demographics and pre-informed difficulty levels.

5.3 Discussion

This study advances our understanding of cognitive load in video gaming by assessing various methods for measuring cognitive load, including EEG, GSR, eye-tracking, pupilometry, and subjective surveys. Each method has proven effective in capturing different facets of cognitive load during gameplay, offering insights into how players react under varying conditions of game difficulty.

The EEG and GSR techniques offered direct measures of neural activity and physiological responses, highlighting players' real-time cognitive and emotional states. However, it is important to note that only specific frequencies (alpha and delta) in EEG showed significant differences across difficulty levels, which aligns with their known roles in cognitive processing. Similarly, GSR's utility was primarily indicated by variations in peak count among different difficulty levels. Eye-tracking and pupilometry provided valuable insights into players' focus and attentional engagement, though eye-tracking data proved challenging for classification due to the dynamic visual demands of racing games. Not all pupilometry measures were equally informative. Only specific metrics

Table 5. F1 score achieved by used classifiers divided into three categories depending on a used dataset for classification: physiological, physiological and performance with subjective measures

Classifier	Phys.	Phys. and perf.	All measures
XGBoost	0,73	0,817	0,826
Random Forest	0,73	0,843	0,824
AdaBoost	0,775	0,706	0,739
SVM	0,775	0,877	0,877
KNN	0,774	0,877	0,866
GBM	0,746	0,78	0,826
Naive Bayes	0,451	0,636	0,666
Bagging	0,423	0,708	0,658
QDA	0,475	0,601	0,668
MLP	0,776	0,867	0,869

showed significant differences, such as mean and median pupil diameter, peak dilatation, PCPS variance, LHIPA, and z-scored pupil diameter standard deviation. These objective measures were effectively complemented by subjective evaluations from surveys that captured players' self-reported experiences of effort and stress, thereby providing a holistic view of experienced cognitive load.

Our findings demonstrated a direct relationship between the levels of game difficulty and the cognitive load experienced by players. As the difficulty increased, so did the cognitive load, as evidenced by physiological responses and performance metrics. This understanding is pivotal for developers designing engaging games to not overwhelm the player, thus maintaining an optimal balance between challenge and player capability.

We identified the most effective measures for classifying cognitive load levels within arcade racing games. Performance-based metrics such as advancement and retries per minute, along with pupilometry measures like LHIPA and z-scored LHIPA, emerged as particularly informative.

Despite the promising findings, we must consider the study's limitations while interpreting results. The relatively homogeneous and small participant group studied in a controlled laboratory setting limits the impact of the findings. Future studies should aim to include a broader, more diverse participant pool and possibly extend the research into more natural settings.

6 Conlusions

We will conclude the paper relating to the research questions from section 3. The first question is tricky to answer since physiological measures may be affected by the nature of the medium that the participant is interacting with. However, in most cases, there were statistical differences that allowed for the automatic classification of different cognitive loads in correspondence to game difficulty.

In our research, biosignals were the base and most thorough data we obtained; saying so, one must also note that in intrusive methods of data harvesting, even with only stuck electrodes to the skin, it can, in some individuals, generate minor levels of agitation or stress. Thus the answer to the second question is positive, meaning the biosignals are relevant and should be used for cognitive load classification.

Within our study, the most useful measures were performance-based ones, mostly those considering pupilometry and GSR. (For pupilometric measures, it was LHIPA and z-scored LHIPA; for performance-based measures, it was advancement and retries per minute).

Although the results are promising there are a couple of issues that should be mentioned. Our participant pool lacked any individuals on an expert level in racing video games, and the number of participants was also low. The fact that participants were from very different nationalities also could have an effect on the data; due to the nature of the study and its limitations, strict and unnatural game time constraints were put into place. In future works, these variables could be accounted for to further develop our understanding of CL modeling for gamers.

References

1. Borysiuk, M.: Modeling cognitive load of computer game users using pupillometry and other biometric techniques. Master's thesis, Wrocław University of Science and Technology, Wrocław, Poland (June 2022)
2. Charles, R., Nixon, J.: Measuring mental workload using physiological measures: A systematic review. Applied Ergono. **74**, 221–232 (08 2018). https://doi.org/10.1016/j.apergo.2018.08.028
3. Drabek, M.: Modeling cognitive load of computer game users using electroencephalography and other biometric techniques. Master's thesis, Wrocław University of Science and Technology, Wrocław, Poland (June 2022)
4. Duchowski, A.T., Krejtz, K., Gehrer, N.A., Bafna, T., Bækgaard, P.: The low/high index of pupillary activity. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. pp. 1–12 (2020)
5. Fortin-Cote, A., Chamberland, C., Parent, M., Tremblay, S., Jackson, P., Beaudoin-Gagnon, N., Campeau-Lecours, A., Bergeron-Boucher, J., Lefebvre, L.: Predicting Video Game Players' Fun from Physiological and Behavioural Data: Proceedings of the 2018 Future of Information and Communication Conference (FICC), Vol. 1, pp. 479–495. Springer (01 2019)
6. Gutierrez, G., Lunsky, I.O., Van Heer, S., Szulewski, A., Chaplin, T.: Cognitive load theory in action: e-learning modules improve performance in simulation-based education. a pilot study. Can. J. Emerg. Med. **25**(11), 893–901 (2023)
7. Lecoutre, L., Lini, S., Bey, C., Lebour, Q., Favier, P.A.: Evaluating EEG Measures as a Workload Assessment in an Operational Video Game Setup. In: PhyCS 2015: 2nd International Conference on Physiological Computing Systems. Angers, France (Feb 2015). https://doi.org/10.5220/0005318901120117, https://hal.science/hal-02506017
8. Liu, Y., Gong, Q., Gong, X.: Research on interface design of meteorological cloud radars based on cognitive load theory. In: 2020 IEEE 3rd International Conference of Safe Production and Informatization (IICSPI). pp. 88–92. IEEE (2020)
9. Mathôt, S., Fabius, J., Van Heusden, E., Van der Stigchel, S.: Safe and sensible preprocessing and baseline correction of pupil-size data. Behav. Res. Methods **50**(1), 94–106 (2018). https://doi.org/10.3758/s13428-017-1007-2

10. Mitre-Hernandez, H., Carrillo, R.C., Lara-Alvarez, C., et al.: Pupillary responses for cognitive load measurement to classify difficulty levels in an educational video game: Empirical study. JMIR Serious Games **9**(1), e21620 (2021)
11. Orru, G., Longo, L.: The evolution of cognitive load theory and the measurement of its intrinsic, extraneous and germane loads: A review. In: Longo, L., Leva, M.C. (eds.) Human Mental Workload: Models and Applications, pp. 23–48. Springer International Publishing, Cham (2019)
12. Paas, F., Tuovinen, J.E., Tabbers, H., Gerven, P.W.M.V.: Cognitive load measurement as a means to advance cognitive load theory. Cognitive Load Theory (2016), https://www.taylorfrancis.com/chapters/edit/10.4324/9780203764770-8/cognitive-load-measurement-means-advance-cognitive-load-theory-fred-paas-juhani-tuovinen-huib-tabbers-pascal-van-gerven
13. Pusey, M., Wong, K.W., Rappa, N.A.: The puzzle challenge analysis tool. a tool for analysing the cognitive challenge level of puzzles in video games. Proc. ACM Hum.-Comput. Interact. **5**(CHI PLAY) (oct 2021). https://doi.org/10.1145/3474703, https://doi.org/10.1145/3474703
14. Sweller, J.: Cognitive load theory, learning difficulty, and instructional design. Learn. Instr. **4**, 295–312 (1994), https://www.sciencedirect.com/science/article/pii/0959475294900035
15. Sweller, J.: Cognitive load during problem solving: Effects on learning. Cognitive Sci. **12**(2), 257–285 (1988). https://doi.org/10.1016/0364-0213(88)90023-7, https://www.sciencedirect.com/science/article/pii/0364021388900237
16. Zmysłowski, J.W.: Modeling cognitive load of computer game users using eye tracking and other biometric techniques. Master's thesis, Wrocław University of Science and Technology, Wrocław, Poland (June 2022)

Behavioral Intention and Use of ChatGPT Among Accounting and Finance Students: A Two-Year Comparative Study

Ewa Wanda Ziemba(✉), Dariusz Grabara, Katarzyna Renik, and Ewa Wanda Maruszewska

University of Economics in Katowice, 40-287 Katowice, Poland
{ewa.ziemba,ewa.maruszewska}@ue.katowice.pl, {dariusz.grabara, katarzyna.renik}@uekat.pl

Abstract. This study assesses the behavioral intention and use of ChatGPT among accounting and finance higher education students, highlighting differences over the years 2023 and 2024. Utilizing survey data from 237 students in 2023 and 288 students in 2024, the study employed frequency analysis and the independent-samples Mann-Whitney U test. The results confirmed that both the levels of behavioral intention and use of ChatGPT increased significantly over the examined years. This research provides valuable insights into the growing fascination and engagement with ChatGPT and its expanding applications in the educational domain. However, using ChatGPT for educational purposes is a relatively new phenomenon that requires further exploration and empirical validation across diverse research samples and disciplines.

Keywords: ChatGPT · Accounting · Finance · Generative AI · Education · Behavioral intention · Use

1 Introduction

Some studies have explored ChatGPT's potential applications in the accounting and finance fields, confirming its ability to reshape various accounting processes by automating repetitive tasks, enhancing financial and managerial reporting and analysis, improving auditing and tax practices, and simplifying client interactions [1–3]. Nonetheless, accounting and finance higher education face challenges in effectively integrating ChatGPT's capabilities into the learning process [4], even though young people seem to benefit from it during the learning process.

Generative artificial intelligence, particularly ChatGPT, has emerged as one of the most promising tools in recent times [5], potentially revolutionizing higher education [6]. ChatGPT presents valuable opportunities for students, such as personalized feedback, enhanced accessibility, interactive conversations, lecture preparation, performance evaluation, and innovative methods for learning and understanding concepts and theories [7]. According to a literature review by Faisal [8], ChatGPT's can reshape the educational

M. Hernes et al. (Eds.): ECAI 2024, LNNS 1217, pp. 64–76, 2025.
https://doi.org/10.1007/978-3-031-78465-1_6

ecosystem, improve research, writing, and bridge language gaps, while Arista et al. [9] emphasized its capabilities in assisting students with academic writing, language instruction, and classroom learning. Rahman and Watanobe [10] highlighted ChatGPT's role as a valuable assistant for learners. It can help students develop their reading and writing skills, create explanations and step-by-step solutions to given problems, enhance problem-solving abilities, and foster analytical and out-of-the-box thinking. Additionally, it supports learners with disabilities by providing services such as speech-to-text and text-to-speech.

The literature review findings by Bhullar et al. [11] indicated that the use of ChatGPT in higher education has been studied extensively by scholars in the United States but less so in other countries. Adams et al. [12] found that a large majority of students perceive ChatGPT as useful for academic purposes and augmenting their learning experience. Costa et al. [5] examined Portuguese graduate students. More than half reported using ChatGPT in an academic context, primarily to search for information and generate initial ideas for tackling topics. Based on the literature synthesis, Bhullar et al. [11] revealed that the use of ChatGPT in higher education has primarily focused on the medical and tourism sectors, indicating a need for researchers to explore its applications in other disciplines. We found only a few studies regarding the use of ChatGPT by accounting and finance students. Wood et al. [13] presented findings on the performance of ChatGPT in the accounting field. They found that ChatGPT provided accurate responses for 56.5% of the questions and partially correct answers for an additional 9.4%. However, students performed significantly better, with an average score of 76.7% on assessments compared to ChatGPT's 56.5%. Despite this, ChatGPT outperformed the student average in 15.8% of the assessments. Abeysekera [4] explored how ChatGPT can support students in accounting course units. Based on the findings, while ChatGPT can support students in learning accounting by providing solutions to both numerical-based and narrative-based questions, it often produces incorrect answers due to misunderstanding underlying assumptions and the complexity of the tasks. However, the improved performance of ChatGPT4, scoring in the 90th percentile for Introductory Financial Accounting and the 70th percentile for Advanced Financial Accounting, indicates its potential as a valuable, albeit imperfect, learning aid. Based on the literature review, it can be concluded that there is limited direct information on whether accounting and finance students use ChatGPT in learning.

The use of ChatGPT among higher education students is influenced by various predictors such as habit, performance expectancy, effort expectancy, learning value, behavioral intention, and hedonic motivation [14–16]. According to Cai et al. [16], behavioral intention predicts learning effectiveness in ChatGPT-assisted language learning better than other predictors like perceived satisfaction and performance expectancy. Furthermore, Strzelecki [14], Acosta-Enriquez et al. [17], and Ziemba et al. [18] also confirmed that behavioral intention has the most significant effect on the use of ChatGPT by students.

However, behavioral intention to use technology does not always lead to actual use, leading to the 'intention-behavior gap' [19]. Therefore, it is interesting to examine the levels of behavioral intention and usage of ChatGPT among accounting and finance higher education students and how these levels have changed over the years. It seems

particularly interesting when because the possibilities of ChatGPT in supporting students' learning overall [8], including in accounting and finance [1], and the influence of students' behavioral intention on the use of ChatGPT [14] are being explored, there is a lack of studies on the levels of behavioral intention and usage of ChatGPT among accounting and finance higher education students and how these have changed over the years.

The main objectives of this paper are to assess the levels of behavioral intention and use of ChatGPT among accounting and finance higher education students and to indicate the differences in these levels over the years. Using survey questionnaires conducted in 2023 and 2024 and statistical analyses, the study focuses on addressing the following four research questions:

RQ1: What is the level of behavioral intention to use ChatGPT among accounting and finance higher education students?
RQ2: What is the level of ChatGPT use among accounting and finance higher education students?
RQ3: How have the levels of behavioral intention and use of ChatGPT among accounting and finance higher education students changed over the past two years?

The remainder of the paper is organized as follows. The next section discusses the behavioral intention and use of ChatGPT in accounting and finance learning by higher education students. Following that, the methodology employed in this study is presented. The subsequent sections report the research findings along with their discussion. Finally, the last section concludes the paper.

2 Theoretical Background and Research Hypotheses Development

Behavioral intention is a core concept in technology adoption [20]. The behavioral intention to use technology refers to an individual's conscious plan and readiness to use a given technology for a specific task or purpose [19, 20]. This concept is also defined as the likelihood of an individual using a particular technology [21]. In this research, the behavioral intention to use ChatGPT is defined as the willingness of higher education students to use ChatGPT to learn accounting and finance. It represents their inclination and desire to use ChatGPT. Behavioral intention to use ChatGPT among accounting and finance higher education students can be measured using a three-item scale adapted from Davis [22] and Venkatesh et al. [23] and extensively used in much of the previous individual technology acceptance research. Based on these items for measuring behavioral intention, the following three hypotheses were developed:

H1: There are statistically significant differences in the behavioral intention to continue using ChatGPT among accounting and finance higher education students between 2023 and 2024
H2: There are statistically significant differences in the behavioral intention always to try using ChatGPT among accounting and finance higher education students between 2023 and 2024
H3: There are statistically significant differences in the behavioral intention to use ChatGPT frequently among accounting and finance higher education students between 2023 and 2024

Numerous researchers have found that behavioral intention significantly impacts technology use [20], including its use in learning [14, 24]. ChatGPT can be used by students in various learning activities. It can enhance academic writing by providing immediate feedback, generating content ideas, and ensuring cohesive organization within students' writing [25]. Additionally, it provides personalized feedback, interactive learning, and individualized assistance, enhancing students' understanding and exploration of complex concepts [26, 27]. Ziemba et al. [18] identified several student learning activities supported by ChatGPT, including writing theses, working on class assignments, doing homework, preparing projects, writing essays and papers, self-studying, and gaining knowledge and skills on a given topic. Based on these items for measuring the use of ChatGPT, the following four hypotheses were developed:

H4: There are statistically significant differences in using ChatGPT among accounting and finance higher education students for writing theses between 2023 and 2024
H5: There are statistically significant differences in using ChatGPT among accounting and finance higher education students for working on classes between 2023 and 2024
H6: There are statistically significant differences in using ChatGPT among accounting and finance higher education students for completing homework between 2023 and 2024
H7: There are statistically significant differences in using ChatGPT among accounting and finance higher education students for self-studying between 2023 and 2024

3 Material and Methods

3.1 Questionnaire

The Likert-type instrument—a survey questionnaire—was developed using the method proposed by Venkatesh et al. [23] and Ziemba et al. [18]. Questions were organized to reflect theoretical constructs, as shown in Table 1.

For the questions related to behavioral intention to use ChatGPT, the scale's descriptions were: 7 – strongly agree, 6 – agree, 5 – somewhat agree, 4 – neither agree nor disagree, 3 – somewhat disagree, 2 – disagree, 1 – strongly disagree. For the questions related to ChatGPT use, the scale's descriptions were: 6 – once a day or more often, 5 – once a week or more often, 4 – several times a month, 3 – several times, 2 – only once, 1 – never.

3.2 Sample and Data Collection

The questionnaire was distributed via Google Forms and sent directly by the Google Classroom platform to accounting and finance students at the University of Economics in Katowice, Poland. The data were collected during two periods: between May 8th, 2023, and June 8th, 2023, as well as between March 20th, 2024, and April 4th, 2024. The studies were designed to be treated as independent to account for the dynamic changes in the subject matter over the two years. While we recognize the value of identifying recurring respondents, the core objective was to capture the evolution and variability of the data across different cohorts. Therefore, the impact of recurring respondents, if any, was not considered a primary factor in our analysis. Additionally, the study was treated

Table 1. Survey questions.

Id	Question	Variable	Source
Behavioral intention to use ChatGPT (BI)			
1	I intend to continue using ChatGPT in my future education	BI1	[23]
2	I will always try to use ChatGPT in my daily education	BI2	
3	I plan to continue to use GenAI in my education frequently	BI3	
Use ChatGPT (USE)			
4	I use ChatGPT for writing my thesis (undergraduate/master/doctoral/postgraduate)	USE1	[18]
5	I use ChatGPT while working on classes at the university	USE2	
6	I use ChatGPT for completing homework, preparing projects, composing papers, writing essays	USE3	
7	I use ChatGPT for self-study, acquiring knowledge and skills in an interesting topic related to my accounting and finance education	USE4	

as independent due to the changing student population, which may not have included students who graduated the previous year and could no longer be surveyed.

According to Kock [28], a minimum sample size of 100–200 observations is required to ensure the accuracy and validity of the findings. A total of 237 valid responses were collected in 2023, and 288 responses were collected in 2024, resulting in response rates of 14% and 17%, respectively. This means that 14% and 17% of all accounting and finance students at the university participated in 2023 and 2024, respectively. The sampling error for an infinite population was approximately 6% in 2023 and 5% in 2024 at a confidence level of 95% (assuming $p = q = 0.5$). Hence, the sample sizes of 237 and 288 students were considered appropriate.

3.3 Respondents Profile

The basic descriptive profile of respondents is shown in Table 2.

The results indicate that most answers were from bachelor's degree students (56.5% in 2023, 62.2% in 2024), while the rest accounted for master's and higher degree students (43.5% in 2023, 37.8% in 2024). The group was also representative due to the greater involvement of females (70% in 2023, 63.9% in 2024) than males (28.7% in 2023, 31.6% in 2024), as it is registered that more women are attending in the presented field of study.

3.4 Methods

The database has been processed to avoid incorrect answers. The process involved identifying and addressing missing or incomplete values and detecting automated answers. Automated answers, such as choosing the same answer for all questions, resulted in the rejection of those responses.

Table 2. Respondents' profile

Characteristics		Students in 2023		Students in 2024	
		Number	%	Number	%
Level of Study	1st degree (bachelor)	134	56.5%	179	62.2%
	2nd degree (master) or higher	103	43.5%	109	37.8%
Gender	Female	166	70.0%	184	63.9%
	Male	68	28.7%	91	31.6%
	Other	3	1.3%	13	4.5%

The data were stored in Microsoft Excel format and analyzed using IBM SPSS Statistics ver.29 and Microsoft Excel. The frequency analysis was used to assess the levels of behavioral intention and use of ChatGPT among accounting and finance higher education students. Frequency analysis was used to evaluate the levels of behavioral intention and use of ChatGPT among accounting and finance higher education students. Furthermore, the independent-sample Mann-Whitney U test was employed to determine if there were statistically significant differences between the distributions of scores for these levels in 2023 and 2024. The Mann-Whitney U test was used to find statistical differences between the two groups [29]. It is important to note that the test results hold true without continuity assumptions [30], making it useful for comparing ordinal data.

4 Results

4.1 Behavioral Intention to Use ChatGPT

Frequency analysis was employed to answer research question **RQ1**. The results presented in Table 3 clearly illustrate that the behavioral intention levels to use ChatGPT among accounting and finance higher education students differ across the measured periods.

In 2023, 60% of students, and in 2024, 73% of students confirmed that they strongly agree, agree, or somewhat agree with the behavioral intention to continue using ChatGPT (BI1). For the behavioral intention to always try using ChatGPT (BI2), there were 25% of students in 2023 and 32% of students in 2024. Regarding the behavioral intention to use ChatGPT frequently (BI3), there were 42% of students in 2023 and 45% of students in 2024.

In summary, it can be stated that the behavioral intention to use ChatGPT among accounting and finance higher education students for educational purposes increased over the two years. These findings extend previous research that determined the impact of student behavioral intention to use ChatGPT [14, 16] by assessing the level of students' behavioral intention.

Table 3. Level of behavioral intention to use ChatGPT

Variable	Year	Level						
		1	2	3	4	5	6	7
BI1	2023	8%	9%	8%	16%	19%	13%	28%
	2024	3%	3%	8%	12%	28%	20%	25%
BI2	2023	24%	17%	17%	17%	12%	8%	5%
	2024	17%	15%	18%	18%	19%	4%	9%
BI3	2023	16%	12%	14%	17%	17%	14%	11%
	2024	11%	9%	16%	19%	22%	9%	14%

4.2 Use ChatGPT

Regarding research question **RQ2**, frequency analysis was employed. The results presented in Table 4 clearly illustrate that the ChatGPT use levels among accounting and finance higher education students differ across the measured periods.

Most accounting and finance higher education students confirmed that they used ChatGPT for educational purposes in both periods, except for using ChatGPT for writing theses (USE1). This may be due to the structure of the research sample, in which a significant number of students have not yet started writing their theses. Concerning the use of ChatGPT while working on classes at the university (USE2), in 2023, 49% of students had never used ChatGPT, whereas, in 2024, only 21% of students had not used it. Regarding the use of ChatGPT for completing homework (USE3), in 2023, 45% of students had never used ChatGPT for this purpose, while in 2024, only 17% of students had not used it. Regarding using ChatGPT for self-study (USE4), in 2023, 33% of students had never used ChatGPT for this, whereas in 2024, only 19% of students had not used it.

Table 4. Level of ChatGPT use

Variable	Year	Level					
		1	2	3	4	5	6
USE1	2023	78%	8%	8%	2%	2%	3%
	2024	67%	5%	15%	7%	5%	2%
USE2	2023	49%	6%	27%	7%	8%	3%
	2024	21%	6%	38%	20%	13%	2%
USE3	2023	45%	10%	29%	7%	6%	3%
	2024	17%	7%	44%	19%	10%	3%
USE4	2023	33%	8%	31%	6%	11%	11%
	2024	19%	7%	36%	24%	10%	4%

In summary, the research findings suggest that ChatGPT is gaining popularity among students for solving educational tasks during university classes, completing homework such as preparing projects, composing papers, and writing essays, as well as for self-study, such as acquiring knowledge and skills in topics related to university accounting and finance education. These findings extend previous research that determined how ChatGPT can become a human agent for accounting students in financial accounting course units [4] and compared ChatGPT and student performance in accounting [13]. Neither of these studies assessed the level of use of ChatGPT by accounting and finance higher education students.

4.3 Differences in Behavioral Intention and Use of ChatGPT

Addressing research question **RQ3**, the Mann-Whitney U Test was employed. The research results allowed us to identify differences in the levels of behavioral intention and use of ChatGPT among accounting and finance higher education students over the past two years (Table 5). It was revealed that the asymptotic significance is >0.05 for the intention to use ChatGPT frequently among accounting and finance higher education students (BI3), whereas this measure is $<=0.05$ for the remaining variables.

In summary, significant differences were observed in the intention to continue using ChatGPT and always try using ChatGPT among accounting and finance higher education students between 2023 and 2024. However, such differences were not confirmed in the intention to use ChatGPT frequently among accounting and finance higher education students between 2023 and 2024. Overall, the behavioral intention to use ChatGPT among accounting and finance higher education students increased over the two years. Furthermore, the use of ChatGPT has also increased over the years, and the difference in using ChatGPT among accounting and finance higher education students was statistically significant between 2023 and 2024. Therefore, hypotheses H1, H2, H4, H5, H6, and H7 are supported, while H3 is not supported.

5 Discussion

Our results indicate that ChatGPT, the new AI technology, is accepted and increasingly used by students for educational purposes. Consequently, recommendations for higher education are necessary. The following are particularly important:

1. Integrating ChatGPT into curricula:
   - incorporating ChatGPT as a supplementary tool in accounting and finance courses can help students solve complex problems, understand difficult concepts, and gain additional insights;
   - developing specific assignments and projects that leverage ChatGPT to enhance learning outcomes and encourage students to explore its capabilities.
2. Providing training and resources:
   - offering training sessions for students and faculty on effectively using ChatGPT for educational purposes, including best practices, potential limitations, and ethical considerations;

Table 5. Mann-Whitney U Test Summary

Statistics	Variables						
	BI1	BI2	BI3	USE1	USE2	USE3	USE4
Number of students	525	525	525	525	525	525	525
Mann-Whitney U	37501.5	37977.5	36459.0	38444.5	45009.5	45818.0	37962.5
Wilcoxon W	79117.5	79593.5	78075.0	80060.5	86625.5	87434.0	79578.5
Test Statistic	37501.5	37977.5	36459.0	38444.5	45009.5	45818.0	37962.5
Standard Error	1694.472	1705.64	1709.012	1370.363	1660.188	1657.624	1676.781
Standardized Test Statistic	1.991	2.257	1.364	3.150	6.554	7.052	2.287
Asymptotic Sig. (2-sided test)[a]	0.046	0.024	0.173	0.002	<.001	<.001	.022
Summary of hypotheses test	Reject the null hypothesis	Reject the null hypothesis	Retain the null hypothesis	Reject the null hypothesis	Reject the null hypothesis	Reject the null hypothesis	Reject the null hypothesis
Research hypotheses verification result	H1 is supported	H2 is supported	H3 is not supported	H4 is supported	H5 is supported	H6 is supported	H7 is supported

[a]Notes: The significance level is .05.

- creating resources and guides that outline how ChatGPT can be used for various academic tasks, such as preparing projects, composing papers, writing essays, and dissertations.

3. Enhancing self-study programs:

- encouraging the use of ChatGPT for self-study by integrating it into independent learning modules can help students acquire knowledge and skills in topics related to accounting and finance outside of regular class hours;
- providing students with access to ChatGPT as a study aid, helping them explore topics in greater depth and from different perspectives.

4. Monitoring and evaluating usage:

- continuously monitoring and evaluating the use of ChatGPT in educational settings to assess its impact on learning outcomes and gather feedback from students and teachers to identify areas for improvement;
- conducting regular surveys and studies to understand how the use of ChatGPT evolves over time and adjust educational strategies accordingly.

5. Promoting ethical use:

- educating students and teachers about the ethical implications of using AI tools like ChatGPT and raising awareness of academic integrity and the responsible use of such technology;
- developing guidelines and policies that address the ethical use of ChatGPT, ensuring that it enhances learning rather than replaces critical thinking and original work.

By implementing these recommendations, higher education institutions can effectively leverage ChatGPT to improve educational outcomes, foster innovation, and prepare students for the evolving demands of the accounting and finance professions.

6 Conclusion

This study assesses the levels of behavioral intention and use of ChatGPT among accounting and finance higher education students and indicates the differences between 2023 and 2024. Overall, this study confirmed that the levels of behavioral intention and use of ChatGPT increased over the examined years, and the differences between the years were statistically significant. From year to year, students more often use ChatGPT for solving educational tasks during university classes, completing homework such as preparing projects, composing papers, and writing essays, writing their dissertations as well as for self-study, such as acquiring knowledge and skills in topics related to university accounting and finance education.

Our findings complement recent studies by Strzelecki [14] and Cai et al. [16], which investigated the impact of students' behavioral intention to use ChatGPT at a single point in time. Our findings also added new insights into cross-sectional study studies on the use Chat GPT for accounting and finance, e.g., Abeysekera's [4] determined how ChatGPT can become a human agent for students in financial accounting course

units and Wood et al. [14] compared ChatGPT and student performance for questions from accounting assessments and textbook test banks. Compared to these studies, our research assessed the levels of behavioral intention and use of ChatGPT. Additionally, our research was longitudinal in nature and conducted over two periods. In summary, our study contributes to the existing body of knowledge by providing a longitudinal analysis of the behavioral intention and use of ChatGPT among accounting and finance higher education students. It confirms the increasing acceptance and ChatGPT use over time and offers a comprehensive view of how this technology is integrated into various educational tasks during and outside of regular class hours. This research underscores the importance of continuous monitoring and adaptation of AI tools like ChatGPT in academic settings to enhance learning outcomes and support educational innovation.

Like any exploratory study, this research has several limitations. First, ChatGPT and its application for educational purposes are relatively new phenomena that require further exploration and repeated empirical validation. Second, the sample consisted of students from the University of Economics in Katowice, Poland. Therefore, caution should be exercised when generalizing the findings to other fields of study, regions, and countries.

This study confirms the recent surge in interest and use of ChatGPT globally, indicating that students' interaction with AI is rising. However, this new phenomenon requires further exploration and empirical validation. Future research could evaluate the use of ChatGPT among higher education students and teachers across different disciplines within the social sciences and in various countries. Subsequent studies might aim to develop guidelines and best practices for incorporating AI tools like ChatGPT into academic settings, ensuring a balance between technological innovation and maintaining academic rigor.

References

1. Zhao, J., Wang, X.: Unleashing efficiency and insights: exploring the potential applications and challenges of ChatGPT in accounting. J. Corp. Account. Finance **35**(10), 269–276 (2023). https://doi.org/10.1002/jcaf.22663
2. Biancone, P., Chmet, F.: Role of ChatGPT in the accounting field. In: de Bem Machado, A., Sousa, M.J., Dal Mas, F., Secinaro, S., Calandra, D. (eds.) Digital Transformation in Higher Education Institutions. EAI/SICC, pp. 139–153. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-52296-3_8
3. Khan, M.S., Umer, H.: ChatGPT in finance: applications, challenges, and solutions. Heliyon **10**(2), e24890 (2024). https://doi.org/10.1016/j.heliyon.2024.e24890
4. Abeysekera, I.: ChatGPT and academia on accounting assessments. J. Open Innov. Technol. Mark. Complex. **10**(1), 100213 (2024). https://doi.org/10.1016/j.joitmc.2024.100213
5. Costa, R., Costa, A.L., Carvalho, A.A.: Use of ChatGPT in higher education: a study with graduate students. In: de Bem Machado, A., Sousa, M.J., Dal Mas, F., Secinaro, S., Calandra, D. (eds.) Digital Transformation in Higher Education Institutions. EAI/SICC, pp. 121–137. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-52296-3_7
6. Rawas, S.: ChatGPT: empowering lifelong learning in the digital age of higher education. Educ. Inf. Technol. **29**, 6895–6908 (2024). https://doi.org/10.1007/s10639-023-12114-8
7. Xie, X., Ding, S.: Opportunities, challenges, strategies, and reforms for ChatGPT in higher education. In: International Conference on Educational Knowledge and Informatization (EKI), Guangzhou, China, pp. 14–18 (2023). https://doi.org/10.1109/EKI61071.2023.00010

8. Faisal, E.: Unlock the potential for Saudi Arabian higher education: a systematic review of the benefits of ChatGPT. Front. Educ. **9**, 1325601 (2024). https://doi.org/10.3389/feduc.2024.1325601
9. Arista, A., Shuib, L., Ismail, M.A.: A glimpse of ChatGPT: an introduction of features, challenges, and threads in higher education. In: International Conference on Informatics, Multimedia, Cyber and Informations System (ICIMCIS), Jakarta, Selatan, Indonesia, pp. 694–698 (2023). https://doi.org/10.1109/ICIMCIS60089.2023.10349057
10. Rahman, M.M., Watanobe, Y.: ChatGPT for education and research: opportunities, threats, and strategies. Appl. Sci. **13**, 5783 (2023). https://doi.org/10.3390/app13095783
11. Bhullar, P.S., Joshi, M., Chugh, R.: ChatGPT in higher education - a synthesis of the literature and a future research agenda. Educ. Inf. Technol. (2024). https://doi.org/10.1007/s10639-024-12723-x
12. Adams, D., Chuah, K.M., Devadason, E., et al.: From novice to navigator: students' academic help-seeking behaviour, readiness, and perceived usefulness of ChatGPT in learning. Educ. Inf. Technol. (2023). https://doi.org/10.1007/s10639-023-12427-8
13. Wood, D.A., et al.: The ChatGPT artificial intelligence chatbot: how well does it answer accounting assessment questions? Issues Account. Educ. **38**(4), 1–28 (2023). https://doi.org/10.2308/ISSUES-2023-013
14. Strzelecki, A.: Students' acceptance of ChatGPT in higher education: an extended unified theory of acceptance and use of technology. Innov. High. Educ. **49**, 223–245 (2024). https://doi.org/10.1007/s10755-023-09686-1
15. Foroughi, B., et al.: Determinants of intention to use ChatGPT for educational purposes: findings from PLS-SEM and fsQCA. Int. J. Hum. Comput. Interact., 1–20 (2023). https://doi.org/10.1080/10447318.2023.2226495
16. Cai, Q., Lin, Y., Yu, Z.: Factors influencing learner attitudes towards ChatGPT-assisted language learning in higher education. Int. J. Hum. Comput. Interact. 1–15 (2023). https://doi.org/10.1080/10447318.2023.2226495
17. Acosta-Enriquez, B.G., Arbulú Ballesteros, M.A., Huamaní Jordan, O., et al.: Analysis of college students' attitudes toward the use of ChatGPT in their academic activities: effect of intent to use, verification of information and responsible use. BMC Psychol. **12**, 255 (2024). https://doi.org/10.1186/s40359-024-01764-z
18. Ziemba, E.W., Maruszewska, E.W., Grabara, D., Renik, K.: Acceptance and Use of ChatGPT among accounting and finance higher education students. In: Hernes, M., Wątróbski, J. (eds.) ECAI 2023. LNNS, vol. 1079, pp. 185–202. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-66761-9_16
19. Doleck, T., Bazelais, P., Lemay, D.J.: The role of behavioral expectation in technology acceptance: a CEGEP case study. J. Comput. High. Educ. **30**, 407–425 (2018). https://doi.org/10.1007/s12528-017-9158-9
20. Venkatesh, V., Brown, S.A., Maruping, L.M., Bala, H.: Predicting different conceptualizations of system use: the competing roles of behavioral intention, facilitating conditions, and behavioral expectation. MIS Q. **32**(3), 483–502 (2008)
21. Songkram, N., Chootongchai, S., Osuwan, H., et al.: Students' adoption towards behavioral intention of digital learning platform. Educ. Inf. Technol. **28**, 11655–11677 (2023). https://doi.org/10.1007/s10639-023-11637-4
22. Davis, F.D.: A technology acceptance model for empirically testing new end-user information systems: theory and results [Massachusetts Institute of Technology] (1986). https://dspace.mit.edu/handle/1721.1/15192
23. Venkatesh, V., Thong, J.Y.L., Xu, X.: Consumer acceptance and use of information technology: extending the unified theory of acceptance and use of technology. MIS Q. **36**(1), 157–178 (2012). https://doi.org/10.2307/41410412

24. Ain, N., Kaur, K., Waheed, M.: The influence of learning value on learning management system use: an extension of UTAUT2. Inf. Dev. **32**(5), 1306–1321 (2016). https://doi.org/10.1177/0266666915597546
25. Tseng, Y.-C., Lin, Y.-H.: Enhancing english as a foreign language (EFL) learners' writing with ChatGPT: a university-level course design. Electron. J. e-Learn. **22**(2), 78–90 (2024). https://doi.org/10.34190/ejel.21.5.3329
26. Al Shloul, T., et al.: Role of activity-based learning and ChatGPT on students' performance in education. Comput. Educ. Artif. Intell. **6**, 100219 (2024). https://doi.org/10.1016/j.caeai.2024.100219
27. Binsztok, A., Butryn, B., Hołowińska, K., Owoc, M.L., Sobińska, M.: ChatGPT as a learning tool in business education. Research on students' motivation. In: Mercier-Laurent, E., Kayakutlu, G., Owoc, M.L., Wahid, A., Mason, K. (eds.) AI4KMES 2023. IFIPAICT, vol. 693, pp. 80–91. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-61069-1_7
28. Kock, N.: Minimum sample size estimation in PLS-SEM: an application in tourism and hospitality research. In: Applying Partial Least Squares in Tourism and Hospitality Research, pp. 1–16 (2018). https://doi.org/10.1108/978-1-78756-699-620181001
29. MacFarland, T.W., Yates, J.M.: Mann–Whitney U test . In: MacFarland, T.W., Yates, J.M. (eds.) Introduction to Nonparametric Statistics for the Biological Sciences Using R, pp. 103–132. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-30634-6_4
30. Fay, M.P., Proschan, M.A.: Wilcoxon-Mann-Whitney or t-test? On assumptions for hypothesis tests and multiple interpretations of decision rules. Stat. Surv., 41–39 (2010). https://doi.org/10.1214/09-SS051

Artificial Intelligence as a Tool to Support the Translator's Work. Bibliometric Analysis

Andrzej Greńczuk[1(✉)], Katarzyna Tryczyńska[2], William Sahinguvy[3], Hilaire Nkunzimana[3], Piotr Ogonowski[4], and Iwona Chomiak-Orsa[1]

[1] Wroclaw University of Economics and Business, 53-345 Wroclaw, Poland
{andrzej.grenczuk,iwona.chomiak-orsa}@ue.wroc.pl
[2] University of Wroclaw, 50-137 Wroclaw, Poland
katarzyna.tryczynska@uwr.edu.pl
[3] University of Burundi, Avenue de l'UNESCO No 2, B.P 1550 Bujumbura, Burundi
{william.sahinguvu,hilaire.nkunzimana}@ub.edu.bi
[4] University of Szczecin, 71-004 Szczecin, Poland
piotr.ogonowski@usz.edu.pl

Abstract. Artificial intelligence tools are currently being implemented in most IT solutions. The range of their capabilities and their ever-improving quality are leading to increasing confidence in these solutions. One such area where the use of generative artificial intelligence is becoming increasingly widespread is translation science. The development of free tools to support translation processes is influencing the spread of their use even in professional text translation processes. The research on the scope and validity of the use of AI tools in translation processes was inspired by the cooperation of the authors of this article in Dutch translation. This article presents the first stage of the author's research dedicated to identifying and assessing the quality of translation processes supported by AI tools. The article aims to perform a bibliometric analysis of the scholarly interest in the topic of the use of AI in translation processes.

Keywords: artificial intelligence · translation processes · bibliometric analysis

1 Introduction

The general availability of IT tools supporting the translation of texts, including free tools, is becoming a reason for increasing interest in the development of these tools and in their use for professional purposes, such as the translation of business documents, scientific papers, or other documents that were previously translated only by professional translators. The implementation of generative artificial intelligence in IT tools in the field of translation has significantly accelerated and made the use of automatic translation even more widespread. The increased need to translate texts from different languages is one of the effects of globalization. Business relations, social relations, education, and other human communication processes have assumed a global dimension. In this dimension, language diversity has become one of the key barriers to effective

M. Hernes et al. (Eds.): ECAI 2024, LNNS 1217, pp. 77–89, 2025.
https://doi.org/10.1007/978-3-031-78465-1_7

communication. Unfortunately, the level of proficiency in English, which is generally recognized as an international language, is not always satisfactory for the parties to the communication processes so that the transfer of information is correct. On the other hand, it is not always possible to involve an interpreter. This contributes to distortions in communication processes and hinders business processes, especially on a global scale. Modern IT tools have been supporting translation processes for many years now, but this has so far been done passively.

The development of AI opens new opportunities for the improvement of real-time translation processes with a concomitant increase in creative translation support.

The above considerations contributed to the authors' formulation of research questions:

RQ 1. What trends have been noted in the context of the use of AI in translation in the academic publishing environment?
RQ 2. Which countries have published most frequently in the context of using AI in translation?
RQ 3. Which authors have published articles on the perspective of using AI in translation?
RQ 4. What journals publish papers specifically addressing the use of AI in translation?

The above-mentioned research questions became the determinant for the formulation of the aim of this article, which is to investigate the degree of interest in the above-mentioned problem among scientists and how often this problem is addressed in the literature. The authors of this article will carry out a bibliometric analysis of the above scientific area to diagnose the state of the literature in this scientific area. To carry out the bibliometric analysis, publications from 2010 to 2024 from scientific databases such as Scopus were examined. In the research, the authors used IT tools such as VOSviewer [1], and Bibliometrix and R programming language package [2].

2 Theoretical Background

2.1 Problems in Translation Processes

Carrying out a text translation is a resource- and time-consuming activity. It requires quite a lot of precision to properly choose the right words or phraseological compounds. Depending on the text to be translated, an inadequate translation can cause problems not only in the communicative layer but can also cause legal problems.

While these approaches are beneficial for classification and providing insights into series, they may encounter difficulties in preserving the integrity of local patterns during translation processes. Additionally, the use of machine translation has been commercially successful but has also faced criticism, indicating a notable level of skepticism and concern regarding the accuracy and reliability of automated translation tools [3]. The translation of artificial intelligence (AI) models into clinical settings presents significant challenges. Although deep learning networks have shown exceptional performance in tasks like speech recognition and image captioning, translating these models into practical clinical applications necessitates thoughtful consideration [4]. Challenges such as unintended negative consequences and the necessity for transparency in decision-making processes within AI systems can impede the seamless integration of AI technologies into

healthcare settings [5]. Moreover, translating behavioral data into actionable suggestions for promoting healthier lifestyles through automated systems encounters obstacles due to the lack of human involvement in the process [6]. Developing systems capable of accurately translating behavioral data into personalized recommendations without human intervention remains a complex challenge in leveraging technology to enhance health outcomes. Additionally, translating machine learning algorithms into clinical practice for predicting diabetes complications underscores the importance of integrating these algorithms with classical statistical strategies to derive meaningful insights from data [7]. Furthermore, translating AI ethics principles into practical tools and methods for ensuring ethical AI development and deployment involves a rationalization process heavily reliant on language and discussion to establish agreed-upon norms and standards [8]. Bridging the gap between ethical principles and their implementation in AI systems demands a comprehensive understanding of the practical implications of these principles and their translation into actionable practices within the AI development process. In the domain of machine translation, translating knowledge graph embeddings through dynamic mapping matrices underscores the importance of representing relations as translations between entities to achieve state-of-the-art performance [9]. This approach emphasizes the significance of effectively capturing and representing relationships within knowledge graphs to enhance the performance of embedding models. Additionally, translating semantic matching techniques in search tasks highlights the reliance on semantic matching for various natural language processing applications, including question answering and cross-language information retrieval [10].

The level of difficulty increases when dealing with highly specialized texts or simultaneous interpreting. In the case of the translation of specialized, industry-specific texts, especially those with sophisticated terminology as well as specific phraseological compounds or numerous idiomatic expressions, a special role was played by sworn interpreters and those cooperating with a specific environment, e.g. medicine, law, etc. However, the turnaround time and costs were correspondingly much longer and higher. A similar situation applies to simultaneous interpreting, which requires interpreters not only a high degree of linguistic competence, including a specialized vocabulary, but also other predispositions such as creative and imaginative interpreting, following the context and intentions of the sender.

Therefore, in the traditional form, the translator had to have the necessary skills and competencies, and, in addition, the translation process had to be supported using auxiliary materials such as dictionaries, lexicons, specialized dictionaries, or appropriate documentation.

Figure 1 presents a visualization of the traditional translation process.

In the traditional approach, the person doing the translation must determine the scope of the translation (how long, whether the text is formal in nature or not), and then identify the scope of the translation. This is followed by the execution of the translation. It is worth noting that this task is burdened with a time criterion, which on the one hand can be predetermined or measured. Once the translation is done, the formal processing of the document with the translation takes place, and then we can consider this as the end of it.

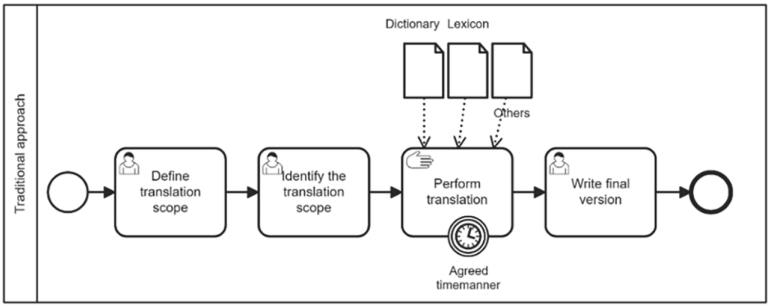


Fig. 1. The traditional translation process.

Although the process presented in the diagram appears relatively straightforward, a number of difficulties may arise during implementation.

In the literature as well as in business, the following aspects have been identified as the biggest problems:

- Low availability of specialized translation services especially in the case of languages with low penetration,
- Long delivery times, especially for long and highly specialized texts,
- extremely low availability of simultaneous interpreting services, or even none at all for rare languages,
- High prices of translation services.

The above-mentioned aspects mainly concern problems classified as the economic availability of translation. In addition, problems resulting from cultural differences, the lack of consideration for the context of the translated text, and the different levels of knowledge between the parties for whom the text is being translated are very important problems that are often overlooked. These difficulties, although much less frequently pointed out in the literature, are nevertheless important, especially in translation processes carried out in the diplomatic, social, legal, or other spheres, which primarily involve the need to empathize with the translation process.

The concept of using IT solutions based on artificial intelligence can significantly improve the efficiency of these processes and, above all, increase their accessibility while reducing costs.

The development of IT tools supporting translation processes available on a mass scale or specialized contributes to their increasingly widespread use.

2.2 AI in Translations

With the development of ICT and the globalization of business processes, there has been a growing interest in IT tools supporting both passive translation (of written texts) and active translation (in face-to-face communication). The thesis process has been further intensified by the popularization of AI tools, whose quality and availability are both increasing. The purpose of these tools is to carry out translation processes so that language barriers between communication subjects are as low as possible. These tools

can not only translate texts into a dozen or even several dozen different languages, but their task is to carry out these processes in real-time, considering the context of the conversation. Using machine and generative learning techniques powered by an ever-increasing amount of text, they are learning and developing their potential. The popularity of these tools is growing, which determines an ever-widening list of programs that can be used in translation processes. Among the most popular are tools such as:

- Google Translate - "Google Translate is an instant translation tool that can be accessed via a web browser or a software application. Using both TTS and ASR, it can translate words, phrases, and full texts from one language to another" [11].
- Smartling Translate – is an example of the CAT tool. The tool itself is a web-based application, which allows bilingual translation. "CAT tools contain no program based on machine translation and no ready-made bilingual dictionary. The "dictionary" is created by the translator with each translation and revision. She explained further that looking for a term with a CAT tool means searching through the previously created translation memory" [12].
- DeepL - is a translator. DeepL has the functionality of DeepL Write, which is an AI-based writing assistant. It corrects the entered text and checks its grammatical syntax. It also enriches the text with phrases typical of the language [13].
- QuillBot AI - is a text paraphrasing tool that uses AI. Its goal is to preserve the intended content while restructuring sentences and using synonyms [14].
- Copy.ai – Copy.ai is an advanced platform for generating high-quality texts that use machine learning algorithms [15]. It offers different pricing options depending on user preferences.
- Microsoft Bing Translator - is a free tool based on machine translation (MT). Its definite advantages are speed of translation, cost, as well as convenience [16].
- Grammarly – is a grammar checker tool available in basic functionality for free. It is a widely used tool around the world [17]. Considering the use of AI, Grammarly offers functionalities for correcting text according to indicated criteria or simplifying it.

When delving into AI in translation, it's important to recognize the wide-ranging effects and difficulties linked with this technology. AI has displayed promise in improving processes, increasing automation, and efficiently communicating with humans [18]. Nevertheless, even with AI progress, human translators are still essential for providing precise, detailed, and culturally sensitive translations [19]. Incorporating AI models in healthcare, like for histopathologic categorization, emphasizes the need to address possible unintended side effects and guarantee the dependability of AI-supported tools [5]. Deep learning, a branch of artificial intelligence, has shown impressive results in activities such as speech recognition, image captioning, and language translation [4]. The transformative potential of AI in bridging communication gaps across diverse linguistic contexts is highlighted by the high levels of performance in recognizing speech and translating text between languages achieved by deep learning networks. Additionally, it is essential to transform AI ethics principles into tangible tools and approaches to guarantee the ethical progress and implementation of AI [8]. This process of translating involves turning ethical principles into practical actions in AI systems to maintain ethical standards and encourage responsible use of AI. Sophisticated algorithms are needed to effectively

learn and represent mappings between different visual domains in image-to-image (I2I) translation [20]. Approaches like varied image-to-image translation through disentangled representations intend to produce a variety of results by utilizing encoded content features and attribute vectors [21]. These methods highlight the intricacy of converting visual data between different fields and the need for sophisticated algorithms to guarantee precise and authentic changes. Moreover, the potential of AI-enabled technologies to enhance communication for individuals with hearing impairments is exemplified by the use of AI in sign language recognition and bidirectional communication through smart gloves [22]. Deep learning structures are essential for identifying sign language movements and facilitating efficient communication in virtual reality environments. Moreover, emphasizing transparency in AI decision-making is crucial when utilizing explainable AI models to predict acute critical illnesses from electronic health records [23]. Explicable AI improves the translation of clinical care by offering explanations for AI predictions, thus building trust and comprehension in healthcare environments. The incorporation of IoT, big data, and AI technologies in agriculture and the food industry signals how AI is revolutionizing agricultural processes and improving food production [24]. Machine learning algorithms are essential for extracting information from unstructured data to complete tasks like notifying users or improving important agricultural procedures. Also, the creation of artificial intelligence tools for predicting demand and optimizing prices in e-commerce showcases the real-world uses of AI in managing revenue and analyzing business data [25]. Machine learning methods allow for precise calculation of past sales records and forecasting of upcoming product demand, improving decision-making in retail environments.

The list of tools using artificial intelligence is much longer and will continue to grow. Scientific studies most often draw attention to the development of AI tools and their increasing effectiveness, accuracy, and speed in translation. The features to be evaluated usually depend on the needs of the users, who decide to choose IT solutions rather than a traditional translator. The choice of a tool that can be used is determined by many factors, but the most common criteria are:

- accuracy of translation – the accuracy and completeness of translations are assessed while considering the context in which the text is translated,
- number of supported languages – some of the tools support only a few languages, other tools can support up to several dozen languages,
- the ability to provide real-time translations – solutions that enable the translation of not only written texts, but also simultaneous interpreters are becoming more and more popular,
- security and privacy – they are of particular importance for formal, business, political, and legal translations, where confidential information is subject to translation,
- access prices – more and more tools have free versions with the possibility of purchasing extended functionalities,
- the ability to integrate with other programs and applications.

The criteria mentioned above determine the choice of an AI solution that can be used in translation processes. However, other criteria may also be considered for specific translations.

3 Methodology

In accordance with the research questions posed, the authors decided to conduct a bibliometric analysis in the first stage of the research, to identify key trends and views on the use of AI in translation processes. In addition, the authors want to identify the authors, research centers, and journals in which the largest number of publications on this subject can be found. Figure 2 presents the research procedure.

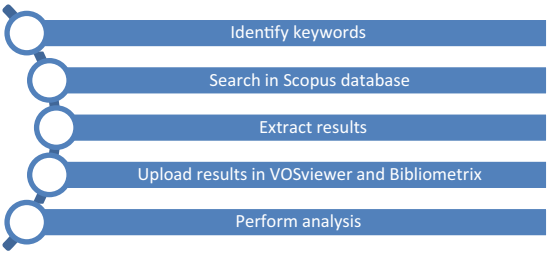


Fig. 2. Research procedure.

Table 1 presents the query search. The table contains information related to inclusion/exclusion criteria.

The final results are 7125 records, which were analyzed with VOSviewer and Bibliometrix.

4 Results

The scope of the obtained collection begins in the 1960s. It can be observed that the largest increase in the number of publications began around 2008. The detailed distribution of publications between 2010 and 2024 is shown in Fig. 3.

As can be seen from the analyzed collection - one can see quite strongly an upward trend. The number of publications indicates interest in the topic. Therefore, it can be assumed that the topic will continue to be updated.

Keywords for translation based on the query used show the relationship between translation keywords and artificial intelligence methods. This relationship is shown in Fig. 4.

From the keyword analysis presented, it can be pointed out that there is a stronger relationship between the keyword's "translation" "artificial intelligence" and "machine learning." In addition, there is also a weaker relationship between the words "natural language processing" and "deep learning." It can be said that the strong relationship occurring with the word "artificial intelligence" is related to the fact that the other keywords are components of this word. Recent years have shown that the development of artificial intelligence uses natural language processing solutions - these solutions are used in the development of ChatGPT.

The next step is to analyze the countries that have the most frequently published papers on the subject under analysis. The detailed distribution of countries is shown in Fig. 5.

Table 1. The query search.

Query	Inclusion/exclusion criteria	Publication number
TITLE-ABS-KEY ((“AI” OR “artificial intelligence”) AND “translat*”)	Inclusion: none Exclusion: none	11904
TITLE-ABS-KEY ((“AI” OR “artificial intelligence”) AND “translat*”) AND (LIMIT-TO (LANGUAGE, “English”))	Inclusion: language to english	11587
TITLE-ABS-KEY ((“AI” OR “artificial intelligence”) AND “translat*”) AND (LIMIT-TO (LANGUAGE, “English”)) AND (LIMIT-TO (DOCTYPE, “cp”) OR LIMIT-TO (DOCTYPE, “ar”))	Inclusion: publication type as article and conference paper	9540
TITLE-ABS-KEY ((“AI” OR “artificial intelligence”) AND “translat*”) AND PUBYEAR > 2009 AND PUBYEAR < 2025 AND (LIMIT-TO (LANGUAGE, “English”)) AND (LIMIT-TO (DOCTYPE, “cp”) OR LIMIT-TO (DOCTYPE, “ar”))	Inclusion: publication year between 2010 and 2024	7125

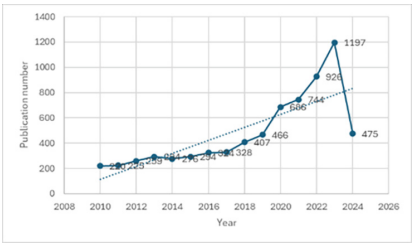


Fig. 3. Distribution of publications between 2010 and 2024.

The figure shows that the largest number of publications comes from countries such as the US, and China - these countries had a frequency of more than 5,000, followed by countries with a frequency between 1,000 and 2,000, namely India, Germany, the UK, Italy. The remaining countries have frequencies below 1,000.

The 10 authors who published most frequently are shown in Fig. 6.

As can be seen, some of them publish continuously, and some of them have been publishing since a certain point. It can be indicated that the analyzed authors continue to publish in the analyzed field.

An analysis of the 10 journals with the most frequent publications is shown in Fig. 7.

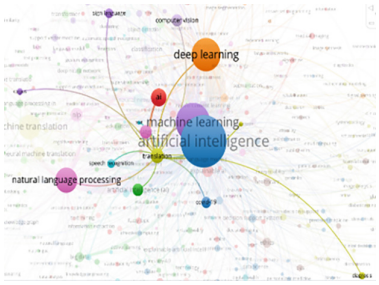


Fig. 4. The relationship between translation keywords and artificial intelligence.

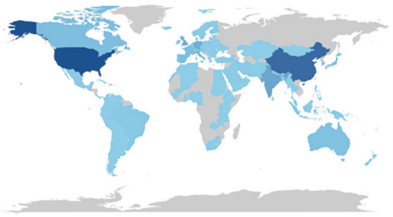


Fig. 5. The detailed distribution of countries.

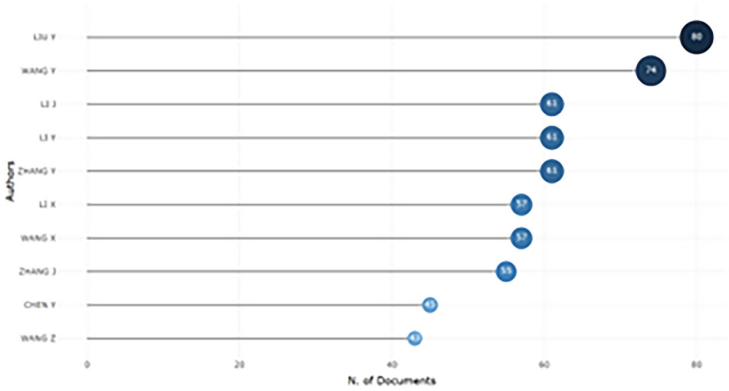


Fig. 6. The 10 authors who published most frequently.

Analyzing the results, it can be noted that:

1. The most frequently used journals are those that publish research results from conferences, such as Lectures Notes In Computer Science (Springer publications), ACM International Conference Proceedings, Ceur Workshop Proceedings, and Lecture Notes In Networks and Systems (Springer publications).
2. Journals that are not strictly related to translation but focus on issues related to computer science.

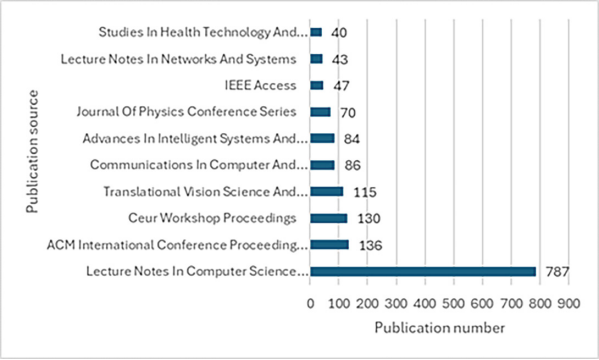


Fig. 7. The 10 journals with the most frequent publications.

5 Conclusion

The bibliometric analysis carried out showed that the subject concerning the use of artificial intelligence methods in translation is a topic of interest in the literature. In addition, it can be pointed out that recent years, including the development of new tools, have made it possible to quickly translate texts (in the form of files, loose passages, and real-time statements) at a good level. Of course, it should be noted that such a translation should be additionally checked for translation noises. Synthetic conclusions providing answers to the research questions posed are presented in Table 2.

As further research, the authors plan to carry out a systematic literature review that will make explicit the current knowledge, to research the comparison of tools based on artificial intelligence methods in translation (including the analysis of legal texts of different languages - a study of machine translation of legal texts in Dutch is planned).

Table 2. Answer to research questions.

Research question	Answer
RQ 1. What trends have been noted in the context of the use of AI in translation in the academic publishing environment?	It was noted that the topics in the analyzed set of publications show a strongly upward trend. This means that the topics are relevant and important. In addition, it can be pointed out that the development of artificial intelligence will intensify the interest of researchers in the possibility of applying it in the field of translation
RQ 2. Which countries have published most frequently in the context of using AI in translation?	It should be indicated that the countries that show a high frequency of publication are the US and China, whose frequency is between 2000 and 5000
RQ 3. Which authors have published articles on the perspective of using AI in translation?	We can observe that the most frequent publishers were Liu Y (80) and Wang Y. (74). Next in number of 61 publications were written by Li J., Li Y., Zhang Y. Below 60 publications were written by Li X., Wang X., Zhang J., Chen Y. and Wang Z
RQ 4. What journals publish papers specifically addressing the use of AI in translation?	The journals in which the results of the research were published include titles that are strictly related to conferences. In most cases, these journals are aimed at publishing computer science research

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. van Eck, N.J., Waltman, L.: Visualizing bibliometric networks. In: Ding, Y., Rousseau, R., Wolfram, D. (eds.) Measuring Scholarly Impact, pp. 285–320. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10377-8_13
2. Aria, M., Cuccurullo, C.: Bibliometrix: an R-tool for comprehensive science mapping analysis. J. Informetr. **11**, 959–975 (2017). https://doi.org/10.1016/j.joi.2017.08.007
3. Prates, M.O.R., Avelar, P.H., Lamb, L.C.: Assessing gender bias in machine translation: a case study with Google Translate. Neural Comput. Appl. **32**, 6363–6381 (2020). https://doi.org/10.1007/s00521-019-04144-6
4. Sejnowski, T.J.: The unreasonable effectiveness of deep learning in artificial intelligence. Proc. Natl. Acad. Sci. U. S. A. **117**, 30033–30038 (2020). https://doi.org/10.1073/pnas.1907373117
5. Kiani, A., et al.: Impact of a deep learning assistant on the histopathologic classification of liver cancer. Npj Digit. Med. **3** (2020). https://doi.org/10.1038/s41746-020-0232-8

6. Rabbi, M., Pfammatter, A., Zhang, I., Spring, B., Choudhury, T.: Automated personalized feedback for physical activity and dietary behavior change with mobile phones: a randomized controlled trial on adults. JMIR MHealth UHealth **3** (2015). https://doi.org/10.2196/mhealth.4160
7. Dagliati, A., et al.: Machine learning methods to predict diabetes complications. J. Diabetes Sci. Technol. **12**, 295–302 (2018). https://doi.org/10.1177/1932296817706375
8. Morley, J., Floridi, L., Kinsey, L., Elhalal, A.: From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices. Sci. Eng. Ethics **26**, 2141–2168 (2020). https://doi.org/10.1007/s11948-019-00165-5
9. Ji, G., He, S., Xu, L., Liu, K., Zhao, J.: Knowledge graph embedding via dynamic mapping matrix. In: ACL-IJCNLP - Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 687–696. Association for Computational Linguistics (ACL) (2015). https://doi.org/10.3115/v1/p15-1067
10. Li, H., Xu, J.: Semantic matching in search. Found. Trends Inf. Retr. **7**, 343–469 (2013). https://doi.org/10.1561/1500000035
11. van Lieshout, C., Cardoso, W.: Google translate as a tool for self-directed language learning. Lang. Learn. Technol. **26**, 1–19 (2022)
12. Apriliana, F., Kurniawan, A., Ferianda, S., Kastuhandani, F.C.: Introducing a cat tool to translate: Wordfast. Indones. J. Engl. Lang. Stud. IJELS **2**, 60–75 (2017). https://doi.org/10.24071/ijels.v2i1.351
13. Chen, T.-J.: ChatGPT and other artificial intelligence applications speed up scientific writing. J. Chin. Med. Assoc. **86**, 351–353 (2023). https://doi.org/10.1097/JCMA.0000000000000900
14. Amyatun, R.L., Kholis, A.: Can artificial intelligence (AI) like QuillBot AI assist students' writing Skills? Assisting learning to write texts using AI. ELE Rev. Engl. Lang. Educ. Rev. **3**, 135–154 (2023). https://doi.org/10.22515/elereviews.v3i2.7533
15. Marquis, Y.A., Oladoyinbo, T.O., Olabanji, S.O., Olaniyi, O.O., Ajayi, S.A.: Proliferation of AI tools: a multifaceted evaluation of user perceptions and emerging trend. Asian J. Adv. Res. Rep. **18**, 30–35 (2024). https://doi.org/10.9734/ajarr/2024/v18i1596
16. Moneus, A.M., Sahari, Y.: Artificial intelligence and human translation: a contrastive study based on legal texts. Heliyon **10**, e28106 (2024). https://doi.org/10.1016/j.heliyon.2024.e28106
17. Fitria, T.N.: Grammarly as AI-powered English writing assistant: students' alternative for writing English. Metathesis J. Engl. Lang. Lit. Teach. **5**, 65 (2021). https://doi.org/10.31002/metathesis.v5i1.3519
18. Wamba-Taguimdje, S.-L., Fosso Wamba, S., Kala Kamdjoug, J.R., Tchatchouang Wanko, C.E.: Influence of artificial intelligence (AI) on firm performance: the business value of AI-based transformation projects. Bus. Process. Manag. J. **26**, 1893–1924 (2020). https://doi.org/10.1108/BPMJ-10-2019-0411
19. Heer, J.: Agency plus automation: designing artificial intelligence into interactive systems. Proc. Natl. Acad. Sci. U. S. A. **116**, 1844–1850 (2019). https://doi.org/10.1073/pnas.1807184115
20. Lee, HY., Tseng, HY., Huang, JB., Singh, M., Yang, M.H.: Diverse image-to-image translation via disentangled representations. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11205, pp. 36–52. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01246-5_3
21. Lee, H.-Y., et al.: DRIT++: diverse image-to-image translation via disentangled representations. Int. J. Comput. Vis. **128**, 2402–2417 (2020). https://doi.org/10.1007/s11263-019-01284-z

22. Wen, F., Zhang, Z., He, T., Lee, C.: AI enabled sign language recognition and VR space bidirectional communication using triboelectric smart glove. Nat. Commun. **12** (2021). https://doi.org/10.1038/s41467-021-25637-w
23. Lauritsen, S.M., et al.: Explainable artificial intelligence model to predict acute critical illness from electronic health records. Nat. Commun. **11** (2020). https://doi.org/10.1038/s41467-020-17431-x
24. Misra, N.N., Dixit, Y., Al-Mallahi, A., Bhullar, M.S., Upadhyay, R., Martynenko, A.: IoT, big data, and artificial intelligence in agriculture and food industry. IEEE Internet Things J. **9**, 6305–6324 (2022). https://doi.org/10.1109/JIOT.2020.2998584
25. Ferreira, K.J., Lee, B.H.A., Simchi-Levi, D.: Analytics for an online retailer: demand forecasting and price optimization. Manuf. Serv. Oper. Manag. **18**, 69–88 (2016). https://doi.org/10.1287/msom.2015.0561

Cognitive Load and Online Customer Decisions: The Role of Visual Perception, Memory, and Brain Activity

Adrianna Mateja[1(✉)], Dawid Subocz[2], Marta Stępień-Słodkowska[3], and Małgorzata Nermend[4]

[1] University of Szczecin, Cukrowa 8, 71-004 Szczecin, Poland
adrianna.mateja@usz.edu.pl
[2] University of Szczecin, Krakowska 69, 71-017 Szczecin, Poland
[3] University of Szczecin, Piastów 40b, 71-065 Szczecin, Poland
[4] University of Szczecin, Ogińskiego 16/17, 71-431 Szczecin, Poland

Abstract. Due to the use of new technologies and ongoing digitization, there is a significant increase in the popularity of online stores. More and more, consumers appreciate the ease of online shopping, enjoying the ability to explore a diverse array of products from the comfort of their own homes. The wide range of products available in online stores allows consumers to precisely tailor their purchases to their individual preferences and needs. However, too much information on a page can lead to cognitive load, affecting visual attention and purchasing decisions. The article focuses on an innovative approach to measuring the impact of cognitive load on the purchasing decision-making process in online stores. The authors focus on psychological aspects, including reducing the influence of external factors, to obtain more reliable results. Understanding how cognitive load affects online shopping decisions is crucial for designing effective, user-friendly e-commerce platforms.

Keywords: Information overload on websites · cognitive neuroscience tools · cognitive neuroscience in user experience

1 Introduction

In today's dynamically developing world, evolution in innovation information and communication technology (ICT) and information technology (IT) play a key role in developing modern information systems [1]. The ability to make expected use of a system is defined as "usability" and includes human interaction with the system [2]. Increasingly, the term usability is complemented by the concept of user experience (UX), which takes into account all the emotions, beliefs, preferences and behaviors that appear before, during and after using a given product [3, 4]. Functionality and usability are key aspects of information system activity, which is why UX research is increasingly an integral part of the customer-centric design process [5]. UX focuses primarily on understanding users' unconscious needs and motivations, enabling the creation of ready-made systems that meet real consumer expectations.

M. Hernes et al. (Eds.): ECAI 2024, LNNS 1217, pp. 90–101, 2025.
https://doi.org/10.1007/978-3-031-78465-1_8

Due to new technologies and ongoing digitization, online stores are increasingly popular. Consumers appreciate the convenience of shopping online, which has become integral to modern lifestyles. E-commerce enables shopping anytime, eliminating time constraints of traditional stores [6]. It provides access to a wide range of products, allowing customers to tailor purchases precisely to their preferences and needs from the comfort of home. However, the vast array of choices and information available in online stores can overwhelm users, highlighting the critical importance of intuitive access to products and effective presentation of information [7–9]. The success of an online store heavily relies on user perception and the design's ability to facilitate seamless navigation and decision-making processes [10, 11].

Cognitive Load Theory (CLT), proposed by John Sweller, posits that human cognitive abilities are limited and can only process a finite amount of information at a time. When the cognitive load exceeds this capacity, users may experience cognitive overload, which can impair decision-making and user experience [12]. Understanding cognitive load in the context of online shopping environments is crucial for optimizing user experience and decision-making processes [13].

The article aims to outline an experimental procedure exploring how visual perception, memory processes, and brain activity relate to decisions made by busy online consumers. Key research questions include:

1. Do cognitively loaded individuals better recall emotional stimuli presented to them compared to neutral stimuli?
2. Is the visual perception of cognitively loaded individuals related to their recall of presented stimuli?
3. Does brain bioelectrical activity during the perception of presented symbols correlate with their subsequent recall?
4. Does the visual perception of stimuli presented during cognitive load relate to preferences in selecting them in decision-making situations?
5. Does the ability to recall stimuli presented during cognitive load relate to preferences in selecting them in decision-making situations?

Based on the research questions posed above, the following hypotheses were formulated:

H1: *Cognitively loaded individuals better recall emotional stimuli presented to them compared to neutral stimuli.*
H2: *There is a relationship between the duration of fixation on stimuli and their subsequent recall.*
H3: *Brain bioelectrical activity during the perception of presented symbols correlates with their subsequent recall.*
H4a: *The duration of fixation on stimuli presented during cognitive load correlates with a preference for selecting them in decision-making situations.*
H4b: *The number of fixations on stimuli presented during cognitive load correlates with a preference for selecting them in decision-making situations.*
H5: *The ability to recall stimuli presented during cognitive load correlates with preferences in selecting them in decision-making situations.*

This research proposal aims to measure the impact of cognitive load on purchasing decisions using psychophysiological methods. By minimizing the influence of factors like knowledge and preferences, we aim for more reliable results. Focusing on psychological aspects enhances measurement validity, shedding light on how cognitive load relates to online purchasing decisions. This study could advance understanding of online consumer behavior and offer new insights into the role of psychological factors in purchasing decisions.

2 Literature Review

Recent research has increasingly focused on how website complexity influences user attitudes and behavior [7–9], particularly concerning the abundance of information available during purchasing decisions. Online stores offer vast options, often overwhelming consumers with stimuli. Therefore, users expect easy and intuitive access to products and information. The success of a website hinges on visitor perception [11], significantly shaping online shopping behavior. Crucial factors for an effective online store include presenting offerings clearly and delivering product information efficiently [10]. Studies consistently indicate that information overload and website complexity affect visual attention and lead to cognitive overload [12–14].

The increasing number of scientific articles on cognitive load over the last 30 years shows a clear interest in this topic [15, 16]. Mental workload is very complex and is closely related to attention [17] and human abilities and effort [18]. From the perspective of John Sweller's [19] cognitive load theory (CLT), load refers to the effort used in working memory. According to CLT, human cognitive abilities are limited and can only process a few items at a time. If the amount of information processed exceeds the limit of working memory, then people become overwhelmed.

Most studies [16, 20–23] distinguish three categories of cognitive load: internal, external and germane. The internal cognitive load is related to the content of the material, the external cognitive load is based on the forms of presentation, and the germane load is based on consolidating information. The complexity of an online store generates external cognitive load, which is modifiable and can be minimized through appropriate visual design of the website in line with UX design [13]. Despite this, there is great variability in how people can handle the load depending on the person's previous experience and general skills. Factors such as knowledge, personality, attitude, stress, fatigue, level of motivation, time constraints, environment or acquired competences may cause differences in the level of perceived load [24–26].

Various tools assess and predict cognitive load, categorized into three main types in the literature [16, 21, 27]: performance measurements, subjective measurements, and physiological measurements. Performance methods analyze task completion time, error rates, and work quality, offering precise but limited results when handling multiple tasks simultaneously [28]. Subjective techniques involve self-assessment using numerical or graphic scales, commonly employing Paas rating scales [29] and NASA-TLX [30]. Despite widespread popularity [31, 32], these methods are not ideal for real-time measurement [33, 34] and may be prone to errors related to recency effects [35] and scale interpretation [36]. Additionally, subjective judgments may be influenced by anchored

experiences [35, 37, 38]. Modern researchers caution against relying solely on subjective assessments to measure mental workload, advocating for supplementing these with other techniques for a more comprehensive understanding [12, 39–41]. Physiological metrics measure mental workload through parameters such as heart rate, skin response, brain activity, and eye reactions [28, 42], providing detailed insights into workload intensity. These measurements are highly sensitive, capable of continuous monitoring [21]. While historically underutilized in UX research due to equipment costs, their increasing affordability has led to more frequent adoption [43].

Cognitive neuroscience research has revolutionized the field of user experience design by enabling a better understanding of user behavior and preferences [44]. In turn, assessing a customer's cognitive load during online shopping is crucial, influencing customer experience and purchasing decisions [45]. Moreover, awareness of cognitive load allows you to identify areas for optimization, which improves the efficiency of online shopping and builds customer trust [44]. Modern information systems should adapt to the individual needs of users and integrate both cognitive and emotional aspects [46].

Previous research has demonstrated that product information, such as brand or price, can shape expectations about product quality, influencing choices significantly [47]. The choice overload hypothesis suggests that having more options can lead to negative outcomes like reduced motivation to choose or satisfaction with the chosen option [48]. Decision-making is influenced by various factors, including environmental cues like product placement on a website and the number of available options [49], as well as biological factors such as memory capacity for product details [50]. In today's world, where cognitive load from abundant online information is prevalent, there is a need for innovative approaches to minimize these influences for more precise and reliable results. Our study aims to explore how cognitive load impacts decision-making processes, integrating psychological and neurocognitive perspectives.

3 Research Framework

In the proposed study, we will define cognitive load in accordance with John Sweller's [19] cognitive load theory. According to this author, cognitive load is the effort used in working memory. In the proposed experiment, we will load the working memory of research participants, while verifying how the type of cognitive load is related to their perception and the decision-making process. Descending to such a reductionist level of analysis of decision-making processes is necessary to draw conclusions about human behavior in a specific environment, including when using websites. The experimental framework is presented in Fig. 1.

Before starting the research project, participants will be informed about the research protocol as well as the measurement devices used. All participants will be required to give written consent to participate in the study and to process personal data. Then, it is planned to collect sociodemographic data about the tested participant and install and calibrate the research equipment.

The Counting Span Test will be used to test working memory. The tested people will perform the task while sitting at the computer. Two types of grains will appear on the monitor screen. These will be green and yellow grains. The subject will have to count the

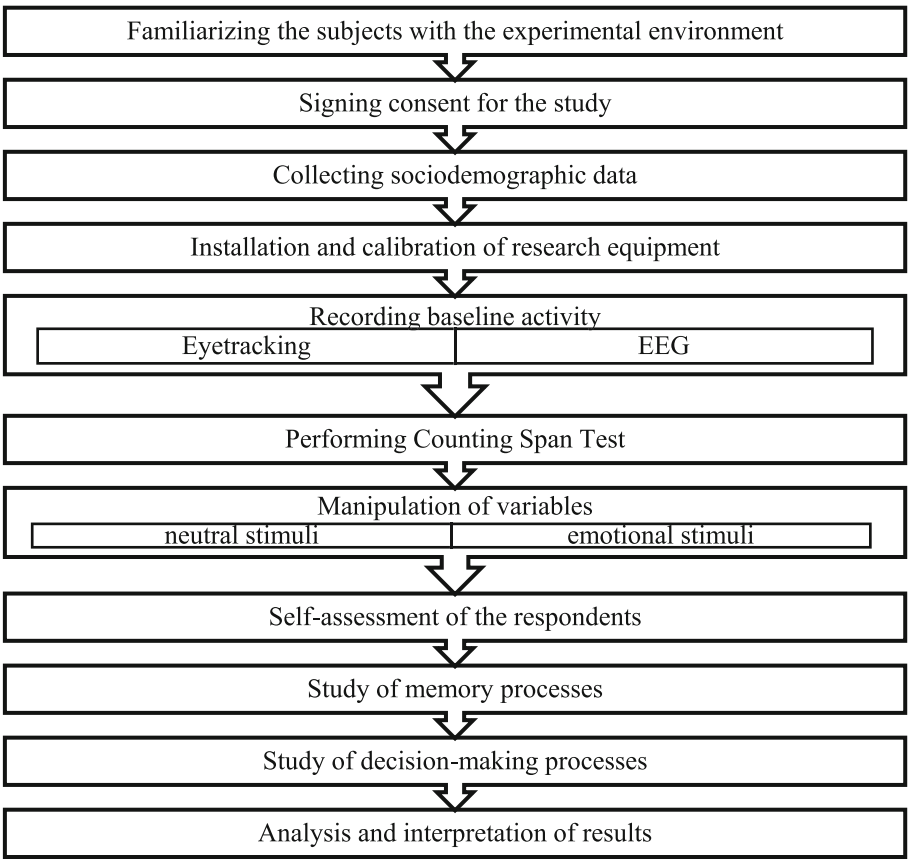


Fig. 1. Research framework

green grains and ignore the yellow grains. Once he has counted, he selects the number on the keyboard corresponding to the green grains counted. At this point, another board with seeds appears on the screen. The subject must remember the number of grains from the previous board and at the same time count the green grains on the next one. The more boards, the more grains to remember and the more loaded the working memory of the subjects. At this stage of the study, we intend to manipulate variables such as exposed neutral stimuli and emotional stimuli. These will be symbols and pictures integrated between green and yellow grains. People will then assess their cognitive load using a questionnaire.

In the last stage of the experiment, we intend to examine the memory and decision-making processes of the subjects. The question will appear on the monitor screen: "Did you see any objects on the monitor screen while performing the task?" If the respondents answer affirmatively, another question will appear: "What were these objects?" Here the respondent will be able to enter the answer. We will also present pairs of symbols and ask you to choose one preferred symbol from each pair. In each of the presented pairs, only one of the symbols will be previously exposed on the Counting Span Test boards.

Both symbols in each pair, will belong to the same category of symbols: emotional or neutral. This will allow us to assess the relationship between the exposed symbols and the respondents' decision-making processes. At the theoretical level, the project refers to the phenomenon of priming. It is an increase in the probability of using a specific cognitive category in perceptual and thought processes because of previous exposure to a stimulus related to this category. We will also assess whether stimuli that were not remembered by the subjects but were exposed to them were related to subsequent decisions when choosing preferred symbols from the exposed pairs. Therefore, in the research project we take unconscious processes into account. We refer here to research on the implicit memory. It is memory that contains content that the individual is unaware of, and is usually acquired through involuntary learning [51]. Numerous scientific studies show that most events reach consciousness only after approximately 300 ms after the stimulus occurs. The processes that take place in the subconscious, even though they are not consciously controlled by a person, have a significant impact on his functioning and the decisions he makes. To better understand some psychological processes, it is necessary to analyze the reaction of the human brain and body to external stimuli. Studying human physiological parameters (including neurophysiological ones) that influence the response to presented stimuli will allow us to expand the scope of information obtained on the reception of a given message from online stores. This is particularly important in relation to declarative research, as it allows obtaining information that is not available in this type of research. As scientific research shows, the gender distinction is also key in these studies, because men and women differ in brain activation during cognitive tasks. It is related to mental rotation, visual stimulation, emotion recognition, working memory, and verbal processing [52, 53]. Therefore, in this study we will also focus on gender-wise analysis. Differences may not appear in every situation, but if they do, they may constitute key information for online store designers.

4 Experimental Case Study

The study will be conducted at the Cognitive Neuroscience Laboratory with approval from the Bioethics Committee, ensuring participant data protection.

4.1 Stimuli

In the study, we plan to utilize pre-made Counting Span Test boards obtained from the Milisecond Test Library [54], a platform providing software for psychological tests, which has been repeatedly used in peer-reviewed scientific research [55–57]. The experimental materials will include a set of Counting Span Test boards (Fig. 2). Before commencing the actual study, participants will perform a practice task consisting of 4 boards. They will then proceed to the main task, which will consist of five trials. In the first trial, 3 boards will be presented. In each subsequent trial, we plan to present three additional boards. As a result, in the last, fifth trial, 15 boards will appear on the monitor screen. In total, across all five trials, 45 boards will be presented. We estimate the execution time of the main task to be approximately 8 min.

The symbols embedded in the boards will appear from the 8th board in the fourth trial. This moment will be crucial as participants will not expect such exposure, and the

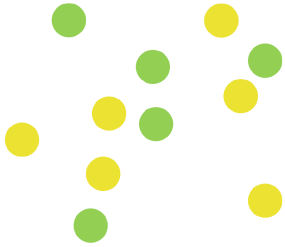


Fig. 2. One of the Counting Span Test boards

task will require significantly greater attentional resources from them than previously. In total, we plan to utilize 20 symbols, one on each board, until the last one. Half of the symbols will be emotionally charged, while the other half will be neutral. Table 1 presents the symbols we intend to use in the study. These symbols were chosen through expert judgment, which is a popular method in the literature [58–60].

Table 1. Examples of neutral and emotional symbols

	Examples of symbols
Neutral symbols	⌂ ∂ $ Ý = ¾
Emotional symbols	

4.2 Participants

We plan to invite students who use the services of online stores to participate in the study. Regarding the number of participants to be tested during the experiment, we will follow the literature recommendations, which suggest using a sample size of N = 30 [61] for each of the examined groups, resulting in a total of 60 participants in the entire study. Additionally, for comparison purposes, the second part of the experiment will be performed by a group of rested people from the control group. This will allow us to assess whether cognitive fatigue and symbol exposure during the load significantly differentiate the decisions of the subjects compared to the decisions of the control group. The study will control variables such as: age of the respondents, gender, intellectual abilities, education, and psychoactive substances taken. These variables are related to cognitive processes, including working memory [52, 62]. The compared groups should not differ statistically significantly in these respects.

4.3 Measures

Based on the literature review, we plan to use traditional and psychophysiological methods in the study. To subjectively assess cognitive load, we intend to use the NASA-TLX questionnaire [30], and from neuroscientific methods we plan to use eye tracking and EEG. Analysis of data collected by the eye tracker will include measurements such as: saccades (frequency, amplitude, duration), fixations (duration, number, frequency), pupils (pupil diameter), and blinks (frequency, duration, number). Regarding EEG analysis, we plan to examine the power spectral density (PSD) in different frequency ranges such as Theta (4–7 Hz), Alpha (8–12 Hz), Beta (13–25 Hz), and Gamma (26–40 Hz). Most researchers agree that the selected measures correlate well with various aspects of workload and promise to objectively measure mental overload [37, 63].

5 Conclusion

Making decisions in online stores is a highly relevant topic today. However, the vast array of stimuli encountered by Internet users complicates the identification of decisive variables influencing their behavior. In our article, we proposed a plan and framework for studying cognitive load among consumers.

The study aims to elucidate the neurocognitive basis of customer behavior in online stores. A reductionist understanding of these mechanisms is crucial for applying this knowledge across information technology. Despite focusing on cognitive processes, our study adopts a holistic approach, encompassing measurements of psychological, oculographic, and neurophysiological variables. By integrating consumer psychology, neurocognitive science, and internet marketing, our research offers an innovative interdisciplinary perspective. This comprehensive analysis seeks to deepen our understanding of the factors shaping consumer choices and how cognitive load impacts them. Additionally, the study will explore whether peripheral exposure to product-related stimuli on websites influences customers' purchasing decisions.

The research proposes the plan and framework for performing research in the problem of cognitive load on consumers. To address practical needs effectively, we need more detailed empirical studies to provide specific design recommendations for different types of customers. This includes those who prefer quick purchases and others seeking unique finds. While the current study's framework is foundational, further empirical support is necessary for direct practical application. Therefore, expanding research is crucial to develop more concrete and practical guidelines on how to design websites that enhance the shopping experience for diverse customer preferences. Additionally future studies should prioritize empirical validation to affirm the practical applicability of our framework. Bridging the theory-practice gap will enhance the reliability of our findings and provide actionable insights for internet marketers and store owners. This validation process aims not only to strengthen theoretical foundations but also to advance interdisciplinary research on cognitive load and decision-making in e-commerce. Ultimately, the study's findings are poised to offer practical guidance for online store designers and owners, facilitating the reduction of cognitive load and enhancing the overall shopping

experience for customers. This contribution is pivotal for advancing marketing strategies in e-commerce and addressing critical gaps in research on cognitive load's role in decision-making processes.

References

1. Zhu, Z.-Y., Xie, H.-M., Chen, L.: ICT industry innovation: knowledge structure and research agenda. Technol. Forecast. Soc. Change **189**, 122361 (2023)
2. Quiñones, D., Rusu, C., Rusu, V.: A methodology to develop usability/user experience heuristics. Comput. Stand. Interfaces **59**, 109–129 (2018)
3. Borawska, A., Mateja, A.: Full paper: incorporating cognitive neuroscience techniques to enhance user experience research practices. In: International Conference on Information Systems Development ISD (2023)
4. Hinderks, A., Domínguez Mayo, F.J., Thomaschewski, J., Escalona, M.J.: Approaches to manage the user experience process in Agile software development: a systematic literature review. Inf. Softw. Technol. **150**, 106957 (2022)
5. Pengnate, S. (Fone), Sarathy, R.: An experimental investigation of the influence of website emotional design features on trust in unfamiliar online vendors. Comput. Hum. Behav. **67**, 49–60 (2017)
6. Srivastava, A., Thaichon, P.: What motivates consumers to be in line with online shopping?: a systematic literature review and discussion of future research perspectives. Asia Pac. J. Mark. Logist. **35**, 687–725 (2023)
7. Agarwal, R., Venkatesh, V.: Assessing a firm's web presence: a heuristic evaluation procedure for the measurement of usability. Inf. Syst. Res. **13**, 168–186 (2002)
8. Geissler, G.L., Zinkhan, G.M., Watson, R.T.: The influence of home page complexity on consumer attention, attitudes, and purchase intent. J. Advert. **35**, 69–80 (2006)
9. Tuch, A.N., Bargas-Avila, J.A., Opwis, K., Wilhelm, F.H.: Visual complexity of websites: effects on users' experience, physiology, performance, and memory. Int. J. Hum. Comput. Stud. **67**, 703–715 (2009)
10. Sarkar, A.R., Ahmad, S.: A new approach to expert reviewer detection and product rating derivation from online experiential product reviews. Heliyon **7**, e07409 (2021)
11. Gürkut, C., Elçi, A., Nat, M.: An enriched decision-making satisfaction model for student information management systems. Int. J. Inf. Manag. Data Insights **3**, 100195 (2023)
12. Borawska, A., Mateja, A.: The use of cognitive neuroscience tools for evaluating the cognitive overload caused by social advertising. In: AMCIS 2023 Proceedings (2023)
13. Wang, Q., Yang, S., Liu, M., Cao, Z., Ma, Q.: An eye-tracking study of website complexity from cognitive load perspective. Decis. Support Syst. **62**, 1–10 (2014)
14. Wlekły, P., Piwowarski, M.: The usability of eye tracking in the design of digital training materials. Procedia Comput. Sci. **207**, 4180–4189 (2022)
15. Wickens, C.D.: Multiple resources and mental workload. Hum. Factors J. Hum. Factors Ergon. Soc. **50**, 449–455 (2008)
16. Young, M.S., Brookhuis, K.A., Wickens, C.D., Hancock, P.A.: State of science: mental workload in ergonomics. Ergonomics **58**, 1–17 (2015)
17. Kantowitz, B.H.: Attention and mental workload. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 44, pp. 3-456–3-459 (2000)
18. Kahneman, D.: Attention and Effort. Prentice-Hall, Englewood Cliffs (1973)
19. Sweller, J.: Cognitive load during problem solving: effects on learning. Cogn. Sci. **12**, 257–285 (1988)

20. Antonenko, P., Paas, F., Grabner, R., Van Gog, T.: Using electroencephalography to measure cognitive load. Educ. Psychol. Rev. **22**, 425–438 (2010)
21. Longo, L., Wickens, C.D., Hancock, G., Hancock, P.A.: Human mental workload: a survey and a novel inclusive definition. Front. Psychol. **13**, 883321 (2022)
22. Sweller, J.: Element interactivity and intrinsic, extraneous, and germane cognitive load. Educ. Psychol. Rev. **22**, 123–138 (2010)
23. Van Merriënboer, J.J.G., Sweller, J.: Cognitive load theory in health professional education: design principles and strategies: cognitive load theory. Med. Educ. **44**, 85–93 (2010)
24. Gao, Q., Wang, Y., Song, F., Li, Z., Dong, X.: Mental workload measurement for emergency operating procedures in digital nuclear power plants. Ergonomics **56**, 1070–1085 (2013)
25. Meshkati, N.: Toward development of a cohesive model of workload. In: Advances in Psychology, pp. 305–314. Elsevier (1988)
26. Orru, G., Longo, L.: The evolution of cognitive load theory and the measurement of its intrinsic, extraneous and germane loads: a review. In: Longo, L., Leva, M. (eds.) H-WORKLOAD 2018. CCIS, vol. 1012, pp. 23–48. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-14273-5_3
27. Eggemeier, F.T., Wilson, G.F.: Performance-based and subjective assessment of workload in multi-task environments. In: Damos, D.L. (ed.) Multiple-Task Performance, pp. 217–278. CRC Press (2020)
28. Louis, L.-E.L., et al.: Cognitive tasks and combined statistical methods to evaluate, model, and predict mental workload. Front. Psychol. **14**, 1122793 (2023)
29. Krieglstein, F., Beege, M., Rey, G.D., Sanchez-Stockhammer, C., Schneider, S.: Development and validation of a theory-based questionnaire to measure different types of cognitive load. Educ. Psychol. Rev. **35**, 9 (2023)
30. Buchner, J., Buntins, K., Kerres, M.: A systematic map of research characteristics in studies on augmented reality and cognitive load. Comput. Educ. Open **2**, 100036 (2021)
31. De Winter, J.C.F.: Controversy in human factors constructs and the explosive use of the NASA-TLX: a measurement perspective. Cogn. Technol. Work **16**, 289–297 (2014)
32. Matthews, G., Reinerman-Jones, L.: Workload Assessment: How to Diagnose Workload Issues and Enhance Performance. Human Factors and Ergonomics Society, Santa Monica, CA (2017)
33. Evans, D.C., Fendley, M.: A multi-measure approach for connecting cognitive workload and automation. Int. J. Hum. Comput. Stud. **97**, 182–189 (2017)
34. Foy, H.J., Chapman, P.: Mental workload is reflected in driver behaviour, physiology, eye movements and prefrontal cortex activation. Appl. Ergon. **73**, 90–99 (2018)
35. Peterson, D.A., Kozhokar, D.: Peak-end effects for subjective mental workload ratings. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 61, pp. 2052–2056 (2017)
36. Longo, L.: A defeasible reasoning framework for human mental workload representation and assessment. Behav. Inf. Technol. **34**, 758–786 (2015)
37. Charles, R.L., Nixon, J.: Measuring mental workload using physiological measures: a systematic review. Appl. Ergon. **74**, 221–232 (2019)
38. Hart, S.G.: Nasa-task load index (NASA-TLX); 20 years later. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 50, pp. 904–908 (2006)
39. Ausin-Azofra, J.M., Bigne, E., Ruiz, C., Marín-Morales, J., Guixeres, J., Alcañiz, M.: Do you see what I see? Effectiveness of 360-degree vs. 2D video ads using a neuroscience approach. Front. Psychol. **12**, 612717 (2021)
40. Curtin, A., Ayaz, H.: Neural efficiency metrics in neuroergonomics. In: Neuroergonomics, pp. 133–140. Elsevier (2019)
41. Guixeres, J., et al.: Consumer neuroscience-based metrics predict recall, liking and viewing rates in online advertising. Front. Psychol. **8**, 1808 (2017)

42. Bell, L., Vogt, J., Willemse, C., Routledge, T., Butler, L.T., Sakaki, M.: Beyond self-report: a review of physiological and neuroscientific methods to investigate consumer behavior. Front. Psychol. **9**, 1655 (2018)
43. Borawska, A., Mateja, A.: Metody neuronauki poznawczej w badaniu doświadczenia użytkowników. In: Dudycz, H., Hernes, M., Pondel, M., Rot, A. (eds.) Informatyka w zarządzaniu, pp. 13–33. Wroclaw University of Economics and Business (2023)
44. Gill, R., Singh, J.: A study of neuromarketing techniques for proposing cost effective information driven framework for decision making. Mater. Today Proc. **49**, 2969–2981 (2022)
45. Jin, J., Lin, C., Wang, F., Xu, T., Zhang, W.: A study of cognitive effort involved in the framing effect of summary descriptions of online product reviews for search vs. experience products. Electron. Commer. Res. **23**, 785–806 (2023)
46. Longo, L.: Experienced mental workload, perception of usability, their interaction and impact on task performance. PLoS ONE **13**, e0199661 (2018)
47. Hilke, P., Tor, D.W.: How expectancies shape consumption experiences. In: Preston, S.D., Kringelbach, M.L., Knutson, B. (eds.) The Interdisciplinary Science of Consumption, pp. 219–240. The MIT Press (2014)
48. Scheibehenne, B., Greifeneder, R., Todd, P.M.: Can there ever be too many options? A meta-analytic review of choice overload. J. Consum. Res. **37**, 409–425 (2010)
49. Durgin, F.H., Doyle, E., Egan, L.: Upper-left gaze bias reveals competing search strategies in a reverse Stroop task. Acta Psychol. **127**, 428–448 (2008)
50. Skurnik, I., Yoon, C., Park, D.C., Schwarz, N.: How warnings about false claims become recommendations. J. Consum. Res. **31**, 713–724 (2005)
51. Graf, P., Schacter, D.L.: Implicit and explicit memory for new associations in normal and amnesic subjects. J. Exp. Psychol. Learn. Mem. Cogn. **11**, 501–518 (1985)
52. Halari, R., Sharma, T., Hines, M., Andrew, C., Simmons, A., Kumari, V.: Comparable fMRI activity with differential behavioural performance on mental rotation and overt verbal fluency tasks in healthy men and women. Exp. Brain Res. **169**, 1–14 (2006)
53. Hill, A.C., Laird, A.R., Robinson, J.L.: Gender differences in working memory networks: a BrainMap meta-analysis. Biol. Psychol. **102**, 18–29 (2014)
54. Counting Span Task – Millisecond. https://www.millisecond.com/download/library/countingspan. Accessed 08 May 2024
55. Conway, A.R.A., Kane, M.J., Bunting, M.F., Hambrick, D.Z., Wilhelm, O., Engle, R.W.: Working memory span tasks: a methodological review and user's guide. Psychon. Bull. Rev. **12**, 769–786 (2005)
56. Weingarten, N.: The effect of information provision on attitude, intention, and food choice behavior (2023). https://bonndoc.ulb.uni-bonn.de/xmlui/handle/20.500.11811/10828
57. García, C.S., Morales-Sánchez, V., Garrido, R.E.R., Hernández-Mendo, A.: Relationships between type of sport played and hot and cold executive functions in children and adolescents: a systematic review. Cuad. Psicol. Deporte **24**, 1–19 (2024)
58. Czarnecki, M.: Metodyka pracy z sędziami kompetentnymi w procesie opracowywania skali do pomiaru konstruktów: rekomendacje i egzemplifikacja. Przegląd Organ, 13–18 (2020). https://doi.org/10.33141/po.2020.12.02
59. Field, A.: Kendall's Coefficient of Concordance. Presented at the October 15 (2005)
60. Brzeziński, J.M.: Metodologia badań psychologicznych, Warszawa (2007)
61. Ramsøy, T.Z.: Building a foundation for neuromarketing and consumer neuroscience research: how researchers can apply academic rigor to the neuroscientific study of advertising effects. J. Advert. Res. **59**, 281–294 (2019)
62. Kyllonen, P.C., Christal, R.E.: Reasoning ability is (little more than) working-memory capacity? Intelligence **14**, 389–433 (1990)

63. Hebbar, P.A., Bhattacharya, K., Prabhakar, G., Pashilkar, A.A., Biswas, P.: Correlation between physiological and performance-based metrics to estimate pilots' cognitive workload. Front. Psychol. **12** (2021)

Advanced Machine Learning and Data Processing Methods for Support Business Processes

Machine Learning-Based Exploration of Eye-Tracking Data to Predict Offer Selection

Mateusz Piwowarski[1(✉)], Paweł Ziemba[2], and Jacek Cypryjański[3]

[1] Institute of Management, University of Szczecin, Cukrowa 8, 71-004 Szczecin, Poland
mateusz.piwowarski@usz.edu.pl
[2] Faculty of Computer Science and Telecommunications, Maritime University of Szczecin, Waly Chrobrego 1-2, 70-500 Szczecin, Poland
p.ziemba@pm.szczecin.pl
[3] Institute of Economics and Finance, University of Szczecin, Mickiewicza 64, 71-101 Szczecin, Poland
jacek.cypryjanski@usz.edu.pl

Abstract. Visualisations of sales offers have a significant impact on their reception. Given the difficulty in assessing the impact of visualizations on consumer decisions, this study attempted to determine whether a consumer's decision can be recognized by how they look at offers. Eye-tracking data obtained from viewing house sales offers was used to develop a classifier model for predicting decisions. The research results show that the use of appropriate classification algorithms (e.g., C4.5, DT, kNN, ANN), methods to reduce the impact of data imbalance on results (e.g., SMOTE), or attribute filtering methods (e.g., CFS) provide effective methodological approaches to develop such a classifier.

Keywords: Machine learning · Eye-tracking Data · Decision Support · Predict Offer Selection

1 Introduction

A review of research on consumer behaviour in recent decades shows a clear shift from rational to psychological and social decision factors and attempts to take into account the role of goals, emotions as well as the nonconscious nature of much of human behaviour [1, 2]. The complexity of the impact of such a broad spectrum of factors can be clearly seen when making a decision to purchase such strategically important goods as houses. In addition to basic and easily comparable parameters such as price, square footage, number and layout of rooms, etc., house sales offers contain visualisations which cannot be easily measured in terms of their impact on the consumer. There are great opportunities to control the consumer's feelings, arouse emotions and influence their attitude to the presented object, both through the way the object itself is presented (using unusual shots, manipulating the lighting, exposing selected details), as well as by placing additional elements such as family, unique plants, cars, etc. in the image [3]. In view of the difficulty of assessing the impact of visualisation on consumer decisions, this research attempted to find out whether a consumer's decision can be recognised by the way the consumer looks at the offers.

M. Hernes et al. (Eds.): ECAI 2024, LNNS 1217, pp. 105–117, 2025.
https://doi.org/10.1007/978-3-031-78465-1_9

Eye-tracking makes it possible to monitor the position of the eyes during stimuli in the course of specific activities. It offers a non-invasive, yet objective determination of gaze direction, the location of the gaze at a given time. Essentially, eye movements and gaze directions are linked to various post-cognitive and perceptual processes. This provides a number of possibilities for interpreting human behaviour, ranging from interest in stimuli to, for example, processes for creating coherent memory representations [4–7]. Visual fixation points (fixations), rapid jumps between fixations (saccades) and the visual pathways they create provide opportunities to extract visual information about, for example, different motor activities, information processing related to reading, visual search and even about one's own preferences [8–11]. As research shows, analyses of consumers' conscious choices have their limitations in capturing important factors influencing actual decision-making. Analysis of visual data, visual attention, metrics based on gaze points, fixations and saccades allows access to at least some of the information processed unconsciously [10, 12].

Research shows that eye-tracking data can be effectively used to predict which offers will attract attention and induce purchase. For example, Castilla et al. [13] showed that eye-tracking data can be used to create predictive models that predict which product offers users will choose based on their gaze patterns. In contrast, Pina et al. [14] used eye movement data and machine learning to predict whether a recruiter will classify a CV for further recruitment. Lim et al. [15], on the other hand, analyse what machine learning features can be extracted from eye-tracking data for classification tasks. However, there is a lack of research in which methods for predicting purchase decisions are proposed based on the analysis of visualizations of viewed house sales offers. Therefore, the authors conducted eye-tracking studies on sales offers from actual catalogs, identified interesting segments of these offers, and calculated metrics based on fixations. The prepared data served as input for machine learning algorithms.

This article aims to present the research results on developing a classification model to predict decisions on selecting sales offers, based on specific eye-tracking metrics calculated for designated areas of these offers. This paper is structured as follows: Sect. 2 presents the description of the research experiment using eye tracking, the research methodology, and the data collection and analysis methods. The data mining procedure used in the study was also characterized. Section 3 presents the results of research using different classification algorithms, and Sect. 4 presents conclusions.

2 Methodology

2.1 Eye-Tracking Research

Protocol and Stimuli. The research experiment consisted of presenting respondents with 10 different single-family house sales offers. The house designs were obtained from a catalogue provided at: https://projekty.muratordom.pl. Each offer was shown for 20 s and the order in which they were displayed was randomised. After the next offers were presented to the respondents, all the visualisations of the houses were shown once again. Respondents were asked to choose which offer they preferred. In this case, the selection time was not limited. In addition, information on the importance of the evaluation criteria was also obtained. It was assumed that the respondents had sufficient financial resources

to potentially purchase the property. Each offer had the same structure and contained the same information (visual and textual) respectively. The structure of the offer consisted of 5 main elements: an external appearance of the single-family house, a graphic with the layout of the rooms, information on the price of the house, the number of rooms and information on the location of the property (e.g. in the city, near a shopping centre; outside the city limits, in a small town; on the outskirts of the city, in a closed estate of houses). A schematic representation of the structure of the offers is presented in Fig. 1.

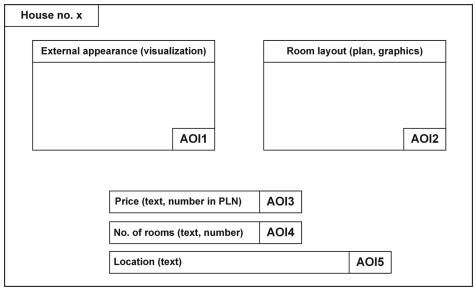


Fig. 1. Structure of offer (stimuli) with marked AOIs.

The selected visualisations of the houses varied enough to take into account the different preferences of the interested parties. As did the layout of the rooms, where the offers differed in terms of the layout and number of rooms, presence of terraces, garage, etc. The house prices proposed in the offers ranged from PLN 381 000 to PLN 540 000, where the average price and standard deviation were: M = 479 600 PLN; SD = 49 967,10 PLN. The number of rooms, depending on the offer variant, was 3 or 4. For each of the elements of the offer structure, areas of interest were created (AOI1-AOI5), which were further subjected to detailed eye-tracking analyses.

Research Methods and Measuring Apparatus. In the research experiment, eye-tracking was used as a research technique. Metrics based on fixations and saccades were used as data analysis methods for the areas of interest created in the single-family house sales offers under consideration.

Eye tracking, as a research technique, makes it possible to accurately record, visualise and analyse the way of looking at different types of visual stimuli. Gaze points and fixation and saccadic eye movements are recorded. Most often heat maps and scanpaths are used to visualise the data. The analysis process is about data processing and covers several main areas: data preprocessing (including noise filtering), detection of fixations and saccades, interpretation of the visual representation of data (heat maps), analysis of gaze sequences (gaze plots), defining areas of interest (AOI), statistical data analysis (quantitative and qualitative). In the context of quantitative data analysis, AOIs are of particular interest. These are fragments of the stimulus of interest selected by the researcher (e.g. graphic elements, text fragments, buttons) to be analysed in detail. Various metrics are calculated for the defined AOIs, which are based on fixations, saccades or gaze points. These allow interpretation and inference about cognitive and behavioural processes [7, 16].

The measurement device used was a stationary eye tracker (based on a monitor screen) Tobii Pro X3–120. The sampling frequency was 120 Hz. The I-HMM (Hidden Markov Model) fixation filtering algorithm was used for fixation detection. The minimum fixation classification time was set at 60 ms, with a maximum inter-fixation interval of 75 ms. Noise reduction was performed using the Moving Average algorithm, and interpolation was performed with gap filling of a maximum of 75 ms.

Analyzes and Indicators of Evaluation. In order to search for and understand the visual processing of specific elements of sales offers (AOI1-AOI5), AOI metrics were determined. They are based on fixations and saccades (Table 1). They are a good analytical tool for understanding the visual interaction of people viewing offers with different types of content (text, graphics), with different meanings.

Table 1. AOI metrics (fixation-based and saccade-based) used in the study.

AOI metrics	Metric description
Fixation count (FC)	Average number of fixations detected inside the AOI [17]
Duration of average fixation (ms) (DAF)	Average duration of all fixations detected inside the AOI [18]
Time to first fixation (ms) (TTFF)	Average time elapsed before the first fixation was detected inside the AOI [18]
First fixation duration (ms) (FFD)	Average duration of respondents' first fixation inside the AOI [19]
Dispersion of average fixation (DSAF)	Average dispersion of all the respondent's fixations detected inside the AOI [18]
Dwell time (%) (DT)	The average amount of time respondents fixated on the AOI relative to the time the AOI was active [18]
Revisits (REV)	Number of AOI return visits. The average of how often respondents looked back at an AOI they had already noticed [18, 20]
Saccade count (SC)	Count of saccades with start and end points detected inside AOI [18]

In order to obtain more complete and reliable interpretations of cognitive and perceptual processes, multiple metrics are analysed simultaneously.

Participants. There were 45 participants in the experiment: 23 women and 22 men. They were of different ages (ranging from 20 to 58 years of age), of which 23 were students and 22 were employed. The average age of the participants and the standard deviation were respectively: $M = 31.44$; $SD = 10.81$. All participants were informed about the course of the experiment, the research apparatus used, the processing of data for scientific work and the possibility of withdrawing from the study at any time. Each person gave free and informed consent to participate in the study, which was confirmed

by written consent. Details of the subject matter, the content of the study (the stimuli presented) were not revealed so as not to influence the way they were perceived. The research experiment was conducted in a specialised laboratory located at the university. It followed the guidelines set out in the Declaration of Helsinki (2013 ver.) [21]. It was reviewed by the Bioethics Committees at Regional Chamber of Physicians in Szczecin.

2.2 Data Mining and Machine Learning

The data on the selection of house offers was subjected to a process of mining based on machine learning algorithms. Given that the data was extracted from 45 respondents and each respondent was presented with 10 offers, a total of 450 data instances were obtained. In addition, it should be noted that each respondent selected only one offer. Therefore, among the 450 instances, only 45 had a positive value for the decision attribute "offer selection" and the rest had a negative value for this attribute. In view of this, the data were strongly imbalanced, i.e. unevenly distributed across decision classes [22]. All AOI metrics acted as conditional attributes. Given that 8 metrics were generated for each AOI, a total of 40 conditional attributes were considered. All conditional attributes were quantitative attributes, while the decision attribute was nominal.

The objective of the data mining process was to develop a classifier model to predict the bidding decision based on the individual AOI metrics. The objective was achieved by using different approaches to the classification process and applying several different classifiers on a pre-prepared dataset. The classifiers used were C4.5 (C4.5 Decision Tree) [23], DT (Decision Table) [24], ANN (Artificial Neural Networks) [25], kNN (k-Nearest Neighbors) [26], and SVM (Support Vector Machines) [27].

It should be noted that data imbalancing usually results in a decrease in the predictive ability of the classifier, worsening the classification performance [28]. Therefore, solutions taken from the literature have been applied in the classification process to reduce the negative impact of data imbalance on classification performance. In particular, three approaches to the classification process were used: (1) modified class weights, (2) modified probability of the class occurrence, (3) SMOTE (Synthetic Minority Oversampling Technique). Modified class weights allow to give higher weight to minority class instances [29]. As the ratio of majority to minority class in the dataset was 9:1, it was this modified weighting that was applied to the minority class instances. The modified probability of the class occurrence allows to reduce the probability threshold determining the membership of an instance to a minority class [30]. In the dataset, minority class instances represent 1/10 of all instances, so the probability threshold for them was set to 0.1. SMOTE is a method for generating synthetic instances from real instances [31]. The dataset was expanded to include 360 synthetic minority class instances, resulting in equal numbers of positive and negative class instances.

Each approach has been used in a stand-alone variant and also in combination with the Correlation-based Feature Selection (CFS) attribute filtering method [32]. CFS reduces the number of conditional attributes describing each data instance. Based on the internal correlations between the conditional attributes and the correlation of the conditional attributes with the decision attribute, CFS selects a subset of attributes that provides good prediction performance. For the CFS filter, the Best First approach was used as the algorithm for searching subsets of attributes. In addition, the study found that the CFS

filter applied to the primary dataset performed incorrectly due to strong data imbalance (CFS selected only the first weighted attribute, AOI1_FC). Therefore, in the study, a CFS filter was applied to the imbalanced data using modified class weights during attribute selection. In this way, CFS selected attributes that were important in terms of both majority and minority class description. In contrast, for SMOTE-balanced data, the CFS filter was applied without any modification.

For comparative reasons, the results obtained were compared with the results of a standard classification of the imbalanced dataset. After building the individual classifier models, each model was tested and evaluated using different evaluation measures. Classifier training and testing was performed using 10-fold cross-validation [33]. The following were used as measures of classifier evaluation: accuracy, MCC (Matthews correlation coefficient), AUROC (area under receiver operating characteristic curve), precision, specificity [34]. On this basis, the classifiers giving the best predictive performance for the selection of house listings were indicated.

The data mining procedure for the selection of house listings is shown in Fig. 2. By analysing Fig. 2, it can be seen that seven different paths were used to process the source dataset. For each path, five classifiers were tested with five measures of classifier evaluation.

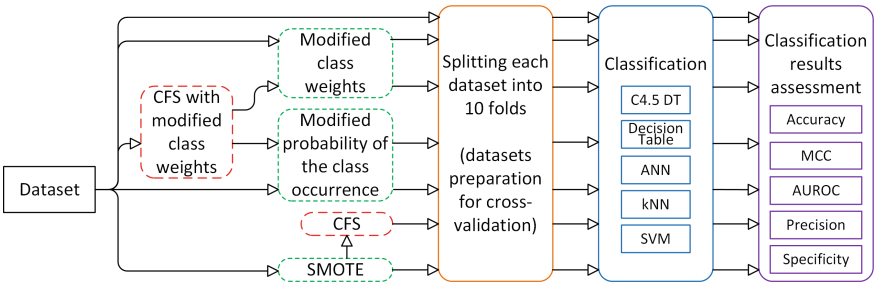


Fig. 2. The data mining procedure used in the study.

3 Results

The basic path of the data mining procedure involved the simple application of classifiers on a non-preprocessed and imbalanced dataset. The results of the classification carried out in this way are shown in Table 2. These results indicate that for the raw dataset, the best classification results are given by the kNN classifier. It is followed by the results of the following classifiers: SVM, C4.5 and ANN. In contrast, the worst results are given by DT, which classifies almost all cases as negative (precision is 0). The classifiers are ranked in this order according to the ROC and MCC measures. A ranking based on precision also gives a similar order, but in this case the SVM classifier performs marginally better than kNN. It is also worth noting that the measures of accuracy, precision, and specificity do not reflect the actual quality of the classifiers, which is a known problem with imbalanced datasets.

Table 2. Classification results: standard classifiers.

Classifier	Accuracy [%]	MCC	ROC	Precision (TPR)	Specificity (TNR)
C4.5	86.44	0.157	0.571	0.200 (9/45)	0.938 (380/405)
DT	89.56	−0.022	0.438	0.000 (0/45)	0.995 (403/405)
kNN	87.33	0.289	0.643	0.356 (16/45)	0.931 (377/405)
ANN	86.22	0.114	0.530	0.156 (7/45)	0.941 (381/405)
SVM	79.56	0.172	0.610	0.378 (17/45)	0.842 (341/405)

A further two variants for processing the dataset involved the use of modified class weights. Table 3 shows the classification results with the modified class weights. Table 4 contains the results of the classification with the modified class weights, but in this case the classification was performed with a set of conditional attributes (features) reduced by a CFS filter using the modified class weights.

Table 3. Classification results: modified class weights.

Classifier	Accuracy [%]	MCC	ROC	Precision (TPR)	Specificity (TNR)
C4.5	84	0.145	0.575	0.244 (11/45)	0.906 (367/405)
DT	59.33	−0.003	0.523	0.378 (17/45)	0.617 (250/405)
kNN	87.33	0.289	0.647	0.356 (16/45)	0.931 (377/405)
ANN	85.56	0.236	0.578	0.333 (15/45)	0.914 (370/405)
SVM	72.89	0.114	0.583	0.400 (18/45)	0.765 (310/405)

Table 4. Classification results: modified class weights and CFS with modified class weights.

Classifier	Accuracy [%]	MCC	ROC	Precision (TPR)	Specificity (TNR)
C4.5	80.44	0.063	0.539	0.200 (9/45)	0.872 (353/405)
DT	66	0.032	0.597	0.356 (16/45)	0.694 (281/405)
kNN	88.44	0.319	0.653	0.356 (16/45)	0.943(382/405)
ANN	70.67	0.095	0.593	0.400 (18/45)	0.741 (300/405)
SVM	46.67	−0.028	0.477	0.489 (22/45)	0.464 (188/405)

The analysis of Table 3 shows that the use of modified class weights improved the MCC, ROC, and precision of the DT and ANN classifiers. For the C4.5 and SVM classifiers, precision improved slightly, but MCC and specificity worsened. In contrast, for the kNN classifier, the use of increased weights for the minority class has no effect on the results obtained. On the other hand, the results in Table 4 show that for the

set of conditional attributes reduced by the CFS filter, with respect to Tables 2 and 3, the precision of the ANN and SVM classifiers improved, but the values of the MCC and specificity indices worsened. The DT classifier achieved results similar to those in Table 3, and the C4.5 classifier performed worse than in the earlier data processing paths. For the kNN classifier, as before, modifications to the dataset and the classification approach did not significantly affect the classification results obtained, but it can be observed that the reduction of conditional attributes using CFS improved the classifier's specificity somewhat.

The next variants of the data mining procedure consisted of applying a modified probability of the class occurrence. Table 5 contains the classification results obtained using the modified probability of the minority class occurrence. Table 6 shows the classification results using the same technique, but with a prior reduction of the set of conditional attributes using CFS with modified class weights.

Table 5. Classification results: classifier with probability threshold.

Classifier	Accuracy [%]	MCC	ROC	Precision (TPR)	Specificity (TNR)
C4.5	85.78	0.178	0.571	0.244 (11/45)	0.926 (375/405)
DT	37.11	−0.034	0.438	0.600 (27/45)	0.346 (140/405)
kNN	87.33	0.289	0.643	0.356 (16/45)	0.931 (377/405)
ANN	80.67	0.082	0.530	0.222 (10/45)	0.872 (353/405)
SVM	79.56	0.172	0.610	0.378 (17/45)	0.842 (341/405)

Table 6. Classification results: classifier with probability threshold and CFS with modified class weights.

Classifier	Accuracy [%]	MCC	ROC	Precision (TPR)	Specificity (TNR)
C4.5	47.33	−0.047	0.452	0.444 (20/45)	0.477 (193/405)
DT	10	–	0.475	1.000 (45/45)	0.000 (0/405)
kNN	88.44	0.319	0.649	0.356 (16/45)	0.943 (382/405)
ANN	77.11	0.073	0.546	0.267 (12/45)	0.827 (335/405)
SVM	49.78	0.016	0.514	0.533 (24/45)	0.494 (200/405)

The results in Table 5 show that, compared to the standard classification results in Table 2, the precision of classifiers C4.5, DT, and ANN improved. However, classifier specificity decreased significantly for DT. Also for the C4.5 and ANN classifiers, the improvement in precision was associated with some decrease in specificity, but this was not as drastic as for DT. In contrast, the modification of the probability of the class occurrence did not affect the classification results using the kNN and SVM classifiers. Comparing the results in Table 5 with those of the modified class weights (Table 3), it can be seen that the performance of the C4.5 and kNN classifiers remained essentially

unchanged, while the performance of ANN worsened. The results of the DT and SVM classifiers, on the other hand, are inconclusive (some indicators improved and others deteriorated). Table 6 shows that the combination of the CFS and modified probability of the class occurrence filter, compared to the standard classification, improves the precision of all classifiers except kNN (in its case the precision remained unchanged). However, the improvement in precision is associated with a large decrease in specificity for each classifier. In the case of kNN, on the other hand, the use of CFS improved specificity, as did classification using modified class weights (Table 4). Summarising the results presented so far, it should be noted that the kNN classifier produces the highest values for classification measures. However, even these values are not satisfactory, due to the relatively low precision and therefore poor classification coefficients of the minority class instances. The last two paths of the data processing procedure involved aligning class counts (improving balancing) using SMOTE. Table 7 shows the classification results of the dataset extended with new instances generated using SMOTE. Table 8 shows the results of the classification of the extended dataset where conditional attribute reduction was simultaneously performed using the CFS filter.

Table 7. Classification results: SMOTE and standard classifiers.

Classifier	Accuracy [%]	MCC	ROC	Precision (TPR)	Specificity (TNR)
C4.5	85.31	0.706	0.867	0.854 (346/405)	0.852 (345/405)
DT	88.52	0.779	0.948	0.812 (329/405)	0.958 (388/405)
kNN	88.76	0.778	0.888	0.931 (377/405)	0.844 (342/405)
ANN	86.79	0.736	0.936	0.877 (355/405)	0.859 (348/405)
SVM	79.51	0.592	0.795	0.830 (336/405)	0.760 (308/405)

Table 8. Classification results: SMOTE, CFS and standard classifiers.

Classifier	Accuracy [%]	MCC	ROC	Precision (TPR)	Specificity (TNR)
C4.5	86.91	0.738	0.905	0.867 (351/405)	0.872 (353/405)
DT	88.52	0.776	0.945	0.825 (334/405)	0.946 (383/405)
kNN	86.05	0.721	0.860	0.872 (353/405)	0.849 (344/405)
ANN	79.14	0.583	0.827	0.788 (319/405)	0.795 (322/405)
SVM	51.61	0.033	0.516	0.598 (242/4050	0.435 (176/405)

The comparison of Table 7 with Tables 2–6 clearly shows that the use of SMOTE has so far produced the best classification results for each of the classifiers studied. The worst results among the classifiers considered were obtained by SVM, but these are still better than the best classifier kNN so far (compare Tables 2–6). Among the classifiers in Table 7, ANN and C4.5 are the most balanced in terms of precision and specificity, but ANN obtained better classification results than C4.5. KNN has the highest precision

and DT dominates in terms of specificity. An analysis of Table 8 shows that the use of CFS improved the results of the C4.5 classifier, essentially did not change the results of the DT classifier, and negatively affected the other classifiers. However, this does not change the fact that all the results presented in Table 8 are still better than the results obtained previously and presented in Tables 2–6.

4 Conclusion

Eye-tracking studies can play an important role in analyzing the sales offer selection process, providing valuable information about consumer behaviour and preferences. The research presented here used eye-tracking data extracted from a selection of home listings from 45 respondents. They evaluated 10 offers by selecting one, resulting in a total of 450 data instances. Various methodological approaches were used to develop a classifier model for predicting the choice of a house offer (based on the determined AOI metrics of the offers). C4.5, DT, ANN, kNN and SVM classifiers were used, as well as techniques to reduce the impact of imbalanced data: modified class weights, modified probability of the class occurrence and SMOTE. Attribute filtering using CFS was also used. Results were compared using different measures of classifier evaluation to identify the best models for predicting the selection of house listings.

For the imbalanced dataset, the kNN classifier performed best for most measures and DT performed worst. Modifying the class weights improved the MCC, ROC and precision indices for the DT and ANN classifiers. Additional application of the CFS filter raised the precision of ANN and SVM, while worsening the MCC and specificity values. The application of modified probability of the class occurrence improved precision for C4.5, DT and ANN classifiers, but lowered DT specificity. Using the CFS filter in this case improved precision for most classifiers, but with a significant decrease in specificity. Precision for kNN remained unchanged, but specificity increased. The best classification results were achieved after using SMOTE, which generates synthetic minority class instances, significantly improving all classification measures. KNN achieved the highest precision, while DT dominated in terms of specificity. Using SMOTE in combination with CFS improved the performance of the C4.5 classifier and maintained the high specificity of DT. Overall, SMOTE significantly improved classification performance for all classifiers compared to previous methods. Research indicates that imbalanced datasets can be effectively balanced using methods such as SMOTE, leading to better performance of classification models.

To sum up, it can be concluded that the use of machine learning methods in combination with eye-tracking data makes it possible to significantly increase the effectiveness of consumer choice predictions. The results show that models based on these methodological approaches can effectively identify patterns of behaviour that traditional data analysis methods could miss. In practice, the research results can be used, for example, by companies in the real estate industry to develop more effective marketing strategies that better engage customers by optimizing visual presentations of home sale offers based on eye-tracking data.

As for the limitations of the conducted research, it should be noted that other classifiers, which may have potential applications to the problem under consideration, were not

verified. These include, among others, Random Forest, AdaBoost, Logistic Regression, etc. The use of other data preprocessing techniques, which may be useful in the case of an imbalanced dataset, should also be considered, such as discretization and over- and undersampling. Another important issue omitted in the study is the identification of those AOIs and, consequently, those metrics that may have the greatest impact on the decision to select or reject an offer. All these limitations indicate further research directions that should be undertaken in future work.

Funding:. Co-financed by the Minister of Science under the "Regional Excellence Initiative".

References

1. Bargh, J.A.: Losing consciousness: automatic influences on consumer judgment, behavior, and motivation. J. Consum. Res. **29**, 280–285 (2002). https://doi.org/10.1086/341577
2. Koklic, M., Vida, I.: A strategic household purchase: consumer house buying behavior. Manag. Global Trans. **7**, 75–96 (2009)
3. Bogdanowicz, B., Cypryjański, J.: Czynniki determinujące pozytywny odbiór wizualizacji architektonicznych w opinii grafików. Architecturae et Artibus. **31**, 5 (2017)
4. Broers, N., Bainbridge, W.A., Michel, R., Balestrieri, E., Busch, N.A.: The extent and specificity of visual exploration determines the formation of recollected memories in complex scenes. J. Vis. **22**, 9 (2022). https://doi.org/10.1167/jov.22.11.9
5. Mikhailova, A., Raposo, A., Sala, S.D., Coco, M.I.: Eye-movements reveal semantic interference effects during the encoding of naturalistic scenes in long-term memory. Psychon. Bull. Rev. **28**, 1601–1614 (2021). https://doi.org/10.3758/s13423-021-01920-1
6. Hessels, R.S., Hooge, I.T.C.: Eye-tracking in developmental cognitive neuroscience—the good, the bad and the ugly. Dev. Cogn. Neurosci. **40**, 100710 (2019). https://doi.org/10.1016/j.dcn.2019.100710
7. Hu, T., Wang, X., Xu, H.: Eye-tracking in interpreting studies: a review of four decades of empirical studies. Front. Psychol. **13** (2022). https://doi.org/10.3389/fpsyg.2022.872247
8. Maioli, C., Falciati, L., Gianesini, T.: Pursuit eye movements involve a covert motor plan for manual tracking. J. Neurosci. **27**, 7168–7173 (2007). https://doi.org/10.1523/JNEUROSCI.1832-07.2007
9. Kowler, E.: Eye movements: the past 25 years. Vision. Res. **51**, 1457–1483 (2011). https://doi.org/10.1016/j.visres.2010.12.014
10. Kim, J.Y., Kim, M.J.: Identifying customer preferences through the eye-tracking in travel websites focusing on neuromarketing. J. Asian Architect. Build. Eng. **23**, 515–527 (2024). https://doi.org/10.1080/13467581.2023.2244566
11. Wlekły, P., Piwowarski, M.: The usability of eye-tracking in the design of digital training materials. Procedia Comput. Sci. **207**, 4180–4189 (2022). https://doi.org/10.1016/j.procs.2022.09.481
12. Fisher, C.E., Chin, L., Klitzman, R.: Defining neuromarketing: practices and professional challenges. Harv. Rev. Psychiatry **18**, 230 (2010). https://doi.org/10.3109/10673229.2010.496623
13. Castilla, D., et al.: Improving the understanding of web user behaviors through machine learning analysis of eye-tracking data. User Model. User-Adap. Inter. **34**, 293–322 (2024). https://doi.org/10.1007/s11257-023-09373-y
14. Pina, A., Petersheim, C., Cherian, J., Lahey, J.N., Alexander, G., Hammond, T.: Using machine learning with eye-tracking data to predict if a recruiter will approve a resume. Mach. Learn. Knowl. Extr. **5**, 713–724 (2023). https://doi.org/10.3390/make5030038

15. Lim, J.Z., Mountstephens, J., Teo, J.: Eye-tracking feature extraction for biometric machine learning. Front. Neurorobot. 15, (2022). https://doi.org/10.3389/fnbot.2021.796895
16. Novák, J.Š., Masner, J., Benda, P., Šimek, P., Merunka, V.: Eye tracking, usability, and user experience: a systematic review. Int. J. Hum. Comput. Interact. **40**(17), 4484–4500 (2024).https://doi.org/10.1080/10447318.2023.2221600
17. Brunyé, T.T., Drew, T., Weaver, D.L., Elmore, J.G.: A review of eye-tracking for understanding and improving diagnostic interpretation. Cogn. Res. Princ. Implic. **4**, 7 (2019). https://doi.org/10.1186/s41235-019-0159-2
18. Mahanama, B., et al.: Eye movement and pupil measures: a review. Front. Comput. Sci. **3** (2022). https://doi.org/10.3389/fcomp.2021.733531
19. van der Laan, L.N., Hooge, I.T.C., de Ridder, D.T.D., Viergever, M.A., Smeets, P.A.M.: Do you like what you see? The role of first fixation and total fixation duration in consumer choice. Food Qual. Prefer. **39**, 46–55 (2015). https://doi.org/10.1016/j.foodqual.2014.06.015
20. Guo, F., Ding, Y., Liu, W., Liu, C., Zhang, X.: Can eye-tracking data be measured to assess product design?: Visual attention mechanism should be considered. Int. J. Ind. Ergon. **53**, 229–235 (2016). https://doi.org/10.1016/j.ergon.2015.12.001
21. World Medical Association: World medical association declaration of Helsinki: ethical principles for medical research involving human subjects. JAMA **310**, 2191 (2013). https://doi.org/10.1001/jama.2013.281053
22. Ziemba, P., Becker, J., Becker, A., Radomska-Zalas, A., Pawluk, M., Wierzba, D.: Credit decision support based on real set of cash loans using integrated machine learning algorithms. Electronics **10**, 2099 (2021). https://doi.org/10.3390/electronics10172099
23. Quinlan, J.R.: Improved use of continuous attributes in C4.5. J. Artif. Intell. Res **4**, 77–90 (1996). https://doi.org/10.1613/jair.279
24. Kohavi, R.: The power of decision tables. In: Proceedings of the 8th European Conference on Machine Learning. Springer-Verlag, Berlin, Heidelberg, pp. 174–189 (1995). https://doi.org/10.1007/3-540-59286-5_57
25. Park, Y.-S., Lek, S.: Chapter 7—Artificial neural networks: multilayer perceptron for ecological modeling. In: Jørgensen, S.E. (ed.) Developments in Environmental Modelling. Elsevier, pp. 123–140 (2016). https://doi.org/10.1016/B978-0-444-63623-2.00007-4
26. Yuk Carrie Lin, K.: Optimizing variable selection and neighbourhood size in the *K*-nearest neighbour algorithm. Comput. Ind. Eng. **191**, 110142 (2024). https://doi.org/10.1016/j.cie.2024.110142
27. Wang, Q., et al.: A hybrid SVM and kernel function-based sparse representation classification for automated epilepsy detection in EEG signals. Neurocomputing **562**, 126874 (2023). https://doi.org/10.1016/j.neucom.2023.126874
28. Chen, Y., Calabrese, R., Martin-Barragan, B.: Interpretable machine learning for imbalanced credit scoring datasets. Eur. J. Oper. Res. **312**, 357–372 (2024). https://doi.org/10.1016/j.ejor.2023.06.036
29. Hwang, J.P., Park, S., Kim, E.: A new weighted approach to imbalanced data classification problem via support vector machine with quadratic cost function. Expert Syst. Appl. **38**, 8580–8585 (2011). https://doi.org/10.1016/j.eswa.2011.01.061
30. Freeman, E.A., Moisen, G.G.: A comparison of the performance of threshold criteria for binary classification in terms of predicted prevalence and kappa. Ecol. Model. **217**, 48–58 (2008). https://doi.org/10.1016/j.ecolmodel.2008.05.015
31. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. **16**, 321–357 (2002). https://doi.org/10.1613/jair.953

32. Pushpalatha, K.R., Karegowda, A.G.: CFS based feature subset selection for enhancing classification of similar looking food grains—a filter approach. In: 2017 2nd International Conference On Emerging Computation and Information Technologies (ICECIT), pp. 1–6 (2017). https://doi.org/10.1109/ICECIT.2017.8453403
33. Malakouti, S.M., Menhaj, M.B., Suratgar, A.A.: The usage of 10-fold cross-validation and grid search to enhance ML methods performance in solar farm power generation prediction. Cleaner Eng. Technol. **15**, 100664 (2023). https://doi.org/10.1016/j.clet.2023.100664
34. Szabó, S., Holb, I.J., Abriha-Molnár, V.É., Szatmári, G., Singh, S.K., Abriha, D.: Classification assessment tool: a program to measure the uncertainty of classification models in terms of class-level metrics. Appl. Soft Comput. **155**, 111468 (2024). https://doi.org/10.1016/j.asoc.2024.111468

Competencies of the Future as a Criterion for Segmentation of Generation Z Candidates: Machine Learning and the CART Model

Dorota Jelonek[1(✉)], Magdalena Graczyk-Kucharska[2], Robert Olszewski[3], Maciej Szafrański[2], and Magdalena Rzemieniak[4]

[1] Czestochowa University of Technology, ul. Dabrowskiego 69, 42-201 Czestochowa, Poland
dorota.jelonek@pcz.pl
[2] Faculty of Engineering Management, Poznan University of Technology, ul. Jacka Rychlewskiego 2, 60-965 Poznan, Poland
[3] Department of Geodesy and Cartography, Warsaw University of Technology, Pl. Politechniki 1, 00-661 Warsaw, Poland
[4] Lublin University of Technology, Nadbystrzycka 38D, 20-618 Lublin, Poland

Abstract. In the era of the digital revolution, there is a great emphasis on implementing new technologies including AI. In the context of the changes strategic resource planning becomes necessary. This also applies to the provision of human resources and strategic competencies, and in the context of the implementation of new, future technologies also to have the company's competence if the future to implement them. The new approach in HR that the authors propose, which allows planning and forecasting the availability of competencies of the future relating to AI, is the well-known customer segmentation used here for the segmentation of potential candidates. The purpose of the article is to explore the possibility of modeling the segmentation of candidates with selected competencies of the future. Data on young generation Z's characteristics, competencies possessed, candidates' job expectations, or location were used for analysis. In this article ML and CART are used as examples of AI methods for data analysis. The analysis was conducted on a sample of 2,197 16–19 year olds, male and female students studying 12 technical subjects. The key finding is that ML and AI are widely applicable in segmenting candidates, and that the competencies of the future can be linked to their expectations of where and how they work, including such things as work atmosphere, company innovation, and the ability to work remotely. The implementation of similar modeling in practice can contribute in organizations to the optimization of management decision-making including increasing the efficiency of resources and recruitment processes.

Keywords: Competence Segmentation · competence management · human resources · Generation Z · Machine Learning · CART · AI

M. Hernes et al. (Eds.): ECAI 2024, LNNS 1217, pp. 118–130, 2025.
https://doi.org/10.1007/978-3-031-78465-1_10

1 Introduction

Artificial intelligence (AI) is transforming modern management [1], offering breakthrough opportunities to optimize processes [2], personalize the customer experience [3], and increase operational efficiency across a variety of business sectors [4]. Changes from a management perspective range from the different public sector [5], education [6], to management practices [8] like marketing [9], manufacturing [9], supply chain [10], or human resource management [11].

In management and quality sciences, AI is seen as a complex ecosystem consisting of three key components: data collection and storage, statistical and computational methods, and advanced output systems [12]. These components support companies in automating processes that typically require human intelligence and autonomous decision-making, such as quality management, process control or operations improvement [13]. AI, which includes technologies such as natural language processing (NLP), machine learning (ML), deep learning (DL), and computer vision [14], enables machines to perform cognitive functions that are key to effective management and raising quality standards. ML and DL are two of the most effective AI techniques that find application in optimizing management, quality, enterprise operational processes, and the ability to leverage human potential in an organization [15, 16].

ML methods, including the classification and regression tree algorithm (CART), were used in the study. CART is a predictive technique in the field of ML, which allows predicting the value of a target variable based on other variables.

AI technologies, are helping to automate and streamline human resource (HR) processes, such as recruitment, employee performance evaluation, talent management and career development [17], which have traditionally depended on human analysis and decision-making. AI is commonly used in the recruitment process [18, 19], primarily at the candidate selection stage. While AI is still not widely used in other areas and HR functions, through data analytics, it can address the challenges faced by recruiters with regard to finding the best candidates for certain jobs. Also, they can be supportive in the decision-making process regarding the competency potential of candidates in the region, which can add analytics to management decisions, for example, where to expand or locate a particular business [20]. Such an approach can be crucial, especially in such cases where there is low unemployment in the region and recruitment must be accomplished efficiently. Candidates' preferences regarding expectations of where and how they want to work can then increase the likelihood of finding a candidate [21], thereby minimizing the company's recruitment and operational costs. By integrating AI into a company's strategy, including its HR management (HRM) strategy, entities can effectively use available data and reach out to potential candidates with job offers. HR professionals, based on the individual preferences of potential candidates, can match job offers, increasing the effectiveness of recruitment. This is crucial in building the company's potential based on HR, knowledge, skills and attitudes of Generation Z entering the labor market.

It is already widely indicated in both academic and industry literature that AI will be used in many areas, both in the life of society and in businesses, government institutions and other organizations. In order for people to effectively use AI tools, it is necessary to have selected skills such as analytical skills [22, 23]. Therefore, for the purpose of

this article, one of the competencies of the future, "the ability to think analytically and draw conclusions," commonly identified in future competency reports as the one directly related to artificial intelligence [24], was selected.

Providing competencies such as analytical skills can contribute significantly to the implementation of AI in organizations and revolutionize the way management is done in the HR area as well. This includes how companies interact with their employees and potential candidates.

The research problem posed by the authors in this study concerns the availability of Generation Z candidates and the possession of the "ability to think analytically and draw conclusions" competencies required in the digital transformation. To this end, the following research questions were posed: (*i*) what are the trends in the self-assessment of "ability to think analytically and draw conclusions" competencies by Generation Z candidates? (*ii*) For the purpose of recruiting sought-after candidates with future competencies, can a candidate segmentation approach using ML and CART be used? This article examines the use of large datasets of potential Generation Z candidates entering the labor market trying to segment potential candidates and predict their behavior using ML and CART. In the context of the presented research directions on the use of AI in HRM, it is reasonable to present the results of the conducted analysis, which shows the wide application of ML and CART in the field of HR, including in the new area of segmentation of potential candidates in the labor market. The example of the analysis refers to Wielkopolska (a region of Poland), with a population of nearly 3.5 million people, which has a low unemployment rate of 3.3%.

2 Literature Review

We live in an era of dynamic changes, and future competencies are key in adapting to the rapidly changing technological and market landscapes. According to [25], the most sought-after future competencies will include analytical thinking and innovation, active learning and learning strategies, comprehensive problem solving, critical thinking and analysis, creativity, originality, and initiative/entrepreneurship. Subsequent studies have confirmed and supplemented the above list of future competencies: analytical thinking, creative thinking, flexibility and agility, motivation and self-awareness, curiosity and lifelong learning [26]; critical-thinking skills that blend creativity and analytics for effective decision-making, digital fluency and a continuous-learning mindset, empathy, compassion, and emotional intelligence, ethical, authentic, and inclusive resilience [27]; digital and media literacies and problem-solving skills [28], and abstract thinking as a manager's competence [29]. The Trends Survey [27] also emphasized that in the world of new technologies and a sea of information, one must be able to navigate, and then future competencies will include: creativity, analytical thinking, and empathy, while in communicating with AI, analytical skills are primarily required [22, 24, 30].

In summary, across various nomenclatures such as analytical thinking and innovation [25], analytical thinking [26], critical-thinking skills [25, 27], digital and media literacies [28], the competence "ability to think analytically and draw conclusions" is emphasized. Moreover, researchers highlight that this competency is directly linked with AI [22, 24, 30], which we are already using, and Generation Z will increasingly use.

For a better understanding of Generation Z and its digitization, it should be highlighted that they were born in the 1990s and raised in the 2000s during the most profound changes in the century, existing in a world with the web, internet, smartphones, laptops, freely available networks, and digital media [31]. No previous generations have as many labels as Generation Z, for example, Online Generation, iGeneration, Gen Tech, Facebook Generation, all defining this generation through the prism of high civilizational competencies, especially related to the use of virtual space. Researchers characterize Generation Z as digital natives [32], highly achievement-oriented [33], and seeking interesting and meaningful work [34]. Generation Z plans a career path based on needs, aligned with their lifestyle, emphasizing the need for a psychological contract with the organization to maintain a balance between work and personal life, expecting freedom and flexibility at work to meet their personal and family needs [35]. Combining AI capabilities with recognizing the preferences and expectations of Generation Z candidates enhances the efficiency of recruitment processes [21, 36] and the effectiveness of HRM [36].

An element that connects the future competencies of Generation Z and the possibilities of AI in relation to certain communities is segmentation. Market segmentation is most often performed as it is key to accommodating the tastes and preferences of collective groups, rather than focusing on individual customers [37]. Segmentation is also implemented in the labor market [38], in political marketing [39] and in healthcare [40].

A detailed analysis of research indicates that although statistical techniques can be used for segmentation, their effectiveness decreases when handling large data sets. An alternative approach involves the use of ML, including decision trees, which are capable of achieving high performance without relying on pre-existing hypotheses. ML revolutionizes HRM practices [41] and automates the segmentation of candidate data, e.g., method for segmenting CVs into basic information, education, and work experience [42]. In the segmentation of Generation Z candidates, ML and CART were employed. The CART model is particularly suitable for the segmentation of Generation Z candidates, as it can effectively handle various types of data, including the competency "ability to think analytically and draw conclusions" as a segmentation criterion.

3 Materials and Methods

3.1 Data

In the conducted research, 2197 questionnaires obtained between 2018 and 2022 were analysed. The study group was technical school students in the Wielkopolska Province (1460 males and 737 females) studying 12 fields of study (Table 1):

Respondents provided a range of qualitative as well as quantitative information in the questionnaires regarding both the course they attended and their preferences for future work, their desire to seek a job/internship and their own skills in terms of analytical thinking skills, problem diagnosis, openness to new technological solutions, programming skills (both in general and Java), awareness of cyber security risks, etc. Those surveyed also indicated their place of residence and the location of the school they attend. Geocoding the questionnaires allowed us to obtain a spatial distribution of the

Table 1. Survey sample divided by respondents' occupation

No	Profession	Number of respondents
1	IT technician	765
2	economic specialist	465
3	logistics technician	302
4	mechatronics technician	244
5	advertising organisation technician	145
6	mechanical engineering technician	80
7	electronics technician	66
8	ICT technician	64
9	forwarding technician	26
10	electrical engineering technician	16
11	digital graphic processes technician	13
12	trade technician	11

respondents' place of residence and place of study, accurate to the postcode of the place of study. The geospatial data was then aggregated to individual districts. In the questionnaires, respondents indicated the year of registration (Table 2).

Table 2. Research sample in divided by years of collected data among respondents

No	Year of registration	Number of respondents
1	2018	685
2	2019	784
3	2020	213
4	2021	368
5	2022	147

Each respondent completed a scale self-assessment of his or her skills and, in addition, the manner of work (individual or group; in a group of Polish employees or in a group of international at the company's headquarters or remotely; 8 h a day or on a task basis) and the priorities that motivate candidates to work (atmosphere at work or salary; work with passion or just work; work in an innovative or traditional company).

3.2 Methods

The study used CART tree methods [43] regression and classification. In addition to these methods for dividing data into classes, there are many others such as linear multiple regression [44], or the more traditional cluster analysis [45]. In the data mining

performed, however, the use of CART trees was crucial, as it allowed both the classification and regression of nonlinear models and the use of data expressed in different measurement scales: qualitative, rank and quantitative. Neural networks and other tree-based models, e.g. random forest, also allow similar processing of broad data sets [46]. These other models allow high quality coefficients for classification and regression. However, the choice of CART methods was justified insofar as they allowed (albeit at the cost of slightly lower accuracy coefficients) for quick interpretation of results given in a transparent manner. Such an approach is particularly useful, given the complexity and intuitive ambiguity of the links between a wide variety of independent variables describing students' preferences and functioning, which can affect the dependent variable studied, adopted in the article, and whose possession and level students only declared. The use of CART trees makes it possible to obtain rules that "translate" the explanatory variable. These rules can be extracted directly from the CART model as SQL language formulas, which allows to automate the process of object selection.

4 Results

4.1 CART Model

In the research conducted, "Ability to think analytically and draw conclusions" was used as the explanatory variable. This variable was coded in the research as an integer value with an interval of <0.5>. In the research conducted, this value was interpreted both as a rank variable aggregated to three intervals: low, average, high, and as a quantitative variable (Table 3). This approach allowed the development of two explanatory models in the form of so-called CART (Classification And Regression Trees) decision trees - a classification tree and a regression tree. In aggregating the data, it was assumed that a response with a rank value higher than 1 indicated a significant level of belief by the respondent in his or her ability to think analytically and draw conclusions.

Table 3. Research sample in divided by years of collected data among respondents

Value	Number of respondents	Rank scale	Number of respondents
0	701	low	701
1	853	average	853
2	26	high	643
3	137		
4	262		
5	218		
TOTAL	2197		2197

Analysing the above data in the context of a continuous quantitative scale, it can be concluded that the mean value of the respondents' answers is relatively low and

amounts to 1.57 with a significant standard deviation value of 1.71. Thanks to the use of geocoding algorithms for the data provided by the respondents, the results of the questionnaires were also analysed spatially (Fig. 1). The map visualises the average value of the respondents' answers in the respective district. The intensity of the red colour indicates the level of belief of those surveyed in the possession of the "Ability to think analytically and draw conclusions". The map shows in grey those counties for which the number of questionnaires collected was too small (less than 10) to make reliable inferences about respondents' beliefs. The map indicates that the spatial distribution of the results given by the respondents is not random. Within the voivodeship, it is possible to distinguish zones in which young people's belief in their own analytical skills is relatively high, and zones in which respondents consider themselves to have low analytical skills. Interpretation of these results may indicate a differentiated level of education in particular regions of Wielkopolska (especially in terms of supporting students in creative thinking and independence). However, this interpretation, due to the relatively large values of the standard deviation treated as a measure of the uncertainty of the result in a given district, should be treated with caution.

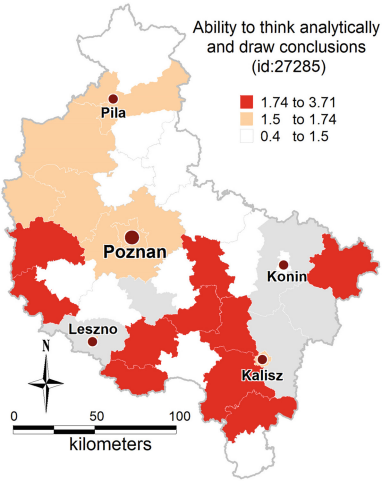


Fig. 1. Spatial distribution of respondents' results

In further research, a regression-type CART tree model was developed on the basis of the analysis of the collected data to explain the respondents' "Ability to think analytically and draw conclusions" values. In this model, the information most strongly correlated with the explanatory variable was used as independent variables: year of registration, gender of the respondent, preference for individual or group work, relevance of the atmosphere in the potential job, and distance from Poznań. The latter variable was determined through a process of spatial data enrichment as the distance in a straight line from the school the respondent attends to the centre of the provincial capital. The results obtained (Fig. 2) indicate that the highest value for analytical thinking skills (2.30) is predicted by the CART model for women who were registered in the system

after 2020. This is indicated by the so-called decision tree leaf with ID = 3, for which the number of respondents meeting these criteria is 146. In contrast, the lowest value is estimated by the leaf ID = 10, which corresponds to the generalised characteristics of 125 respondents. These conditions are met by women registered before 2019 who live relatively close to Poznań (less than 65 km). For the 56 men who meet the identical conditions and, in addition, consider the atmosphere in a potential job as important (at least 3.5 on a scale of 0 to 5), the predicted value of analytical skills by the model is 1.55 (leaf with ID = 13). When analysing the model developed, it is important to remember that the number of levels, "branches" and, above all, "leaves" determines the level of detail of the inference. A model that consisted of a number of leaves equal to the number of respondents (in this case 2197) would describe the individual preferences of each respondent. In contrast, a model containing just one leaf would show the averaged characteristics of all respondents. In the research conducted, 31 terminal nodes were taken as the level of generalisation. Leaf ID = 24 describes the largest number of 381 male respondents, who study in a school at least 80 km away from Poznań, were registered in the system before 2021 and prefer individual to group work.

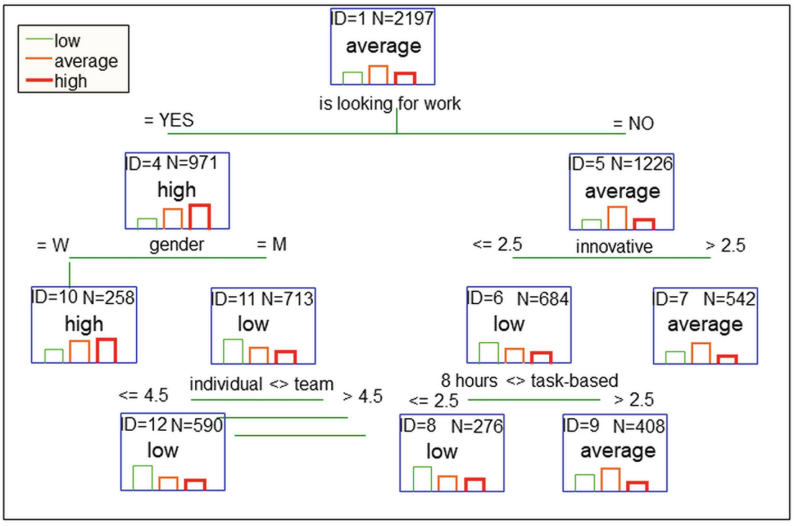


Fig. 2. Regression tree for the explanatory variable "Ability to think analytically and draw conclusions."

In an analogous way, a CART classification tree was developed in which three classes of analytical thinking skills were explained: low, average and high. The results (Fig. 3) indicate that the explanatory variables in this model are willingness to look for a job, gender and respondents' opinion of innovative work vs. "simply" work, task-based work vs. "8 h a day", and individual work vs. team work.

Respondents claiming to have high analytical thinking skills as determined by the CART model are:

– women looking for a job (leaf ID = 10, 258 respondents,

– men looking for jobs of a team nature (at min. 4.5 on a scale of 0–5) - leaf ID = 13, 123 respondents.

Respondents declaring to have average analytical thinking skills:

– unisex not looking for work and valuing innovative challenges (at a minimum level of 2.5 on a scale of 0–5) - leaf ID = 7, 542 respondents,
– unisex people not looking for a job and valuing innovative work, but preferring task-based work to fixed working hours - leaf ID = 9, 408 respondents.

Respondents declaring to have low analytical thinking skills:

– male jobseekers who prefer working alone (leaf ID = 12, 590 respondents),
– unisex jobseekers who do not value work of an innovative nature, but prefer to work fixed hours in cities with daytime challenges - leaf ID = 8, 276 respondents.

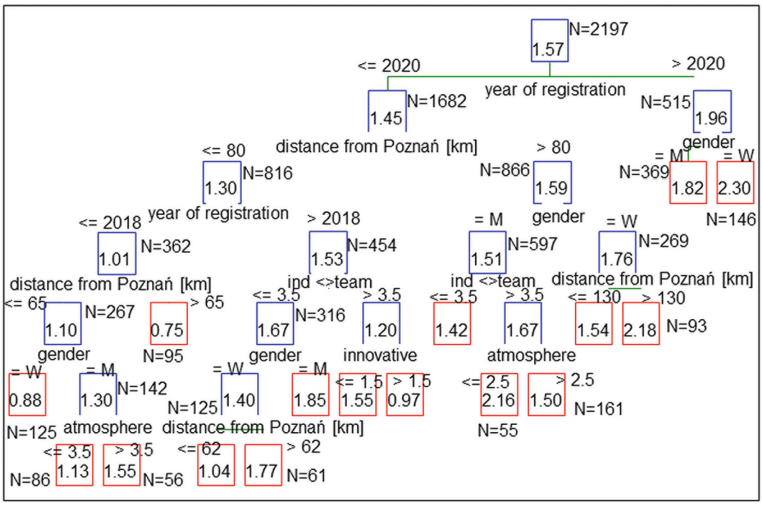


Fig. 3. Regression tree for the explanatory variable "Ability to think analytically and draw conclusions".

5 Discussion and Conclusions

The conducted research makes it possible to extend the application of CART methods in HRM to the segmentation of technical high school students. The interest in this target group - representatives of Generation Z - stems from the search for methods of accelerating the preparation of employees for work, even at the stage of educating future candidates [48], especially in the area of competencies of the future, which "Ability to think analytically and draw conclusions" was taken as a representative in this article.

The study conducted represents a novel and unique approach of AI's impact on HRM with the necessary future competencies required to adapt to technological advances.

The article draws a marketing concept for segmenting potential candidates using ML methods. The study shows (Fig. 2) that newer vintages of students rate their level of analytical thinking competence higher than earlier vintages, which may indicate that, in general, the level of this competence is increasing in successive vintages, or that students from successive vintages are characterized by greater self-confidence. It has been noted that the highest self-assessments of the level of the competency under study are indicated by students from schools located closer to the regional capital, while the lowest assessments are indicated by students studying in schools further away. Quite surprisingly, it turns out that the highest self-assessments are most often given by female students. The above answers the first research question. Answering the second research question, it can be concluded that ML with the support of the CART method can be an effective tool for segmenting potential candidates, thus can be effectively used in the digital transformation of HR departments.

The research was conducted under certain constraints, such as regionality and non-representativeness of the foundational data, self-assessment of competencies by students. Nevertheless, the results achieved are very promising in both theoretical and utilitarian dimensions. The article answers the research question posed: how to search for classes of students - future job candidates, with similar characteristics, who will be characterized by similar levels of the competence selected for analysis? Identification of classes of students with the lowest levels of competence makes it possible to target training activities, diversify training programs, shorten training processes, and thus increase the effectiveness and efficiency of the activities of schools, local governments, training companies and entities that are responsible for the distribution of funds, including EU funds. Quick identification of groups of students with the highest levels of competence will allow companies to speed up access to them and establish cooperation with them (internships, apprenticeships) more quickly, getting ahead of competitors in reaching them and reducing the cost of putting future graduates to work by shortening recruitment and onboarding processes.

Further avenues of research may include consideration of regional and demographic context variation, with a diverse sample in different regions, analysis of other core competencies for digital transformation or green transformations. It is also possible to use and analyze other AI and ML methods, among others, for prediction purposes, also to compare and verify the results obtained.

References

1. Kitsios, F., Kamariotou, M.: Artificial intelligence and business strategy towards digital transformation: a research agenda. Sustainability **13**(4), 2025 (2021)
2. Dwivedi, Y.K., et al.: Artificial Intelligence (AI): multi-disciplinary perspectives on emerging challenges, opportunities, and agenda for re-search, practice and policy. Int. J. Inf. Manage. **57**, 101994 (2021)
3. Ameen, N., Tarhini, A., Reppel, A., Anand, A.: Customer experiences in the age of artificial intelligence. Comput. Hum. Behav. **114**, 106548 (2021)
4. Wamba-Taguimdje, S.L., Wamba, S.F., Kamdjoug, J.R.K., Wanko, C.E.T.: Influence of artificial intelligence (AI) on firm performance: the business value of AI-based transformation projects. Bus. Process. Manag. J. **26**(7), 1893–1924 (2020)

5. Zuiderwijk, A., Chen, Y.C., Salem, F.: Implications of the use of artificial intelligence in public governance: a systematic literature review and a research agenda. Gov. Inf. Q. **38**(3), 101577 (2021)
6. Chen, L., Chen, P., Lin, Z.: Artificial intelligence in education: a review. IEEE Access **8**, 75264–75278 (2020)
7. Rakova, B., Yang, J., Cramer, H., Chowdhury, R.: Where responsible ai meets reality: practitioner perspectives on enablers for shifting organizational practices. Proc. ACM on Hum. Comput. Interact. **5**(CSCW1), 1–23 (2021)
8. Haleem, A., Javaid, M., Qadri, M.A., Singh, R.P., Suman, R.: Artificial intelligence (AI) applications for marketing: a literature-based study. Int. J. Intell. Netw. **3**, 119–132 (2022)
9. Kinkel, S., Baumgartner, M., Cherubini, E.: Prerequisites for the adoption of AI technologies in manufacturing—evidence from a worldwide sample of manufacturing companies. Technovation **110**, 102375 (2022)
10. Helo, P., Hao, Y.: Artificial intelligence in operations management and supply chain management: an exploratory case study. Prod. Plann. Control **33**(16), 1573–1590 (2022)
11. Vrontis, D., Christofi, M., Pereira, V., Tarba, S., Makrides, A., Trichina, E.: Artificial intelligence, robotics, ad-vanced technologies and human resource management: a systematic review. Int. J. Hum. Resour. Manage. **33**(6), 1237–1266 (2022)
12. Agrawal, A., Gans J., Goldfarb, A.: Prediction machines: the simple economics of artificial intelligence, In: Agrawal, A., Gans, J., Goldfarb A. (eds.) The Economics of Artificial Intelligence: An Agenda, pp. 1–19. University of Chicago Press (2018)
13. Senoner, J., Netland, T., Feuerriegel, S.: Using Explainable Artificial Intelligence to Improve Process Quality: Evidence from Semiconductor Manufacturing. Manage. Sci. **68**(8), 5704–5723 (2022)
14. Lauriola, I., Lavelli, A., Aiolli, F.: An introduction to deep learning in natural language processing: models, techniques, and tools. Neurocomputing **470**, 443–456 (2022)
15. Enholm, I.M., Papagiannidis, E., Mikalef, P., Krogstie, J.: Artificial intelligence and business value: a literature review. Inf. Syst. Front. **24**(5), 1709–1734 (2022)
16. Garg, S., Sinha, S., Kar, A.K., Mani, M.: A review of machine learning applications in human resource management. Int. J. Product. Perform. Manag. **71**(5), 1590–1610 (2021)
17. Kambur, E., Yildirim, T.: From traditional to smart human resources management. Int. J. Manpow. **44**(3), 422–452 (2022)
18. Van Esch, P., Black, J.S., Ferolie, J.: Marketing AI recruitment: the next phase in job application and selection. Comput. Hum. Behav. **90**, 215–222 (2019)
19. Pan, Y., Froese, F., Liu, N., Hu, Y., Ye, M.: The adoption of artificial intelligence in employee recruitment: The influence of contextual factors. Int. J. Hum. Resour. Manage. **33**(6), 1125–1147 (2022)
20. Nong, T.N.M.: A hybrid model for distribution center location selection. Asian J. Shipp. Logist. **38**(1), 40–49 (2022)
21. Swider, B.W., Steed, L.B.: Applicant initial preferences: the relationship with job choices. Pers. Psychol. **75**(2), 321–346 (2022)
22. Dolan, E., Kosasi, S., Sari. S.N.: Implementation of competence-based human resources management in the digital era. Startupreneur Business Digital (SABDA J.) **1**(2), 167–175 (2022)
23. Graczyk-Kucharska, M., Szafranski, M., Golinski, M., Spychala, M. Borsekova, K.: Model of competency management in the network of production enterprises in industry 4.0—Assumptions. In: Hamrol A., Ciszak O., Legutko S., Jurczyk M. (eds.) Advances in Manufacturing, pp. 195–204. Springer Cham (2018)
24. Santana, M., Díaz-Fernández, M.: Competencies for the artificial intelligence age: visualisation of the state of the art and future perspectives. RMS **17**, 1971–2004 (2023)

25. World Economic Forum.: The Future of Jobs Report 2020. World Economic Forum, Geneva (2020). https://www.weforum.org/reports/the-future-of-jobs-report-2020
26. Di Battista, A., Grayling, S., Hasselaar, E., Leopold, T., Li, R., Rayner, M., Zahidi, S.: Future of jobs report 2023. In World Economic Forum, Geneva, Switzerland, (2023), https://www.weforum.org/reports/the-future-of-jobs-report-2023
27. HR/L&D Trends Survey 2024. Navigating the High-Expectation Work Environment, Blanchard (2023). https://www.blanchard.com/about-us/pressroom/2024-trends-survey-release
28. Okros A.: Cognitive capacities and competencies. In: Okros A. (eds.) Harnessing the Potential of Digital Post-Millennials in the Future Workplace. Management for Professionals, pp. 155–169, Springer Cham (2020)
29. Jelonek, D., Stepniak, C.: Evaluation of the usefulness of abstract thinking as a manager's competence. Polish J. Manage Stud. **9**, 62–71 (2014)
30. Jelonek, D., Nitkiewicz, T., Koomsap, P.: Soft skills of engineers in view of industry 4.0 challenges. In: Proceedings Conference Quality Production Improvement–CQPI, vol. 2, no. 1, pp. 107–116. De Gruyter, Warszawa (2020)
31. Singh, A.P., Dangmei, J.: Understanding the generation Z: the future workforce. South-Asian J. Multidiscip. Stud. **3**(3), 1–5 (2016)
32. Twenge, J.M.: iGen: Why Today's Super-Connected Kids Are Growing Up Less Rebellious, More Tolerant, Less Happy and Completely Unprepared for Adulthood And What That Means for the Rest of Us. Atria Books (2017)
33. Barna Group.: Is Gen Z the most success oriented generation? (2018). www.barna.com/research/is-gen-z-the-most-success-oriented-generation/
34. Schroth, H.: Are you ready for Gen Z in the workplace? Calif. Manage. Rev. **61**(3), 5–18 (2019)
35. Bohdziewicz, P.: Career anchors of representatives of Generation Z: some conclusions for managing the younger generation of employees. Hum. Resour. Manage. **6**(113), 57–74 (2016)
36. Pan, Y., Froese, F.J.: An interdisciplinary review of AI and HRM: challenges and future directions. Hum. Resour. Manag. Rev. **33**(1), 100924 (2023)
37. Khan, R.H., Dofadar, D.F., Alam, M.G.R.: Explainable customer segmentation using k-means clustering. In: 2021 IEEE 12th Annual Ubiquitous Computing. Electronics & Mobile Communication Conference (UEMCON), pp. 0639–0643. IEEE, New York (2021)
38. Colombo, E., Mercorio, F., Mezzanzanica, M.: AI meets labor market: exploring the link between automation and skills. Inf. Econ. Policy **47**, 27–37 (2019)
39. Davidson, S., Binstock, R.H.: Political marketing and segmentation in aging democracies. In: Lees-Marshment, J. (ed.) Routledge Handbook of Political Marketing, Routledge, pp. 36–49. Routledge, London (2012)
40. Swenson, E.R., Bastian, N.D., Nembhard, H.B.: Healthcare market segmentation and data mining: A systematic review. Health Mark. Q. **35**(3), 186–208 (2018)
41. Garg, S., Sinha, S., Kar, A.K., Mani, M.: A review of machine learning applications in human resource management. Int. J. Product. Perform. Manag. **71**(5), 1590–1610 (2022)
42. Tikhonova, M., Gavrishchuk, A.: NLP methods for automatic candidate's CV segmentation. In: 2019 International Conference on Engineering and Telecommunication (EnT), pp. 1–5. IEEE, Dolgoprudny, Russia (2019)
43. Wei-Yin, L.: Classification and regression trees. WIREs Data Min. Knowl. Discovery **1**(1), 14–23 (2011)
44. Gülden, K.U., Neşe, G.: A study on multiple linear regression analysis. Procedia. Soc. Behav. Sci. **106**, 234–240 (2013)
45. Henning, C., Meila, M.: Cluster analysis: An Overview, In: Henning, C., Meila, M., Murtagh, F., Rocci, R. (eds.) Handbook of Cluster Analysis, pp. 1–21. Taylor & Francis Group, Boca Raton, London, New York (2015)

46. Biau, G., Scornet, E.: A random forest guided tour. TEST **25**, 197–227 (2016)
47. Szafrański, M.: Acceleration of educating as an external factor supporting preventive and improving actions in businesses. Procedia Manuf. **3**, 4948–4955 (2015)

The Impact of Current Temperatures on Forecasting Natural Gas Futures Prices in the USA Using Machine Learning Algorithms

Grzegorz Wojarnik(✉)

Institute of Economics and Finance, University of Szczecin, Mickiewicza 64, 71-101 Szczecin, Poland
grzegorz.wojarnik@usz.edu.pl

Abstract. The aim of this study was to analyze the impact of current temperatures in major US cities on the accuracy of forecasts for natural gas futures prices using various machine learning algorithms. The study involved comparing the results of six algorithms (decision trees, random forests, support vector machines, gradient boosting, k-NN, and neural networks) in two versions: without temperature data (Study A) and with this data (Study B). The results indicate that the inclusion of temperature data significantly improves prediction accuracy in all tested algorithms. The best results were achieved by gradient boosting, showing the highest accuracy (0.581) after adding weather data. The analysis of variable importance revealed that minimum and maximum temperatures, as well as the "day_of_year" variable, were key predictors of gas prices. The conclusions of the study suggest that incorporating temperatures into predictive models may be beneficial for "end-of-day trading" investment strategies, leading to more accurate investment decisions. The study's limitations include the limited scope of weather data and the need for further testing under various market conditions. These results can help investors and market analysts develop more effective investment strategies.

Keywords: Artificial Intelligence · Machine Learning · Stock Exchange · Natural Gas Futures

1 Introduction

The analysis of the relationship between energy commodity prices and various environmental and economic factors plays a crucial role in decision-making processes in financial markets. One such commodity is natural gas, the prices of which are significantly dependent on many variables, including air temperature. Temperature fluctuations affect gas demand, particularly in the context of heating during winter months and cooling during summer periods.

The objective of this paper is to conduct a detailed analysis of the impact of current temperatures on the prices of natural gas futures in major US cities. The main research question concerns whether the inclusion of current daily temperatures in predictive models based on various machine learning algorithms can improve the quality of forecasts regarding natural gas prices.

M. Hernes et al. (Eds.): ECAI 2024, LNNS 1217, pp. 131–139, 2025.
https://doi.org/10.1007/978-3-031-78465-1_11

The working hypothesis suggests that introducing data on current temperatures into the data model with technical analysis indicators used to train a classifier based on various machine learning algorithms allows for increased prediction quality related to possible investment decisions (buy, sell, stay out of the market) made based on the classifier's predictions.

The analysis will utilize different machine learning algorithms such as decision trees, random forests, support vector machines, gradient boosting, k-NN, and neural networks. They will be tested in two versions: considering only historical price data and technical indicators and with additional temperature data.

An important aspect of the study is also assessing which of the applied algorithms best responds to the introduction of temperature data, allowing for conclusions regarding optimal investment strategies in the context of natural gas futures markets.

Research on the impact of temperature on natural gas prices is extensive, as shown by studies such as [22], which modeled gas prices considering demand, supply, and economic growth indicators, or [24], which demonstrated a positive correlation between winter temperatures and natural gas consumption in residential buildings in the USA. This study fits into this research trend, extending it by analyzing the potential use of "end-of-day trading" strategies using machine learning for forecasting natural gas futures prices based on current weather data.

The next section will discuss the research methodology and the results of the analyses in detail.

2 Related Works

The analysis of natural gas prices concerning temperatures is the subject of numerous studies. [22] in their study model natural gas prices considering demand, supply, and economic growth indicators. They state that gas prices are mainly shaped by market demand and supply functions, with factors such as weather and temperature playing a key role in demand changes. In [27], a comparative analysis of industrial heat supply processes was conducted, assessing their thermodynamic and economic efficiency, finding that fluctuations in gas and electricity prices influence the choice of the most favorable heat production methods. The study [20] identifies the relationships between fossil fuel prices, including natural gas, and climate change variables such as temperature and precipitation, proving that natural gas prices are strongly linked to climate changes.

The analysis of natural gas prices concerning temperatures is also present in other works, for example, in [1], a comprehensive thermo-economic analysis of a modern Power-to-Gas system was conducted, pointing out the influence of temperature and pressure on the costs of producing synthetic natural gas. In [24], it was shown that winter temperatures in central and eastern United States are positively correlated with natural gas consumption in residential buildings, suggesting that residential demand dependent on temperature is more sensitive to market changes. In [4], a joint model of natural gas and temperature prices was proposed, showing their mutual dependencies in the US market. In [11], the impact of CO2 prices on electricity prices was studied, considering changes in natural gas prices and temperature. It is also worth noting the study [8], which developed a stochastic model that can effectively manage temperature-related risks in

natural gas demand, helping to assess the exchange of energy quanta. Meanwhile, in [3], the authors propose a simple joint model for natural gas and temperature prices, considering time delay affecting their mutual dependencies, allowing them to demonstrate a model accurately describing spot prices of natural gas and temperatures with a time delay parameter affecting the temperature forecast in mutual dependency.

Regarding the impact of current temperatures on future natural gas prices, works such as [19] examined the influence of climate change and natural gas prices on electricity transmission revenues using climate change forecasts and various natural gas price scenarios. They found that future natural gas prices will be the main factor influencing future transmission line revenues with a minor relative impact from climate changes. Furthermore, in [26], energy price projections for various fuels, including natural gas, were analyzed, considering climate change. They examined price trends for natural gas and other fuels in the context of the competitiveness of low-temperature geothermal resources. Additionally, in [7], the impact of future greenhouse gas emissions from natural gas production was assessed using a dynamic techno-economic model, focusing also on the effects of different methane emission metrics on the value of various projects and identifying resources that may become unprofitable under various carbon price scenarios. Interesting conclusions are also found in the earlier cited publication [24], where the relationships between winter temperature and natural gas consumption in households in central and eastern USA were examined. They analyzed gas consumption data at the state level, aggregating it for various subregions, which showed that winter temperatures in central and eastern United States are positively correlated with natural gas consumption in residential buildings, suggesting that residential demand dependent on temperature is more sensitive to market changes than previously thought.

There are also studies on short-term forecasts of natural gas prices and the impact of various factors, including temperature, on these prices, for example, in [15], a new deep learning model was proposed for short-term forecasting of natural gas prices using convolutional and LSTM layers to analyze natural gas data and learn short- and long-term dependencies. In [10], the impact of mild winter temperatures in the first months of 2020 on the decline in natural gas demand for heating, which contributed to lower natural gas prices, was analyzed. In [17], a structural vector autoregression (VAR) and Markov models were used to examine the short-term impact of temperature deviations on natural gas prices in the USA, showing that the US natural gas market is sensitive to temperature changes in the short term.

As the above studies show, natural gas prices are related to temperatures, but it is worth considering whether knowledge of current temperatures can be a more or less significant parameter influencing investment decisions in the natural gas futures market in the USA.

Assuming the influence of temperature on gas prices, it is worth looking at "end-of-day trading" strategies, which involve making investment decisions (actions) for the next trading session based on data and information available at the end of the current day. "End-of-day trading" strategies can be effectively used as a market entry method when long-term strategies indicate the validity of purchasing shares. This combination of a short-term entry technique with long-term investment analysis is a good strategy

for investors who want to optimize the timing of their transactions, which can impact investment efficiency.

The effectiveness of "end-of-day trading" strategies has been demonstrated by numerous studies, for example, in [9], it was noted that the last 5 min of the trading day account for a significant part of the variability of daily returns, which is consistent with the hypothesis that institutional interest affects the common components of stock returns at the end of the day. In the study [6], it was found that higher day trading activity leads to greater return volatility and that their short-term strategies contribute to positive future returns. In [16], it was discovered that returns and the number of traded shares have a U-shaped pattern during the trading day, confirming that high end-of-day returns are partly due to an increase in the proportion of transactions at the sale price. The authors of the study [5] presented a stock trading system that uses multi-objective particle swarm optimization (MOPSO) of technical indicators based on end-of-day market data, demonstrating that this system achieves better results than traditional methods. In [13], it was found that there are clear advantages to buying shares at the close of the day and selling them at the opening of the next trading day, especially when it is the next calendar day with the highest ratio of average daily return to standard deviation. Meanwhile, the study [12] shows that securities that achieve high relative returns in a given half-hour of a given day exhibit similar outperformance in the same half-hour on subsequent days, suggesting that strategically changing transaction times can significantly reduce costs. These studies show that the "end-of-day trading" strategy can bring benefits such as better use of institutional information, higher transaction efficiency, and strategic use of short-term return patterns.

The "end-of-day trading" strategy has been successfully used with machine learning, for example, in [25], it was shown how a machine learning-based statistical arbitrage strategy was successfully transferred from equity markets to futures markets using data from recent months to predict returns. In [2], a hybrid model combining technical analysis indicators with machine learning techniques for generating trading signals was presented, improving trading effectiveness on major indices. The study [14] applied various machine learning algorithms to forecast stock prices based on technical indicators, generating buy and sell signals. Deep neural networks were used in [28] to predict the direction of daily stock market returns based on financial and economic data. In [18], a decision model for daily investments in the stock market was proposed, using a combination of a machine learning-based classifier and a mean-variance portfolio selection method. The study [21] demonstrated the predictability and profitability of machine learning-based trading strategies in cryptocurrency markets using techniques such as linear models, random forests, and support vector machines. The article [23] includes trading strategies on the SSE-50 index, using machine learning to predict intraday price movements based on opening data. These studies show that "end-of-day trading" strategies using machine learning can be effective in predicting price movements and generating profits in various market conditions.

3 Methodology of the Research Experiment

The research procedure involves the following steps:

1. Data Acquisition: Retrieve Gas Futures data from the investing.com platform.
2. Calculation of Basic Indicators: Compute the simple moving averages (SMA) for 5, 20, 40, and 200 days for closing prices, as well as the 5-day SMA for volume (SMA_Vol_5).
3. Determination of Dependent Variable: Define the dependent variable as the action to be taken based on data from the following 5 days:
   a. Buy: When the price increases by more than 6% from the closing price.
   b. Sell: When the price decreases by more than 6% from the closing price.
   c. Hold: When neither of the above conditions is met.
4. Weather Data Acquisition: Collect daily weather data (minimum and maximum temperatures) from 8 weather stations located in the largest cities in the USA.
5. Data Integration: Merge the financial data and the weather data to form the training dataset.
6. Classifier Evaluation: Evaluate the classifiers for two versions of the study trained using various machine learning algorithms (decision tree, Random Forest, SVM, Gradient Boosting, k-NN, Neural Network):
   A. Study without Temperature Data
   B. Study with Temperature Data: Include two additional attributes: day of the year and day of the week.
7. Performance Metrics Calculation: Compute the evaluation metrics for both versions of the study for comparison purposes: Accuracy, Confusion Matrix, and Classification Report (including precision, recall, and f1-score).

The selection of machine learning algorithms for this study was guided by their established effectiveness in handling complex and non-linear relationships in financial data, particularly relevant for forecasting natural gas futures prices. Each algorithm's inclusion is justified by its unique strengths:

Decision Trees are renowned for their simplicity and interpretability, decision trees offer a clear visual representation of the decision-making process. They can capture non-linear relationships and interactions between features without requiring data normalization. Random Forests are known as an ensemble method based on decision trees, random forests enhance predictive accuracy and robustness by aggregating the results of multiple trees. This aggregation reduces the risk of overfitting and improves generalization. Support Vector Machines (SVM) are effective in high-dimensional spaces and are particularly suitable for classification tasks with clear margins of separation. They are resilient to overfitting, especially in scenarios where the number of dimensions exceeds the number of samples. Gradient Boosting is powerful ensemble technique builds models sequentially, with each new model correcting errors made by the previous ones. Gradient boosting is adept at handling complex data patterns and often achieves higher predictive accuracy than individual models. k-Nearest Neighbors (k-NN) is a simple, instance-based learning algorithm that makes predictions based on the closest training examples in the feature space. It is intuitive and effective for small to medium-sized datasets. With their ability to learn and model non-linear and complex relationships,

neural networks are highly versatile. They can capture intricate patterns in data through multiple layers of processing units.

These algorithms were selected to provide a comprehensive evaluation of different machine learning approaches, ensuring that the study leverages a diverse set of techniques to identify the most effective model for forecasting natural gas futures prices.

4 Results

The following are the results of the study aimed at determining whether the inclusion of temperature data from the most populous cities in the USA in machine learning models improves the predictive ability of these models. The evaluation methods used include: Accuracy, Classification Report (precision, recall, f1-score), and for Gradient Boosting, an additional evaluation method called 'importance'.

For each method in Study A, the classifier was trained using data on natural gas futures prices and basic technical analysis indicators, whereas in Study B, data on the minimum and maximum temperatures of the day from eight weather stations located in the largest US cities were additionally included.

The results for each machine learning method are summarized in Table 1.

Table 1. A set of data on the basis of which the analysis was carried out

Classifier	Study	Accuracy	Precision	Recall	F1-Score
Decision Tree	A	49.4%	50.0%	45.0%	45.0%
	B	52.6%	51.0%	53.0%	51.0%
Random Forest	A	50.9%	54.0%	51.0%	45.0%
	B	53.4%	54.0%	53.0%	49.0%
SVM	A	49.3%	50.0%	49.0%	43.0%
	B	53.2%	52.0%	53.0%	51.0%
Gradient Boosting	A	55.6%	56.0%	56.0%	53.0%
	B	58.1%	57.0%	58.0%	57.0%
k-NN	A	46.8%	46.0%	47.0%	46.0%
	B	54.0%	53.0%	54.0%	53.0%
Neural Network	A	49.3%	48.0%	49.0%	48.0%
	B	52.4%	51.0%	52.0%	52.0%

Additionally, using Gradient Boosting, the evaluated attributes were ranked according to their importance:

```
            importance
SMA_200       0.166728
day_of_year   0.111566
```

```
SMA_40 0.097743
High 0.079219
SMA_20 0.056694
USW00023183_TMIN 0.046271
SMA_Vol_5 0.044201
Low 0.042345
Change % 0.035783
USC00285728_TMAX 0.028167
Vol. 0.026458
USW00023183_TMAX 0.025339
USC00285728_TMIN 0.022827
USW00012918_TMAX 0.021427
USW00094728_TMAX 0.018707
SMA_5 0.018216
Close 0.017070
USC00410902_TMIN 0.016668
USW00023188_TMAX 0.015936
USW00093134_TMIN 0.015803
USW00023188_TMIN 0.014437
USW00012918_TMIN 0.013306
USC00111577_TMAX 0.013272
USW00094728_TMIN 0.012591
USC00111577_TMIN 0.011149
USC00410902_TMAX 0.009755
Open 0.008814
USW00093134_TMAX 0.008306
day_of_week 0.001203
```

5 Conclusion

The conducted study aimed to assess whether including current temperature data from the largest cities in the USA in predictive models based on various machine learning algorithms improves the quality of forecasts for natural gas futures prices. The obtained results indicate that adding temperature data to the models significantly enhanced prediction accuracy in most of the applied algorithms. All tested classifiers, including decision trees, random forests, support vector machines, Gradient Boosting, k-NN, and neural networks, showed higher accuracy in the version with temperature data (Study B) compared to the version without these data (Study A).

Among the tested algorithms, Gradient Boosting exhibited the highest accuracy in both Study A (0.556) and Study B (0.581), proving to be the most effective in considering complex relationships between variables and responding best to the inclusion of temperature data. The variable importance analysis in the Gradient Boosting models indicated that minimum and maximum temperature data from individual cities were significant predictors of natural gas prices. The variable "day_of_year" was particularly significant, suggesting a seasonal nature of natural gas demand.

Incorporating temperature data into predictive models can be beneficial for "end-of-day trading" investment strategies. Models that included weather data demonstrated a

better ability to predict price movements, which could lead to more accurate investment decisions. Despite the positive results, the study has its limitations. The temperature data were sourced only from the eight largest cities in the USA, which may not reflect the full impact of temperature on gas prices across the country. Additionally, the predictive models were tested on historical data, which does not guarantee their effectiveness in future market conditions.

Further research should include a broader range of weather data from more locations and consider other factors affecting natural gas prices, such as political or economic changes. Testing models on data from different periods and under various market conditions could provide more comprehensive results. While the study demonstrates the benefits of incorporating temperature data, it also highlights the need for broader weather data coverage and consideration of additional factors such as political or economic changes. Investors should remain aware of these limitations and continuously update and validate their models with new data under varying market conditions.

By leveraging these findings, investors and analysts can develop more robust and effective investment strategies, ultimately enhancing their ability to forecast natural gas futures prices and achieve better returns on their investments.

In summary, the study demonstrated that incorporating current temperatures in predictive models based on machine learning algorithms can significantly improve the quality of forecasts for natural gas futures prices. These findings may be useful for investors and market analysts in developing more effective investment strategies.

References

1. Ancona, M., et al.: Parametric thermo-economic analysis of a power-to-gas energy system with renewable input. High Temperat. Co-Electroly. Methanat. Energ. (2022). https://doi.org/10.3390/en15051791
2. Ayala, J., García-Torres, M., Noguera, J., Gómez-Vela, F., Divina, F.: Technical analysis strategy optimization using a machine learning approach in stock market indices. Knowl. Based Syst. **225**, 107119 (2021). https://doi.org/10.1016/J.KNOSYS.2021.107119
3. Baviera, R., Mainetti, T.: Going hybrid: A joint model for temperature and natural gas. Environ. Econom. eJ. (2015). https://doi.org/10.2139/ssrn.2585927
4. Baviera, R., Mainetti, T.: A joint model for temperature and natural gas with an application to the US market. Quantitat. Fin. **17**, 927–941 (2017). https://doi.org/10.1080/14697688.2016.1247981
5. Briza, A., Naval, P.: Stock trading system based on the multi-objective particle swarm optimization of technical indicators on end-of-day market data. Appl. Soft Comput. **11**, 1191–1201 (2011). https://doi.org/10.1016/j.asoc.2010.02.017
6. Chung, J., Choe, H., Kho, B.: The impact of day-trading on volatility and liquidity. ERN: Econometric Modeling in Financial Economics (Topic) (2008) https://doi.org/10.2139/ssrn.1855759
7. Crow, D., Balcombe, P., Brandon, N., Hawkes, A.: Assessing the impact of future greenhouse gas emissions from natural gas production. Sci. Tot. Environ. **668**, 1242–1258 (2019). https://doi.org/10.1016/J.SCITOTENV.2019.03.048
8. Cucu, L., Döttling, R., Heider, P., Maina, S.: Managing temperature-driven volume risks. J. Ener. Mark. (2016). https://doi.org/10.21314/jem.2016.145
9. Cushing, D., Madhavan, A.: Stock returns and trading at the close. J. Fin. Mark. **3**, 45–67 (2000). https://doi.org/10.1016/S1386-4181(99)00012-9

10. Ekwunife, I.: Technology focus: Natural gas processing and handling (April 2021). J. Petrol. Technol. **73**, 34 (2021). https://doi.org/10.2118/0421-0034-JPT
11. Freitas, C., Silva, P.: Evaluation of dynamic pass-through of carbon prices into electricity prices - a cointegrated VECM analysis. Int. J. Pub. Poli. **9**, 65–85 (2013). https://doi.org/10.1504/IJPP.2013.053440
12. Heston, S., Korajczyk, R., Sadka, R., Thorson, L.: Are you trading predictably? Financ. Anal. J. (2011). https://doi.org/10.2469/faj.v67.n2.6
13. Jacobsen, B.: Stock price patterns. Appl. Fin. Econ. Lett. **3**, 301–306 (2007). https://doi.org/10.1080/17446540701222375
14. Khandelwal, S., et al.: Machine learning-based probabilistic profitable model in algorithmic trading. J. Electron. Imaging **32**, 013039 (2023). https://doi.org/10.1117/1.JEI.32.1.013039
15. Livieris, I., Pintelas, E., Kiriakidou, N., Stavroyiannis, S.: An advanced deep learning model for short-term forecasting U.S. natural gas price and movement. Artif. Intell. Appl. Innov. AIAI 2020 IFIP WG 12.5 International Workshops, **585**, 165–176 (2020) https://doi.org/10.1007/978-3-030-49190-1_15
16. Mcinish, T., Wood, R.: An analysis of transactions data for the Toronto stock exchange: Return patterns and end-of-the-day effect. J. Bank. Finance **14**, 441–458 (1990). https://doi.org/10.1016/0378-4266(90)90058-A
17. Nkwantabisa, D.: Determinants of natural gas prices in the United States – A structural VAR approach. Develop. Econ. Agricult. (2021).https://doi.org/10.2139/ssrn.3905560
18. Paiva, F., Cardoso, R., Hanaoka, G., Duarte, W.: Decision-making for financial trading: A fusion approach of machine learning and portfolio selection. Expert Syst. Appl. **115**, 635–655 (2019). https://doi.org/10.1016/j.eswa.2018.08.003
19. Pineau, P., Dupuis, D., Cenesizoglu, T.: Assessing the value of power interconnections under climate and natural gas price risks. Energy **82**, 128–137 (2015). https://doi.org/10.1016/J.ENERGY.2014.12.078
20. Saeed, T., Tularam, G.: Relations between fossil fuel returns and climate change variables using canonical correlation analysis. Ener. Sour. Part B **12**, 675–684 (2017). https://doi.org/10.1080/15567249.2016.1265615
21. Sebastião, H., Godinho, P.: Forecasting and trading cryptocurrencies with machine learning under changing market conditions. Fin. Innov. **7** (2021). https://doi.org/10.1186/s40854-020-00217-x
22. Singh, D., Kumar, P.: Statistical modeling of the natural gas prices in relation to demand, supply and economic growth indicators. J. Stat. Manag. Syst. **23**, 713–735 (2020). https://doi.org/10.1080/09720510.2019.1662628
23. Tang, P., Tang, X., Yu, W.: Intraday trend prediction of stock indices with machine learning approaches. Eng. Econ. **68**, 60–81 (2023). https://doi.org/10.1080/0013791X.2023.2205841
24. Timmer, R., Lamb, P.: Relations between temperature and residential natural gas consumption in the Central and Eastern United States. J. Appl. Meteorol. Climatol. **46**, 1993–2013 (2007). https://doi.org/10.1175/2007JAMC1552.1
25. Waldow, F., Schnaubelt, M., Krauss, C., Fischer, T.: Machine Learning in Futures Markets. **14**, 119 (2021).https://doi.org/10.3390/JRFM14030119
26. Weissbrod, R., Barron, W.: Geothermal energy market study on the Atlantic coastal plain. A Review of Recent Energy Price Projections for Traditional Space Heating Fuel 1985–2000 (1979). https://doi.org/10.2172/894654
27. Wolf, M., Pröll, T.: A comparative study of sustainable industrial heat supply based on economic and thermodynamic factors. Die Bodenkultur J. Land Managem. Food Environ. **68**, 145–156 (2018). https://doi.org/10.1515/boku-2017-0013
28. Zhong, X., Enke, D.: Predicting the daily return direction of the stock market using hybrid machine learning algorithms. Fin. Innov. **5** (2019) https://doi.org/10.1186/s40854-019-0138-0

The Role of Data Normalization Algorithm in Measuring of EU Countries Digitalization Degree

Anna Borawska(✉), Mariusz Borawski, and Małgorzata Łatuszyńska

University of Szczecin, Cukrowa 8, 71-004 Szczecin, Poland
anna.borawska@usz.edu.pl

Abstract. The article addresses the issue of measuring the degree of digitalization in European Union countries using the Digital Economy and Society Index (DESI). The aim is to analyze the impact of selected normalization algorithms on the DESI ranking and to identify the algorithm that is least sensitive to unusual indicator values. An analysis of eleven selected normalization algorithms was conducted using computer programs in Python, specifically developed for this study. The findings indicate that the linear normalization sum-based algorithm is the most effective, providing the most stable ranking where changes in the DESI index of some countries are minimally affected by changes in the indicator values of other countries in the ranking. The authors' main contribution is the analysis of data normalization algorithms and the identification of the most suitable one for DESI.

Keywords: Data normalization · Digitalization · DESI

1 Introduction

A new type of economy, the digital economy, is emerging before our eyes. It is driven by the intensification of digitization processes over the past two decades, including the application of digital technologies by companies, public institutions, NGOs, employees, consumers, and citizens [1–4].

Digital technologies are critical to achieving economic and sustainability goals [5–8], and they contribute to economic growth [9]. They create new opportunities in areas such as knowledge sharing, communication, management, information transfer, and employee coordination [5]. Additionally, digitization drives economic and social innovation [7, 10] and is beneficial in giving countries a competitive advantage [8, 11]. Consequently, the EU's efforts to develop and disseminate digital technologies play a crucial role [3]. In this context, action programs like the 2030 Digital Compass [12] and the accompanying Digital Decade Policy Programme [13] are highly relevant to Europe's future. These programs contain measurable targets and mechanisms for monitoring the progress of digitization in individual EU countries [3, 14].

The primary tool used by the European Commission to assess the digitization levels in EU countries is the Digital Economy and Society Index (DESI). Published annually

M. Hernes et al. (Eds.): ECAI 2024, LNNS 1217, pp. 140–151, 2025.
https://doi.org/10.1007/978-3-031-78465-1_12

since 2015, it considers indicators (KPIs -Key Performance Indicators) measuring key dimensions of digitization: infrastructure, digital skills, use of technology by businesses and citizens, and the level of digitization in public administration [15]. Decisions on the implementation of digitization policies at various levels are based on DESI index values and its indicators.

DESI's annual reports include country profiles and comparisons, enabling member states to identify strengths and weaknesses. The rankings based on this index guide funding for the digital decade goals using EU funds [3, 14]. Therefore, a country's position in this ranking is important. However, it may depend not only on actual performance but also on the methodology used to calculate the DESI index, including the normalization algorithm applied to the data.

Literature indicates that different normalization approaches can lead to varying rankings of alternatives [16–19], with some methods being more suitable for specific decision-making scenarios [20]. Selecting the most appropriate normalization algorithm is crucial, as it impacts the final ranking list [21] and affects the relative importance of indicators [22]. Therefore, choosing the right normalization algorithm is essential to ensure accurate and reliable ranking results. Key questions in this context include: (1) How significant is the impact of the normalization algorithm on the DESI ranking? (2) Which normalization algorithm provides the most reliable representation of DESI results?

This article aims to analyze the impact of selected normalization algorithms on the DESI ranking and to identify the algorithm that is least sensitive to abnormal indicator values. The authors' main contribution is the analysis of data normalization algorithms and the identification of the most suitable one for DESI.

2 Materials and Methods

The survey was conducted based on the DESI 2022 methodology described in [23]. DESI 2022 was chosen despite the availability of DESI 2023 data and rankings, due to the lack of a detailed description of the DESI 2023 methodology. The DESI 2023 methodological note [24] does not indicate the data normalization algorithm or the weights used to build the index. It can be presumed they are the same as DESI 2022, but this information is not included in the document. Additionally, https://digital-decade-desi.digital-strategy.ec.europa.eu/datasets/desi/indicators lists 34 indicators, whereas [24] lists only 32. Therefore, the survey used the indicators specified in the DESI 2022 methodology [23].

Data on indicators for individual EU countries were downloaded from https://digital-decade-desi.digital-strategy.ec.europa.eu/datasets/desi-2022/indicators. Due to incomplete data for all indicators, countries such as Finland (FI), Luxembourg (LU), Croatia (HR), Malta (MT), and Cyprus (MT) were not included in the study. According to the DESI 2022 methodology [25], missing data in the EU country rankings presented at https://digital-strategy.ec.europa.eu/en/policies/desi were filled using data from other years and approximations based on available data from other years. Since the DESI 2022 methodology does not specify the approximation methods used, the previously mentioned countries were removed from this survey.

For DESI 2022, the linear max-min algorithm was used to normalize the data. This algorithm is sensitive to unusual indicator values, potentially leading to errors in the ranking results. To select alternative normalization algorithms for the study, the literature on normalization was reviewed.

Normalization is a scaling process used to ensure data comparability by eliminating the optimization orientation (benefit or cost), unit of measurement, and range of variability [20]. It involves preliminary processing to convert heterogeneous input data into dimensionless data [16]. Normalization algorithms can be divided into two basic classes based on whether they consider the optimization orientation of the indicators used to calculate the ranking value [20]. Indicators can be either cost-oriented (destimulant), where lower values are preferable, or benefit-oriented (stimulant), where higher values are preferable. Algorithms that use different formulas for benefit and cost type indicators are listed in Table 1. The alternatives mentioned in the table refer to countries.

Table 1. Normalization techniques depending on the optimization orientation

Normalization method	Benefit type indicator	Cost type indicator
Vector normalization [25]	$n_{ij} = \frac{r_{ij}}{\sqrt{\sum_{i=1}^{m} r_{ij}^2}}$	$n_{ij} = 1 - \frac{r_{ij}}{\sqrt{\sum_{i=1}^{m} r_{ij}^2}}$
Linear normalization sum-based [26]	$n_{ij} = \frac{r_{ij}}{\sum_{i=1}^{m} r_{ij}}$	$n_{ij} = \frac{1/r_{ij}}{\sum_{i=1}^{m} 1/r_{ij}}$
Logarithmic normalization [27]	$n_{ij} = \frac{\ln(r_{ij})}{\ln(\prod_{i=1}^{m} r_{ij})}$	$n_{ij} = \frac{1 - \frac{\ln(r_{ij})}{\ln(\prod_{i=1}^{m} r_{ij})}}{m-1}$
Enhanced accuracy normalization [28]	$n_{ij} = 1 - \frac{r_j^{max} - r_{ij}}{\sum_{i=1}^{m}(r_j^{max} - r_{ij})}$	$n_{ij} = 1 - \frac{r_{ij} - r_j^{min}}{\sum_{i=1}^{m}(r_{ij} - r_j^{min})}$
Maximum – linear normalization [29]	$n_{ij} = \frac{r_{ij}}{r_j^{max}}$	$n_{ij} = 1 - \frac{r_{ij}}{r_j^{max}}$
Minimum – linear normalization [20]	$n_{ij} = 1 - \frac{r_j^{min}}{r_{ij}}$	$n_{ij} = \frac{r_j^{min}}{r_{ij}}$
Non-linear normalization [27]	$n_{ij} = \left(\frac{r_{ij}}{r_j^{max}}\right)^2$	$n_{ij} = \left(\frac{r_j^{min}}{r_{ij}}\right)^3$
0–1 interval normalization using max-min [30]	$n_{ij} = \frac{r_{ij}}{r_j^{max}}$	$n_{ij} = \frac{r_j^{min}}{r_{ij}}$

(*continued*)

Table 1. (*continued*)

Normalization method	Benefit type indicator	Cost type indicator
Jüttler's and Körth's normalization [31]	$n_{ij} = 1 - \left\lvert \frac{r_j^{max} - r_{ij}}{r_j^{max}} \right\rvert$	$n_{ij} = 1 - \left\lvert \frac{r_j^{min} - r_{ij}}{r_j^{min}} \right\rvert$
Stopp normalization [30]	$n_{ij} = \frac{100 r_{ij}}{r_j^{max}}$	$n_{ij} = \frac{100 r_j^{min}}{r_{ij}}$
Lai and Hwang normalization [32]	$n_{ij} = \frac{r_{ij}}{r_j^{max} - r_j^{min}}$	$n_{ij} = \frac{r_{ij}}{r_j^{min} - r_j^{max}}$
Linear max-min normalization [33]	$n_{ij} = \frac{r_{ij} - r_j^{min}}{r_j^{max} - r_j^{min}}$	$n_{ij} = \frac{r_j^{max} - r_{ij}}{r_j^{max} - r_j^{min}}$
Z-transformation [34]	$\mu_j = \frac{\sum_{i=1}^{m} r_{ij}}{m} \sigma_j = \sqrt{\frac{\sum_{i=1}^{m} (r_{ij} - \mu_j)^2}{m}}$	
	$n_{ij} = \frac{r_{ij} - \mu_j}{\sigma_j}$	$n_{ij} = -\frac{r_{ij} - \mu_j}{\sigma_j}$
n_{ij}– normalized value of alternative i in indicator j	r_j^{min}– minimum value of indicator j	
r_{ij}– original value of alternative i in indicator j	μ_j– arithmetic mean of indicator j	
m – number of alternatives	σ_j– standard deviation of indicator j	
r_j^{max}– maximum value of indicator j		

Source: Own elaboration based on [20, 31, 34]

In contrast, algorithms that apply the same formula regardless of indicator type are shown in Table 2.

The study excluded normalization algorithms that required a reference value, as such a value does not exist in the DESI methodology. For further calculations, we selected the following methods:

- N1 – DESI (linear max-min),
- N2 – Z-transformation,
- N3 – Vector normalization,
- N4 – Linear normalization sum based,
- N5 – Enhanced accuracy normalization,
- N6 – Maximum linear normalization,
- N7 – Non-linear normalization,
- N8 – Jüttler's and Körth's normalization,
- N9 – Lai and Hwang normalization,
- N10 – Normalization equalizing the average to 1,
- N11 – Normalization equalizing the standard deviation to 1.

The following methods listed in Tables 1 and 2 were omitted:

- Minimum-linear normalization, as normalization could not be performed for indicator values equal to zero (division by zero);

Table 2. Normalization techniques independent of the optimization orientation

Normalization method	All types indicators
Normalization equalizing the average to 1 [20]	$n_{ij} = \frac{r_{ij}}{\mu_j}$
Normalization equalizing the standard deviation to 1 [20]	$n_{ij} = \frac{r_{ij}}{\sigma_j}$
Range normalization between -1 and + 1 [20]	$n_{ij} = \frac{r_{ij} - \left(\frac{r_j^{max} + r_j^{min}}{2}\right)}{\frac{r_j^{max} - r_j^{min}}{2}}$
Range normalization between 0 and + 1 [20]	$n_{ij} = \frac{r_{ij} - r_j^{min}}{r_j^{max} - r_j^{min}}$
n_{ij}– normalized value of alternative i in indicator j	r_j^{min}– minimum value of indicator j
r_{ij}– original value of alternative i in indicator j	μ_j– arithmetic mean of indicator j
m – number of alternatives	σ_j– standard deviation of indicator j
r_j^{max}– maximum value of indicator j	

Source: Own elaboration based on [20, 31, 34]

- 0–1 interval normalization using max-min and Stopp normalization, as the formula for stimulants is identical (or represents a rescaling) to maximum-linear normalization. In the study, only one indicator was a destimulant, making differences difficult to observe.

The next steps of the study are presented in the "Results and Discussion" section to create a coherent narrative. The study was conducted using specially written Python programs.

3 Results and Discussion

The study identified the positions of selected countries in the DESI ranking based on the normalization algorithm used. The ranking was created based on the DESI procedure [23]. Initially, normalization was performed. Then, for each object, a weighted average was calculated with weights corresponding to the values proposed in the DESI procedure. The weighted average result determines the position of objects in the ranking. For the basic DESI procedure, the linear max-min normalization was used, which was replaced for the study with the methods listed in Sect. 2. The results are shown in Table 3.

It should be noted that the ranking obtained in the survey for DESI 2022 (column N1) differs from the official ranking presented at https://digital-strategy.ec.europa.eu/en/policies/desi - especially in terms of the countries in the middle and near the end of the ranking. This may be due to two reasons. This may be due to two reasons. First, the DESI 2022 methodology [23] does not clearly indicate what values of weights should be adopted for each indicator. In this study, the weights of groups and subgroups of indicators were adopted according to the [23, p. 12]. For specific indicators, [23]

indicates that most have the same weights, but it does not specify which indicators have deviations. Therefore, this article assumes that all indicator weights are the same, which could explain the different results obtained. The second reason could be the removal of several countries from the ranking due to incomplete data, which may have altered the order of other countries. However, it should be emphasized that the differences in the ranking order do not affect the overall results obtained in the article, as the properties of DESI calculation were studied, which remain consistent regardless of the data used.

Table 3. DESI ranking with different normalization algorithms

Code	Normalization algorithm										
	N1	N2	N3	N4	N5	N6	N7	N8	N9	N10	N11
DK	1	1	1	1	1	1	1	1	1	1	1
NL	2	2	2	3	2	2	2	2	2	3	2
SE	3	3	3	2	3	3	3	3	3	2	3
IE	4	5	4	4	4	4	4	4	4	4	5
ES	5	4	5	6	5	5	5	5	5	6	4
EE	6	7	9	9	6	6	6	6	6	9	7
AT	7	8	7	8	7	7	8	7	7	8	8
SI	8	6	8	7	8	9	11	9	8	7	6
BE	9	15	14	12	12	10	7	10	9	13	15
FR	10	10	6	5	9	8	9	8	10	5	10
PT	11	9	11	11	10	11	10	11	11	11	9
EU	12	11	10	10	14	12	15	12	12	10	11
LT	13	12	15	15	11	14	12	14	13	15	12
DE	14	14	12	13	15	13	13	13	14	12	14
LV	15	13	16	18	13	15	14	15	15	18	13
CZ	16	17	17	16	16	16	16	16	16	16	17
IT	17	16	13	14	17	17	17	17	17	14	16
HU	18	18	18	17	18	18	18	18	18	17	18
SK	19	19	20	20	19	19	20	20	19	20	19
EL	20	20	19	19	20	20	19	19	20	19	20
PL	21	21	21	21	21	21	21	21	21	21	21
BG	22	22	22	22	22	22	22	22	22	22	22
RO	23	23	23	23	23	23	23	23	23	23	23

Table 3 shows that the best and worst positions are occupied by the same countries, regardless of the normalization algorithm used. The closer to the middle, the greater the diversity in ranking positions. This observation is confirmed by the standard deviations and ranges of variation in ranking positions shown in Table 4 and Fig. 1.

Table 4. Ranking statistics

Code	Stand. Deviation	Interval		Code	Stand. Deviation	Interval
DK	0.00	0		*cont*		
NL	0.40	1		LT	1.42	4
SE	0.40	1		DE	0.92	3
IE	0.40	1		LV	1.79	5
ES	0.63	2		CZ	0.47	1
EE	1.34	3		IT	1.51	4
AT	0.52	1		HU	0.40	1
SI	1.45	5		SK	0.52	1
BE	2.66	8		EL	0.52	1
FR	1.99	5		PL	0.00	0
PT	0.82	2		BG	0.00	0
EU	1.62	5		RO	0.00	0

Figure 1 shows that Belgium (BE) has the highest variability in ranking position, ranking 9th out of 23 possible positions according to DESI. The volatility decreases as one moves away from the central position, but it is subject to cyclical fluctuations of roughly similar periods.

The normalization algorithms tested were grouped using the k-means method to determine which produced similar results. The number of classes, determined using the Elbow method [35] (Fig. 2), was five.

Classification showed that the maximum linear normalization (N6), Jüttler's and Körth's normalization (N8), and Lai and Hwang normalization (N9) algorithms produced results closest to DESI (N1). They were found in Class 0. Class 1 included algorithms that normalize data only by dividing by a value determined from the data. Class 2 included all algorithms where the divisor is the standard deviation. Class 3 included the only nonlinear algorithm in the list. Class 4 included algorithms where the divisor is the maximum value of the data set (Table 5).

A good normalization algorithm ensures the stability of ranking results. Changing the index value for any country should not affect the differences in index values between other countries. For example, if one of the indicators for Denmark (DK) was incorrectly measured, affecting its index value, correcting this value should not change the difference in index values for Sweden and Norway.

The index values for each country were treated as a vector pattern. Denmark (DK) was selected because it ranked first. For Denmark, the value of the "1a2 Above basic

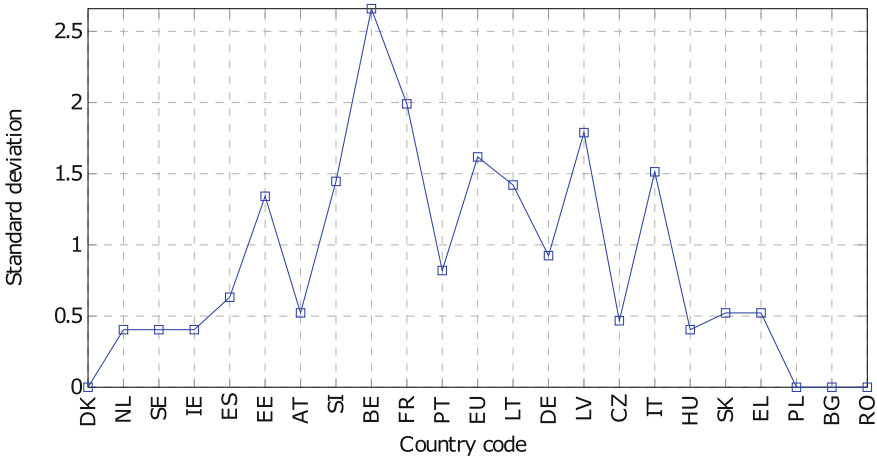


Fig. 1. Standard deviations of positions in rankings

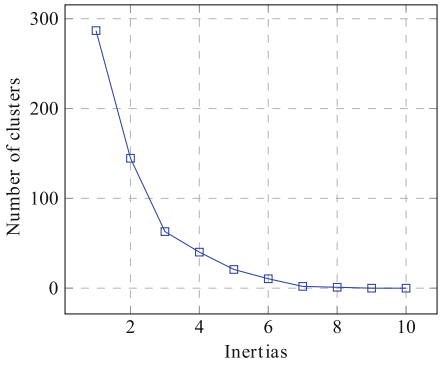


Fig. 2. Elbow method

digital skills" index was changed from −50 to +50 of the original value in increments of 1. For each increment, the DESI index values for each country were calculated, creating a vector. Each vector was then compared with the benchmark vector. In this procedure, Denmark's index values were removed from the vectors, and their average values were subtracted. The distances between the resulting points were calculated, and this value should be zero. Alternatively, the index values for all countries (except Denmark) could change by the same amount, which would not affect the interpretation of the result.

Figure 3a shows the calculated distance changes for the zero-class normalization algorithms. The graphs for DESI, Lai and Hwang normalization, maximum linear normalization, and Jüttler's and Körth's normalization overlap. Changes in the index values within a certain range do not affect the index values for other countries. However, outside this range, the graphs rise sharply, resulting in increasingly significant changes in index values. Maximum linear normalization and Jüttler's and Körth's normalization methods

Table 5. Classification of normalization agorithms

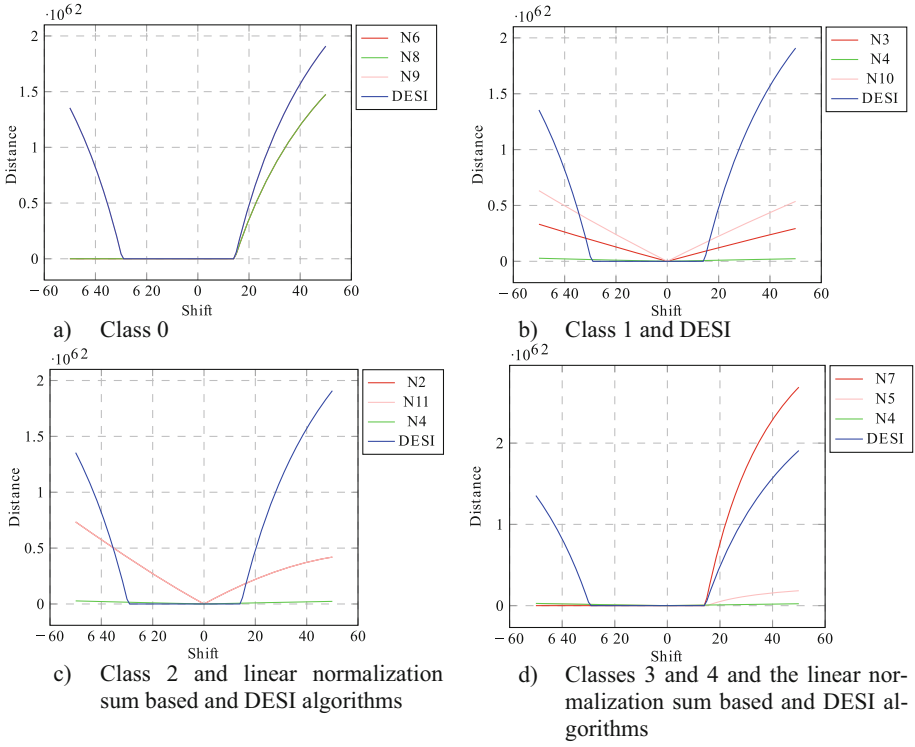
Normalization algorithm		Class
N1	DESI	0
N6	Maximum linear normalization	0
N8	Juttlers and Korths Normalization	0
N9	Lai and Hwang Normalization	0
N3	Vector normalization	1
N4	Linear normalization sum based	1
N10	Normalization equalizing the average to 1	1
N2	Z transformation	2
N11	Normalization equalizing the standard deviation to 1	2
N7	Non-linear normalization	3
N5	Enhanced accuracy normalization	4

perform more favorably than DESI here, as the increase in distance values is only on the right part of the graph and is less steep.

Figure 3b shows the calculated distance changes for the normalization algorithms from class 1 and DESI The impact of changing the index value for one country here does not significantly affect the index values for other countries. However, the impact appears with even the slightest change in the index value. By far, the most favorable is the linear normalization sum-based algorithm, for which the distances change the least.

Figure 3c shows the calculated distance changes for the class 2 and linear normalization sum-based algorithms and DESI. The waveforms for all the algorithms from class 2 overlap. They have a similar character to the waveforms from class 1, with the right side of the graph being non-linear. The linear normalization sum-based algorithm performs best in this comparison.

Figure 3d shows the calculated distance changes for the normalization algorithms from class 3, class 4, linear normalization sum-based, and DESI. The non-linear normalization algorithm from class 3 is characterized by a very rapid increase in distance values on the right side of the graph, much larger than DESI. On the left side of the graph, the values obtained for it are zero. The enhanced accuracy normalization algorithm on the left side of the graph has values very close to the linear normalization sum-based algorithm. On the right, the graph is clearly non-linear, and the values are higher than those for the linear normalization sum-based algorithm. The enhanced accuracy normalization algorithm performs best right after the linear normalization sum-based algorithm.



a) Class 0

b) Class 1 and DESI

c) Class 2 and linear normalization sum based and DESI algorithms

d) Classes 3 and 4 and the linear normalization sum based and DESI algorithms

Fig. 3. Changes in distance between vectors representing DESI index values for classes

4 Conclusion

The normalization algorithm used in DESI produces results most similar to the Lai and Hwang normalization, maximum linear normalization, and Jüttler's and Körth's normalization algorithms. These algorithms are characterized by the lack of impact of a change in one country's index value on the differences in the index values of other countries within a certain range. Outside this range, the impact grows rapidly. The linear sum-based normalization algorithm performs best, as its influence is always present but very weak. The second-best algorithm, though slightly worse than the linear sum-based normalization, is enhanced accuracy normalization. The strength of the impact of a change in an indicator for one country on the index values of other countries depends on the weight of that indicator. The smaller the weight of the indicator, the weaker the impact. In the case of DESI, the large number of indicators means that individual indicators have small weights. Ultimately, the impact will be small, but it can be cumulative. A common issue is the lack of data for individual countries, which may come in late. Adding a country to the ranking after obtaining its missing data will cause changes in all indicators after normalization, potentially significantly impacting the index values for other countries.

The susceptibility of the entire ranking order to a change in the value of one country's indicator is one of the major problems of ranking methods. Another problem worth studying is the sensitivity of the ranking to data precision. Few data are known precisely. However, some data are subject to measurement errors, some are estimated with a certain precision, and some change over the course of the year, making the indicator values different at the end of the year compared to the beginning. Therefore, it would be worthwhile to check how sensitive the DESI ranking is to data fluctuations and to supplement the index with a measure indicating how much it can change due to imprecision and fluctuations in the indicators. This would make it possible to indicate which countries are better than others, and which are ranked equally due to similar DESI index values and the imprecision and fluctuations in its value.

Acknowledgments. Co-financed by the Minister of Science under the "Regional Excellence Initiative" Program.

References

1. Śledziewska, K., Śledziewska, K., Włoch, R.: Gospodarka cyfrowa: jak nowe technologie zmieniają świat. Wydawnictwa Uniwersytetu Warszawskiego, Warszawa (2020)
2. Graetz, G., Michaels, G.: Robots at work. Rev. Econ. Stat. **100**, 753–768 (2018)
3. European Commission. Joint Research Centre.: Mapping EU level funding instruments to Digital Decade targets: application to main digital instruments in 2014 2027. Publications Office, LU (2023)
4. Tratkowska, K.: Digital transformation: theoretical backgrounds of digital change. Manag. Sci. **24**, 32–37 (2020)
5. Di Vaio, A., Palladino, R., Hassan, R., Escobar, O.: Artificial intelligence and business models in the sustainable development goals perspective: a systematic literature review. J. Bus. Res. **121**, 283–314 (2020)
6. Nambisan, S., Wright, M., Feldman, M.: The digital transformation of innovation and entrepreneurship: progress, challenges and key themes. Res. Policy **48**, 103773 (2019)
7. Guandalini, I.: Sustainability through digital transformation: a systematic literature review for research guidance. J. Bus. Res. **148**, 456–471 (2022)
8. George, G., Merrill, R.K., Schillebeeckx, S.J.D.: Digital sustainability and entrepreneurship: how digital innovations are helping tackle climate change and sustainable development. Entrep. Theory Pract. **45**, 999–1027 (2021)
9. Alakeson, V., Wilsdon, J.: Digital sustainability in Europe. J. Ind. Ecol. **6**, 10–12 (2002)
10. Bocean, C.G., Vărzaru, A.A.: EU countries' digital transformation, economic performance, and sustainability analysis. Humanit. Soc. Sci. Commun. **10**, 875 (2023)
11. Laitsou, E., Kargas, A., Varoutas, D.: Digital competitiveness in the european union era: the Greek case. Economies. **8**, 85 (2020)
12. Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and The Committee of the Regions. 2030 Digital Compass: the European way for the Digital Decade (2021). https://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX%3A52021DC0118
13. Decision (EU) 2022/2481 of the European Parliament and of the Council of 14 December 2022 establishing the Digital Decade Policy Programme 2030 (2022). https://eur-lex.europa.eu/eli/dec/2022/2481/oj

14. Święcicki, I., Witczak, J.: Jak osiągnąć cele cyfrowej dekady w Polsce? Polish Economic Institute, Warsaw (2023)
15. Święcicki, I.: How to measure the Digital Decade – recommendations for an evolution of the DESI index. Polish Economic Institute, Warsaw (2022)
16. Vafaei, N., Ribeiro, R.A., Camarinha-Matos, L.M.: Selecting normalization techniques for the analytical hierarchy process. In: Camarinha-Matos, L.M., Farhadi, N., Lopes, F., and Pereira, H. (eds.) Technological innovation for life improvement, pp. 43–52. Springer International Publishing, Cham (2020). https://doi.org/10.1007/978-3-030-45124-0_4
17. Walesiak, M.: The choice of normalization method and rankings of the set of objects based on composite indicator values. Stat. Transit. New Ser. **19**, 693–710 (2018)
18. Leoneti, A.B.: Considerations regarding the choice of ranking multiple criteria decision making methods. Pesqui. Oper. **36**, 259–277 (2016)
19. Więckowski', J., Sałabun, W.: How the normalization of the decision matrix influences the results in the VIKOR method? Procedia Comput. Sci. **176**, 2222–2231 (2020)
20. Aytekin, A.: Comparative analysis of the normalization techniques in the context of MCDM problems. Decis. Mak. Appl. Manag. Eng. **4**, 1–25 (2021)
21. Mhlanga, S.T., Lall, M.: Influence of normalization techniques on multi-criteria decision-making methods. J. Phys. Conf. Ser. **2224**, 012076 (2022)
22. Jiang, R.: A novel normalization method for using in multiple criteria decision analysis. In: 2019 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), pp. 268–272. IEEE, Macao, Macao (2019)
23. Digital Economy and Society Index (DESI) 2022. Methodological Note. https://digital-strategy.ec.europa.eu/en/policies/desi
24. DESI 2023 methodological note (2023). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52023SC0574
25. Delft, A. van, Nijkamp, P.: Multi-criteria analysis and regional decision-making. Leiden : Martinus Nijhoff Social Sciences Division (1977)
26. Wang, Y.-M., Luo, Y.: Integration of correlations with standard deviations for determining attribute weights in multiple attribute decision making. Math. Comput. Model. **51**, 1–12 (2010)
27. Zavadskas, E.K., Turskis, Z.: A new logarithmic normalization method in games theory. Informatica **19**, 303–314 (2008)
28. Zeng, Q.-L., Li, D.-D., Yang, Y.-B.: VIKOR method with enhanced accuracy for multiple criteria decision making in healthcare management. J. Med. Syst. **37**, 9908 (2013)
29. Milani, A.S., Shanian, A., Madoliat, R., Nemes, J.A.: The effect of normalization norms in multiple attribute decision making models: a case study in gear material selection. Struct. Multidiscip. Optim. **29**, 312–318 (2005)
30. Brauers, W., Zavadskas, E.: The MOORA method and its application to privatization in a transition economy. Control. Cybern. **35**, 445–469 (2006)
31. Gardziejczyk, W., Zabicki, P.: Normalization and variant assessment methods in selection of road alignment variants – case study. J. Civ. Eng. Manag. **23**, 510–523 (2017)
32. Lai, Y.-J., Hwang, C.-L.: Fuzzy multiple objective decision making. In: Fuzzy multiple objective decision making, pp. 139–262. Springer Berlin Heidelberg, Berlin, Heidelberg (1994)
33. Tzeng, G.-H., Huang, J.-J.: Multiple attribute decision making. Chapman and Hall/CRC (2011)
34. Jahan, A., Edwards, K.L.: A state-of-the-art survey on the influence of normalization techniques in ranking: improving the materials selection process in engineering design. Mater. Des. **1980–2015**(65), 335–342 (2015)
35. Thorndike, R.L.: Who belongs in the family? Psychometrika **18**, 267–276 (1953)

Enhancing Examination Question Quality Through Data-Driven Analysis and Systematic Maintenance

Grzegorz Szyjewski(✉)

Institute of Management, University of Szczecin, Szczecin, Poland
grzegorz.szyjewski@usz.edu.pl

Abstract. This article describes an approach for maintaining a high-quality question test database, implemented in a real-world case study. The approach leverages answering history from past tests to identify and address problematic records, thus enhancing the overall quality of the database. Key criteria for selection include the percentage of incorrect answers, the proportion of choices for single distractors, and the time spent on each question. These criteria effectively highlight questions needing revision.

The method demonstrated simplicity and efficacy in managing a large dataset from past examinations, ensuring continuous database quality. Almost all records flagged for revision were modified by reviewers, leading to significant improvements in the percentage of correct answers in subsequent tests. This indicates that the primary issue was often the question construction rather than test-taker knowledge.

An automated module was developed using PHP and MySQL, integrated with frontend frameworks for enhanced usability. This module assists in identifying problematic questions, with final decisions made by reviewers. The study concludes that this tool is valuable for systematic improvements, maintaining the integrity and reliability of the assessment process.

Keywords: Automated Testing · Question Database · Exam Quality · Test Analysis · Answer History · Question Revision · Test Integrity · Response Patterns · Test Efficiency

1 Introduction

Tests have long been a fundamental component of educational assessment, providing a systematic approach to evaluating knowledge and skills. Beyond academic environments, tests are essential for professional certification and licensure, ensuring that individuals meet the requisite standards to perform their roles proficiently. Fundamentally, tests serve several critical functions in both educational and professional contexts. They offer a quantifiable measure of an individual's knowledge and comprehension of a particular subject or topic, functioning as a vital assessment tool. By identifying strengths and weaknesses, tests can pinpoint areas where a learner requires additional support or

M. Hernes et al. (Eds.): ECAI 2024, LNNS 1217, pp. 152–160, 2025.
https://doi.org/10.1007/978-3-031-78465-1_13

instruction. The prospect of being tested motivates learners to engage more deeply with the material and assume responsibility for their learning [1, 2]. In professional settings, tests verify that individuals possess the necessary knowledge and skills to perform their duties safely and effectively.

The format of a test significantly influences its efficacy and the type of information it yields. Multiple-choice questions (MCQs) present several answer choices, only one of which is correct, offering an efficient means to assess a broad range of content [3]. Free-response questions require brief written answers, allowing for more nuanced responses than MCQs. Administering tests to large groups necessitates automation, which is challenging for written-response tests. MCQs are more appropriate to automated testing systems, making them a popular assessment tool in both educational and professional contexts due to their versatility and ease of grading. With technological advancements, automating MCQ tests has become increasingly feasible, offering benefits such as efficiency, accuracy, and scalability. These tests provide a structured approach to assessing a wide range of knowledge and skills, particularly effective for evaluating recall, comprehension, and application of information.

Deploying tests on online platforms facilitates easy access and administration. These platforms can manage various administrative tasks such as scheduling, proctoring, and providing immediate feedback. One of the most significant advantages of test automation is the provision of instant grading. Automated systems can grade tests by comparing the selected answers to the correct answers stored in a database, eliminating the need for manual grading and significantly reducing the turnaround time for results. Any testing method requires a question test database, which is central to an automated testing system. Each question should include the stem (the question itself), the correct answer, and several distractors (incorrect options). Questions must be unambiguous, with one correct answer and reasonable distractors. It is crucial to avoid complex wording and ensure that all options are grammatically consistent with the question stem. Automated systems can generate tests by randomly selecting questions from the database, ensuring that each test is unique and reducing the chances of cheating or receiving the same test during a retake. Questions can be selected based on predefined criteria such as difficulty level or knowledge scope.

2 Question Test Base

Developing a comprehensive question bank encompassing all relevant topics and difficulty levels is crucial when implementing an examination system. The question database should be regularly updated to ensure it remains current and relevant. Maintaining high-quality questions within a large question database can be challenging. To maximize the effectiveness of automated testing, several ongoing actions are necessary. One key action is ensuring the quality of questions, which involves creating high-quality items that accurately measure the intended knowledge or skills. It is also important to avoid ambiguous or misleading questions within the database [4, 5]. Another critical element that enhances automated testing is the test question structure. Each test should include a range of difficulty levels, combining easy, moderate, and difficult questions to accurately evaluate the range of test-takers abilities. Additionally, individual test questions should

cover all testing areas. This means that the question database structure should categorize questions not only by difficulty level but also by knowledge area. Regular updates and improvements are essential for maintaining high question quality. This process requires questions to be reviewed and updated continually, which can be time-consuming, especially for large question databases [6]. Identifying elements that need improvement or updating is a demanding task, with the criteria for review varying depending on the purpose of the update.

Constant changes in various fields necessitate the regular updating of questions. These changes may include legal amendments, product or procedure updates, or time-based questions that reference specific dates [7]. Such updates are not the only challenge in maintaining a high-quality question database. In some cases, questions prepared by subject matter experts may lack clarity. Experts might use mental shortcuts or jargon familiar to those within the same field, but these can be unclear when used in test questions. Therefore, questions should be written using clear and generally understandable language [8]. Locating such unclear questions in an extensive database is even more difficult than identifying those with substantive errors. The challenge, therefore, is how to maintain continuous question quality in large question databases. One possible approach was presented in this article, based on a case study where automated testing is used to maintain high service quality for customers.

3 Periodic Testing as an Effective Tool for Maintaining High-Quality Staff

The research was conducted in a real production environment where automated tests are used to maintain high service quality. The quality of service depends primarily on the individuals responsible for executing specific procedures and actions. The knowledge and competencies of staff play a crucial role in shaping a company's performance and long-term success. It is important to understand how these factors influence various aspects of a business. One of the most direct impacts of staff knowledge and competence is on productivity and efficiency. Employees with a deep understanding of their roles and organizational processes are more adept at identifying inefficiencies and suggesting improvements. Competent employees tend to make fewer mistakes, thereby reducing the time and resources needed for error correction and enhancing overall operational efficiency. Knowledgeable employees are often the source of new ideas and creative solutions, leveraging their expertise to develop innovative approaches to business challenges. Furthermore, skilled employees excel in problem-solving, effectively addressing complex issues, and contributing to continuous improvement within the company. All these factors positively impact the quality of service provided, thereby enhancing customer satisfaction [9, 10]. Well-trained employees are pivotal in delivering high-quality customer service. Employees with a thorough understanding of the company's products and services can better meet customer needs and expectations, leading to increased satisfaction and loyalty.

Well-trained staff also positively impact other areas. The company's reputation is significantly influenced by the professionalism and competence of its staff. Competent employees help maintain a high level of professionalism, which enhances the company's

image. Knowledgeable employees are better equipped to handle changes in the market, technology, or organizational structure. Their ability to adapt quickly and efficiently ensures that the company remains agile and responsive to external pressures. A highly skilled workforce provides a strategic advantage in a competitive market [11]. Skilled employees are better equipped to execute the company's strategic plans effectively, aligning their efforts with the organization's long-term goals and ensuring successful implementation [12]. The impact of staff knowledge and competencies on a company's success cannot be overstated. From enhancing productivity and driving innovation to improving customer satisfaction and strengthening the company's reputation, skilled employees are the backbone of any successful organization. By investing in workforce development, companies can build a resilient, adaptable, and highly competitive business poised for long-term success. To achieve this, the subject company decided to invest not only in training and staff development but also in an automated testing tool that continuously assesses current knowledge and skills. This tool also identifies areas where further training is needed and measures the effectiveness of completed training programs.

The service provider in this case study chose to qualify and maintain the credentials of their representatives through periodic competence tests. Given that the service is provided in person across various locations nationally, and with a workforce of approximately 2,500 representatives, traditional and regular testing would be impractical. Hence, an automated online testing system was implemented. Consequently, each staff member representing the company in front of clients must maintain current knowledge of the company's products and services. Employees work through the company's online system, where appropriate credentials are assigned. To keep their accounts active, which is necessary for their daily tasks, they must pass periodic verification tests. These tests include questions and problems designed to assess whether representatives are capable and competent to perform their jobs safely and effectively.

This approach has another rationale. The type of service offered by the company may be provided by representatives periodically. This means an employee may be active for a specific period when the service is needed in a region, otherwise, they may be in a standby state, with their service provision suspended. It is well known that if skills and knowledge are not actively used, they may diminish over time. Continuous learning and application of skills are necessary to remain a competent and competitive employee. Technologies, methodologies, and industry standards evolve rapidly, and professionals need to keep up with these changes to maintain relevance in the job market [13]. Regular training, attending workshops, and staying updated with industry trends are critical for career growth and effectiveness. In this case, after a break, a representative is required to take a test to confirm their ability to deliver the service properly. Thus, the testing system in this environment not only examines active employees periodically but also ensures that returning employees have updated and adequate knowledge to perform their jobs effectively. That is another argument to keep question database updated and reviewed.

4 Maintaining High-Quality Exam Questions Through Automated Analysis

Maintaining the quality of examination questions is crucial to ensure that assessments are fair, reliable, and valid measures of an examined person's performance. Examination questions are the foundation of any assessment system. High-quality questions not only accurately measure knowledge and skills but also enhance the purposefulness of conducted training and tutorials. Ensuring the quality of the questions involves a systematic approach to their development, review, and maintenance. As stated before company that is a subject of the case study implemented an individual approach to questions database maintenance based on constant monitoring of the percentage of incorrect answers to questions. The main problem of massive question databases is how to select those that are the most likely subject to error. Another issue is to locate questions that are just outdated. This may be records concerning procedures or rules that are no longer in use or were depreciated. Finding and filtering those records is half the battle. So this approach aims to find such questions and verify if they need to be updated, revised, or just deleted. What is more complicated this process needs to be executed constantly because the database needs to be updated all the time.

To address the problem, the approach involved investigating questions that received the highest number of incorrect answers. When a particular question presents difficulty in obtaining the correct answer, it becomes a prime candidate for review and revision. Accordingly, the existing testing system was enhanced with an answer analyzer that tracked the responses of the individuals being tested. Furthermore, the application designed to create and maintain the question database was augmented with additional analytical capabilities for each question. Consequently, each question in the database was evaluated based on its difficulty, indicated by the frequency of incorrect responses. This enhancement enabled the identification of questions that could be subject to poor-quality data. A low number of correct answers does not automatically imply that a question is flawed. In some instances, the question may simply be challenging, resulting in many test-takers not knowing the correct answer. However, it is equally possible that the question is poorly constructed or requires updating. In the most critical situations, there may be an issue with the correct answer's definition, where the incorrect answer is mistakenly marked as correct [14]. To further elaborate, the extended system's answer analyzer operates by compiling response data and generating statistical reports on the performance of each question. This data-driven approach allows for a more precise identification of problematic questions. The analyzer considers factors such as the distribution of responses, time taken to answer each question, and patterns of repeated errors among different groups of test-takers.

To implement the described approach, a proprietary solution in the form of an ICT (Information and Communication Technology) system was specifically designed. This system includes an expanded version of the previously developed question test base environment. The expansion features an authors' module that links each question to an extensive set of examination data, facilitating the identification of elements useful for maintaining and ensuring the high quality of the testing data. The ICT system's architecture integrates several key components to support this approach. Firstly, the question test base environment serves as the central repository for all test questions.

This environment is equipped with advanced data analytics tools that analyze response patterns and error frequencies taking into account the number of times in which the question was used in a test. The authors' module is a significant addition to this environment. It provides a comprehensive interface for subject matter experts to access and review detailed examination data associated with each question. By examining trends in incorrect answers, the module helps identify questions that may require revision due to ambiguity, outdated content, or potential errors in the correct answer key. Moreover, the system incorporates basic machine learning algorithms to automatically flag questions that consistently yield high rates of incorrect responses. These algorithms analyze large datasets to detect patterns and anomalies that might indicate issues with question quality. This automated analysis assists authors in prioritizing their review efforts, ensuring that the most problematic questions are addressed first. Additionally, the ICT system features robust reporting capabilities. It may generate detailed reports that summarize the performance of each question, highlighting areas of concern and providing candidates for improvement. These reports are essential for maintaining transparency and accountability in the question review process.

In this particular case, multiple-choice questions in the test were displayed in a way where the order of answers was randomized in each test. That is a widely used method to enhance the fairness and integrity of assessments. This approach involves presenting the same set of questions to different test-takers but varying the sequence in which the possible answers are listed. Randomizing answers helps balance any unintentional preference that might occur if correct answers were always positioned in a particular order. This ensures that no specific answer position (e.g., always 'A' or 'C') is perceived as more likely to be correct. By reducing predictability, randomizing answer choices contributes to the overall reliability and validity of the test. It ensures that test scores more accurately reflect a test-taker's knowledge and understanding rather than their test-taking strategies. Using such a method of answer order randomization makes analysis a little more complicated because the algorithms need first to investigate what was the answer order in that particular case. On which position each answer was located and which one was chosen by the test taker? After that all records need to be unified to the original question and answer sequence, to display an analysis that matches the appearance displayed in the editor.

5 Question Analysis Online Module

To perform the analyses described in this article, an author's module was developed. This module was part of an internet-based system used to maintain the question database. Initially, the original system was only used to manage questions by creating new ones, editing, or deleting existing entries. These management activities were limited to manual and occasional reviews, typically triggered by information or complaints from tested individuals. Such an approach does not lead to a high-quality question database. To identify records that need review, a new approach was introduced, involving a module that acted as an analyzer of test history concerning the existing question database. The primary function of the module was to search for suspicious questions by analyzing test response records. When atypical behavior was detected, it was linked to a specific

question in the database, which was then marked for review. The analyses were based on elements such as incorrect answer rates, time spent on answering questions, and a high percentage of incorrect responses pointing to a specific answer, regardless of its position in each test occurrence (Fig. 1).

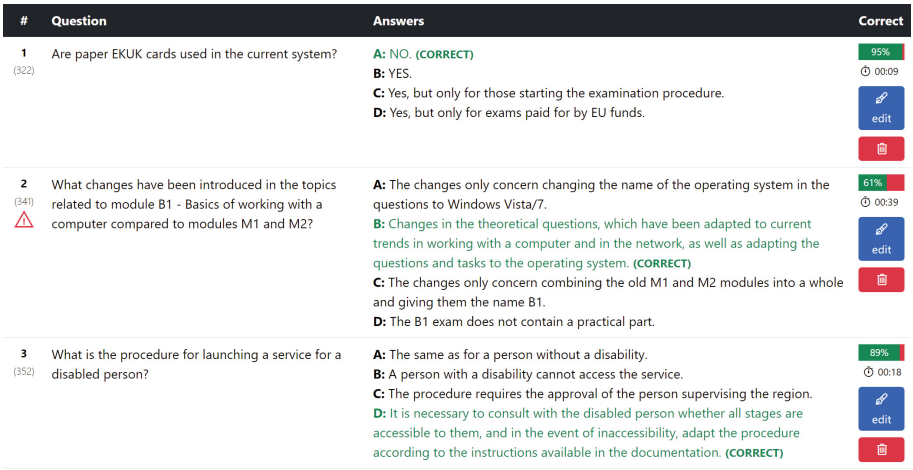


Fig. 1. General view of a question list with basic module analysis and markers.

The module was developed using the existing application programming language PHP, and the system utilized a MySQL database. To enhance readability and user engagement, additional frontend frameworks were employed. The Bootstrap 5 framework was used for responsiveness, Icons 8 provided graphic elements, and Google Charts facilitated the visual presentation of data for better understanding. It is important to note that the module primarily assists in identifying problematic questions; the final decision on whether a question should be modified or deleted is made by a reviewer. This research aimed to determine how such a mistake detection module can help locate questions that require review and to assess the accuracy of these suggestions. Screenshots containing sample data were presented to demonstrate the module's operation.

The primary interface for identifying questionable records was the general list of questions in the test database. Each question and its possible answers were displayed in a row. Each row included a question identification number, linking it to a record in the database. The question stem was displayed in the second column, while distractors and the correct answer were in the third column, with the correct answer marked in green and labeled "CORRECT." The last column displayed the percentage of correct answers for the question in past tests, without analyzing the breakdown of incorrect answers. A stopwatch icon indicated the average time spent on the question, helping to estimate if the question was excessively lengthy or problematic. The most critical indicator was the red exclamation mark in a triangle, marking questions that required review due to high incorrect answer rates and longer response times compared to similar questions (Fig. 2).

For further analysis, reviewers could access detailed data through the record editing option. This screen displayed the full questions and answers, along with key elements

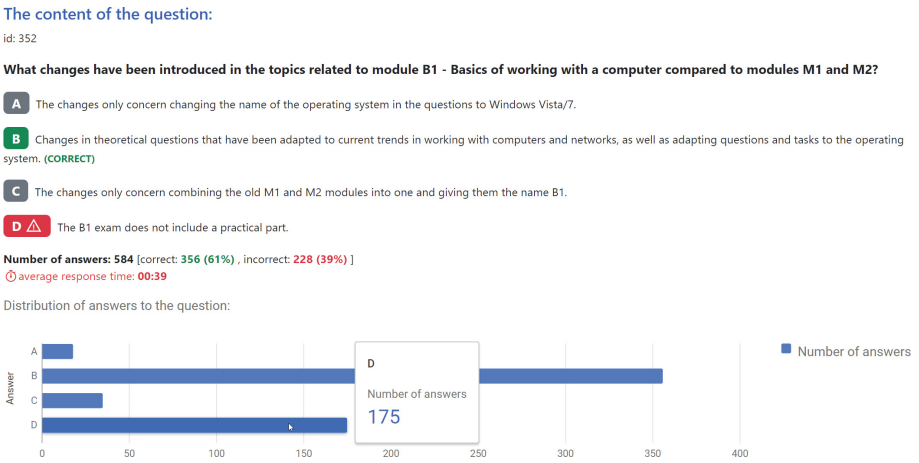


Fig. 2. Full content of the question and detailed analysis of past given answers.

needing attention. It presented hard numbers on the frequency of the question's use, the number of correct and incorrect responses, and the average response time. The most valuable data were the detailed response histories, shown as horizontal bar charts where each bar represented one answer, and its length indicated the number of times it was chosen. For instance, in this example, the distractor marked as answer D caused the highest percentage of incorrect answers. Using active charts from Google Charts allowed for displaying exact numbers when hovering over each bar. In this case, answer D was chosen 175 out of 584 times, indicating that nearly 30% of test-takers believed it was correct. This was highlighted with a red badge and warning icon. Distractors A and C were also chosen as correct answers by some, but their selection numbers were relatively low, likely due to a lack of knowledge in the tested area.

With this decision-making panel, reviewers have access to multiple data points based on real test data. They can decide whether to edit or delete a question to improve the overall quality of the question database. By editing a question, the reviewer archives the old record (to maintain past test integrity) and replaces it with the updated content. From that moment, the question will appear in tests in its new form, and the history analysis will reset to begin tracking new responses.

6 Summary

The article presents an approach implemented in a case study to maintain a large question test database effectively. By leveraging answering history from past tests, the method allows for the identification of records that compromise the quality of the question database. The criteria used for selection include the percentage of incorrect answers, the proportion of choices for single distractors, and the time spent on each question. These criteria proved to be effective in identifying inaccurate records.

Applying these criteria to a large dataset from past examinations demonstrated that this method is both simple and effective for maintaining continuous database quality.

Among the records flagged for revision, nearly all were deemed necessary for modification by the reviewers. Following these modifications, which involved altering either the question or the possible answers, there was a significant improvement in the percentage of correct answers in subsequent tests. This suggests that the issues in past tests were often due to poorly constructed questions rather than a lack of knowledge among the test-takers.

The module developed for this purpose has proven to be a valuable and efficient tool. Its use in the described case study indicates its potential for ongoing application to ensure the question database remains well-maintained and accurate. This approach allows for systematic improvements and helps maintain the integrity and reliability of the assessment process. The described company plans to continue using this tool in the future to uphold the quality of its question database.

Acknowledgments. Co-financed by the Minister of Science under the "Regional Excellence Initiative" Program.

References

1. Cheng, L., Spaling, M.A., Song, X.: Barriers and facilitators to professional licensure and certification testing in canada: Perspectives of internationally educated professionals. J. Int. Migr. Integr. **14**, 733–750 (2013)
2. Laduca, A.: Validation of professional licensure examinations. Eval. Health Prof. **17**, 178–197 (1994)
3. Ch, D., Saha, S.: Generation of multiple-choice questions from textbook contents of school-level subjects. IEEE Trans. Learn. Technol. **16**, 40–52 (2023)
4. Costello, E., Holland, J., Kirwan, C.: The future of online testing and assessment: question quality in MOOCs. Int. J. Educ. Technol. High. Educ. **15**, 42 (2018)
5. Presser, S., et al.: Methods for testing and evaluating survey questions. Public Opin. Q. **68**(1), 109–130 (2004)
6. Malau-Aduli, B.S., Zimitat, C.: Peer review improves the quality of MCQ examinations. Assess. Eval. High. Educ. **37**(8), 919–931 (2012)
7. Gottlieb M., Bailitz J., Fix M., Shappell E., Wagner M.: Educator's blueprint: a how-to guide for developing high-quality multiple-choice questions. AEM Education and Training **7** (2023)
8. Stevens S., Palocsay S.W., Novoa L.J.: Practical guidance for writing multiple-choice test questions in introductory analytics courses. INFORMS Transactions on Education (2022)
9. Shashovec Y.: Improving Approaches to evaluation of efficiency of enterprise staff in modern conditions. Market Infrastructure (2022)
10. Zongping Z.: Effective staff training: efficiency & effect based on cost-benefit analyses method. Science and Technology Management Research (2010)
11. Luo C.-C., Wang Y.-C., Tai Y.-F.: Effective training methods for fostering exceptional service employees. J. Hosp. Tour. Insights (2019)
12. Sturman M., Ford R.: Motivating your staff to provide outstanding service (2011)
13. Almotairi M., Alyami M., Song Y.-T.: Enhance employees' competences through customized learning. In: Proceedings of the 10th International Conference on E-Education, E-Business, E-Management and E-Learning (2019)
14. Webb, E., Phuong, J.S., Naeger, D.: Does educator training or experience affect the quality of multiple-choice questions? Acad. Radiol. **22**, 1317–1322 (2015)

Analyzing Data Science Labor Market Trends in Poland Using NLP Techniques

Ryszard Zygała, Wiesława Gryncewicz(✉), and Agnieszka Rosa

Wroclaw University of Economics and Business, Komandorska 118-120, 53-345 Wroclaw, Poland
{ryszard.zygala,wieslawa.gryncewicz,agnieszka.rosa}@ue.wroc.pl

Abstract. This paper responds to the need for in-depth analyzis of trends in data science and the requirements of data science professionals in the long term. The primary goal of the research is to analyze trends within the Polish data science job market. The study covers job openings in Poland from 2015 to 2023, containing more than one million job advertisements. This extensive dataset, which spans nine years, allows us to present changes and trends in the demand for data science specialists in the Polish labor market. This study utilized a variety of data analyzis tools and techniques at various stages of the data workflow: (1) data acquisition—Beautiful Soup and Apache Spark were used to extract data from the website; (2) data cleaning and preparation—Pandas, NLTK, and Spacy were used to clean and prepare the data for analyzis; (3) exploratory analyzis—LLM Mistral, Pandas and Matplotlib were used to conduct exploratory analyzis of the data. Understanding this evolution is critical not only for data science professionals, but also for educators who develop relevant curricula, policymakers who design skills training programs and HR departments that deal with talent acquisition in the dynamic labor market.

Keywords: NLP · job offers analyzis · labor market trends · Polish labor market · data science

1 Introduction

In today's rapidly changing technological landscape, data science has emerged a major driving force behind innovation and data-driven decision-making across industries [6]. By leveraging powerful algorithms and vast datasets (big data), data science unlocks valuable insights that can significantly improve customer service and operational effectiveness. As a result, businesses are experiencing a surging demand for skilled data scientists who can interpret and analyze this information. This demand has reached unprecedented levels worldwide in recent years [5].

In Poland, the technology sector is also growing strongly, but the data science job market exhibits a developmental gap compared to established markets in Western nations like the United States and the United Kingdom [3]. This discrepancy can be partly attributed to a longer history of investment in technology and data analytics in these

M. Hernes et al. (Eds.): ECAI 2024, LNNS 1217, pp. 161–172, 2025.
https://doi.org/10.1007/978-3-031-78465-1_14

countries and greater financial support. However, Poland demonstrates a trajectory of rapid growth [18]. Companies are increasingly recognizing the value of data-driven knowledge, leading to an increase in demand for data science specialists. This trend is fuelled by both domestic companies and Western firms operating in Poland, who are actively recruiting local talent [16]. As a result, the data science field in the Polish job market is experiencing a process of convergence with the standards observed in Western countries. This research aims to analyze trends in the Polish data science job market. We focus on job postings advertised on pracuj.pl, a prominent Polish job portal, between 2015 and 2023. Pracuj.pl boasts a significant user base with nearly 6.5 million registered users and 40,000 active recruiters, providing access to an extensive dataset of job listings for our investigation [20].

By analyzing these postings, we aimed to provide insights into the growth and trends of the demand for data science professionals in Poland. Understanding this evolution is crucial not only for data scientists navigating career opportunities but also for educators developing relevant curricula, policymakers designing skills training programs, and industry leaders refining their talent acquisition strategies in the dynamic tech job market.

This study focuses on analyzing job postings for data science positions in Poland between 2015 and 2023. We found this timeframe particularly relevant because it covers a period of very significant growth and development in the field of data science. We employed a quantitative approach, utilizing NLP (Natural Language Processing) techniques to examine a vast dataset of job listings. Our analyzis aims to identify trends in the demand for data science professionals, focusing on the specific skills, tools, and qualifications sought after by employers. Additionally, we explore potential regional variations within the Polish job market for data science specialists. Our study underscores the interdisciplinary nature of the field of data science and recognizes that various economic, technological, and social factors can influence observed trends.

In addition, while conducting literature research, we discovered a research gap. Among about 50 papers dealing with issues related to analyzing job opportunities for IT professionals, none dealt with the situation in Poland. This finding became an additional motivation and confirmation of the importance of undertaking the research.

The paper is structured as follows. In the next section, we provide an overview of global trends in employment analyzis for data science experts in Poland. The Methodology part describes how the dataset was prepared. In the Results section, we examined the number of job posts for the three roles of data analyst, data engineer, and data scientist across the country and by region, and then compared the tools, technologies, and soft skills required of candidates for each position. Finally, we shared our findings on trends in the Polish data science employment market, as well as our future research goals.

2 Related Work

To contextualize our investigation of the Polish data science job market, we conducted a comprehensive literature review of global trends in job offer analyzis. We utilized the Scopus database as a primary source, supplemented by additional credible sources. The search encompassed scholarly publications dating from 2003 to 2023. This extensive

review identified approximately 500 relevant papers on job offer analyzis. Within this corpus, roughly 10% pertained specifically to the data science field. Notably, only three studies focused on the Polish job market; however, none directly addressed our specific area of inquiry, necessitating their exclusion from further analyzis.

To examine trends in the data science labor market, the authors of the related papers used various methodologies. These include data mining techniques for analyzing job postings [7], quantitative content analyzis for identifying required skills and tools [13], and surveys or placement reports to understand industry demand [9]. Existing research highlights the ever-evolving nature of the data science job market, with a constant shift in required skill sets. Consequently, studies across various domains, such as job performance optimization [8], sentiment analyzis for startup success [6], and student placement prediction [19], emphasize the critical need for data science professionals. While these studies encompass diverse topics, a common thread lies in their focus on methodologies like skills profiling, job title identification, skill mapping, and skill recommendation. This focus underscores the importance of understanding employment trends and addressing challenges related to job matching and classification within the data science field. In this context, the results on the situation in Poland will make a valuable contribution to science and will be of interest.

The data science domain is undergoing a significant shift. Studies indicate a growing demand for skills related to artificial intelligence (AI), cloud computing, and SQL, while the reliance on traditional tools like Excel and SPSS is diminishing [13]. This trend reflects the increasing prominence of data in various sectors and the evolving nature of data-related professions. Artificial intelligence (AI) has had a profound impact on the job market, automating tasks and creating new job opportunities [4]. This necessitates adaptation within the job market to accommodate emerging skillsets and needs [17] and [11]. While Lerman (2015) previously suggested that the job market, particularly IT occupations, might not face a significant skills shortage despite the rising demand for skilled workers, the landscape has demonstrably changed. Technological advancements have propelled the job market towards a greater emphasis on automation and AI. Organizations are recognizing the need to equip their workforce with the necessary skills (reskilling and upskilling) to navigate these new roles demanding advanced capabilities [12]. This trend aligns with projections from the US Bureau of Labor Statistics, which anticipates a substantial growth (35%) in data analytics jobs between 2022 and 2032, significantly exceeding the average growth rate across other industries (3%) [2]. As new digital technologies emerge to optimize company operations, data analysts are poised to become even more crucial and sought-after.

The report "The Data Scientist Job Market in 2024" noted that Python was referenced in 78% of job postings for data scientists in 2023 and in 57% in 2024. The demand for natural language processing abilities has risen from 5% in 2023 to 19% in 2024. Excel remains an important tool in the data scientist's arsenal, as listed in 10% of job postings—though to a lesser extent than in prior years (26% in 2023). The development of artificial intelligence underscores the essential nature of data science skills. Understanding, developing, and innovating in AI projects requires mastery of statistics, probability, Python, APIs, machine and deep learning [21].

It is worth noting that in 2023, data scientists accounted for only 3% of individuals laid off by large IT businesses. Other tech workers, such as software engineers, were hit far harder, accounting for more than 22% of all layoffs. This figure emphasizes that data scientists play an important role in technical progress and business, especially in the face of enormous layoffs. [21].

The researchers address also the importance of aligning employers and academia, highlighting the need for collaboration in curriculum development to ensure that the skills taught align with the demands of the job market [19] and [14]. This collaboration is essential to bridge the gap between theory and practice, preparing graduates with the relevant skills needed to succeed in the real world. When UC Berkeley added a new data science major in 2018, and it quickly became one of the school's most popular majors [15].

3 Methodology

Our study analyzed online job postings from pracuj.pl, for the period 2015–2023. The process of downloading data from the website involved several steps. The algorithm followed a specific procedure: systematically scanned through all available subpages, captured the URLs of job offers, and stored them in a list. Subsequently, each URL was accessed, and the corresponding web content was saved as a pickle file. PySpark was employed for data extraction because of its effective handling of massive volumes of data. This is especially crucial when dealing with Windows, as it can be challenging to manage a lot of little, large-capacity files and to make use of all the available CPUs. The data that was downloaded included the job title, job language, date of first publication, company description, experience, requirements, capabilities, employer name, responsibilities, etc. The data was collected in a data frame format, containing 1,049,560 records and 59 columns, which posed a challenge in extracting observations relevant to the data science domain. Irrelevant features and records with missing values exceeding 75% were excluded to improve data quality. Additionally, only entries with an 'en' flag were considered, indicating English language postings to align with the research focus. Initially, we filtered job postings containing the keywords "data" or "analyst" to discern the breadth of job titles under scrutiny. From the query we received 29,521 unique job titles, indicating a substantial diversity potentially hindering detailed analyzes. Among the 10 most frequently obtained job names were also those unrelated to data science, such as business analyst, financial analyst, senior business analyst, etc. Consequently, we refined our search scope to encompass job titles featuring the keyword "data".

This narrowing of the search met our expectations because among the most frequently occurring job titles were those that reflect the main areas of activity in the field of data science and align well with our research goals (see Fig. 1). It became justified to focus on job postings containing the three key phrases: "data scientist", "data engineer", and "data analyst". This allowed for a more focused and in-depth analyzis of the data science labor market in Poland. Despite narrowing down the scope, the diversity of job titles remained substantial, with 378 unique titles for "data scientist," 529 for "data engineer," and 1498 for "data analyst," respectively. To better understand the data science labor market in Poland and compare job postings, we unified the job title values into three professional groups:

- data scientist—with 2634 records,
- data engineer—with 5227 records,
- data analyst—with 6259 records.

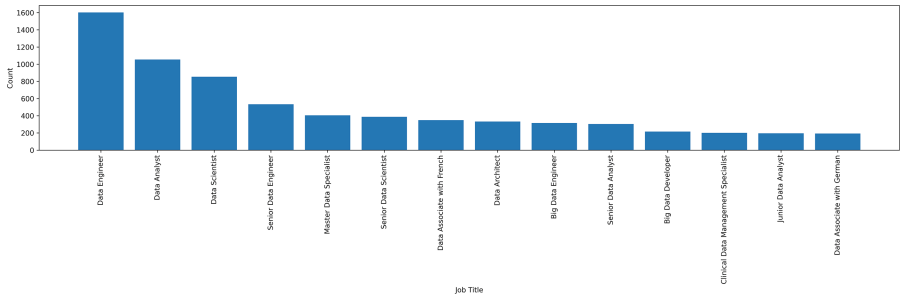


Fig. 1. The most common job titles that include the term "data"

In summary, the main filtering criterion targeted occupations within the "job title" column containing the terms "data scientist," "data engineer," and "data analyst." This selection aimed to restrict the dataset to postings relevant to the research objectives within the data science domain. The final analyzis focused on a cleaned dataset of 14,120 records and 7 relevant columns.

4 Results

The results of the analyzis of the number of job offers in Poland for three positions: Data Analysts, Data Engineers and Data Scientists in 2015–2023 were presented in Fig. 2. The number of job openings for Data Analysts remains relatively low through 2020. From mid-2021 to the end of 2022, there was a slight increase in the number of announcements, but in 2023 we again see a gentle decline. At the beginning of the analyzed period, job postings for Data Engineer positions were low compared to Data Analysts roles.

There were only single job offers or none. However, in the middle of 2021, there was a noticeable increase in the number of ads, and then the number of ads for this position equaled the number of ads for Data Analysts. Initially, the situation with Data Scientist was similar to Data Engineer, there were only a few job offers. In 2021–2022, there was a marked increase in interest, as with the other two positions. However, in 2023, the number of offers fell again, even more so than for Data Analysts. Currently, Data Engineer dominates in terms of job offers, it is the most sought-after role in data science.

The significant spike in job postings for all three positions was around 2021 to early 2022. After that, there is a noticeable decline in job postings for all three roles. The first release of ChatGPT was in late November 2022, which is likely the reason for this decline. Technologies like ChatGPT and other language models have the ability to automate and minimize the time of some operations, especially those involving large volumes of data processing, such as predictive modelling, image and natural language recognition, and recommendations. Employers had been holding off on hiring new employees,

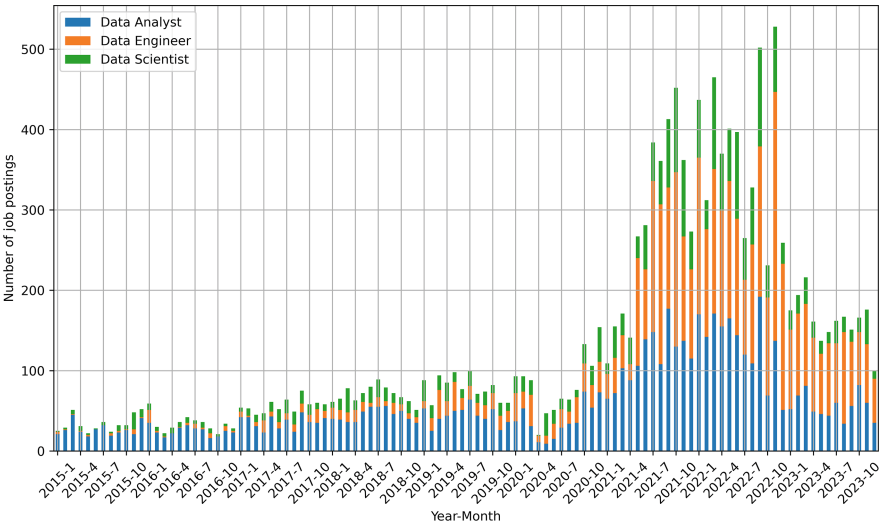


Fig. 2. Job postings for different positions in different years and months 2015–2023

expecting artificial intelligence to help reduce the number of workers and cut operating costs. However, such technology is not intended to replace human occupations, but to improve the efficiency and precision of their tasks [1]. Furthermore, AI development and deployment necessitate the involvement of professionals such as data scientists, researchers, data analysts, and engineers. This can be seen in the presented chart: after a gentle decline in the number of job openings in the fall/winter of 2022, we see an upward trend again in 2023, especially for Data Analyst.

The changes observed in the chart could also potentially be linked to the COVID-19 pandemic. The pandemic had a serious impact on global labor markets. The effects were sudden and often surprising: millions of people lost their jobs or took vacations, and others quickly had to adapt to working from home while offices were closed [10].The COVID-19 pandemic lasted in Poland from 20.03.2020 to 01.07.2023. In the context of the data industry, the significant spike in job postings around 2021 to 2022 could be a result of increased digital transformation efforts during the pandemic. Companies may have accelerated their data initiatives in response to the new challenges posed by the pandemic, leading to a higher demand for data professionals.

In the next step, the distribution of the number of offers in each region was examined. It was noted that the demand for data analyzis and data science specialists is not evenly distributed across different regions of Poland. The highest demand for these professions is found in the largest urban areas of the country, primarily in regions that symbolize the development of technological companies and higher education institutions that educate specialists in data-related professions, namely Mazowieckie (Warsaw), Małopolskie (Cracow), Dolnośląskie (Wroclaw), and Pomorskie (Gdansk) (see Fig. 3).

Another analyzis was focused on job offers in terms of the required tools and technologies for candidates in the field of data science. Based on the research of the dataset content, a list of leading technologies and tools expected from data science professionals

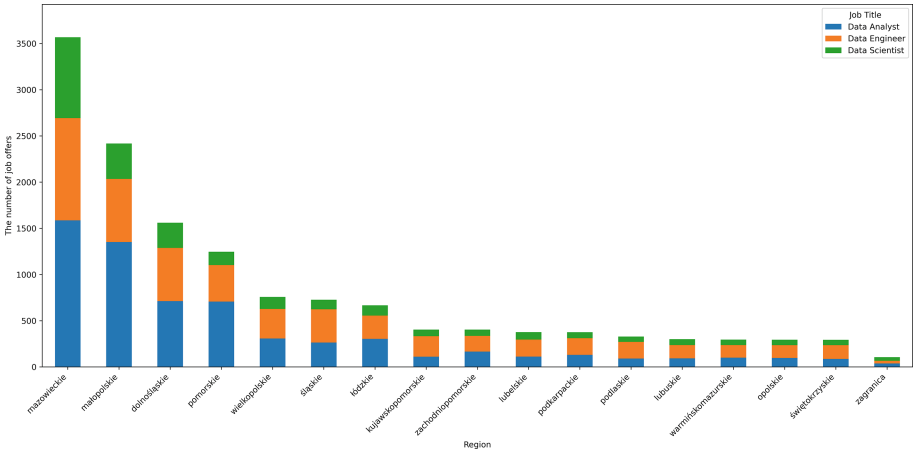


Fig. 3. Job offers for various positions in different regions of Poland

has been compiled. The list includes programming languages (Python, R, SQL, Scala), data analyzis applications (Excel, Power BI), big data technologies (Spark, Hadoop), and cloud computing technologies (AWS, Azure, Databricks for Spark).

Changes in the technologies required over the studied time were shown in Fig. 4. Analyzis of this figure provides several interesting conclusions:

- Since 2015, a strong trend illustrating the increasing importance of programming tools for data analyzis can be observed. At the beginning of the period under consideration, programming languages (Python, R, and SQL) and Excel were already present, but Excel was the primary tool for data science. From 2019 onwards, a predominance of Python and SQL over Excel can be observed.
- Python's dominance over R is increasing rapidly as of 2017, and this may be related to the versatile nature of Python as a tool used at every stage of data analyzis application development. Two decades ago, statistical data analyzis dominated, and R was the primary tool for data scientists. With the advent of TensorFlow in 2015, an open-source library for deep learning, Python's importance in data science has systematically increased.
- Nowadays a significant portion of the data science world is shifting to cloud computing. While there was no demand for cloud computing specialists in data analyzis in the early years of the period under consideration, recent years indicate that they are becoming a key corporate resource for data analyzis. This is reflected in employers' growing interest in technologies such as Azure, Databricks and AWS.
- In the big data domain, it is also evident that Apache Spark is gaining popularity over Apache Hadoop. With this increase in demand for Spark specialists, there is also a growing need for programming languages such as Scala and SQL for implementing Spark in cloud computing (e.g. Databricks).

With a high probability, the emergence of ChatGPT and the development of LLMs (large language models) are influencing the job market and the structure of requirements for data science specialists. These initial signs could be observed in 2023, where job

postings began to include requirements related to LLM technology. However, the number of these advertisements was so small that they were not captured in the analyzis.

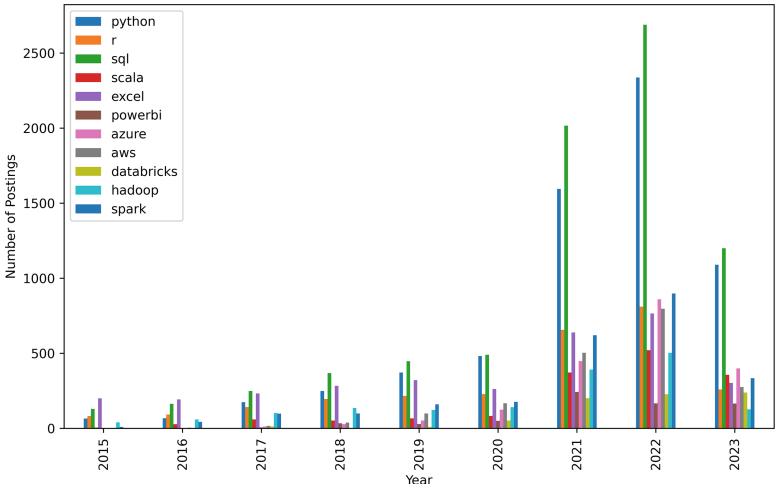


Fig. 4. Technologies for data science in the years 2015–2023

To complement the analyzis of technological requirements for data science specialists, a study of so-called soft skills was also conducted. In the prepared dataset, job postings contain a description of requirements in the form of unstructured text, and it is very difficult to identify the key expectations of the future employer. To gain insight into soft requirements, it was assumed that an effective way to achieve this goal would be to look at the list of the most frequently occurring phrases, in the form of n-grams, which are the basic method of NLP. After obtaining a list of the most common bigrams and trigrams, the focus of the rest of the study was on trigrams, as they were considered more informative. The results were presented in Fig. 5.

The analyzis of these trigrams allows for formulating several important conclusions. The highest requirements regarding experience in each role are placed on candidates for the position of data engineer. These candidates are also expected to have experience with big data technologies. Surprisingly, strong analytical skills and problem-solving skills are not equally important for data scientists as they are for data analysts.

Detecting key skills of job candidates using uni-, bi- and tri-grams proves to be a simple and effective method, providing these most common expectations of job companies. The disadvantage of this method is that this process takes place with a large human participation to extract the elements of interest for the analyzis. The appearance of Large Language Models on the market of data analyzis tools prompted the authors of the article to ask themselves whether these types of models can better cope with the extraction of the features of the analyzed text that are of interest to the researcher. An analysis environment was prepared for this study, using the open-source Ollama platform to run LLM Mistral on our local Linux-based machine. We decided to use Mistral locally to gain more control over the configuration and customization of the environment using the

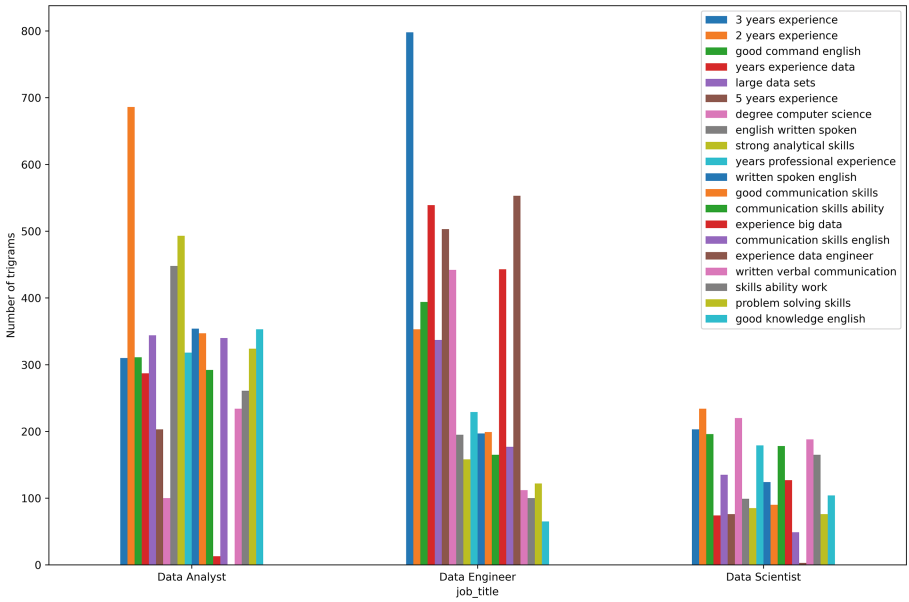


Fig. 5. 20 most frequently occurring trigrams for requirements

LLM for analyzis purposes. The task of the LLM-based analyzis model was to analyze the text of the job advertisement and extract keywords relating to soft and technical skills that should be included as separate features of the Pandas dataframe. Table 1 shows an example of the first row of the dataframe after the procedure.

Most of the features were detected correctly, e.g. among the technical skills appeared: data mining, statistical analysis, machine learning, python, R, SQL, etc., and among the soft skills: communication, meticulousness, and patience. Some attributes were even additionally formatted correctly, e.g. instead of 'fluent' and 'english' we got 'fluency in English'. Unfortunately, there were also incorrect answers, e.g. among technical skills there appeared: excellent communication skills, organizational skills, or self-discipline. The above example shows that using an LLM to extract keywords from text and classify them semanticly is a very promising approach, despite the imperfections observed in the table. We believe that LLMs, with certain modifications, can be very successful in this type of task due to the automation and the significant time savings for researchers.

5 Conclusion and Future Work

The analyzis of job offers presents valuable research material not only for labor market specialists but also for analysts examining developmental trends across various industries and professions. At the beginning of the analyzis period, the number of offers for each position did not change significantly over time. An increase in the number of job offers for all three positions occurred between 2021 and early 2022, but a decrease in the number of offers was observed thereafter. In our opinion, these changes may have been related to the covid-19 pandemic and the emergence of ChatGPT. In-depth research

Table 1. Skills detected by LLM Mistral

Oryginal requirements (tokens)	Skills detected by LLM Mistral	
	Technical	Soft
[‘least’, ‘2’, ‘years’, ‘professional’, ‘experience’, ‘working’, ‘data’, ‘scientist’, ‘andor’, ‘data’, ‘analyst’, ‘strong’, ‘experience’, ‘data’, ‘mining’, ‘statistical’, ‘analysis’, ‘good’, ‘knowledge’, ‘statistical’, ‘modeling’, ‘machine’, ‘learning’, ‘techniques’, ‘tools’, ‘experience’, ‘python’, ‘andor’, ‘r’, ‘proficiency’, ‘sql’, ‘andor’, ‘pyspark’, ‘andor’, ‘scala’, ‘skills’, ‘ability’, ‘work’, ‘big’, ‘amounts’, ‘data’, ‘ability’, ‘clearly’, ‘easily’, ‘present’, ‘analysis’, ‘outcome’, ‘excellent’, ‘communication’, ‘skills’, ‘meticulousness’, ‘patience’, ‘attention’, ‘details’, ‘great’, ‘organizational’, ‘skills’, ‘selfdiscipline’, ‘sense’, ‘ownership’, ‘fluent’, ‘english’, ‘b2c1’]	data mining, statistical analysis, machine learning, python, R, SQL, pyspark, Scala, big data, excellent communication skills, organizational skills, self-discipline, fluent English	communication, excellence, meticulousness, patience, attention to details, great organizational skills, self-discipline, fluency in English

(e.g., interviews, surveys) is needed to verify these conclusions. We also conducted a regional analyzis of job offers. It confirmed that the greatest demand for data science specialists is in major urban areas. In the third step, an analyzis of the required tools and technologies for data science candidates was undertaken. Noticeable over the years has become the dominance of Python and SQL while the demand for specialists in cloud technologies has increased. Based on the evaluation of the trigrams, it was noted that the highest experience requirements are placed on candidates for the position of data engineer.

In summary, our study points to hypotheses and theories about the growing demand for data science occupations, the impact of artificial intelligence on job creation and skill requirements, and the influence of social processes on the observed changes. However, further research would be highly recommended to provide conclusive evidence and delve into the long-term trends driving this transformation.

What is worth mentioning, the collected dataset contains more than one million jobs from 9 years, so it offers great opportunities for in-depth analyzis using advanced natural language processing and machine/deep learning techniques.

In future research, we plan to focus on an extended analyzis of soft requirements for data science specialists. We plan to use tools such as LLMs to perform automated detection of these requirements. We also plan to compare the requirements of the analyzed labor market with the educational offerings of universities in Poland. We would like to investigate whether these universities are preparing graduates of the kind the labor market expects.

References

1. Arindam-talentintellect: How ChatGPT is changing the Job Market? In: Asia-Pacific's Leading HR Management Company Based in Thailand |Talent Intellect (2023). https://www.talentintellect.com/how-chatgpt-is-changing-the-job-market/. Accessed 3 Apr 2024
2. Bureau of Labor Statistics: U.S. Department of Labor, Occupational Outlook Handbook, Operations Research Analysts (2023). https://www.bls.gov/ooh/math/operations-research-analysts.htm#tab-6. Accessed 3 Apr 2024
3. Frankowska, A., Pawlik, B., Feldy, M., Sobestjańska, A.: Sztuczna inteligencja: osiągnięcia publikacyjne z zakresu nauk ścisłych i technicznych w latach 2010–2021
4. Gryncewicz, W., Zygała, R., Pilch, A.: AI in HRM: case study analysis. Prelim. Res. Procedia Comput. Sci. **225**, 2351–2360 (2023)
5. IABAC®: The growing demand for data analytics jobs. In: IABAC® (2024). https://iabac.org/blog/the-growing-demand-for-data-analytics-jobs. Accessed 3 Apr 2024
6. Iskamto, D.: Data science: trends and its role in various fields. Adpebi Int. J. Multidiscip. Sci. **2**, 165–172 (2023)
7. Karakatsanis, I., et al.: Data mining approach to monitoring the requirements of the job market: a case study. Inf. Syst. **65**, 1–6 (2017). https://doi.org/10.1016/j.is.2016.10.009
8. Lerman, R.I.: Emerging trends in the Information Technology Job Market: how should the public and private sectors respond? (2015)
9. Lopez, L.: Placement report: political science Ph.D.s and ABDs on the job market in 2001–2002. APSC **36**, 835–841 (2003). https://doi.org/10.1017/S1049096503003263
10. Lund, S., Madgavkar, A., Manyika, J., Smit, S., Ellingrud, K., Meaney, M., Robinson, O., et al.: The future of work after COVID-19. McKinsey global institute **18** (2021)
11. Manresa Matas, A., Berbegal-Mirabent, J., Gil-Domenech, D.: Challenging students to develop work-based skills: A PBL experience. In: 6th International Conference on Higher Education Advances (HEAd'20). Editorial Universitat Politècnica de València, pp. 561–568 (2020)
12. Miranda, M., León, J.: An Analysis of the Job Offer for Professionals in Information Technology. In: 2023 IEEE Seventh Ecuador Technical Chapters Meeting (ECTM), pp. 1–4 (2023)
13. Musazade, N.: Tools and technologies utilized in data-related positions: an empirical study of job advertisements. 36th Bled eConference Digital Economy and Society: The Balancing Act for Digital Innovation in Times of Instability **155** (2023)
14. Ong, X.Q., Lim, K.H.: SkillRec: a data-driven approach to job skill recommendation for career insights (2023)
15. Oostendorp, N.: Radical change is coming to data science jobs. In: Forbes (2019). https://www.forbes.com/sites/forbestechcouncil/2019/03/01/radical-change-is-coming-to-data-science-jobs/. Accessed 3 Apr 2024

16. Pasławski, K.: (2022) Zagraniczne firmy idą po polskich fachowców IT. In: CRN. https://crn.pl/aktualnosci/zagraniczne-firmy-ida-po-polskich-fachowcow-it/. Accessed 18 Apr 2024
17. Ravin, M., Tailor, R., Kamble, A.: A review paper on the impact of artificial intelligence on the job market. Int. J. Adv. Res. Sci., Commun. Technol. **68–73**, (2023). https://doi.org/10.48175/ijarsct-10724
18. Skotnicka, E., Andrzejewska, A., Niedźwiedzki, B.: Poland is another Silicon Valley! https://www.ey.com/en_pl/workforce/poland-is-another-silicon-valley. Accessed 18 Apr 2024
19. Smaldone, F., Ippolito, A., Lagger, J., Pellicano, M.: Employability skills: profiling data scientists in the digital labour market. Eur. Manag. J. **40**, 671–684 (2022). https://doi.org/10.1016/j.emj.2022.05.005
20. About us—Grupa Pracuj: https://grupapracuj.pl/about-us. Accessed 18 Apr 2024
21. Data Scientist Job Market in 2024: Analysis, Trends, and Opportunities. In: 365 Data Science. https://365datascience.com/career-advice/data-scientist-job-market/. Accessed 3 Apr 2024

An Empirical Study of a Dynamic Stop Loss Strategy with Deep Reinforcement Learning on the NASDAQ Stock Market

Mateusz Anders[1], Jozef Zurada[2], and Paweł Weichbroth[3](✉)

[1] Gdansk University of Technology, Narutowicza 11/12, 80-233 Gdansk, Poland
[2] University of Louisville, College of Business, Louisville, USA
[3] Gdansk University of Technology, Faculty of Electronics, Telecommunication and Informatics, Narutowicza 11/12, 80-233 Gdansk, Poland
pawel.weichbroth@pg.edu.pl

Abstract. The objective of this paper is to empirically investigate the efficacy of using Deep Reinforcement Learning (DRL) to maximize investment returns by incorporating expected optimal closing prices of long positions into a daily strategy. This paper extends existing research on the impact of stop-loss orders on investment strategy results and brings contribution of these orders to trading strategies into a completely new perspective. We propose a novel approach using DRL, in contrast to fixed-price stop-loss strategies, trailing-stop strategies, or other machine learning approaches. In the backtesting experiment, daily OHLCV data for stocks from the NASDAQ-100 index (as of May 2024) were used for the period spanning from January 2014 to January 2024. The strategy is compared with buy-and-hold, stop-loss, and trailing stop-loss strategies. Significant effort was made to accurately reflect market conditions in the simulation. We found a positive impact of using DRL compared to other tested strategies when encountering entirely new data, suggesting positive serial market correlations. The results suggest that appropriate closing rules and active management of stop levels can increase investment returns without necessarily reducing portfolio return volatility.

Keywords: NASDAQ-100 · Deep Reinforcement Learning · Stop-Loss · algorithmic trading · backtesting · machine learning in trading · algorithmic strategy · closing rules · trading simulation

1 Introduction

It is common for investors to protect their returns by applying stop-loss (SL) orders, which activate market transactions when a certain threshold in the asset's price is reached, triggering a sell order for long positions and a buy order for short positions [1]. Among investors, the SL order is primarily considered a risk-reducing tool rather than a key part of a profit-taking strategy. Besides the obvious benefit of SL mitigating the disposition effect [2, 3], there is controversy regarding the efficacy of stopping losses using

M. Hernes et al. (Eds.): ECAI 2024, LNNS 1217, pp. 173–183, 2025.
https://doi.org/10.1007/978-3-031-78465-1_15

SL orders. In [4], further extended in [5], researchers provide both empirical and analytical evidence that in an environment governed by the Efficient Markets Hypothesis (EMH) [6] and described by the Random Walk Hypothesis, SL does not effectively stop losses. However, under more realistic conditions with positive market correlations, well-described by Behavioral Finance, SL adds values. On the other hand, some argue that using SL has neither a positive nor negative impact on a portfolio's performance [7]. Additionally, there is evidence suggesting a positive impact of using SL for assets with a high probability of losing and a low probability of large gains [8], as well as for the combination of SL with the momentum strategy [9].

A significant challenge arises when selecting the appropriate level for closing a position [10]. Closing too early generally limits potential profits, while closing too late might result in substantial losses. With the rise in popularity of machine learning algorithmic trading strategies, one approach addressing the SL issue [11] uses supervised learning techniques for SL price prediction to create a profit-taking high-frequency trading strategy.

In this paper, we propose a profit-taking strategy based on dynamic SL order using Deep Reinforcement Learning (DRL), which yields promising results broadening horizons both for investors and scientists. We present a trading agent that learns to choose the expected optimal stop-loss levels through trial and error in a simulated environment after each market close. The environment is deterministic, so the objective of the agent is to maximize immediate rewards. In our study design, we follow the guidelines provided in [12]. First, we set stopping prices relative to the open, rather than the close, to overcome the overnight phenomenon. Second, we adjust the SL thresholds using volatility as a reference.

2 Data Description

The data used in this experiment consists of daily open (O), high (H), low (L), close (C), and volume (V) (OHLCV) data for stocks listed on the NASDAQ-100 index, spanning from January 2014 to January 2024. The NASDAQ-100 index comprises the 100 largest non-financial companies listed on the NASDAQ stock exchange, based on market capitalization. It is important to note that the NASDAQ-100 index list changes over time due to annual rebalancing, typically occurring in December. The index list used in this study is dated May 2024.

Some of the stocks do not have historical data dating back 10 years or less. Therefore, stocks were filtered based on the required minimum data length to conduct an evaluation. The OHLCV data for a given asset is a sequence of numbers, where for each day d within the date range, we can distinguish O_d, H_d, L_d, C_d, V_d. The data is divided into training and test datasets with a ratio of 8:2.

3 Environment

3.1 Observation Space

Before taking an action for day $d + 1$, the agent performs an observation after day d. The agent's state is 15-dimensional. The environment is deterministic, meaning the agent's action does not influence the next state. Each observation is calculated based

on historical data only, specifically after the market closes each day. To predict utilities, the agent has access to the given day's OHLCV data and corresponding information regarding previous days. Data is normalized via described operations (ratio or fractional change) to ensure each observation represents values in a similar range, independent of time and asset, which further stabilizes deep network training and predictions.

The basic elements of the observation for a given day d of the dataset are the open O_d, high H_d, and low L_d prices relative (r) to the corresponding close price (C_d). They are calculated as shown in (1).

$$O_{\mathrm{r},d} = \frac{O_d}{C_d},\ H_{\mathrm{r},d} = \frac{H_d}{C_d},\ L_{\mathrm{r},d} = \frac{L_d}{C_d} \tag{1}$$

The close price relative indicator, denoted as $C_{\mathrm{r},d}$, is the ratio of C_d to C_{d-1}. Conversely, $V_{\mathrm{r},d}$ is the fractional change in day-to-day volume. Both equations are presented below in (2). High $V_{\mathrm{r},d}$ and upward/downward trend might suggest substantial market movements that could potentially cause an asset to increase/decrease further in price. $C_{\mathrm{r},d} > 1$ might be interpreted as a short-term upward trend, while $C_{\mathrm{r},d} < 1$ might indicate a short-term downward trend. This logic is similar to that applied in the on-balance volume indicator.

$$V_{\mathrm{r},d} = \frac{V_d}{V_{d-1}} - 1,\ C_{\mathrm{r},d} = \frac{C_d}{C_{d-1}} \tag{2}$$

Furthermore, feature extraction, an important part of training a deep network, was conducted. This includes calculating the moving average $SMA_{n,d}$ (3) along with the standard deviation $\sigma_{n,d}$ (4) for various periods n including the day d and preceding days. The rationale for choosing standard deviation and moving average is as follows: Based on the moving average calculated for longer and shorter time periods, it is possible to predict whether the trend is upward or downward and the likelihood of the price increasing or decreasing the following day. On the other hand, a low/high standard deviation for a short period relative to a high/low standard deviation for a longer period suggests a decrease/increase in volatility, which might be reflected during the following trading day. Further explanation can be found in [13].

$$SMA_{n,d} = \frac{1}{n}\sum_{i=0}^{n-1} C_{d-i}, \text{ where } n \in \{2, 5, 10, 20, 30\} \tag{3}$$

$$\sigma_{n,d} = \sqrt{\frac{\sum_{i=0}^{n-1}(C_{d-i} - SMA_{n,d})^2}{n}}, \text{ where } n \in \{2, 5, 10, 20, 30\} \tag{4}$$

Both $SMA_{n,d}$ and $\sigma_{n,d}$ are provided to the agent relative to C_d, as shown below in (5).

$$SMA_{\mathrm{r},n,d} = \frac{SMA_{n,d}}{C_d},\ \sigma_{\mathrm{r},n,d} = \frac{\sigma_{n,d}}{C_d} \tag{5}$$

3.2 Action Space

After observing day $d-1$, the agent can choose action a for day d from $a_{\max}+1$ possible actions, where $a_{\max} = \frac{S_{\text{pct,max}} - S_{\text{pct,min}}}{S_{\text{pct,step}}} + 1$. Here, $S_{\text{pct,max}}$ is the maximum stop-loss level expressed as a percentage, $S_{\text{pct,min}}$ is the minimum stop-loss level, and $S_{\text{pct,step}}$ is the step size in the closed-set range. Thus, the action space is $A = \{0, 1, 2, \ldots, a_{\max}\}$ and $a \in A$, where action equal to zero is the additional always occurring action, which translates to sell order. For any action, the corresponding expected optimal closing price is calculated as shown in (6) below.

$$S_{d,a} = \left(\frac{100 - S_{\text{pct,step}} \cdot a}{100}\right) \cdot O_d, \quad \text{where } a \in \{0, 1, 2, \ldots, a_{\max}\} \tag{6}$$

The action translates into a market decision for the next day, $signal_{d,a}$, specifically for the next market open, as shown in (7). Market sell and market buy orders are market-on-open (MOO) orders. A stop-loss order is submitted immediately after the buy order is executed or as an MOO order. In the latter case, One Cancels the Other (OCO) must be applied, meaning the stop order should be canceled if the buy order is canceled. The second option is generally preferred because the first requires action after the market opens and can result in a time gap between the execution of the buy order and the submission of the stop order. The selected option may depend on the broker, particularly whether OCO is supported. If the agent is already in a position, any unexecuted stop-loss order is canceled immediately after C_{d-1}, and a new stop-loss order is submitted immediately after O_d.

$$signal_{d,a} = \begin{cases} \text{sell} & \text{if } a = 0 \text{ and in position} \\ \text{hold} & \text{if } a = 0 \text{ and not in position} \\ \text{buy and stop with } S_d & \text{if } a > 0 \text{ and not in position} \\ \text{update stop with } S_d & \text{if } a > 0 \text{ and in position} \end{cases} \tag{7}$$

To determine $S_{\text{pct,min}}$ and $S_{\text{pct,max}}$, a train data analysis was conducted based on Kernel Density Estimation (KDE) of the price drop from open to low (see Fig. 1), where the drop is the percentage change from market open to the asset's lowest price during a given day.

Additionally, an empirical analysis was conducted by observing the KDE of actions chosen by the agent during training (see Fig. 1). The following logic was applied: If the density of the maximum action $a_{\max}$ was significantly higher than that of other actions, it could indicate that the agent might prefer a stop-loss level higher than the chosen $S_{\text{pct,max}}$ to increase reward, thus $S_{\text{pct,max}}$ should be increased, and vice versa. However, increasing $S_{\text{pct,max}}$ substantially can cause the deep neural network (DNN) to take longer to converge or to converge to a sub-optimal solution, which involves going with the market by applying the $a_{\max}$ action instantly. The action parameters used for the training are as follows: $S_{\text{pct,max}} = 1.9$, $S_{\text{pct,min}} = 0.1$, and $S_{\text{pct,step}} = 0.2$. Furthermore, the agent's choice precision can be improved through chart analysis and tuning the $S_{\text{pct,max}}$, $S_{\text{pct,min}}$, and $S_{\text{pct,step}}$ parameters.

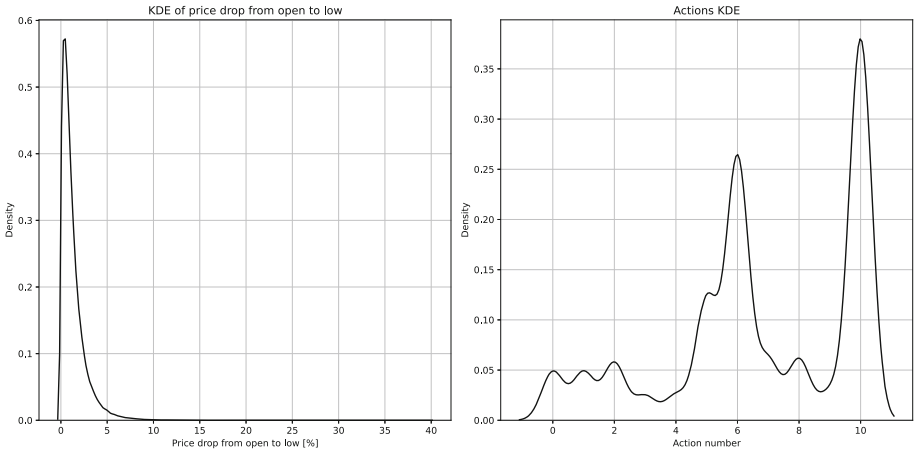


Fig. 1. Actions analysis

3.3 Agent Reward

After day d is over and the market is closed, the agent is rewarded for making a decision based on the observation from day $d-1$. The reward is the difference R_d (10) between the open-to-open (daily) strategy return $DR_{\text{strategy},d}$ (8) and the open-to-open (daily) market return $DR_{\text{market},d}$ (9).

$$DR_{\text{strategy},d} = \begin{cases} 0 & \text{if not in position} \\ \frac{S_d(1-\text{slipp})-O_d}{O_d} \cdot 100 & \text{if SL executed} \quad (L_d \leq S_d) \\ r_{\text{market},d} & \text{if SL not executed} \quad (L_d > S_d) \end{cases} \tag{8}$$

$$DR_{\text{market},d} = \frac{O_{d+1} - O_d}{O_d} \cdot 100 \tag{9}$$

$$R_d = DR_{\text{strategy},d} - DR_{\text{market},d} \tag{10}$$

If the agent is not in a position during the day, the reward is always 0. If the stop-loss was not triggered during the trading day, the reward equals the market daily return. A stop-loss order is triggered when the low price L_d is lower than or equal to the stop-loss order price S_d. The stop-loss cannot trigger at a market opening because the stop-loss order is submitted right after the market opens with a price lower than the opening price and is canceled after the market closes. An arbitrary slippage of slipp $= 0.0001$ was applied to simulate market conditions close to reality. The slippage could be a potential cause of criticism directed at the presented results in this paper and should be taken into consideration when analyzing the results.

4 DRL Algorithm

The algorithm can be seen as improved and adjusted to the problem Deep Q-learning (DQL) algorithm [14] with $\gamma = 0$, where γ is discount factor for future rewards. The environment was intenionally simplified by not providing state elements that would be

action-dependent, such as position or available cash, to expedite the learning process and achieve optimal results in the primary area of interest. However, the environment can be easily converted to reflect Markov process. For example, if the available cash is relatively low compared to an asset's price, the agent should take less risky actions to avoid ending up with a lack of possibility of buying back the asset or drastically lowering expected future rewards. More about using DQL in algorithmic trading can be seen in [15]. The significant benefit of using DRL, although not tested here, is the possibility for the agent to continuously learn when interacting with a live trading environment.

An epsilon greedy policy (see e.g. [16]) was applied to address "exploitation vs exploration" problem. The probability of selecting a random action is ϵ, which decays linearly over the first 1440 episodes from an initial value of $\epsilon = 1$ to a final value of $\epsilon = 0.01$, and then decreases exponentially. To mitigate the network "catastrophic forgetting" problem a replay buffer is used to store all experienced transitions.

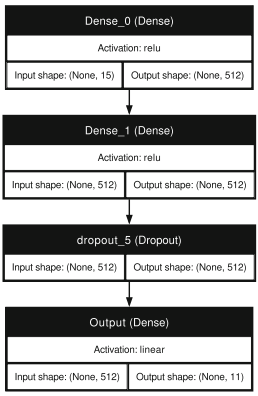


Fig. 2. DNN architecture

4.1 Deep Neural Network

The objective of the agent is to maximize immediate rewards in the investment environment by taking actions with the maximum expected utility resulting from the environment's observation. Therefore, the input to the Deep Neural Network (DNN) is the observation and the output is the action's utilities. The network consists of two hidden layers with 512 units each (see Fig. 2). To achieve better generalization and prevent overfitting, several regularization techniques were applied. Firstly, every episode lasts only 10 days and each episode is a randomly selected slice from the training dataset, which mixes up densely data with different characteristics. Moreover, a dropout layer [18] was used with a dropout rate of 0.1. The training process involves only a subset (approximately. 25%) of the entire training dataset. The Adam optimizer, introduced in [19], was used with a learning rate 0.0001, which has been shown to improve DRL performance.

5 Methodology

The training dataset consists of approximately 200 000 days of data. The training process consists of 4320 episodes, each 10 days long. For each episode, a ticker is randomly selected from NASDAQ-100 tickers and then a slice of the stock's data is randomly selected. Every day, the DNN is fitted for one epoch to a batch sampled from the replay buffer (stored experiences) spanning over 126 days.

The agent is evaluated over a 2-year period with starting cash being $10 \cdot C_0$. The agent can trade with fractional trades of $1 \cdot 10^{-15}$, always using 90% of available cash. In a marginal situation where there is not enough cash to satisfy an MOO buy order, the position is not taken. 10% of the cash must be put aside to be able to repurchase the position in case of a price rise from close to open. Fractional trades with low fractions were used because the fraction and starting cash relative to an asset's price depend on the broker used and cash available for investment. Commission fees were not taken into consideration, as many online brokers now offer commission-free trading. As previously mentioned, slippage with a value of 0.01% was used in the experiment.

Apart from the B&H strategy, we compare the designed strategy with fixed stop-loss and trailing-stop strategies, with stop levels corresponding to each action of the agent, to demonstrate that using dynamic thresholds inevitably leads to better performance than fixed ones. Trades sizing differs for the B&H strategy and is equal to 100%. The fixed stop-loss strategy can be seen as the designed strategy with stop levels that are fixed and not updated across days. In contrast, the trailing-stop strategy resembles the DRL strategy more closely, as the stopping level is adjusted daily to the current opening price. We provide cumulative returns along with risk-adjusted cumulative returns, assuming zero risk-free returns as performance measures.

6 Results

To present the results, a sample of assets with various market returns, was selected (see Tabel 1, with tickers showing lower market returns on the left). The results indicate that trailing stop-loss (SL) is evidently more efficient than fixed SL in terms of returns, but not necessarily when considering risk-adjusted returns (see Table 2). The aggressive trailing stop level $\text{TrailingSL}_{0.1}$ proved to be the best strategy in the context of risk-adjusted returns. Nevertheless, the objective of the agent was to maximize returns, not the Sharpe ratio. The selected starting cash value ($10C_0$) favors assets with a higher price per share, which led to higher overall returns ($Portfolio_{\text{mean}} < Portfolio$), but also demonstrated that both fixed and dynamic trailing SL strategies are more effective for such assets. Further research could focus on extracting features of assets that favor these strategies.

The DRL strategy was shown to effectively adjust stop levels, yielding positive results on average for both tickers with negative market returns and positive market returns (see Fig. 3). None of the strategies outperformed the B&H strategy for assets with a positive trend, which could be attributed to some buy orders not being executed due to significant overnight price increases. Although these instances were rare, the drastic rise in prices could have significantly impacted the overall results. SL strategies with low fixed stop levels were more effective at stopping losses but also tended to cut gains, and vice versa.

Table 1. Cumulative returns benchmark

Ticker	PYPL	IDXX	AEP	LIN	VRTX	Portfolio	$\text{Portfolio}_{\text{mean}}$
Strategy							
DRL	−26.74	−7.71	−18.10	28.49	51.63	28.36	17.55
$\text{TrailingSL}_{0.10}$	−30.06	−1.17	32.22	36.10	59.01	20.56	11.45
$\text{TrailingSL}_{0.70}$	−42.45	11.96	76.53	25.76	50.25	25.56	11.28
$\text{TrailingSL}_{0.30}$	−49.03	−7.68	10.13	16.82	69.96	20.06	9.69
$\text{TrailingSL}_{1.70}$	−47.16	−18.22	29.46	39.57	72.42	15.22	9.45
$\text{TrailingSL}_{1.50}$	−50.26	−11.75	26.37	33.53	56.78	17.63	9.26
$\text{TrailingSL}_{1.90}$	−57.13	−18.37	5.51	26.76	75.37	14.64	8.62
$\text{TrailingSL}_{0.50}$	−47.94	−13.67	65.50	28.50	46.69	24.49	8.49
$\text{TrailingSL}_{0.90}$	−28.38	−9.46	29.76	35.61	51.04	23.19	8.34
B&H	−68.88	−10.31	5.81	25.89	120.28	11.65	8.00
$\text{TrailingSL}_{1.10}$	−43.32	−12.64	46.42	36.91	41.59	17.37	7.30
$\text{TrailingSL}_{1.30}$	−47.28	−28.27	26.31	25.59	59.62	15.36	6.42
$\text{FixedSL}_{1.10}$	−54.80	−9.70	17.28	9.19	109.70	6.98	4.67
$\text{FixedSL}_{0.90}$	−49.87	−19.50	19.16	9.60	108.33	6.92	4.48
$\text{FixedSL}_{1.70}$	−65.93	−19.34	16.64	10.93	109.19	5.92	4.36
$\text{FixedSL}_{1.30}$	−58.76	−9.96	16.01	8.31	102.27	5.28	3.64
$\text{FixedSL}_{1.50}$	−60.58	−12.22	17.84	8.22	109.57	6.04	3.63
$\text{FixedSL}_{0.70}$	−54.34	−35.87	18.73	8.56	109.09	4.46	3.26
$\text{FixedSL}_{0.10}$	−63.78	−10.79	20.64	−10.00	110.52	−1.48	2.15
$\text{FixedSL}_{1.90}$	−66.18	−27.76	15.31	9.92	108.80	4.57	2.10
$\text{FixedSL}_{0.50}$	−54.72	−15.79	19.38	−12.73	109.85	3.07	1.94
$\text{FixedSL}_{0.30}$	−59.12	−13.33	14.69	−13.73	110.33	0.79	0.61

7 Conclusions

As mentioned in [20], SL orders have hidden costs that are revealed by stopping out not only down-trending but also up-trending movements, as shown also on Fig. 3. However, setting dynamic day-to-day stop levels helps mitigate these costs. The usage of DNNs to adjust stop levels is a promising approach, but results indicate that this may have a negative effect on returns volatility. Standard deviation seems to be a good indicator for adjusting stop levels [13], as the agent achieved positive results by incorporating it into its observations. Despite the promising results presented in this paper, they should be viewed with some skepticism. Although the tested assets are expected to be relatively liquid, in real market conditions, slippage could be significantly larger during major market movements that cause an asset to fall in price, potentially impacting the strategy's results negatively. Additionally, the results may not hold across different date

Table 2. Sharpe ratio benchmark

Ticker	PYPL	IDXX	AEP	LIN	VRTX	$\text{Portfolio}_{\text{mean}}$
Strategy						
$\text{TrailingSL}_{0.1}$	−1.08	0.16	0.62	0.98	1.39	0.42
$\text{TrailingSL}_{1.1}$	−1.46	−0.15	−0.04	1.02	1.27	0.32
$\text{FixedSL}_{1.5}$	−0.95	−0.01	0.18	0.24	3.01	0.30
$\text{FixedSL}_{0.9}$	−0.90	−0.07	0.16	0.27	2.90	0.30
$\text{FixedSL}_{0.7}$	−0.83	−0.15	0.19	0.25	2.97	0.29
$\text{FixedSL}_{1.1}$	−0.94	0.01	0.13	0.26	3.03	0.29
$\text{TrailingSL}_{1.5}$	−1.04	−0.22	0.33	1.45	2.26	0.27
$\text{TrailingSL}_{0.5}$	−1.29	−0.31	0.71	0.77	1.64	0.27
$\text{FixedSL}_{0.5}$	−0.97	−0.02	0.22	−0.08	3.04	0.27
$\text{FixedSL}_{1.9}$	−1.01	−0.10	0.10	0.27	2.94	0.27
DRL	−1.46	−0.06	−0.62	0.82	1.07	0.27
$\text{FixedSL}_{1.7}$	−0.97	−0.04	0.12	0.29	2.97	0.27
$\text{TrailingSL}_{1.7}$	−0.70	−0.28	0.48	1.24	2.22	0.26
$\text{TrailingSL}_{0.7}$	−0.78	0.29	0.31	0.80	2.04	0.25
$\text{FixedSL}_{0.1}$	−0.85	0.03	0.27	−0.04	3.10	0.25
$\text{TrailingSL}_{1.3}$	−1.22	−0.30	0.38	1.17	1.97	0.25
$\text{TrailingSL}_{0.9}$	−1.64	−0.03	0.08	1.00	1.51	0.24
B&H	−1.17	0.05	0.22	0.65	3.03	0.24
$\text{FixedSL}_{0.3}$	−0.93	0.01	0.25	−0.09	3.09	0.21
$\text{TrailingSL}_{1.9}$	−1.02	−0.26	0.25	0.72	3.38	0.21
$\text{TrailingSL}_{0.3}$	−1.31	−0.01	−0.27	1.06	1.93	0.13

ranges or for different portfolios. Further research is needed to examine the characteristics of the assets that yield the most profits for the strategy. Moreover, the reason for increased returns volatility in the presented strategy, despite not involving short selling [21], remains unclear. Changing objective of the agent from maximizing daily returns into maximizing risk-adjusted returns calculated over some days period could be the solution for lowering the volatility. Other solution could involve transformation of the environment from deterministic to non-deterministic by adding cash available into the observation space, where the agent would be forced to consider future consequences of its actions and potentially lower the risk. Future research could be conducted in a more realistic simulated environment using higher frequency data, including price spreads, to minimize slippage error. When trading large volumes, it's crucial to recognize that slippage can be significantly impacted and potentially altered drastically. Additionally, consideration of commission fees remains essential in these scenarios.

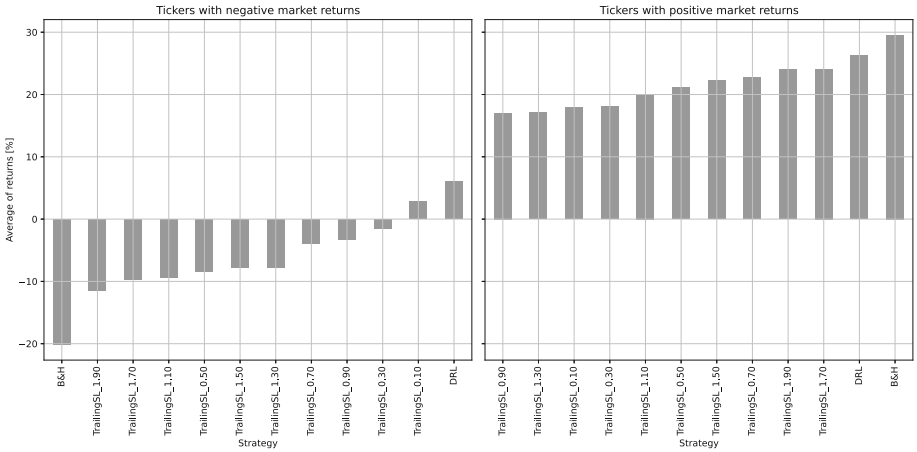


Fig. 3. Returns on average for tickers with negative and positive market returns

References

1. Vezeris, D., Kyrgos, T., Schinas, C.: Take profit and stop loss trading strategies comparison in combination with an MACD trading system. J. Risk Financ. Manag. **11**(3), 56 (2018). https://doi.org/10.3390/jrfm11030056
2. Shefrin, H., Statman, M.: The Disposition to Sell Winners Too Early and Ride Losers Too Long: Theory and Evidence. J. Financ. **40**, 777–790 (1985). https://doi.org/10.1111/j.1540-6261.1985.tb05002.x
3. Richards, D.W., Rutterford, J., Kodwani, D., Fenton-O'Creevy, M.: Stock market investors' use of stop losses and the disposition effect. Eur. J. Financ. (2015). https://doi.org/10.1080/1351847X.2015.1048375
4. Kaminski, K., Lo, A.W.: When Do Stop-Loss Rules Stop Losses? EFA 2007 Ljubljana Meetings Paper (2007). https://doi.org/10.2139/ssrn.96833
5. Snorrason, B., Yusupov, G.: Performance of stop-loss rules vs. Buy-and-hold strategy, Lund University (2009). https://www.lunduniversity.lu.se/lup/publication/1474565
6. Fama, E.F.: Efficient capital markets. J. Financ. **25**(2), 383–417 (1970)
7. Lei, A.Y.C., Li, H.: The Value of Stop Loss Strategies. Financ. Serv. Rev. **18**(1), 23–51 (2009). https://doi.org/10.2139/ssrn.1214737
8. Dai, B., Marshall, B.R., Nguyen, N.H., Visaltanachoti, N.: Lottery Stocks and Stop-loss Rules (2021). https://doi.org/10.2139/ssrn.3836739
9. Han, Y., Zhou, G., Zhu, Y.: Taming Momentum Crashes: A Simple Stop-Loss Strategy (2016). Available at SSRN: https://doi.org/10.2139/ssrn.2407199
10. Kobiela, D., Krefta, D., Król, W., Weichbroth, P.: ARIMA vs LSTM on NASDAQ stock exchange data. Procedia Comput. Sci. **207**, 3836–3845 (2022)
11. Samarasekara, I.K., Mendis, O.K., Ahangama, S., Atukorale, A.S.: Dynamic Stop-Loss Approach for Short Term Trades using Deep Learning. 2022 IEEE International Conference on Big Data (Big Data), Osaka, Japan, pp. 3537–3546 (2022). https://doi.org/10.1109/BigData55660.2022.10020535
12. Thomakos, D.D., Yahlomi, R.: Dynamic stop-loss rules as universal performance enhancers. Investment management & financial innovations, 15, 1–16 (2018). https://doi.org/10.21511/imfi.15(2).2018.01

13. Schalow, D.: Setting Stops with Standard Deviations (1996). https://api.semanticscholar.org/CorpusID:154411338
14. Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., Riedmiller, M.: Playing Atari with Deep Reinforcement Learning (2013). https://doi.org/10.48550/arXiv.1312.5602
15. Théate, T., Ernst, D.: An Application of Deep Reinforcement Learning to Algorithmic Trading. Expert Syst. Appl. **173**, 114632 (2021). https://doi.org/10.1016/j.eswa.2021.114632
16. Langford, J.:Efficient Exploration in Reinforcement Learning. Encyclopedia of Machine Learning and Data Mining (2017). https://doi.org/10.48550/arXiv.2305.18246
17. Catak, F.O., Arslan, E., Sener, F.A.: Generalized Huber Loss for Robust Learning and its Efficient Minimization for Robust Statistics (2021). https://doi.org/10.48550/arXiv.2108.12627
18. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: A Simple Way to Prevent Neural Networks from Overfitting. J. Mach. Learn. Res. **15**, 1929–1958 (2014)
19. Kingma, D., Ba, J.: Adam: A Method for Stochastic Optimization. International Conference on Learning Representations (2014). https://doi.org/10.48550/arXiv.1412.6980
20. Ma, W., Morita, G., Detko, K.: Re-Examining the Hidden Costs of the Stop-Loss (2008). https://doi.org/10.2139/ssrn.1123362
21. Macrae, R.: The Hidden Cost of the StopLoss. Arcus Investments, AIMA Journal (2005). https://www.eurekahedge.com/Research/News/1138/The-Hidden-Cost-of-the-Stoploss

Machine Learning for Prediction of the Importance of Factors Influencing Prosumer Attitudes

Ewa Walaszczyk[1](✉), Michał Nadolny[1], Marcin Hernes[1], Agata Kozina[1], and Bogdan Franczyk[2]

[1] Wroclaw University of Economics and Business, Wroclaw, Poland
{ewa.walaszczyk,michal.nadolny,marcin.hernes,agata.kozina}@ue.wroc.pl
[2] Leipzig University, Leipzig, Germany
franczyk@wifa.uni-leipzig.de

Abstract. A prosumer is a person who plays the role of a consumer and a producer simultaneously. He shares his knowledge and experience from using a product, gives feedback, and helps to improve products and services. Such an attitude is very desired by the producers as they benefit from it. The most common way to get feedback is to conduct a survey. Sometimes, it is hard to convince clients to fill out the questionnaire. Machine learning and deep learning techniques give opportunities to predict survey responses based on a sample of real survey results. The aim of the research was to determine the possibility of classifying prosumer survey respondents depending on their method of assessing each decision-making factor separately using selected machine learning and deep learning methods. Decision trees, LightGBM, and neural network models were used. The input data were the respondents' characteristics and the output data was the importance of four factors in the prosumption decision-making process. The accuracy of built models was the highest for *prosumption destimulants* using DNN, *for information contribution* using Decision trees, for *communication channels* using DNN and for *stimulants* using LightGBM. The presented research shows that there is a possibility of predicting the importance of decision-making factors based on respondents' characteristics. It may be especially important for smaller companies that do not have much money or time to take large market surveys.

Keywords: Prosumer · Decision-making process · Prediction · Neural networks · Machine learning · Deep learning

1 Introduction

Today's consumers are active users of goods and services. They establish relationships with producers using information and communication technologies (ICT) and the Internet. They share knowledge and experience using a product or service and become prosumers. The term refers to customers engaged in the co-creation of products and services.

M. Hernes et al. (Eds.): ECAI 2024, LNNS 1217, pp. 184–196, 2025.
https://doi.org/10.1007/978-3-031-78465-1_16

Prosumers have a dual function in society, both consuming and producing for themselves or others. They may receive different incentives from companies involved in the selling process [1, 2].

Knowledge is the key advantage of modern market competition. It increases efficiency and productivity in every enterprise, enables a better response to consumers' needs, and helps create new and innovative products [3]. The most common way to get knowledge and consumer feedback is to take surveys. The survey response rate is important in marketing because companies want to identify customers' attitudes toward products and services, which helps determine long-term relations and supports enlarging the market share. Companies need valid survey results to make decisions about their efforts to attract new customers and to lower the retention of the old ones. The survey response rate has recently declined [4]. One of the solutions to get missing results might be to predict survey responses based on respondents' characteristics. Companies have databases of certain characteristics of their customers. They also sometimes have survey results conducted on the part of the clients. Based on real survey results, it is possible to predict the attitudes of the rest. Machine learning and deep learning techniques give such opportunities.

The aim of the research was to determine the possibility of classifying prosumer survey respondents depending on their method of assessing each decision-making factor separately using selected machine learning and deep learning methods, including decision trees, LightGBM, and neural networks. Separate estimation of each model for each factor corresponds to how the factors were created: they are assumed to be independent and unrelated to each other (orthogonal rotation). The model's input data were the respondents' characteristics, and the output (classifying variable) was the respondent's assessment of a given factor.

To achieve the aim of the study, two hypotheses were posed. The methodological hypothesis assumes that the machine learning / deep learning model effectively assigns qualitative and quantitative respondents' characteristics to how individual factors are assessed in the prosumption decision-making process. The hypothesis will be verified positively if the model parameters are satisfactory. The second hypothesis is the cognitive hypothesis, which assumes that how a respondent evaluates individual decision-making factors depends on their individual characteristics. This hypothesis will be verified based on the interpretation of the model.

The structure of the paper is as follows. Section 2 presents the related works. Section 3 indicates materials and methods, including data source, data preparation, and modeling. In Sect. 4, the results are presented and discussed. Section 5, the conclusion, summarizes the paper's findings and indicates the future research directions.

2 Related Works

In the scientific literature, one can find various examples of the application of machine learning and deep learning methods for predicting survey results based on respondent characteristics. Predictive models in survey analysis often use various machine learning algorithms to predict results based on survey data. Traditional methods include regression analysis, decision trees, and support vector machines (SVMs). Linear and logistic

regression models are often used to determine the relationship between survey variables and outcomes. They are advantageous because of their interpretability but can lack ac-curacy when dealing with complex, non-linear relationships [5]. Decision tree and random forest models are popular because of their ability to deal with non-linear relationships and interactions between variables [6]. SVMs are used for classification tasks in survey data, especially when the data set is not linearly separable. They have been used to categorize prosumer attitudes based on survey responses [7].

Deep learning techniques, such as neural networks, offer significant co-benefits in processing large-scale and high-dimensional survey data. These methods can uncover complex patterns and interactions that traditional machine-learning techniques may miss. Artificial neural networks have been used to predict survey results related to consumer satisfaction and behavior. They can model complex non-linear relationships but require significant computational resources and data preprocessing. Although traditionally used in image processing, CNNs have been adapted to analyze structured survey data by treating it as a spatial grid. This approach is less common but has shown potential in exploratory research. RNN and LSTM models are suitable for time series data and sequential information. They have been used in longitudinal surveys to predict trends in prosumer behavior over time [8, 9].

Data imputation techniques are methods used to replace missing values in datasets. Data gaps are common in survey analyses and other datasets, and imputation helps maintain the integrity of analyses by preventing problems associated with missing data. Machine learning methods are often used to deal with missing data in surveys. Imputation of missing values is crucial to ensure the durability and consistency of analysis results. The panel survey research used decision trees to predict and model when respondents do not answer. Analysis of respondent characteristics included a variety of demographics, history of participation in previous surveys, and variability in the dataset. Decision trees were used to develop models that flexibly accounted for complex, non-linear relationships between outcomes and explanatory variables. The methods used were found to be more effective than the classical logistic regression model in predicting nonresponse in panel surveys [10].

There are also studies on identifying factors that affect survey response rates and how different factors can affect survey response rates. Decision tree analysis was used to examine the impact of various survey features on response rates in published surveys. Researchers analyzed aspects such as the method of data collection, the survey sponsor, the level of confidentiality, the method of invitation to participate, and the cultural background of respondents [4].

Metamodeling was used to combine two different methods: a survey and a decision tree. Data was collected through surveys, while a decision tree was used to analyze and present decision-making situations. The survey metamodel considered various characteristics of respondents, such as their role, participation in the survey, and degree of involvement in completing the survey. Decision trees were used to model the decision-making process based on the collected survey data [11].

A decision tree analysis method was used to study tourists' purchasing preferences and intentions. Data was collected through questionnaires, and the study focused on tourists' characteristics, travel patterns, and shopping behaviors that may influence

their purchase intentions. Data analysis was carried out using the decision tree analysis method, which allowed the identification of key factors affecting tourists' preferences and purchase intentions [12].

Decision tree algorithms were used to identify factors affecting student success and reducing school failure. Three classifiers were used to analyze questionnaires completed by students. The survey included 161 questionnaires containing 60 questions on health, social activities, relationships and academic performance, most of which were related to student achievement. The analysis helped identify key determinants of educational success and develop strategies to reduce the number of failures in the educational system [13].

Decision trees were used to analyze teachers' behavior in completing online surveys, where the response time to survey questions was studied. Two decision tree techniques and a factor analysis method were used to reduce the number of input variables. Finally, a simplified decision model was developed based on four variables: years of teaching, school level, gender, and area, which maintained statistical equivalence with the original model using eight input variables. The use of decision trees made it possible to classify the timing of survey responses, considering respondents' characteristics such as age, gender, years of schooling, school level, and area of residence [14].

However, no articles were found on examples of the application of machine learning and deep learning methods in the prediction of survey results for predicting the importance of factors influencing prosumer decisions. Only one publication was found for the keywords "prosumption" and "machine learning." The paper proposes a method for strategic day-ahead marketing in multi-period energy markets using a machine-learning approach based on neural networks. An energy aggregator with renewable energy generation facilities must plan for energy production and consumption (prosumption) when the amount of renewable energy produced is not accurately predicted the day before. If there is a difference between the planned and actual prosumption profile, i.e., an imbalance, the aggregator must bear the cost of imbalance penalties. As a scheduling method to avoid paying imbalance penalties, a machine-learning-based scheduling model based on the results of previous transactions is proposed. Specifically, the planning model is represented as a neural network, which gives an advantage in computational cost compared to the kernel method. To develop a training algorithm, it was shown that the gradient of the profit function with respect to the design parameters can be calculated as a solution to a linear programming problem. Finally, the effectiveness of the proposed method is demonstrated using a numerical example [15].

Prosumption here is related to the management of energy production and consumption, which refers to the concept of a prosumer who not only consumes energy but also produces it. Machine learning is used to develop a planning model that helps the aggregator make energy management decisions to avoid the costs associated with the imbalance between planned and actual energy production [16]. Prosumption in the context of machine learning is emerging primarily in the energy field, where it involves both energy production and consumption. It is a concept referring to user involvement in both energy production and consumption, usually in the context of renewable energy sources. Unlike survey data, where research focuses on respondents' subjective answers

to questions, prosumption topics connected to machine learning focus on analyzing hard data, such as transactional data related to energy production and consumption [16].

Prediction of factors influencing prosumer decisions involves the analysis of various attributes, such as economic incentives, environmental impact, and social impact. More interdisciplinary research is needed to integrate behavioral science, economics, and data science knowledge to develop holistic models of prosumer decisions. This is indicated as an important research gap in the scientific literature [17].

3 Materials and Methods

3.1 Data Source

The data used in the research were taken from a survey conducted on a group of 1620 prosumers [18], which was a representative group of respondents for the population of Poland. The questionnaire consisted of questions about respondents' characteristics and of the selected aspects of consumer attitudes connected to prosumption. The answers to respondents' characteristic questions were single-choice descriptive answers, and the answers to the questions about the prosumption attitudes were given on a 7-point Likert scale from 0 to 6.

Based on the answers obtained, the dimensions of the research area were reduced by using EFA with varimax rotation. The method allowed us to isolate key factors in respondents' assessments. These factors were subjected to further statistical analysis. As a result, the key decision-making factors of the prosumer attitudes were identified and indicated with the use of EFA. These factors were: prosumption destimulants, information contribution, information channels, incentives for prosumption, and reactive attitude [18]. The factors loadings were expressed as numbers from 1 to 6 in increments of 0.2.

3.2 Data Preparation

The dataset taken to modeling consisted of five respondents' characteristics (sex, age, education level, place of residence, subjective income satisfaction) and values of four factors influencing prosumption: prosumption destimulants, information contribution, communication channels, and stimulants. The factors were used as classification (target) variables in the models built.

The database needed filling in and transformations. Missing data (relatively few) were filled with the average; the distribution of the variable with missing data was symmetric and close to normal. The answers to respondents' characteristic questions (models' input data) were transformed into numbers:

- quantitative interval data were recoded into an ordinal numerical variable of the mean value from the interval;
- ordinal text data was recoded into an ordinal numerical variable;
- categorizing text data was recoded into consecutive integers.

Factor values (models' output data) were recoded with three levels:

– level 0 was defined when a factor value had a mean value within its confidence interval. This level should be defined as an indeterminate assessment; the respondent cannot decide whether a given factor is important to him or not in relation to the population;
– level -1 when values were below the confidence interval. This means that the respondent assesses a given factor as insignificant in relation to the population;
– level + 1 when values are above the confidence interval. The respondent assesses the impact of a given factor relatively highly.

Factor coding favors the minimization of unmarked states in favor of states that discriminate between respondents. The sample of the dataset is presented in Fig. 1.

	ID	sex	age	education level	place of residence	income satisfaction	prosumption destimulants	information contribution	communication channels	stimulants
0	22276879	1	65	3	650000	4	-1	-1	-1	-1
1	620568031	0	45	2	150000	4	0	-1	-1	1
2	95983054	0	55	3	650000	2	-1	-1	-1	1
3	44722240	1	55	1	350000	2	1	0	1	1
4	209403870	0	35	1	500	3	0	1	0	1
...	...	...	...	...	...	...	...	...	...	...
1615	102839976	1	45	1	650000	2	1	-1	0	-1
1616	664039958	1	55	3	75000	3	1	1	-1	1
1617	653284324	0	21	2	75000	0	-1	1	-1	1
1618	172377433	0	65	2	500	3	1	1	-1	1
1619	28547253	0	45	3	500	4	1	-1	-1	1

1620 rows × 10 columns

Fig. 1. The sample of the dataset.

3.3 Modeling

The dataset consisted of 1620 records and was divided into two sets. 80% of the dataset (1296 records) was used to train and validate the model, and 20% (324 records) for testing the prediction capabilities. The set for training and validation was divided into two parts: 80% was used for training (1036 records), and 20% was used for validation of the model (260 records).

In order to predict the importance of factors, models using three different techniques were created: classification decision trees, LightGBM boosting, and deep neural network (DNN). The deep neural network model was the main model. We developed a fully connected model of DNN. The results of the remaining two models were compared to it, and the best solution was indicated. For each model, accuracy was calculated. Precision and recall were measured for decision trees and neural networks. Additionally, feature importance was measured for decision trees and LightGBM.

4 Results and Discussion

The first models built on the dataset were the decision trees. Decision trees are a widely used technique in research to examine the impact of various survey features on response rates in conducted surveys [4, 6, 11, 12]. For each of the outputs tested in the research

presented in this paper – decision-making factors – another tree had to be built. Among many hyperparameters of decision trees, the *criterion*, which measures the quality of a split, and the *maximum depth* of the tree, which is the number of split levels, were changed in order to get the best results. The *criterion* hyperparameter was changed between *gini* and *entropy*, and the *maximum depth* of trees was tested from 2 to 5. As the measure that defined the best solution, the accuracy of the model validation was calculated. Table 1 presents the hyperparameters and results of the decision trees for which the accuracy of model validation was the highest among tested variants and other selected measures of trees, i.e., the accuracy of validation, precision, recall, and finally, the accuracy of prediction of the values on new data.

Table 1. Hyperparameters and results of decision trees

Parameter	prosumption destimulants	information contribution	communication channels	stimulants
criterion	Gini	Gini	Entropy	Entropy
max_depth	2	2	4	4
Accuracy of validation	0.415	0.481	0.581	0.523
Accuracy of prediction	0.389	0.497	0.534	0.519

As seen in Table 1, the accuracy of the models' validation is not high. The highest value was obtained for the factor *communication channels*, 0.581. The precision and recall of this tree were also the highest among all the factors; both equaled 0.39. In most cases, the accuracy of factors' values prediction on new data was lower than the accuracy of model validation. Only for the *information contribution* tree was it a little higher. The values of accuracy and other parameters (precision, recall) were a bit lower than in other studies [4, 6, 13], but this was probably due to other tree parameters that were not changed in these studies.

Among respondents' characteristics, age seems to be the most important feature influencing decision tree models (Fig. 2). It was the most important for three out of four factors. It was not important at all for *prosumption destimulants* only. For this factor, sex and income satisfaction were the only two characteristics that mattered and were nearly equally important. Place of residence seems to be the least relevant for all tested factors.

The second method used to predict the values of factors was LightGBM boosting for the decision trees. As for classic decision trees, other models were built for each factor. Among hyperparameters for LightGMB, trees' *learning rate* and *maximum depth* were changed to get the best accuracy. The *learning rate* tested values were: 0.1, 0.5, and 1.0; *maximum depth* was changed between 2 and 5, as in decision trees. Table 2 summarises the models' hyperparameters that obtain the highest accuracy of validation.

The highest accuracy, 0.585, was achieved for the *communication channels* factor. A little lower accuracy was measured for *stimulants*. The lowest value for this parameter, 0.388, was calculated for *prosumption destimulants*. Besides *stimulants*, the learning

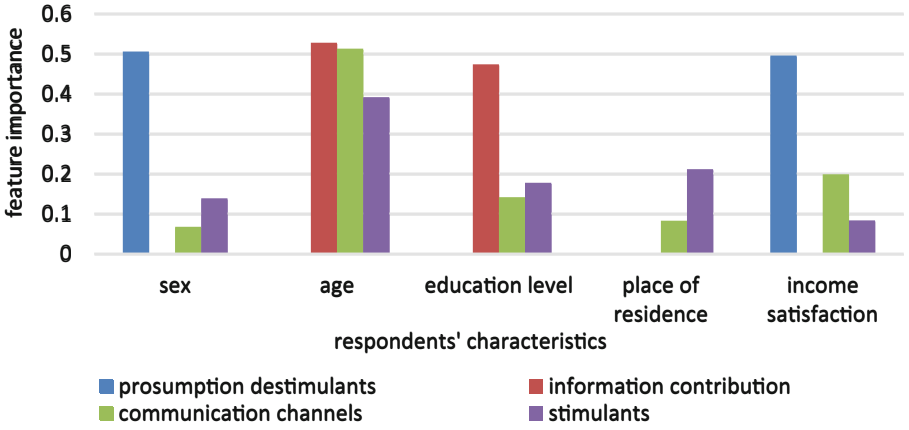


Fig. 2. The feature importance for decision trees.

Table 2. Hyperparameters and results of LightGBM boosted decision trees.

Parameter	prosumption destimulants	information contribution	communication channels	stimulants
Learning rate	0.1	0.1	0.1	0.5
max_depth	2	3	2	2
Accuracy	0.388	0.469	0.585	0.523

rate of 0.1 gave the best results in most cases. The maximum depth of a boosted tree was mostly 2. Only for the *information contribution* factor, the best depth was 3 levels.

The feature importance for LightGBM boosted decision trees is presented in Fig. 3.

Age is the most important respondents' characteristic for almost all factors. Only in the case of the *prosumtion destimulants*, is income satisfaction more important. The least important feature for all factors is sex.

The last developed model was based on a deep neural network. This technique has also been used in survey response prediction [5, 8, 9]. The network was built using Keras Framework based on Tensorflow Engine. The first model was developed with two hidden layers, but the performance of this model was low (accuracy about 0.27). Then, several experiments were performed with different models' architectures, and hyperparameters were tuned. The best results were achieved using the architecture presented in Fig. 4.

The model is a multi-class classifier (three classes). There are 4 hidden layers. The following hyperparameters were set: *relu* activation function for hidden layers and *softmax* activation function for output layer. The BatchNormalization and Dropout were used. Without dropout, the overfitting phenomenon appeared. We used *sparse_categorical_crossentropy* as a loss function and *sparse_categorical_accuracy* as an accuracy metric. The reason for using these functions is to provide labels of classes as integers. The *adam* was used as an optimizer. For the training process, the number

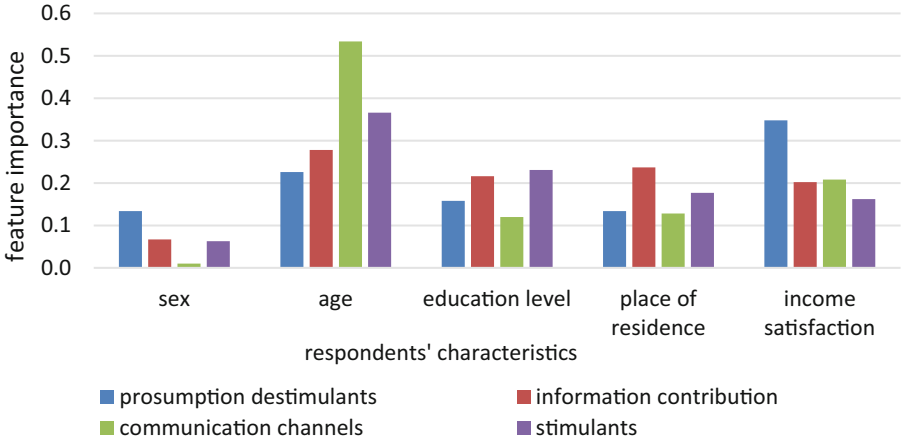


Fig. 3. The feature importance for LightGBM boosted decision trees.

```
Layer (type)                 Output Shape              Param #
=================================================================
=
 batch_normalization (Batch   (None, 5)                 20
 Normalization)
 dense (Dense)                (None, 10)                60
 dense_1 (Dense)              (None, 50)                550
 dropout (Dropout)            (None, 50)                0
 dense_2 (Dense)              (None, 25)                1275
 dropout_1 (Dropout)          (None, 25)                0
 dense_3 (Dense)              (None, 12)                312
 dense_4 (Dense)              (None, 3)                 39

=================================================================
```

Fig. 4. The architecture of the deep neural network model.

of *epochs* equaled 100, and *batchsize* 25. Figure 5. Presents the results of the training process.

The learning process is characterized by a high level of fluctuation. This may result from the fact that training data is not large enough or diverse enough. The results of DLL model are presented in Table 3.

After building all the models, the accuracy of the models' validation was compared to asses which method is the best for the prediction of factors importance in the presumption decision-making process. Table 4 summarizes the accuracy of classic classification decision trees, LightGBM-boosted decision trees, and neural networks.

Based on the information provided in Table 4, the highest accuracy was achieved for *prosumption destimulants* using DNN (0.398), for *information contribution* using Decision trees (0.497), for *communication channels* using DNN (0.586), and for *stimulants* using LightGBM (0.523). It was not as high as for the prediction of hard data because

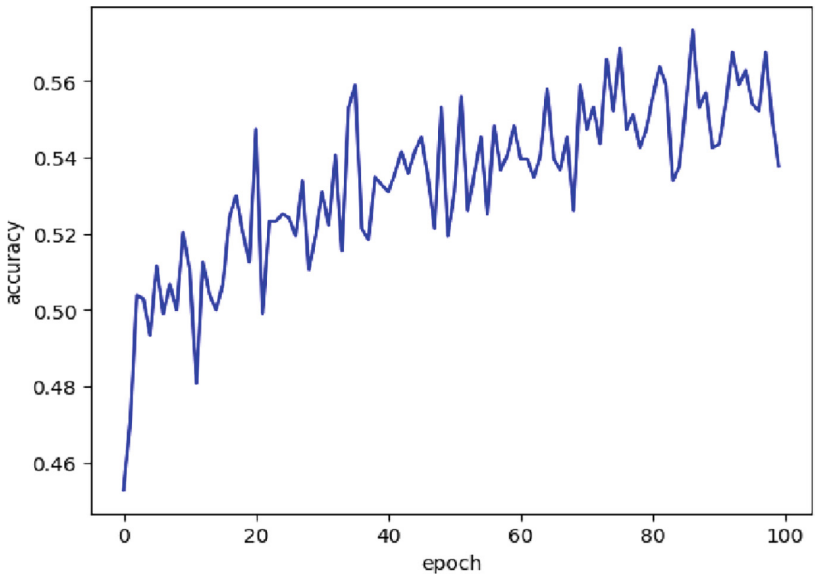


Fig. 5. The training process of DLL.

Table 3. Results of DLL model.

Parameter	prosumption destimulants	information contribution	communication channels	stimulants
Accuracy of validation	0.449	0.508	0.538	0.531
Accuracy of prediction	0.398	0.423	0.586	0.472

Table 4. The accuracy of built models.

Model	prosumption destimulants	information contribution	communication channels	stimulants
Decision trees	0.389	0.497	0.534	0.519
LightGBM	0.388	0.469	0.585	0.523
DNN	0.398	0.423	0.586	0.472

of the subjective nature of the output variable. Human behavior and social preference prediction is more complicated and harder [19].

Supervised methods based on trees were used by other authors for survey results to replace missing data in survey analysis [10], to identify factors that affect survey

response rates [4], or to classify the timing of survey responses [14]. All of these research took respondents' characteristics as input data. The models' accuracy was around 0.8, which was higher than in the research presented in this paper. However, the aim of our research and the output data were different; we aimed to predict the importance of factors influencing decision-making. The goal was much harder than predicting whether a respondent would answer the survey questions or not or how much time he would need to fill in the questionnaire. The complex nature of subjective data used in our research is reflected in the level of the models' measures.

Research considering prosumption and machine learning mainly focuses on optimizing energy production and distribution processes, taking into account various factors such as the variability of renewable energy production, energy demand, or costs associated with imbalances in the energy system. By analyzing transactional data and energy consumption patterns, machine learning models such as decision trees or neural networks can be used to forecast energy production and consumption and optimize energy planning [16, 20]. Unlike survey research, which focuses on the subjective opinions and preferences of respondents, prosumption research using machine learning emphasizes analyzing objective data and using it to make decisions in the energy field. These include the development of predictive, optimization, and classification models that can be used to improve the efficiency of energy systems and reduce energy management costs. In this way, machine learning methods in prosumption are becoming an important area of research in the context of sustainable development and efficient use of energy resources [16]. However, this topic is very far from our research.

Conducted research allows us to verify the hypotheses posed in the Introduction. The first was methodological and assumed that the machine learning or deep learning models effectively assign qualitative and quantitative respondents' characteristics to how individual factors are assessed in the prosumption decision-making process. The verification of this hypothesis is based on the models' parameters. As the parameters are not on a high level, the verification should be rather negative. However, considering that subjective assessments are more complicated to predict, the hypothesis might be verified positively, as the models' measures are quite satisfactory. More research is needed to check if other parameters of built models may influence their accuracy. The second hypothesis was cognitive and assumed that how a respondent evaluates individual decision-making factors depends on their individual characteristics. This hypothesis may be verified positively because the results showed that a respondent's age was the most important feature influencing the models in most cases. Other characteristics did not show such a clear impact on the results.

Positive verification of the posed hypotheses implies that the aim of the research was achieved. Determination of the possibility of classifying prosumer survey respondents depending on their method of assessing each decision-making factor separately using selected machine learning and deep learning methods, including decision trees, LightGBM, and deep neural networks, is possible. Still, the model parameters are not as high as expected when predicting hard data. Taking into account the nature of output data, the parameters of the models are quite satisfactory. More research is needed to improve models or test other machine learning or deep learning techniques.

5 Conclusion

The conducted research showed that it is possible to predict the importance of factors in the decision-making process of prosumers based on respondents' characteristics. The output data are subjective, so the model measures are not as high as when predicting hard data, but they are quite satisfactory. This is the main contribution of the research, as it opens the way to building models that would predict the survey answers when the response rate is low, or the respondents are unwilling to share knowledge and experience.

Connected to the above, there are practical implications for survey conductors. When there is a problem with getting the survey answers, or the answers are needed in a very short time, or the cost of conducting research is high, there is a possibility to build a model that would predict the survey answers based on a sample of collected data. This may be a solution, especially for smaller companies that do not have the money or time to conduct large market surveys.

The main limitation of the research was the use of a dataset that was not large enough and did not have enough diverse data. The augmentation of data will be performed in further research. Another limitation is using only one type of neural network. The NARX neural network can be analyzed in future research. Future research will also cover improving built models, mainly changing different hyperparameters to obtain higher accuracy. Other techniques will also be used to predict factors importance in the decision-making process. Finally, the authors would like to try to build models that would directly connect respondents' answers with factors' importance, simplifying the statistical analysis of the results of such surveys.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Filho, W.L., Trevisan, L.V., Salvia, A.L., Mazutti, J., Dibbern, T., de Maya, S.R., Bernal, E.F., Eustachio, J.H.P.P., Sharifi, A., Alarcón-del-Amo, M. del C., Kushnir, I.: Prosumers and sustainable development: An international assessment in the field of renewable energy. Sustain. Futur. **7**, (2024). https://doi.org/10.1016/j.sftr.2024.100158
2. Maciaszczyk, M., Kocot, M.: Behavior of online prosumers in organic product market as determinant of sustainable consumption. Sustainability (Switzerland). **13**, 1–16 (2021). https://doi.org/10.3390/su13031157
3. Ziemba, E., Eisenbardt, M.: Incentives encouraging prosumers to knowledge sharing-framework based on Polish study. Online J. Appl. Knowl. Manag. **4**, 146–166 (2016)
4. Han, J., Fang, M., Ye, S., Chen, C., Wan, Q., Qian, X.: Using decision tree to predict response rates of consumer satisfaction, attitude, and loyalty surveys. Sustainability (Switzerland). 11, (2019). https://doi.org/10.3390/su11082306
5. Das, A.: Logistic Regression. In: Encyclopedia of Quality of Life and Well-Being Research. pp. 3985–3986. Springer International Publishing, Cham (2023). https://doi.org/10.1007/978-3-031-17299-1_1689
6. Buskirk, T.D.: Surveying the Forests and Sampling the Trees: An overview of Classification and Regression Trees and Random Forests with applications in Survey Research. Surv Pract. 11, 1–13 (2018). https://doi.org/10.29115/SP-2018-0003

7. Shin, H., Cho, S.: Response modeling with support vector machines. Expert Syst. Appl. **30**, 746–760 (2006). https://doi.org/10.1016/j.eswa.2005.07.037
8. Alhudhaif, A., Polat, K.: Spatio-temporal characterisation and compensation method based on CNN and LSTM for residential travel data. PeerJ Comput Sci. **10**, e2035 (2024). https://doi.org/10.7717/peerj-cs.2035
9. Su, D., Zhang, X., He, K., Chen, Y.: Use of machine learning approach to predict depression in the elderly in China: A longitudinal study. J. Affect. Disord. **282**, 289–298 (2021). https://doi.org/10.1016/j.jad.2020.12.160
10. Kern, C., Klausch, T., Kreuter, F.: Tree-based machine learning methods for survey research. Surv Res Methods. **13**, 73–93 (2019)
11. Suljic, M., Osmanbegovic, E., Dobrović, Ž: Common metamodel of questionnaire model and decision tree model. Res. Appl. Econ. **10**, 106 (2018). https://doi.org/10.5296/rae.v10i3.13540
12. Kim, S.S., Timothy, D.J., Hwang, J.: Understanding Japanese tourists' shopping preferences using the Decision Tree Analysis method. Tour. Manag. **32**, 544–554 (2011). https://doi.org/10.1016/j.tourman.2010.04.008
13. Hamoud, A.K., Hashim, A.S., Awadh, W.A.: Predicting student performance in higher education institutions using decision tree analysis. Int. J. Interact. Multimed. Artif. Intell. **5**, 26 (2018). https://doi.org/10.9781/ijimai.2018.02.004
14. Lin, C., Yang, H., Kuo, L.: Behaviour analysis of internet survey completion using decision trees: An exploratory study. Online Inf. Rev. **33**, 117–134 (2009)
15. Watanabe, F., Kawaguchi, T., Ishizaki, T., Takenaka, H., Nakajima, T.Y., Imura, J.: Day-ahead Strategic Marketing of Energy Prosumption: A Machine Learning Approach Based on Neural Networks. In: 2019 18th European Control Conference (ECC). pp. 3910–3915. IEEE (2019). https://doi.org/10.23919/ECC.2019.8796040
16. Silva, W.N., Henrique, L.F., Silva, A.F.P. da C., Dias, B.H., Soares, T.A.: Market models and optimization techniques to support the decision-making on demand response for prosumers. Electr. Power Syst. Res. **210**, 108059 (2022). https://doi.org/10.1016/j.epsr.2022.108059
17. Vona, G.: International comparative analysis of prosumers in selected fields of energy use and further customer preferences in environmental issues. Hungarian Statistical Review. **6**, 3–31 (2023)
18. Nadolny, M., Walaszczyk, E., Lopacinski, K.: Key preferences of IT system users in the context of prosumer attitudes. In: Procedia Computer Science. pp. 3106–3115. Elsevier B.V. (2022). https://doi.org/10.1016/j.procs.2022.09.369
19. Grossmann, I., Rotella, A., Hutcherson, C.A., Sharpinskyi, K. et al.: Insights into the accuracy of social scientists' forecasts of societal change. Nat Hum Behav. **7**, 484–501 (2023). https://doi.org/10.1038/s41562-022-01517-1
20. Ruiz-Abellón, M.C., Fernández-Jiménez, L.A., Guillamón, A., Falces, A., García-Garre, A., Gabaldón, A.: Integration of demand response and short-term forecasting for the management of prosumers' demand and generation. Energies (Basel). **13**, 11 (2019). https://doi.org/10.3390/en13010011

Intelligent Multiple Criteria Decision Analysis Methods

RRF-EDAS An Extended Approach Free from the Rank Reversal Paradox

Bartłomiej Kizielewicz[1(✉)], Arkadiusz Marchewka[2], and Wojciech Sałabun[1,3]

[1] National Institute of Telecommunications, ul. Szachowa 1, 04-894 Warsaw, Poland
{b.kizielewicz,w.salabun}@il-pib.pl
[2] Institute of Management, University of Szczecin, ul. Cukrowa 8, 71-004 Szczecin, Poland
arkadiusz.marchewka@usz.edu.pl
[3] West Pomeranian University of Technology in Szczecin, ul. Żołnierska 49, 71-210 Szczecin, Poland

Abstract. The Evaluation based on Distance from Average Solution (EDAS) method is a prominent Multi-Criteria Decision-Making (MCDM) technique used to rank alternatives based on their distance from an average solution. While EDAS is effective and requires fewer calculations than other MCDM methods, it suffers from the rank reversal paradox. This paradox occurs when minor changes in the input data lead to significant and illogical shifts in the rankings of alternatives, undermining the reliability of the decision-making process. We propose the Rank Reversal Free EDAS (RRF-EDAS) approach to address this issue. RRF-EDAS modifies the traditional EDAS method by introducing static decision boundaries for criteria determined by experts. These boundaries define each criterion's minimum and maximum values, creating a stable reference point for normalizing the decision matrix. This adjustment prevents the rank reversal paradox by ensuring that the normalization process is based on fixed values rather than dynamic changes in the data. The proposed RRF-EDAS approach was tested using Spearman's weighted correlation coefficient to analyze rank stability. Results show that RRF-EDAS effectively eliminates rank reversal, providing more consistent and reliable rankings than the classical EDAS method. While using fixed boundaries has limitations in dynamic environments, RRF-EDAS offers a robust solution for scenarios where stability and trust in the ranking process are crucial.

Keywords: MCDA · Rank reversal · EDAS · Reference point

1 Introduction

Multi-Criteria Decision-Making (MCDM) is the decision-making process in situations with multiple, often conflicting, criteria for evaluating alternative solutions [26]. MCDM plays a key role due to the complexity of decisions that require consideration of multiple aspects of a problem simultaneously. Through the use of MCDM, it is possible to make more balanced and reasoned decisions that better meet diverse needs and priorities. In addition, MCDM is extremely useful when finding an optimal solution that maximizes benefits while minimizing costs, risks, or other undesirable effects is necessary [3]. Using

M. Hernes et al. (Eds.): ECAI 2024, LNNS 1217, pp. 199–212, 2025.
https://doi.org/10.1007/978-3-031-78465-1_17

this approach increases the transparency of the decision-making process, which in turn can lead to greater acceptance and trust from stakeholders.

Many MCDM approaches are widely used in sustainability [31], industry 4.0 [6], medicine [20], green governance [25], and others [34]. The most popular MCDM approaches are echnique of Order Preference Similarity to the Ideal Solution (TOPSIS) [2], ELimination Et Choice Translating REality (ELECTRE) [10], Preference Ranking Organization METhod for Enrichment Evaluations (PROMETHEE) [4], Analytic Hierarchy Process (AHP) [36], Analytic Network Process (ANP) [13], or VlseKriterijuska Optimizacija I Komoromisno Resenje (VIKOR) [18]. These approaches are the classic MCDM approaches. Their logic is mainly based on ranking or classifying alternatives based on their performance on various criteria and selecting the most preferred option. TOPSIS, for example, is based on identifying the option closest to the ideal and farthest from the anti-ideal. AHP and ANP are based on hierarchization and analysis of the relationships between criteria. However, new MCDM approaches such as Stable Preference Ordering Towards Ideal Solution (SPOTIS) [7], Best-Worst Method (BWM) [22], RANking COMparison (RANCOM) [33], Multi-Attributive Border Approximation area Comparison (MABAC) [19], and Evaluation based on Distance from Average Solution (EDAS) [12] are also being developed. These new methods often introduce innovative ways of aggregating assessments, weighting criteria, or analyzing uncertainty, allowing for a more flexible and user-specific approach to solving decision-making problems.

The Evaluation based on Distance from Average Solution (EDAS) method, developed by Keshavarz Ghorabaee et al. in 2015 [12], ranks alternatives in complex multi-criteria decision-making problems. Unlike traditional methods such as TOPSIS and VIKOR, which rely on distances from ideal and anti-ideal solutions, EDAS uses average values as a benchmark. The logic behind EDAS is to calculate positive (PDA) and negative (NDA) distances from average values, which allows for a more realistic and practical ordering of alternatives [28].

However, there needs to be a solution in the EDAS method in the form of the paradox of reversal of rankings. The paradox of reversal of rankings is when a change in the value of one of the criteria leads to a non-intuitive change in the ranking order of the alternatives. This means that an alternative that was initially better can suddenly worsen despite a seemingly minor change in the data. This paradox is encountered in many MCDM methods, such as TOPSIS, VIKOR, AHP, and ELECTRE. The paradox of reversed rankings is undesirable because it undermines confidence in the results of decision analysis and can lead to the selection of less favorable alternatives. It can also introduce uncertainty and complicate the decision-making process, as users cannot be sure of the stability and reliability of ranking results.

Therefore, this paper proposes the Rank Reversal Free EDAS (RRF-EDAS) approach, which addresses the rank reversal paradox. In order to eliminate this problem, an additional parameter has been introduced in the EDAS method in the form of decision boundaries for criteria. These boundaries are defined by decision experts or decision makers, who determine the most minor and most significant value of the selected criteria in the decision problem. Using these boundaries, an average point is determined, which is used to normalize the decision matrix. Using normalization based on static data makes obtaining a model resistant to the rank reversal paradox possible. The proposed

approach has been tested for reversal of rankings using Spearman's weighted correlation coefficient. In addition, the RRF-EDAS approach was compared with the classical EDAS method to show the differences and the susceptibility of EDAS to rank reversal paradox.

The rest of the paper is organized as follows. Section 2 presents a literature review related to the phenomenon of reverse rankings. Section 3 presents preliminary assumptions, i.e., a description of the EDAS method and the weighted Spearaman correlation coefficient. Section 4 presents the proposed RRF-EDAS approach. Section 5 presents a study on the phenomenon of reversal of rankings, where the phenomenon in the EDAS approach is transferred and the absence of the phenomenon in the RRF-EDAS approach is shown. Section 6 presents a discussion of the reversal of rankings phenomenon and the proposed approach. Important aspects of the proposed approach and its limitation are also pointed out in this section. Finally, Sect. 7 presents conclusions and future research directions.

2 Literature Review

Rank reversal paradox is a phenomenon that occurs in a Multi-Criteria Decision Making (MCDM) process when the ranking of an option or decision changes after an alternative is added, removed, or modified [1]. It is particularly problematic because it undermines the consistency and stability of the analysis results, leading to uncertainty and potential errors in decision-making [30]. In practice, this means that decisions based on the analysis results may be unreliable, which is unacceptable in the context of strategic planning or resource management.

Rank reversal occurs in many popular MCDM methods, including the Analytic Hierarchy Process (AHP) [17, 29], where adding or removing alternatives can change the hierarchy of existing options. A similar problem arises in the Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS) method [8], where new alternatives can affect the relative positions of existing options relative to the ideal solution. Multi-Attribute Utility Theory (MAUT) is also susceptible to rank reversal due to changes in preferences or the addition of new options [9], which can lead to changes in ratings and rankings.

To counter the problem of rank reversal, several methods have been developed to minimize its impact or eliminate the phenomenon. One such method is the Reference Ideal Method (RIM) [5], which seeks to preserve rankings' stability by considering decision boundaries and referring to a reference ideal point. Another method is the Stable Preference Order Towards Ideal Solution (SPOTIS) [7], which stabilizes the ranking by comparing alternatives with the ideal solution, thereby reducing the impact of new options on the existing ranking. The Characteristic Objects Method (COMET) [23] is based on distance from characteristic objects and their values, eliminating rank reversal and leading to more realistic results than traditional methods such as TOPSIS.

In addition to the methods mentioned above, specific normalization techniques and hybrid approaches effectively deal with rank reversal. R-normalization, for example, includes procedures such as Max-Min normalization and Max normalization, which minimize the impact of dependence on alternatives by adjusting parameters accordingly

[24]. G-normalization, used in the G-TOPSIS method, uses a cumulative normal distribution function to normalize criteria values, which ensures that values remain constant regardless of the number of alternatives and stabilizes the ranking [27]. Fuzzy normalization [14, 16], based on fuzzy set theory, allows more flexible evaluations, stabilizing rankings and reducing the impact of rank reversal. Finally, hybrid approaches, such as TOPSIS-COMET [15] and VIKOR-COMET (V-COMET) [32], combine the advantages of different methods to increase the stability and consistency of rankings, minimizing the risk of rank reversal and providing more reliable results in multi-criteria analysis.

3 Preliminaries

3.1 The EDAS Method

In the Evaluation based on Distance from Average Solution (EDAS) method, the Positive Distance from Average (PDA) and Negative Distance from Average (NDA) are calculated to determine the alternative preference value [12]. The optimal alternative has a higher distance from the worst solution and a lower distance from the ideal solution [11]. Its advantage lies in requiring fewer calculations to achieve the results [21]. EDAS is designed to evaluate decision alternatives according to the following steps:

Step 1. Define a decision matrix of dimension $n \times m$, where n is the number of alternatives, and m is the number of criteria (1).

$$X_{ij} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1m} \\ x_{21} & x_{22} & \dots & x_{2m} \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{nm} \end{bmatrix} \tag{1}$$

Step 2. Calculate the average solution for each criterion according to the formula (2).

$$AV_j = \frac{\sum_{i=1}^{n} X_{ij}}{n} \tag{2}$$

Step 3. Calculating the positive distance from the mean solution and the negative distance from the mean solution for the alternatives. When the criterion is of profit type, the negative distance and the positive distance are calculated using equations (3) and (4), while when the criterion is of cost type, the distances are calculated using formulas (5) and (6).

$$NDA_{ij} = \frac{\max\left(0, \left(AV_j - X_{ij}\right)\right)}{AV_j} \tag{3}$$

$$PDA_{ij} = \frac{\max\left(0, \left(X_{ij} - AV_j\right)\right)}{AV_j} \tag{4}$$

$$NDA_{ij} = \frac{\max\left(0, \left(X_{ij} - AV_j\right)\right)}{AV_j} \tag{5}$$

$$PDA_{ij} = \frac{\max\left(0, \left(AV_j - X_{ij}\right)\right)}{AV_j} \tag{6}$$

Step 4. Calculate the weighted sums of *PDA* and *NDA* for each decision variant using equations (7) and (8).

$$\mathrm{A}SP_i = \sum_{j=1}^{m} w_j PDA_{ij} \tag{7}$$

$$SN_i = \sum_{j=1}^{m} w_j NDA_{ij} \tag{8}$$

Step 5. Normalize the weighted sums of negative and positive distances using equations (9) and (10).

$$NSN_i = 1 - \frac{SN_i}{\max_i(SN_i)} \tag{9}$$

$$NSP_i = \frac{SP_i}{\max_i(SP_i)} \tag{10}$$

Step 6. Calculate the evaluation score (*AS*) for each alternative using the formula (11). A higher point value determines a higher ranking alternative.

$$AS_i = \frac{1}{2}(NSP_i + NSN_i) \tag{11}$$

3.2 Weighted Spearman's Correlation Coefficient

The weighted Spearman rank correlation coefficient (r_w), introduced by [37], is based on the traditional Spearman coefficient but includes weighting factors. This modification increases the effect of changes in leading positions on the total correlation score. The correlation is calculated between two rankings, each with a size of N, where x_i is the position in the first ranking and y_i in the second (refeq). The coefficient of r_w ranges from -1 to 1: one indicates identical rankings, minus one inverted, and zero no correlation. This approach gives different rankings different levels of significance. The r_w coefficient is mathematically defined as follows (12):

$$r_W = 1 - \frac{6 \cdot \sum (x_i - y_i)^2 ((n - x_i + 1) + (n - y_i + 1))}{n \cdot (n^3 + n^2 - n - 1)} \tag{12}$$

4 Proposed Approach: The RRF-EDAS Method

This section presents the RRF-EDAS approach, which addresses the rank reversal paradox. The approach has been modified to determine the mean solution using static bounds, reducing the rank reversal paradox. In order to eliminate it, the normalization format in the last step of the classic EDAS method was also changed, using maximum normalization on the determined sum distances. The algorithm of the proposed RRF-EDAS approach consists of the following steps:

Step 1. Define a decision matrix of dimension $n \times m$, where n is the number of alternatives, and m is the number of criteria (13).

$$X_{ij} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1m} \\ x_{21} & x_{22} & \dots & x_{2m} \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{nm} \end{bmatrix} \tag{13}$$

Step 2. Provide boundary (B) values for each criterion.

$$B_j = \begin{bmatrix} b_1^{Min} & b_1^{Max} \\ b_2^{Min} & b_2^{Max} \\ \dots & \dots \\ b_m^{Min} & b_m^{Max} \end{bmatrix} \tag{14}$$

Step 3. Calculate the average solution for each criterion based on the decision boundaries according to the formula (15).

$$AV_j = \frac{1}{2}\left(b_j^{Max} - b_j^{Min}\right) \tag{15}$$

Step 4. Calculating the positive distance from the mean solution and the negative distance from the mean solution for the alternatives. When the criterion is of profit type, the negative distance and the positive distance are calculated using equations (16) and (17), while when the criterion is of cost type, the distances are calculated using formulas (18) and (19).

$$NDA_{ij} = \frac{\max\left(0, \left(AV_j - X_{ij}\right)\right)}{AV_j} \tag{16}$$

$$PDA_{ij} = \frac{\max\left(0, \left(X_{ij} - AV_j\right)\right)}{AV_j} \tag{17}$$

$$NDA_{ij} = \frac{\max\left(0, \left(X_{ij} - AV_j\right)\right)}{AV_j} \tag{18}$$

$$PDA_{ij} = \frac{\max\left(0, \left(AV_j - X_{ij}\right)\right)}{AV_j} \tag{19}$$

Step 5. Calculate the weighted sums of PDA and NDA for each decision variant using equations (20) and (21).

$$\mathrm{A}SP_i = \sum_{j=1}^{m} w_j PDA_{ij} \tag{20}$$

$$SN_i = \sum_{j=1}^{m} w_j NDA_{ij} \tag{21}$$

Step 6. Calculate the normalized assessment score (*NAS*) for each alternative using the formula (22). A higher point value determines a higher ranking alternative.

$$NAS_i = \frac{DST_i}{max(DST_i)} \tag{22}$$

where

$$DST_i = (1 - SN_i) + SP_i \tag{23}$$

The entire RRF-EDAS algorithm can be represented using Fig. 1.

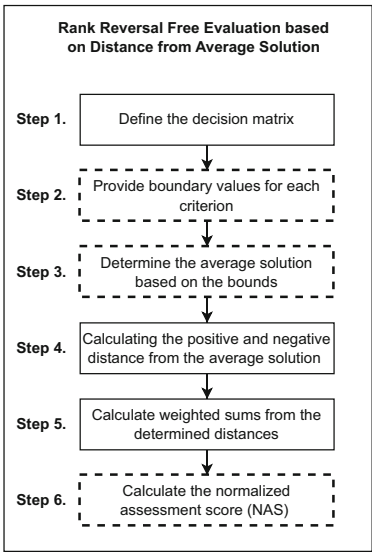


Fig. 1. RRF-EDAS algorithm.

5 Research on the Rank Reversal Paradox

In this paper, a study is presented to resolve the paradox of ranking reversals in the Evaluation based on the Distance from Average Solution (EDAS) method. To do this, a proposed Ranking Reversal Free EDAS (RRF-EDAS) approach, which is based on an assumption related to the boundaries of the decision problem, was used. Suppose we are dealing with a decision problem consisting of two criteria and twelve alternatives. In the framework of the present work, the data for the problem were drawn from the interval [0, 1]. In the work, the reference set means the set of 12 alternatives shown in Table 1.

With the proposed RRF-EDAS approach, decision boundaries for each criterion were decided to be set at values from 0 to 1. This decision was made due to the range of drawing values for the set of alternatives. In real problems, it is recommended to use historical boundary values or values consistent with the model.

Table 1. Drawn decision matrix consisting of 2 criteria ($C_1 - C_2$) and 12 alternatives ($A_1 - A_{12}$).

C_i	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}	A_{11}	A_{12}
C_1	0.409533	0.083141	0.481854	0.162950	0.021360	0.152046	0.691076	0.337563	0.055135	0.300725	0.996651	0.879061
C_2	0.148242	0.738495	0.252571	0.573244	0.752337	0.480851	0.112324	0.091584	0.172909	0.649881	0.070409	0.905634

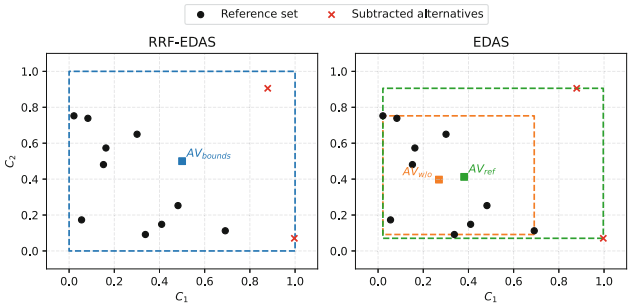
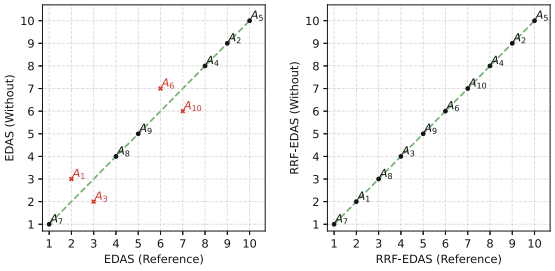


Fig. 2. Boundaries of EDAS and RRF-EDAS models with alternatives and average solutions.

With the help of Fig. 2, the limits of the considered decision models (EDAS and RRF-EDAS) are presented, as well as the average solutions and the set of alternatives. As can be seen, the RRF-EDAS approach uses boundary values to determine the average solution. As a result, alternatives are normalized statically rather than dynamically, as in the EDAS approach. The EDAS approach determines the average solution from a set of alternatives. Because of this, the EDAS model may be more sensitive than approaches such as TOPSIS or VIKOR, which refer to ideal and negatively ideal solutions. Determining the mean solution by EDAS using alternatives is one of the steps leading to the paradox of reversed rankings in this model. By excluding the alternative(s), the reference, i.e., the mean solution, also changes; thus, normalizing the values of the alternatives also changes. Therefore, the RRF-EDAS approach uses fixed/static values.



(a) EDAS. (b) RRF-EDAS.

Fig. 3. Comparison of rankings from EDAS approaches from a reference set and a set with exclusions.

Suppose we exclude two boundary alternatives from the reference set. At this point, the model's boundary will change, as will the average value of the alternatives. This leads to a change in the mean solution and normalized values. By evaluating the entire set of alternatives and ranking only 10 alternatives, we get a different ranking using the EDAS approach than by evaluating 10 alternatives and ranking them. In contrast, with the RRF-EDAS approach based on fixed values, the model evaluates the alternatives

with reference to the same mean solution. The paradox of reversal of rankings in the model is noticeable in the rankings shown by Fig. 3a.

In contrast, the absence of this paradox in the proposed approach is shown using Fig. 3b. In addition, the rankings themselves are presented using Table 2. In it, it can be seen that the 4 positions in the EDAS approach were changed for pairs of alternatives (A_1, A_3) and (A_6, A_{10}). In contrast, these items remained unchanged in the RRF-EDAS approach.

Table 2. Summary of rankings from EDAS and RRF-EDAS for an example with reverse rankings.

A_i	EDAS		RRF-EDAS	
	Reference	Without	Reference	Without
A_1	2	3	2	2
A_2	9	9	9	9
A_3	3	2	4	4
A_4	8	8	8	8
A_5	10	10	10	10
A_6	6	7	6	6
A_7	1	1	1	1
A_8	4	4	3	3
A_9	5	5	5	5
A_{10}	7	6	7	7

Given the broader spectrum of possible paradox reversal of rankings, the EDAS and RRF-EDAS approaches were also examined for exclusions of subsequent alternatives. Using Fig. 4, an analysis of the exclusions of each alternative from the reference set is presented. For each exclusion, the resulting ranking was examined by correlation with the reference rank in which the excluded alternative was not included. As can be seen, only in two cases of exclusion of alternatives, i.e. alternative A_6 and alternative A_9, did there not reversal of rankings. This means the EDAS approach is very sensitive to changing the reference set by adding/removing an alternative. The same analysis was also performed on the RRF-EDAS approach, visualized using Fig. 5. No reversal of rankings was shown for this approach, which means that this approach is more stable than the EDAS approach.

In addition, a simulation was conducted that included various randomly generated decision matrices and excluded 2, 5, 10, 25, and 50 alternatives. The decision matrices were drawn 1,000 times for each scenario, excluding alternatives. The study also included different criteria (2, 5, 10, 20). However, due to the similar results in each criterion studied, it was decided to present only an example for two criteria.

To analyze the reversal of the rankings in the EDAS and RRF-EDAS methods for each alternative exclusion scenario, box plots showing the values of the weighted Spearman correlation coefficient were used in Fig. 6. In addition, the data were described using

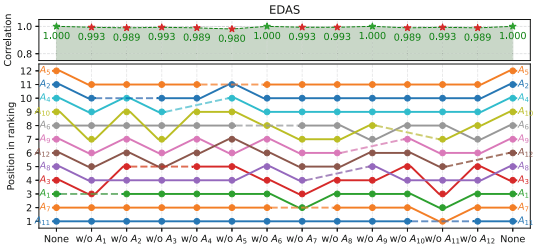


Fig. 4. Analysis of the effect of exclusions on the rank reversal paradox (model : EDAS, measure : r_w).

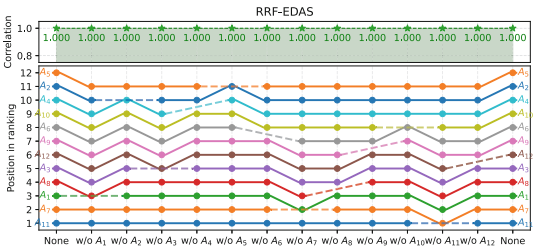


Fig. 5. Analysis of the effect of exclusions on the rank reversal paradox (model : RRF-EDAS, measure : r_w).

a statistical table that included characteristics of the data distribution, such as mean, standard deviation, minimum and maximum values, and quartiles.

Table 3 shows the statistical results for both methods, RRF-EDAS and EDAS. It is worth noting that the EDAS method is prone to reverse rankings. In some cases, the rankings of alternatives may be reversed relative to the decision maker's expectations. In contrast, the RRF-EDAS method is characterized by the absence of this phenomenon. The average value of the weighted Spearman correlation coefficient for the RRF-EDAS method is 1.000, suggesting an excellent correlation between the rankings obtained in the various simulations. In contrast, for the EDAS method, the average is 0.999, indicating a very high correlation but slightly lower than for RRF-EDAS. This means that in the case of the EDAS approach, the phenomenon of reversal of rankings occurs, while in the case of the RRF-EDAS approach, this phenomenon does not occur.

6 Discussion

This study addresses the rank reversal paradox in the EDAS multi-criteria decision-making (MCDM) method and proposes the RRF-EDAS approach as a solution. The rank reversal paradox, where small changes in inputs lead to significant shifts in alternative rankings, is a known issue in many MCDM methods, including TOPSIS and VIKOR. While these methods use normalization techniques to mitigate this problem, such as fuzzy or R-normalization, these solutions are not directly applicable to EDAS due to its reliance on an average solution and end-process normalization.

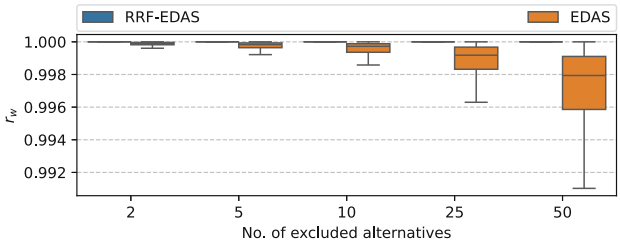


Fig. 6. Analysis of rank reversal in EDAS and RRF-EDAS methods under different numbers of alternative exclusions (random sets of 100 alternatives, two criteria).

Table 3. Summary of statistical metrics for reverse ranking analysis in EDAS and RRF-EDAS methods with different numbers of excluded alternatives (random sets of 100 alternatives, two criteria).

Approach	Mean	STD	Min	25%	50%	75%	Max
RRF-EDAS	1.000000	0.000000	1.000000	1.000000	1.000000	1.000000	1.000000
EDAS	0.998947	0.002064	0.964692	0.998955	0.999691	0.999893	1.000000

The RRF-EDAS approach, by using constant values, eliminates the rank reversal paradox, offering more stable results. However, this method, like other fixed-boundary models such as R-TOPSIS [24], R-VIKOR [35], SPOTIS [7], RIM [5], or COMET [23], has limitations in evaluating alternatives with criteria outside predefined ranges. Although these fixed-boundary models provide greater stability compared to dynamic value-based methods, their applicability may be restricted in evolving real-world scenarios where criteria values fluctuate. Nonetheless, in contexts where result stability is crucial, RRF-EDAS and similar approaches may be preferable.

7 Conclusions and Future Research

This paper introduces the RRF-EDAS approach to address the rank reversal paradox associated with the EDAS method. The analysis indicates that EDAS can produce unexpected and illogical rankings of alternatives, a problem that RRF-EDAS successfully mitigates. Tested on a generic case, RRF-EDAS demonstrated stability while adhering to model constraints, suggesting it effectively prevents reverse rankings and provides more reliable results. Future research should explore the applicability of RRF-EDAS to scenarios with uncertainty, assess its performance under varying inputs, and apply it to real-world decision-making contexts like project management and strategic planning, to evaluate its practical utility and identify potential areas for enhancement.

Acknowledgement. The work was supported by the National Science Centre 2021/41/B/HS4/01296.

References

1. Aires, R.F.D.F., Ferreira, L.: The rank reversal problem in multi-criteria decision making: a literature review. Pesqui. Oper. **38**, 331–362 (2018)
2. Behzadian, M., Otaghsara, S.K., Yazdani, M., Ignatius, J.: A state-of the-art survey of topsis applications. Expert Syst. Appl. **39**(17), 13051–13069 (2012)
3. Bonissone, P.P., Subbu, R., Lizzi, J.: Multicriteria decision making (MCDM): a framework for research and applications. IEEE Comput. Intell. Mag. **4**(3), 48–61 (2009)
4. Brans, J.P., De Smet, Y.: Promethee Methods. Multiple Criteria Decision Analysis: State of the Art Surveys, pp. 187–219 (2016)
5. Cables, E., Lamata, M.T., Verdegay, J.L.: Rim-reference ideal method in multicriteria decision making. Inf. Sci. **337**, 1–10 (2016)
6. Chang, S.C., Chang, H.H., Lu, M.T.: Evaluating industry 4.0 technology application in SMEs: Using a hybrid MCDM approach. Mathematics **9**(4), 414 (2021)
7. Dezert, J., Tchamova, A., Han, D., Tacnet, J.M.: The spotis rank reversal free method for multi-criteria decision-making support. In: 2020 IEEE 23rd International Conference on Information Fusion (FUSION). pp. 1–8. IEEE (2020)
8. García-Cascales, M.S., Lamata, M.T.: On rank reversal and topsis method. Math. Comput. Model. **56**(5–6), 123–132 (2012)
9. Gass, S.I.: Model world: ihe great debate-maut versus ahp. Interfaces **35**(4), 308–312 (2005)
10. Govindan, K., Jepsen, M.B.: Electre: a comprehensive literature review on methodologies and applications. Eur. J. Oper. Res. **250**(1), 1–29 (2016)
11. Huang, Y., Lin, R., Chen, X.: An enhancement EDAS method based on prospect theory. Technol. Econ. Dev. Econ. **27**(5), 1019–1038 (2021)
12. Keshavarz Ghorabaee, M., Zavadskas, E.K., Olfat, L., Turskis, Z.: Multi-criteria inventory classification using a new method of evaluation based on distance from average solution (EDAS). Informatica **26**(3), 435–451 (2015)
13. Khan, A.U., Ali, Y.: Analytical hierarchy process (AHP) and analytic network process methods and their applications: a twenty year review from 2000–2019: Ahp & anp techniques and their applications: twenty years review from 2000 to 2019. Int. J. Anal. Hierarchy Process. **12**(3) (2020)
14. Kizielewicz, B., Dobryakova, L.: Stochastic triangular fuzzy number (S-TFN) normalization: a new approach for nonmonotonic normalization. Procedia Comput. Sci. **225**, 4901–4911 (2023)
15. Kizielewicz, B., Shekhovtsov, A., Sałabun, W.: A new approach to eliminate rank reversal in the MCDA problems. In: International Conference on Computational Science, pp. 338–351. Springer (2021)
16. Kizielewicz, B., Więckowski, J., Paradowski, B., Sałabun, W.: Dealing with nonmonotonic criteria in decision-making problems using fuzzy normalization. In: International Conference on Intelligent and Fuzzy Systems, pp. 27–35. Springer (2022)
17. Maleki, H., Zahir, S.: A comprehensive literature review of the rank reversal phenomenon in the analytic hierarchy process. J. Multi-Criteria Decis. Anal. **20**(3–4), 141–155 (2013)
18. Mardani, A., Zavadskas, E.K., Govindan, K., Amat Senin, A., Jusoh, A.: Vikor technique: a systematic review of the state of the art literature on methodologies and applications. Sustainability **8**(1), 37 (2016)
19. Pamučar, D., Ćirović, G.: The selection of transport and handling resources in logistics centers using multi-attributive border approximation area comparison (MABAC). Expert Syst. Appl. **42**(6), 3016–3028 (2015)
20. Pamučar, D., Puška, A., Stević, Ž, Ćirović, G.: A new intelligent MCDM model for HCW management: The integrated BWM-MABAC model based on d numbers. Expert Syst. Appl. **175**, 114862 (2021)

21. Rashid, T., Ali, A., Chu, Y.M.: Hybrid BW-EDAS MCDM methodology for optimal industrial robot selection. PLoS ONE **16**(2), e0246738 (2021)
22. Rezaei, J.: Best-worst multi-criteria decision-making method. Omega **53**, 49–57 (2015)
23. Sałabun, W.: The characteristic objects method: a new distance-based approach to multicriteria decision-making problems. J. Multi-Criteria Decis. Anal. **22**(1–2), 37–50 (2015)
24. Seiti, H., Hafezalkotob, A.: Developing the R-TOPSIS methodology for risk-based preventive maintenance planning: a case study in rolling mill company. Comput. Ind. Eng. **128**, 622–636 (2019)
25. Soni, V., Dey, P.K., Anand, R., Malhotra, C., Banwet, D.K.: Digitizing grey portions of e-governance. Transform. Gov. People Process. Policy **11**(3), 419–455 (2017)
26. Thakkar, J.J.: Multi-criteria decision making, Vol. 336. Springer (2021)
27. Tiwari, R.K., Kumar, R.: G-topsis: a cloud service selection framework using gaussian topsis for rank reversal problem. J. Supercomput. **77**(1), 523–562 (2021)
28. Torkayesh, A.E., Deveci, M., Karagoz, S., Antucheviciene, J.: A state-of-the-art survey of evaluation based on distance from average solution (EDAS): Developments and applications. Expert Syst. Appl. **221**, 119724 (2023)
29. Tu, J., Wu, Z.: Analytic hierarchy process rank reversals: causes and solutions. Ann. Oper. Res., pp. 1–25 (2023)
30. Wang, Y.M., Luo, Y.: On rank reversal in decision analysis. Math. Comput. Model. **49**(5–6), 1221–1229 (2009)
31. Wątróbski, J., Bączkiewicz, A., Rudawska, I.: A strong sustainability paradigm based analytical hierarchy process (SSP-AHP) method to evaluate sustainable healthcare systems. Ecol. Indic. **154**, 110493 (2023)
32. Wątróbski, J., Bączkiewicz, A., Sałabun, W.: New multi-criteria method for evaluation of sustainable res management. Appl. Energy **324**, 119695 (2022)
33. Więckowski, J., Kizielewicz, B., Shekhovtsov, A., Sałabun, W.: Rancom: A novel approach to identifying criteria relevance based on inaccuracy expert judgments. Eng. Appl. Artif. Intell. **122**, 106114 (2023)
34. Więckowski, J., Wątróbski, J.: How to determine complex MCDM model in the comet method? automotive sport measurement case study. Procedia Comput. Sci. **192**, 376–386 (2021)
35. Yang, W., Wu, Y.: A new improvement method to avoid rank reversal in vikor. IEEE Access **8**, 21261–21271 (2020)
36. Yu, D., Kou, G., Xu, Z., Shi, S.: Analysis of collaboration evolution in AHP research: 1982–2018. Int. J. Inf. Technol. Decis. Mak. **20**(01), 7–36 (2021)
37. Zheng, H., Liu, H., Li, H., Dou, W., Wang, X.: Weighted gene co-expression network analysis identifies a cancer-associated fibroblast signature for predicting prognosis and therapeutic responses in gastric cancer. Front. Mol. Biosci. **8**, 744677 (2021)

MCDM Approach to Quality Assessment of Functioning of e-Commerce Platforms Operating in Poland

Paula Bajdor(✉)

Czestochowa University of Technology, Czestochowa, Poland
paula.bajdor@pcz.pl

Abstract. In the beginning, the popularity of Internet shopping was determined by the ability to make purchases at any time; over time, attention began to be paid to other criteria that influence the assessment of the quality of functioning of e-commerce platforms. The analysis of available literature and own research enabled the identification of a total of 33 criteria grouped into seven groups: functionality, security/privacy, assurance, convenience, personalization, customer satisfaction, and behavioral intentions. Bearing in mind that MCDM methods are widely used in quality assessment processes, this article uses the AHP and Entropy method, which made it possible to carry out analyses regarding the impact of changes in one group of criteria on the others and the overall quality level of the functioning of e-commerce platforms. The obtained results indicated the completeness of criteria set identified, as all of them are relevant to assess the quality of functioning of e-commerce platforms.

Keywords: AHP · quality assessment · e-commerce · management · e-commerce platform

1 Introduction

E-commerce is a solution that brings effects, services, or information between controllers and sellers via electronic devices connected to the Internet [1]. What distinguishes e-commerce from traditional trade is that buyers need the opportunity to physically check the product before purchasing it; they decide to buy it based on its description, photo, ratings, and reviews of other users [2].

In 1979, the first online purchase was made in Great Britain. Currently, the e-commerce market is worth approximately USD 8 trillion and is expected to reach over USD 18 trillion in 2029. This means that online shopping is currently the most common form of shopping. Over the next few years, it will become a primary sales strategy, enabling customers to purchase many goods from different parts of the world [3]. Analyzing research on e-commerce, a significant increase in scientists' interest in the quality of services offered by e-commerce platforms is noticed. This results in the identification of many factors influencing the quality level. These factors are quantitative

M. Hernes et al. (Eds.): ECAI 2024, LNNS 1217, pp. 213–223, 2025.
https://doi.org/10.1007/978-3-031-78465-1_18

and qualitative, making it worth using a multi-criteria approach to assess them. Therefore, the main aim of the article is to presents the possibility of using a multi-criteria approach to evaluate the quality of functioning of e-commerce platforms. The article includes the following sections: Section 2 – Literature review – E-Commerce platforms' quality assessmend, Section 3 – MCDM approach in the quality assessment process, Section 4 – The methodology, Section 5 – Research results and Section 6 – Discussion and Conclusion.

2 Literature Review - E-commerce Platforms' Quality Assessment

The growing popularity of e-commerce platforms has also contributed to conducting many studies on the quality of their functioning. These studies were determined by the fact that the research on service quality conducted so far did not cover the e-commerce market. Moreover, in the beginning, the main advantage of e-commerce was that making purchases no longer required leaving home; they could be made at any time of the day; now, apart from these two main factors, many others can be identified that significantly influence how customers perceive e-commerce platforms. These factors contribute to the popularity of e-commerce platforms and demonstrate the quality of their functioning.

This resulted in much attention being paid to developing models that assess the quality of functioning of e-commerce platforms. The model developed by M. Blut et al. [4] made it possible to identify and then verify the criteria that should be considered when determining the quality of functioning of e-commerce platforms. G. Sharma [5] developed a model covering the factors affecting customers' overall satisfaction level using the e-commerce platform service. The model proposed by R.P. Mohanty et al. [6] distinguishes three main groups of factors - basic-must, performance-expected, and delight-excitement, covering 24 sub-criteria. In turn, the hierarchical model developed by M. Ingaldi [7] makes it possible to identify areas that require improvement to increase the quality of functioning of e-commerce platforms.

Research on e-commerce platforms also covers issues related to identifying factors that arouse customers' willingness to make online purchases [8], problems caused by the COVID-19 pandemic or the impact of technological innovations on the post-pandemic e-commerce economy [9]. The literature studies conducted on the level of quality of functioning of e-commerce platforms enabled the development of many criteria, including a total of 33 criteria in seven main groups: functionality [10], security/privacy [11], assurance [12], convenience [13], personalization [14]), customer satisfaction [15] and behavioral intensities [12].

However, taking into account the multi-criteria nature of the factors contributing to the overall level of quality, in this article, we propose using the Analytical Hierarchy Process and Entropy methods, which belongs to the MCDM group, widely used in assessing the quality of various phenomena.

3 MCDM Approach in the Quality Assessment Process

For many years, the level of service quality was examined using the SERVQUAL method; however, over time, research was undertaken in which, in addition to this method, MCDM methods were used. Such research was carried out, among others, by T-H. Hsu et al. [16]

developed a service quality measurement model using the CV-SQ model in conjunction with the fuzzy linguistic preference relations (Fuzzy LinPreRa), which can be an essential tool in service quality management. M.A.S. Monfared and F. Dadashian [17] presented a new multiple-criteria decision-making (MCDM) methodology for assessing the quality of textile products. C-C. Wei et al. [18] proposed a fuzzy model to analyze the knowledge quality level, including a knowledge quality fuzziness index (KQFI) and a checking gate. The aim of the research conducted by A.K. Garside et al. [19] was to assess the quality of services of online transport service providers using MCDM methods. Jou et al. [20] use the Analytical Hierarchy Process (AHP) and SERVQUAL methods to assess the quality of bus transport services. J. You and T. Xu [21] proposed using MCDM methods to evaluate the quality of data assets by combining the best worst method (BWM), the order preference technique based on similarity to the ideal solution (TOPSIS), and the triangular fuzzy number, which was used to evaluate the quality of commercial bank asset data.

As seen above, multi-criteria MCDM methods are commonly used to assess the quality level in various areas. However, the analysis of available literature in this field made it possible to identify a significant research gap: the need for MCDM methods to assess the quality of e-commerce platforms' functioning. This article is a response and an important complement to the identified research gap.

4 The Methodology

The research aimed to assess the quality of functioning of online platforms operating in Poland - advertising platforms, shopping platforms, price comparison websites, auction websites, and online stores. For the research, a survey questionnaire was developed consisting of issues enabling the assessment of the quality of the functioning of online platforms. The answers to the questions took the form of a 5-point Likert scale, which allowed obtaining the most detailed answers possible. 491 respondents participated in the study, and the research was conducted from March to August 2023. The Analytical Hierarchy Process method was used to carry out the analyses, which enables the decomposition of a complex decision-making problem and the creation of a final ranking for a finite set of variants, as is the case in this case.

4.1 Analytical Hierarchy Process

Modeling using the AHP method has numerous applications in multi-criteria technical, economic, political, and social decision support, e.g., conflicts and negotiations, quality management, resource allocation, variant selection, team decision support, cost-benefit analysis, credit analyses, production management, formulating strategies and evaluation of their principles.

The foundations of the AHP method are the T.L. Saaty's theory [22]. T.L. Satty (1990) states that human judgments are relative, depending on the current system of values, the role held, and the characteristics of the evaluator. Consequently, there are different points of view on evaluation, which results in different values of partial utilities.

The superordinate goal, located at the top of the hierarchical decision-making structure, is the target state resulting from the successful solution of the decision-making problem. Each variant's implementation of the primary objective results from the fulfillment of intermediate objectives (expressed by the corresponding criteria and sub-criteria). The idea of the method is based on presenting the results of comparisons of evaluation criteria and selection alternatives in the form of square matrices. The procedure begins with defining the goal and a coherent family of criteria relevant to a given group of interveners in the decision-making process. The next stage is iterative and compares the designated criteria in unique pairs using the formula:

$$r_{ij} = \frac{w_i}{w_j} \tag{1}$$

while:

$$r_{ij} \in \{1,2,3,\ldots,9\}, r_{ij} = \frac{1}{r_{ji}} for i, j = 1,2,\ldots,n$$

where:

n - size of the square comparison matrix,

r_{ij} - the degree of dominance of the *i*-th criterion over the *j*-th,

wi, wj -coefficients with values from the Saaty scale (absolute ranks of the Ki and Kj criteria).

The degree of mutual dominance is determined using the criteria introduced by T.L. Saaty's binary relationship, distinguishing five basic situations: equivalence situation, a situation of weak preference, situation of significant preference, a situation of clear preference and a situation of absolute preference.

The possibility of indirect relationships is also assumed, resulting in the creation of a nine-point scale. The adoption of such a constructed scale (based on natural numbers) is justified from the point of view of psychology, in particular, the Weber-Fechner theorem (the relationship between a change in the stimulus and the effect) and the statement that a person can compare at most 7 2 objects at the same time (Table 1).

Table 1. Quality criteria for the functioning of purchasing platforms

Equivalence	Weak preference	Strong preference	Very strong preference	Extreme domination	Intermediate grades
1	3	5	7	9	2,4,6,8

The result of the previous step is the obtained character preference matrix:

$$A = \begin{bmatrix} r_{11} & r_{12} & \ldots & r_{1n} \\ r_{21} & r_{22} & \ldots & r_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ r_{n1} & r_{n2} & \cdots & r_{nn} \end{bmatrix} \tag{2}$$

Synthetic indicators are the values of the utility function of subsequent decision variants and are determined using the additive form:

$$U(A_i) = \sum_{j=1}^{n} w_{ij} * e_{ij} \tag{3}$$

where:

- U(Ai) -values of the utility function of the *i*-th decision variant,
- eij - the value of the *i*-th alternative due to the *j*-th attribute (criterion),
- wij - the weight of the *j*-th criterion.

Several methods are used to determine global preference vectors (weights) and local preferences. The maximum eigenvalue method discussed above is prevalent.

Since AHP is based on the use of expert reactive weights, in the following research, it was decided to replace subjective weights with a different method of determining objective weights - the Entropy method, which includes three main steps [23]:

Step 1. Normalization of the decision matrix $X = [x_{ij}]_{mxn}$:

$$p_{ij} - \frac{x_{ij}}{\sum_{i=1}^{m} x_{ij}} \tag{4}$$

Step 2. Calculation of entropy h_j:

$$h_j = -(lnm)^{-1} \sum_{i=1}^{m} p_{ij} * lnp_{ij} \tag{5}$$

Step 3. Calculation of weights values:

$$w_j = \frac{1 - h_j}{\sum_{j=1}^{n} (1 - h_j)} \tag{6}$$

5 Research Results

The evaluation was made using the AHP method; the averaged input data was normalized using the sum normalization method (Table 2). Calculations using equal weights for the seven main criteria groups. The sub-criteria weights were distributed equally based on the weights of the main criteria. For comparison, calculations were performed using criteria weights calculated using the objective Entropy method based on the input data.

The results show that both rankings are identical, but there are differences in the values of the AHP utility function for the evaluated alternatives.

Table 2. Results of AHP and Entropy Evaluation

Place	AHP, Equal weights, utility function value	AHP, Equal weights, ranking	AHP, Equal weights, utility function value	AHP, Equal weights, ranking
Advertising Platforms (AP)	0.1945	5	0.1935	5
Shopping Platforms (SP)	0.2023	2	0.2030	2
Price Comparison Websites (PCW)	0.2048	1	0.2058	1
Auction Websites (AW)	0.1985	4	0.1979	4
Online stores (OS)	0.1998	3	0.1998	3

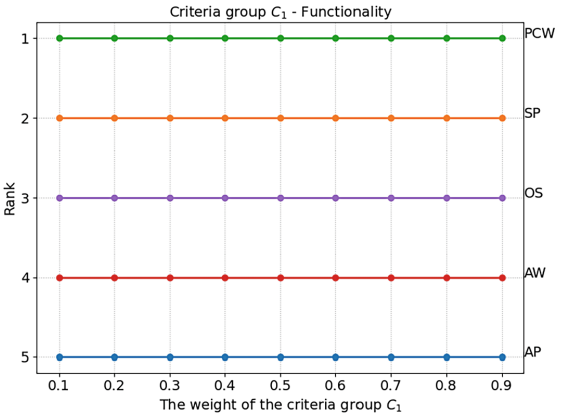


Fig. 1. Sensitivity Analysis for Functionality

5.1 The Sensitivity Analysis

As the weight of one main criterion was increased, the weights of the remaining groups were equally adjusted so that the sum of the weights was 1. The sub-criteria weights were distributed equally based on the weights of the main criteria.

In both cases, despite modifications to the weights of the C1 and C2 criteria groups, there were no shifts in the rankings (Figs. 1 and 2).

If the importance of the C3 criteria group increases from its weight exceeding 0.5 (50%), shopping platforms advance from the second place to the first place, and price comparison websites drop from the first place to the second place (Fig. 3). There were no changes in the rankings for the remaining alternatives.

If the importance of the C4 criteria group increases by more than 0.8 (80%), shopping platforms advance from second place to first place, and price comparison websites drop

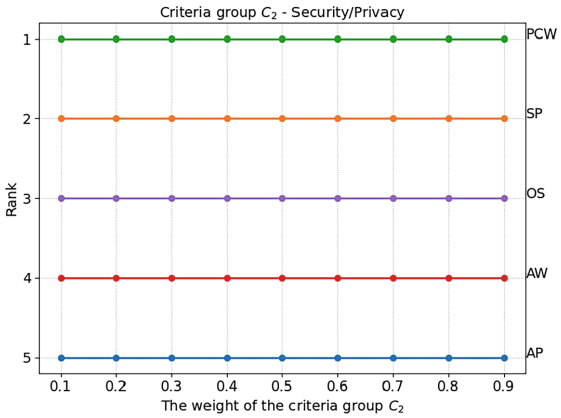


Fig. 2. Sensitivity Analysis for Security/Privacy

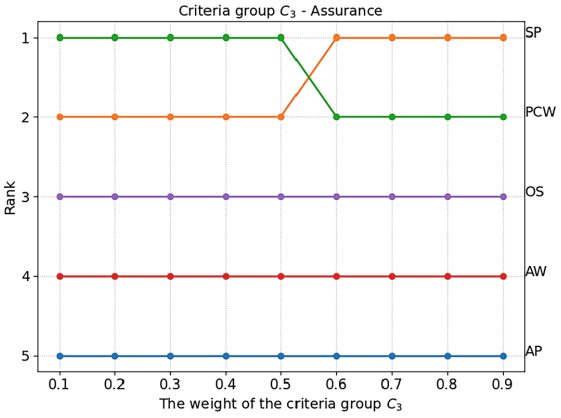


Fig. 3. Sensitivity Analysis for Assurance

from first place to second place (Fig. 4). There were no changes in the rankings for the remaining alternatives.

If the importance of the C5 criteria group increases by more than 0.5 (50%), the auction website advances from fourth place to third place, and the online store drops from third place to fourth place (Fig. 5).

There were no changes in the rankings for the remaining alternatives. Despite modifying the weights of the C6 and C7 criteria groups, the rankings were not changed (Figs. 6 and 7).

6 Discussion and Conclusion

The above-presented research results using the AHP method in the field of MCDM methods constitute an innovative approach to assessing the quality of functioning of online platforms. Previous research involving the use of MCDM in the area of e-commerce

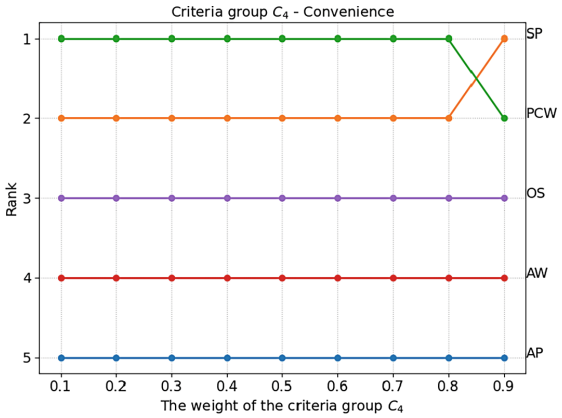


Fig. 4. Sensitivity Analysis for Convenience

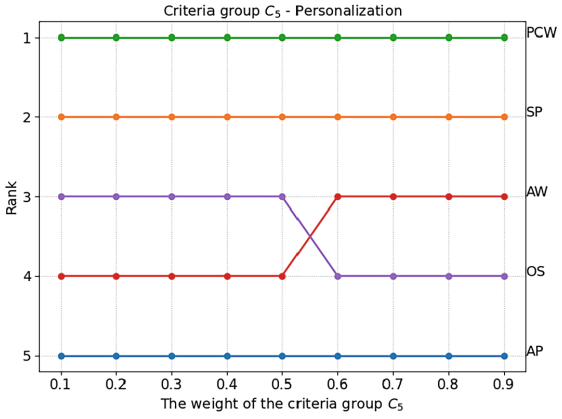


Fig. 5. Sensitivity Analysis for Personalization

focused, among others, on creating a comprehensive and complete framework for assessing m-commerce services by developing a hybrid model tested on the example of food delivery platforms [24]. M-commerce services were the subject of research by G. Kabir and MAA Hasin [25], aimed at elucidating the factors that affect success in m-commerce. Factors influencing the choice of distribution channels among general trade, modern trade, e-commerce, and hyperlocal trade for FMCG companies were also studied [26]. The efficiency of e-commerce supply chains was also measured by identifying key performance metrics (KPM) [27] or assessing the selection of suppliers in fashion supply chains [28].

It follows from the above that many studies have been conducted on e-commerce platforms, but the research context was very diverse. However, the research carried out in this article presents a comprehensive set of criteria consisting of the quality of functioning of e-commerce platforms, taking the form of advertising platforms, shopping platforms,

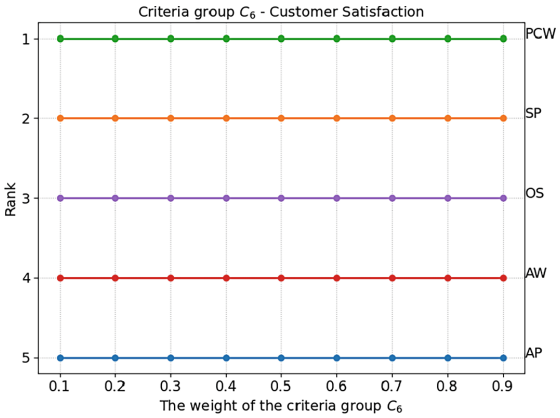


Fig. 6. Sensitivity Analysis for Customer Satisfaction

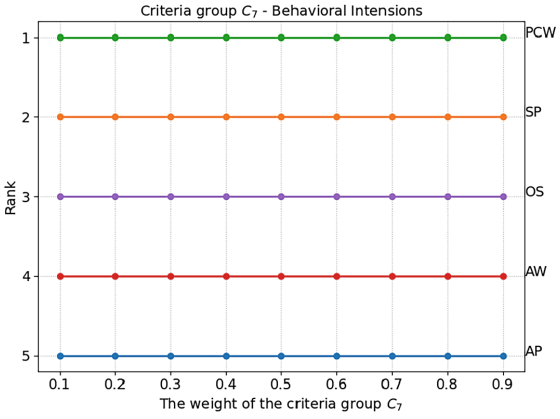


Fig. 7. Sensitivity Analysis for Behavioral Intensions

price comparison websites, auction websites, and online stores operating in Poland and offering an extensive range of products and services.

In the era of universal access to the Internet and the spread of purchases made via online platforms, assessing the quality of their functioning is one of the most important aspects of managing these platforms. As indicated above, so far, many attempts have been made to assess the quality of their functioning, using models previously used to evaluate traditional shopping platforms and now adapted for the needs of e-commerce platforms, or also using MCDM methods; however, limiting the research to a specific type of e-commerce platform, industry of activity or specific products offered via e-commerce platforms. However, more comprehensive research has yet to be conducted to assess the quality of functioning of several types of e-commerce platforms offering various products and services. This highlights the comprehensiveness of the research conducted and its usefulness. To assess the quality of the functioning of e-commerce platforms, the AHP and entropy methods were used, which enabled analyses of the

significance of individual criteria constituting the overall assessment of the quality of the functioning of e-commerce platforms.

The proposed method can be used to conduct a similar assessment but at the international level; it also allows for the examination of other types of e-commerce platforms and allows the development of a ranking based on the main and partial criteria used. The proposed approach can also be a tool for managers to more effectively manage e-commerce platforms, in which quality plays one of the most critical roles in determining their development in the e-commerce market and their further development. Moreover, the research results presented in the article make a significant contribution to the existing literature on the criteria influencing the comprehensive assessment of the quality of functioning of e-commerce platforms, especially those operating in Poland.

References

1. Mainardes, E.W., Coutinho, A.R.S., Alves, H.M.B.: The influence of the ethics of E-retailers on online customer experience and customer satisfaction. J. Retail. Cons. Services **70** (2023)
2. Ambilkar, P., Dohale, V., Gunasekaran, A., Bilolikar, V.: Product returns management: a comprehensive review and future research agenda. Int. J. Prod. Res. **12**, 3920–3944 (2021)
3. Corejova, T., Jucha, P., Padourova, A., Strenitzerova, M., Stalmachova, K., Valicova, A.: E-commerce and last mile delivery technologies in the European countries. Prod. Eng. Arch. **28**(3), 217–224 (2022)
4. Blut, M., Nivriti, C., Vikas, M., Brock, C.: E-Service quality: a meta-analytic review. J. Retail. **91**(4), 679–700 (2015)
5. Sharma, G.: Service quality, satisfaction and loyalty on online marketing: an empirical investigation. Glob. J. Manage. Bus. Res. E Market. **17**(2), 57–66 (2017)
6. Mohanty, R.P., Seth, D., Mukadam, S.: Quality dimensions of e-commerce and their implications. Total Qual. Manag. Bus. Excell. **18**(3), 219–247 (2007)
7. Ingaldi, M.: E-service quality assessment according to hierarchical service quality models. Manage. Syst. Prod. Eng. **30**(4), 311–318 (2022)
8. Sanda, M.-A.: Client's quality assessment of digital transaction platforms interactivenesses in a Covid-19 e-commerce business environment. Lect. Notes Netw. Syst. **265**, 182–190 (2021)
9. Borowski-Beszta, M.: The impact of the COVID-19 pandemic on the behavior of European consumers. In: 3rd Proceedings of International Conference on Management, Economics and Finance (2021)
10. Chmielarz, W., Zborowski, M.: On quality assessment of websites offering mobile applications. In: Hernes, M., Rot, A., Jelonek, D. (eds.) Towards Industry 4.0—current challenges in information systems. Stud. Comput. Intell., p. 887. Springer, Cham (2020)
11. Peráček, T.: E-commerce and its limits in the context of the consumer protection: the case of the Slovak Republic. Jurid. Trib. **12**(1), 35–50 (2022)
12. Springer, M., Tyran, C.K., Ross, S.: Measuring the quality of e-business services. Encyclop. E-Bus. Dev. Manage. Glob. Econ. **1**, 125–134 (2010)
13. Singh, T., Malik, S., Sarkar, D.: E-commerce website quality assessment based on usability. In: Proceedings of IEEE International Conference on Computing, Communication and Automation, ICCCA, 101–105 (2017)
14. Zumstein, D., Kotowski, W.: Success factors of e-commerce–drivers of the conversion rate and basket value. In: Proceedings of the 18th International Conference on E-Society (2020)
15. Tzavlopoulos, I, Gotzamani, K., Andronikidis, A., Vassiliadis, C.: Determining the impact of e-commerce quality on customers' perceived risk, satisfaction, value and loyalty. Int. J. Qual. Serv. Sci. **11**(4), 576–587 (2019)

16. Hsu, T.-H., Her, S.-T., Hou, J.-J.: Developing universally applicable service quality assessment model based on the theory of consumption values, and using fuzzy linguistic preference relations to empirically test three industries. Mathematics **9**(20), 1–29 (2021)
17. Monfared, M.A.S., Dadashian, F.: Development of a new MCDM methodology for textile product quacity assessment using fuzzy-taguchi quality functions. J. Sci. Technol. **15**(58D), 497–520 (2004)
18. Wei, C.-C., Tai, C.-C., Lee, S.-C., Chang, M.-L.: Assessing knowledge quality using fuzzy MCDM model. Mathematics **11**(17) (2023)
19. Garside, A.K., Tyas, R.P.A., Wardana, R.W.: Intuitionistic fuzzy AHP and WASPAS to assess service quality in online transportation. J. Optim. Sist. Ind. **22**(1), 38–51 (2023)
20. Jou, Y.-T., Saflor, C.S., Mariñas, K.A., Young, M.N.: Determining factors affecting perceived customer satisfaction on public utility bus system in occidental Mindoro, Philippines: a case study on service quality assessment during major disruptions. Sustainability **15**(4) (2023)
21. You, J., Xu, T.: A data asset quality evaluation model based on a multi-criteria decision-making method. J. Tongji Univ. **49**(4), 585–590 (2021)
22. Saaty, T.L.: How to make a decision: the analytic hierarchy process. Eur. J. Oper. Res. **48**(1), 9–26 (1990)
23. Wątróbski, J., Bączkiewicz, A., Rudawska, I.: A strong sustainability paradigm based analytical hierarchy process (SSP-AHP) method to evaluate sustainable healthcare systems. Ecol. Ind. **154** (2023)
24. Yao, K.-C., Yang, J.-J., Lo, H.-W., Lin, S.-W., Li, G.-H.: Using a BBWM-PROMETHEE model for evaluating mobile commerce service quality: a case study of food delivery platform. Res. Transp. Bus. Manage. **49** (2023)
25. Kabir, G., Hasin, M.A.A.: Evaluation of customer oriented success factors in mobile commerce using fuzzy AHP. J. Ind. Eng. Manage. **4**(2), 361–386 (2011)
26. Guru, S., Verma, S., Baheti, P., Dagar, V.: Assessing the feasibility of hyperlocal delivery model as an effective distribution channel. Manag. Decis. **61**(6), 1634–1655 (2023)
27. Praneeth, B.B., Nadeem, S.P., Vimal, K.E.K., Kandasamy, J.: Performance measurement of e-commerce supply chains using BWM and fuzzy TOPSIS. Int. J. Qual. Reliab. Manage. **40**(5), 1259–1291 (2023)
28. Harale, N., Thomassey, S., Zeng, X.: Small series fashion supplier selection using MCDM methods. Contrib. Manage. Sci. **4**(2), 203–224 (2021)

Multi-Criteria Evaluation of Floating Photovoltaics Considering a Strong Sustainability Paradigm Regarding Sustainable Solutions

Aleksandra Bączkiewicz[1(✉)], Jarosław Wątróbski[1,2], Anna Shkurina[1], and Wojciech Książek[3]

[1] Institute of Management, University of Szczecin, ul. Cukrowa 8, 71-004 Szczecin, Poland
{aleksandra.baczkiewicz,jaroslaw.watrobski}@usz.edu.pl, anya.shkurina@gmail.com

[2] National Institute of Telecommunications, ul. Szachowa 1, 04-894 Warsaw, Poland

[3] Department of Computer Science, Faculty of Computer Science and Telecommunications, Cracow University of Technology, ul. Warszawska 24, 31-155 Kraków, Poland
wojciech.ksiazek@pk.edu.pl

Abstract. Floating photovoltaic (FPV) systems are a novelty of interest in Poland due to the demand for increasing the share of renewable energy sources in electricity generation. Due to the novelty of FPV systems, insufficient maturity, and limited research available, FPV systems are an area requiring intensive research to evaluate their potential as a competitive solution to classical ground-mounted photovoltaic systems (GMPV). The evaluation of FPVs involves multiple indicators comprising aspects such as cost-effectiveness from an economic point of view and technical performance. This suggests the usefulness of multi-criteria analysis in their evaluation. However, the choice of method should take into account aspects such as sustainability and decision-maker preferences in addition to the multi-criteria analysis. The aim of this paper is to introduce the Strong Sustainability Paradigm based Technique for Order Preference by Similarity to Ideal Solution (SSP-TOPSIS) method for modeling the degree of compensation reduction of criteria according to the strong sustainability paradigm. The method was applied to the practical multi-criteria problem of evaluating FPV systems against a reference GMPV system taking into account technical and economic aspects. A modification of the level of compensation reduction and the preferences of the decision maker were included in the sensitivity analysis. The results confirmed the suitability of the proposed multi-criteria evaluation approach in intelligent decision supporting requiring consideration of compensation reduction and different preferences. The results indicate the potential of FPV systems as an alternative to GMPV, especially in terms of technical performance, and encourage further research into FPV in terms of its suitability and cost-effectiveness.

Keywords: Floating photovoltaic · Intelligent decision supporting · Multi-criteria decision making · Strong sustainability paradigm · Renewable energy sources

M. Hernes et al. (Eds.): ECAI 2024, LNNS 1217, pp. 224–236, 2025.
https://doi.org/10.1007/978-3-031-78465-1_19

1 Introduction

The Polish energy system is mainly based on coal-fired energy and is currently facing a transformation towards an increased share of renewable energy sources in the economy [10]. This necessity is driven by the European Union's goal of achieving climate neutrality by 2050 [7]. The advantages of a transition towards energy systems based on renewable energy sources are to reduce dependence on imports of fossil fuels, to minimize the harmful impact of the economy on the climate and the environment by limiting emissions of greenhouse gases and pollutants into the atmosphere, and to slow down the depletion of natural resources [8]. However, the installed capacity of renewable energy sources (RES) in Poland is not high enough to cover the peak in power demand, resulting in the need to import electricity from neighboring countries. These limitations can be partly solved by the introduction and development of floating photovoltaics (FPV) [1].

To accelerate the development of FPV in Poland, it is worth investigating its cost-effectiveness under Polish conditions [3]. Such a simulation study comparing a ground-mounted photovoltaic system as a reference and four different FPV designs was carried out by the authors of the mentioned article [1]. For the simulation, the authors used PVsyst software. PVsyst gives an analytical solution using well-defined variables. The study aimed to determine whether the application of FPV can be beneficial in Poland. This provides an opportunity to conclude not only on the profitability of such an investment but also on the adaptation of the products available on the market to the specifics of the region and the legal barriers, especially in the context of the fact that floating photovoltaics are currently new in Poland. The study was carried out for the artificially created upper reservoir of the Porąbka-Żar pumped-storage hydropower plant. Because the reservoir is artificially created, the environmental impact is mitigated. As the auction mechanism provides the most cost-effective prices for PV systems below 1 MWp, the installed capacity was chosen to be 983.04 kWp. This paper constitutes a valuable contribution to the rising topic of evaluating floating photovoltaic systems. The authors have provided the parameters of the photovoltaic systems investigated with a group of technical and economic indicators relevant to their evaluation and conclusions confirmed by analysis through simulation. Due to the value of the present study, the authors decided to use it as a reference for investigating the potential of FPV in relation to GMPV using multi-criteria analysis [1].

However, the study identified areas that represent an interesting research gap regarding the method of their evaluation. Firstly, the evaluation is carried out and the results are analyzed for each indicator individually. This method of evaluation forces each aspect to be considered separately, which is difficult for many evaluation dimensions. Secondly, the study did not take into account the preferences of decision-makers regarding the relevance of individual indicators and groups of indicators. Thirdly, the aspect of assessing the sustainability of the alternatives studied was not included. These observations motivated the authors to propose a multi-criteria approach taking into account compensation reduction in line with the strong sustainability paradigm and the preferences of decision-makers [2].

The aim of this paper is to analyze and evaluate FPV in relation to the reference GMPV using a multi-criteria Strong Sustainability Paradigm based Technique for Order Preference by Similarity to Ideal Solution (SSP-TOPSIS) method taking into account

the possibility of modeling a reduction in criteria compensation. Such a study will be a valuable addition to the study already carried out in [1]. This is justified by the fact that the study presented earlier forces the analysis of each indicator separately and does not take into account the decision-maker's preferences regarding the relevance of the criteria. The SSP-TOPSIS method allows all criteria to be considered simultaneously and aggregates the assessment results. In addition, the feature to model the compensation reduction of criteria provides the possibility to assess the sustainability of the systems under consideration. Compensation reduction prevents poor performances for some criteria from being compensated by outstandingly favorable performances of other criteria. This approach provides a new contribution to intelligent decision support and is in line with the strong sustainability paradigm relevant to the assessment of RES issues. In addition, the paper includes a sensitivity analysis with varying values of criteria weights and sustainability coefficient values.

The rest of the paper is organized as follows. Section 2 provides basics and assumptions of the proposed method and presents practical case study. The research results are demonstrated in Sect. 3 and discussed in Sect. 4. Finally, in Sect. 5 conclusions are drawn.

2 Methodology

The SSP-TOPSIS is a novel method incorporating the basics of the well-known and commonly used Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) method, that takes into account the distance of alternatives from ideal and anti-ideal reference solutions [13]. The proposed method includes a novel stage that enables modeling criteria compensation reduction. Compensation is a shortcoming of multi-criteria methods originating from American school like TOPSIS [13], VIKOR (Vlsekriterijumska Optimizacija I Kompromisno Resenje) [12], and AHP (Analytic Hierarchy Process) [11]. This stage incorporates reducing the compensation achieved by subtracting the Mean Deviation (*MD*) from the performance value. *MD* is the difference between the performance of the alternative for a given criterion and the averaged performance within the criterion across all alternatives multiplied by the sustainability coefficient. The sustainability coefficient s is used for modeling compensation reduction and is within the scope of 0 to 1. It may be set as the standard deviation from the normalized decision matrix within each considered criterion. Detailed Steps of the SSP-TOPSIS method are given below.

Step 1. Compute the Mean Deviation MD_{ij} for every performance value contained in the decision matrix by subtraction of the mean value $\bar{x}_j$ from performance values x_{ij} for every C_j criterion. Then, multiply the outcome by s_j coefficient. Coefficient s_j denotes the sustainability coefficient, which reflects the level of criteria performance compensation reduction, and its values are real numbers within the range from 0 to 1. Value 0 of the s coefficient denotes a lack of compensation reduction. On the other hand, the s coefficient equal to 1 represents a full reduction of compensation. Criteria are numbered by $j = 1, 2, \ldots, m$. High values of s_j denote a more significant reduction in the compensation of a j-th criterion performance value. Otherwise, low values of s_j define a low degree of reduction in the compensation of a j-th criterion. The complete procedure of Mean

Deviation calculation is carried out with Equ. (1).

$$MD_{ij} = (x_{ij} - \bar{x}_j)s_j \tag{1}$$

Step 2. Set 0 values to these MD_{+ij} for profit criteria C_{+j} that are lower than 0. When MD_{+ij} is lower than 0 it denotes that x_{+ij} is lower than $\bar{x}_{+j}$. Assign 0 values for these MD_{-ij} for cost criteria C_{-j} that are higher than 0. It denotes that r_{-ij} are higher than $\bar{x}_{-j}$. Mentioned procedure is presented by Equ. (2)

$$MD_{ij} = 0 \; \forall \; MD_{+ij} < 0 \; \vee \; MD_{-ij} > 0 \tag{2}$$

where MD_{ij} defines Mean Deviation values computed for criteria C_j. This stage is relevant since it is intended to prevent improving unmentionable performance values that are outliers from the mean toward the worse.

Step 3. Subtract MD_{ij} values from performance values x_{ij} included in decision matrix x_{ij} according to Equ. (3).

$$t_{ij} = x_{ij} - MD_{ij} \tag{3}$$

The rest of the steps are identical to the classic TOPSIS method.

Step 4. Normalize the decision matrix demonstrated in Equ. (4) with the chosen normalization method, such as the Minimum-Maximum normalization or Vector normalization. In Minimum-Maximum normalization, normalized values represented by r_{ij}^{+} for profit criteria and r_{ij}^{-} for cost criteria are obtained using Equ. (5). After normalization, all criteria are transformed to profit type.

$$T = [t_{ij}]_{m \times n} = \begin{bmatrix} t_{11} & t_{12} & \cdots & t_{1n} \\ t_{21} & t_{22} & \cdots & t_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ t_{m1} & t_{m2} & \cdots & t_{mn} \end{bmatrix} \tag{4}$$

$$r_{ij}^{+} = \frac{t_{ij} - min_j(t_{ij})}{max_j(t_{ij}) - min_j(t_{ij})}, \; r_{ij}^{-} = \frac{max_j(t_{ij}) - t_{ij}}{max_j(t_{ij}) - min_j(t_{ij})} \tag{5}$$

Step 5. Compute the weighted normalized decision matrix. In order to do so, multiply values contained in the normalized decision matrix by the corresponding criteria weights w_j as Equ. (6) shows.

$$v_{ij} = r_{ij} w_j \tag{6}$$

Step 6. Calculate the Positive Ideal Solution (PIS), namely v_j^{+}, and Negative Ideal Solution (NIS), namely v_j^{-} using Equ. (7). PIS contains the maximum values of the weighted normalized decision matrix, while NIS includes its minimum values. Since the decision matrix was normalized in the previous step, the division of criteria into profit and cost is unnecessary.

$$v_j^{+} = \{v_1^{+}, v_2^{+}, \ldots, v_n^{+}\} = \{max_j(v_{ij})\}, \; v_j^{-} = \{v_1^{-}, v_2^{-}, \ldots, v_n^{-}\} = \{min_j(v_{ij})\} \tag{7}$$

Step 7. Compute the Euclidean distance from PIS D_i^+ and NIS D_i^- for each alternative with Equ. (8).

$$D_i^+ = \sqrt{\sum_{j=1}^{n}(v_{ij} - v_j^+)^2},\ D_i^- = \sqrt{\sum_{j=1}^{n}(v_{ij} - v_j^-)^2} \tag{8}$$

Step 8. Compute the score for each alternative included in the analysis according to Equ. (9). The C_i value is always within the range between 0 and 1. The alternative that demonstrates the highest C_i value is considered the best scored. The ranking is built by sorting alternatives according to preference values in descending order.

$$C_i = \frac{D_i^-}{D_i^- + D_i^+} \tag{9}$$

Case study

Table 1 shows the technical parameters and economic indicators proposed in the reference research paper referred by the authors in this research. The individual parameters and indicators can be either profit-driven, where the objective is to maximize their value, or cost-driven, where the objective is to minimize their value. Area is a cost parameter because, due to the design, the usable area of a floating photovoltaic system is often limited. In addition, an appropriate distance from the edge of the reservoir is required. Final PV system yield is a profit-driven parameter, as the aim is to strive for the highest possible efficiency. The final system yield is the net power output of the entire PV system on the AC side per unit of rated installed capacity on the DC side of the system. The capacity factor PR is the ratio of the electricity actually generated and the electricity that would be generated if the modules were always operating at STC (Standard Test Conditions) efficiency and is a profit parameter. Annual Energy Density (AED) is a profit criterion. Irradiance-weighted average temperature, on the other hand, is a cost criterion and achieves a more reliable comparison between different PV systems than average module temperature, as it does not take into account hours without sunshine. Three well-established indicators, namely Net Present Value (NPV), Internal Rate of Return (IRR), and Levelised Cost of Energy (LCOE), were used to evaluate the systems from an economic perspective. LCOE is a measure of the average net present cost of generating electricity by a power plant over its lifetime. It represents the average revenue per unit of electricity generated that would be required to recover the costs of constructing and operating a power plant over its assumed life and operating cycle and is calculated as the ratio of all discounted costs required to generate electricity (including construction) divided by the discounted sum of the quantities of energy actually delivered. Table 2 shows decision matrix including the performances collected for the PV systems considered in the present study against the criteria presented in Table 1.

The alternatives evaluated include four floating PV systems and one reference classic ground-mounted PV system (GMPV S25). Among the FPV systems are two systems produced by Ciel&Terre: C&T S12 and C&T EW12 and two by Solaris Synergy: SolSyn S12 and SolSyn S25. Profit criteria are represented by 1 and cost criteria are denoted by -1.

Table 1. Criteria for photovoltaic systems assessment.

Criteria		Unit	Type
Technical parameters			
C_1	Area	[m^2]	Cost
C_2	Y_f (Final PV system yield)	[kWh/kWp]	Profit
C_3	PR (Performance ratio)	[%]	Profit
C_4	AED (Annual Energy Density)	[kWh/m^2]	Profit
C_5	T_{IWA} (Irradiance-weighted average temperature)	[$^\circ C$]	Cost
Economic indicators			
C_6	NPV (Net Present Value)	[€]	Profit
C_7	IRR (Internal Rate of Return)	[%]	Profit
C_8	LCOE (Levelized Cost of Energy)	[€/MWh]	Cost
C_9	Minimum auction price for which NPV = 0	[€/MWh]	Cost

Table 2. Performance values collected for evaluated photovoltaic systems.

Technology	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9
GMPV S25	10982	1079	86.92	96.5	25.5	82645	13.3	73.2	46.7
SolSyn S25	15220	1104	89.52	71.3	19.1	82426	12.9	74.5	46.9
SolSyn S12	9901	1046	89.26	103.8	19	44062	10.63	79	51.1
C&T S12	9901	1027	87.67	102	22.2	−26787	8	88.8	57.2
C&T EW12	8514	936	87.99	108	21.4	−83161	3.3	96	63.5
Criteria types	−1	1	1	1	−1	1	1	−1	−1

3 Results

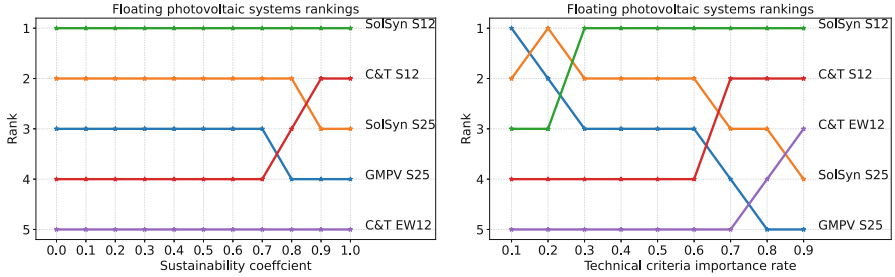
This section provides the results of research conducted in three stages. The first stage includes the assessment of considered PV systems using SSP-TOPSIS with the change of level of criteria compensation reduction by modification of sustainability coefficient values. Table 3 contains scores of SSP-SPOTIS achieved by evaluated alternatives for the following s coefficient values incremented from 0 to 1 by 0.1 for each criterion. The analysis presents the application of the modification of the s coefficient for all criteria simultaneously, as previous studies with modification of the s coefficient only for selected groups of criteria did not report significant shifts in rankings for this problem. Criteria weights were set as equal, which means all considered criteria are equally important to decision-makers.

Based on SSP-TOPSIS scores rankings of PV systems were generated. Shifts in rankings appearing during the incrementation of s coefficient are displayed in Figure 1a. It can be noted that for lack of compensation reduction, which is the same as the use of

Table 3. SSP-TOPSIS scores received for sustainability coefficient modification.

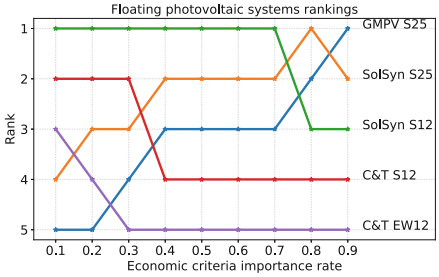
Technology	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
GMPV S25	0.6116	0.6150	0.6185	0.6222	0.6260	0.6300	0.6341	0.6383	0.6426	0.6468	0.6511
SolSyn S25	0.6473	0.6476	0.6479	0.6483	0.6486	0.6490	0.6495	0.6499	0.6505	0.6510	0.6517
SolSyn S12	0.7845	0.7978	0.8124	0.8283	0.8458	0.8650	0.8863	0.9100	0.9365	0.9663	1.0000
C&T S12	0.4969	0.5106	0.5255	0.5416	0.5592	0.5784	0.5995	0.6227	0.6482	0.6764	0.7074
C&T EW12	0.4063	0.4094	0.4129	0.4167	0.4210	0.4257	0.4309	0.4367	0.4433	0.4506	0.4589

TOPSIS the ranking leader is SolSyn S12. SolSyn S12 maintains its leadership position in the ranking regardless of the increase in criterion compensation reduction caused by increasing s, which indicates the sustainable performance values of this system compared to the other alternatives included in this analysis. The second rank is taken by the SolSyn S25 FPV system for s coefficient up to and including 0.8, with higher s values the system falls to third rank.



(a) Shifts in rankings for s coefficient modification.

(b) Shifts in rankings for technical criteria weights modification.



(c) Shifts in rankings for economic criteria weights modification.

Fig. 1. Shifts in rankings appearing during sensitivity analysis with SSP-TOPSIS.

Drops of alternatives in the rankings with an increase in s mean that there are areas where the alternatives have extremely favorable values compensating for poor values appearing within other criteria. An increase in s reduces this compensation, which causes the alternatives to fall. However, SolSyn S25 and GMPV S25 fall for rather high values of s, indicating that they have quite sustainable performances across all criteria. The FPV C&T S12 system emerged as an interesting alternative, ranking penultimate fourth with low compensation reduction, while moving up to second place with s values greater than 0.7, indicating the overall balanced performance of this alternative despite their lack of outstanding values. The worst performer in this analysis is the FPV C&T EW12 system, which is justified by the outlier poor performances for most criteria.

Analysis of the performances allows us to notice that the SolSyn S12 has an extremely favorable value for criterion C_5 (T_{IWA}). In this case, the superiority of the SolSyn S12

and the other FPV systems over the GMPV can be seen in terms of module operating temperature. The water-cooling effect of the FPV shows effectiveness and reduces the T_{IWA} compared to the reference GMPV S25. SolSyn S12 also presents a favorable performance for criteria C_4 (AED) and C_3 (PR). At this stage of the analysis, it can be noted that the SSP-TOPSIS ranking leader SolSyn S12 shows more favorable performance than GMPV S25 in the area of technical criteria.

The aim of the next stage of the research was to check how decision-makers' preferences for each criterion group affect scores. For this purpose, the weights of the technical and then economic criterion groups were modified with the sustainability coefficient set as the standard deviation for the standardized performances within each criterion. The research therefore simultaneously incorporated the sustainability aspect of the alternatives and the preferences of the decision-makers. Firstly, a sensitivity analysis was carried out for the weights of the technical criteria. The weights of the technical criteria group were gradually increased from 0.1 to 0.9 in 0.1 increments, and the weights of the individual criteria in this group were distributed equally. At the same time, the weights of the economic criteria group were reduced and distributed equally for each criterion in this group. Table 4 includes SSP-TOPSIS scores for sensitivity analysis considering technical criteria weights modification.

Table 4. SSP-TOPSIS scores received for technical criteria weights modification.

Technology	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
GMPV S25	0.9390	0.8733	0.8037	0.7319	0.6610	0.5964	0.5440	0.5084	0.4900
SolSyn S25	0.9354	0.8738	0.8077	0.7412	0.6792	0.6268	0.5881	0.5642	0.5526
SolSyn S12	0.8165	0.8181	0.8215	0.8272	0.8350	0.8437	0.8512	0.8564	0.8590
C&T S12	0.4421	0.4508	0.4687	0.4967	0.5325	0.5697	0.6013	0.6227	0.6336
C&T EW12	0.0690	0.1422	0.2191	0.2984	0.3775	0.4518	0.5150	0.5605	0.5852

Figure 1b displays shifts in rankings during increasing of technical criteria weights. It can be observed that when the significance of technical criteria is low, GMPV S25 is the leader of the ranking. As the significance of the technical criteria group increases, GMPV S25 successively drops in the ranking until it reaches the last place for the high significance of technical criteria. The PV systems that move up as the significance of the technical criteria group increases are the FPV systems among which we can mention SolSyn S12, which was the leader in the sensitivity analysis ranking taking into account the modification of s coefficient, C&T S12, and C&T EW12. In the next step, sensitivity analysis considering increasing of the economic criteria weights group was performed analogously to sensitivity analysis considering technical weights modification. SSP-TOPSIS scores achieved by considered PV systems during sensitivity analysis involving economic criteria weights incrementation are shown in Table 5. Shifts in rankings during the sensitivity analysis involving economic criteria weights increasing are visualized in Figure 1c. It can be observed that the results obtained for increasing the importance

of the economic criteria group are opposite to the results obtained when increasing the importance of the technical criteria group.

Table 5. SSP-TOPSIS scores received for economic criteria weights modification.

Technology	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
GMPV S25	0.4900	0.5084	0.5440	0.5964	0.6610	0.7319	0.8037	0.8733	0.9390
SolSyn S25	0.5526	0.5642	0.5881	0.6268	0.6792	0.7412	0.8077	0.8738	0.9354
SolSyn S12	0.8590	0.8564	0.8512	0.8437	0.8350	0.8272	0.8215	0.8181	0.8165
C&T S12	0.6336	0.6227	0.6013	0.5697	0.5325	0.4967	0.4687	0.4508	0.4421
C&T EW12	0.5852	0.5605	0.5150	0.4518	0.3775	0.2984	0.2191	0.1422	0.0690

For the predominant importance of the economic criteria, the reference GMPV S25 system leads the ranking. However, this situation is only the case for a very high importance of the economic criteria. When the importance of the economic criteria is set at 70% or less, the SolSyn S12 FPV system is the leader, as in previous studies. Among the FPV systems, the SolSyn S25 system also achieved a good result in increasing the importance of economic criteria.

4 Discussion

Research on the multi-criteria evaluation of PV systems including four FPV systems and a reference GMPV using the SSP-TOPSIS method shows that the results are influenced by the level of criterion compensation reduction and the relevance of the different criterion groups for the decision maker. The results demonstrated that the most competitive alternative to GMPV is the SolSyn S12 FPV system, which shows an advantage over GMPV, especially in terms of technical criteria. Among the superior technical aspects of FPV over GMPV, it is worth mentioning the operating temperature of the module (C_5 – T_{IWA}), which is influenced by the water-cooling effect in the case of FPV. FPVs achieve higher values than GMPVs for the other technical criteria, especially for C_3 (PR) and C_4 (AED), but also for C_1 (Area) and C_2 (Y_f) for selected FPV systems. The higher PR for FPV is justified by the fact that the modules on the water have a lower temperature, so their power is closer to the nominal power, and therefore the PR increases [6, 9]. However, the results may vary depending on geographical location. For example, at latitudes closer to the Equator, high temperatures affect PV output more significantly than in the Polish climate, so FPV structures are more justified there. It should also be emphasized that the simulations carried out in the study from the reference paper referred to by the authors do not take into account the wind speeds and environmental temperatures that can occur between land and water. These results should therefore be treated with caution as preliminary research on emerging FPV technology [1].

In the case of the economic analysis, systems based on the Ciel&Terre design, that indicate lower profits, were found to be unprofitable. The reference study presented in

the reference article showed that, in the case of C&T EW12, even the highest auction price does not provide a profit, confirming the authors' results. Systems using Solaris Synergy designs proved to be far more profitable than Ciel&Terre and are a competitive alternative to GMPV when technical criteria are also taken into account. However, if only economic criteria are considered, GMPV can provide more profit than FPV systems, at least in an auction mechanism. On the other hand, floating technologies are still immature, hence their cost decline may proceed faster compared to conventional systems, making FPV more financially competitive [4, 5]. Furthermore, according to the authors of the reference article, the results could change in favor of FPV if a longer period was analyzed or other sales channels were taken into account [1]. It is to be expected that as the technology develops and more manufacturers enter the market, the price will come down. In addition, new European or local financial incentives (e.g. fixed prices for floating photovoltaic installations) could increase the profitability of FPV [6].

5 Conclusions

Floating photovoltaics is currently a novelty in Poland, so the reference work, which contains the data on which the authors based their analysis, is the only comprehensive source of knowledge on how such systems can work in Polish conditions. The results obtained confirmed the authors' observations in the reference study, which demonstrates the usefulness of the proposed method. It allowed simultaneous multi-dimensional analysis, the results of which were obtained in an automated manner and are consistent with the reference study, confirming the usefulness and reliability of the proposed approach.

In conclusion, FPV currently appears to have limited profitability in Polish conditions, especially from an economic point of view. The additional efficiency due to the water-cooling effect cannot compensate for the high investment costs. However, additional long-term experiments would need to be conducted to verify whether the results of various studies can be implemented in simulations for colder climates. The results obtained confirm the potential of the selected FPV systems to demonstrate a technical advantage over GMPV. This, combined with the maturation of FPV technology and the prospect of decreasing costs, could make FPV a viable competitive alternative to GMPV.

This paper as preliminary research has several shortcomings such as the limited set of alternatives evaluated and the use of only one MCDA approach. Directions for future work include an examination of the proposed method for other datasets addressing the problems of energy generation from renewable sources and a comparison of the proposed method with other methods considering a strong sustainability paradigm. In order to enhance the practicality and applicability of the findings, it would be advantageous to consider in future analyses a more comprehensive investigation of the impact of FPVs on local aquatic ecosystems and to expand research on the durability and maintenance of these systems in the long term. It is advisable to consider the regulatory and legal issues concerning installations in public and private waters and potential barriers related to infrastructure and integration with the existing energy grid.

Acknowledgments. Publication funded by the state budget under the program of the Minister of Education and Science named Perły Nauki, Poland, project number: PN/01/0022/2022, total

project value: PLN 165 000,00 (A.B.) and Co-financed by the Minister of Science under the "Regional Excellence Initiative" Program RID/SP/0046/2024/01 (J.W.).

Ethics Declarations.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Boduch, A., Mik, K., Castro, R., Zawadzki, P.: Technical and economic assessment of a 1 MWp floating photovoltaic system in Polish conditions. Renew. Energy **196**, 983–994 (2022). https://doi.org/10.1016/j.renene.2022.07.032
2. Etxano, I., Villalba-Eguiluz, U.: Twenty-five years of social multi-criteria evaluation (SMCE) in the search for sustainability: Analysis of case studies. Ecol. Econ. **188**, 107131 (2021). https://doi.org/10.1016/j.ecolecon.2021.107131
3. Exley, G., Hernandez, R., Page, T., Chipps, M., Gambro, S., Hersey, M., Lake, R., Zoannou, K.S., Armstrong, A.: Scientific and stakeholder evidence-based assessment: Ecosystem response to floating solar photovoltaics and implications for sustainability. Renew. Sustain. Energy Rev. **152**, 111639 (2021). https://doi.org/10.1016/j.rser.2021.111639
4. Ghose, D., Pradhan, S., Shabbiruddin: Floating solar plants—Exploring a new dimension of energy generation: A case study. Energy Sour. Part A: Recovery Utilization Environ. Effects, pp. 1–31 (2021). https://doi.org/10.1080/15567036.2021.1965266
5. Kichou, S., Skandalos, N., Wolf, P.: Floating photovoltaics performance simulation approach. Heliyon **8**(12) (2022). https://doi.org/10.1016/j.heliyon.2022.e11896
6. Kumar, M., Niyaz, H.M., Gupta, R.: Challenges and opportunities towards the development of floating photovoltaic systems. Sol. Energy Mater. Sol. Cells **233**, 111408 (2021). https://doi.org/10.1016/j.solmat.2021.111408
7. Maduta, C., D'Agostino, D., Tsemekidi-Tzeiranaki, S., Castellazzi, L., Melica, G., Bertoldi, P.: Towards climate neutrality within the European Union: assessment of the energy performance of buildings directive implementation in member states. Energy Build., 113716 (2023). https://doi.org/10.1016/j.enbuild.2023.113716
8. Potrč, S., Čuček, L., Martin, M., Kravanja, Z.: Sustainable renewable energy supply networks optimization—The gradual transition to a renewable energy system within the European Union by 2050. Renew. Sustain. Energy Rev. **146**, 111186 (2021). https://doi.org/10.1016/j.rser.2021.111186
9. Rahaman, M.A., Chambers, T.L., Fekih, A., Wiecheteck, G., Carranza, G., Possetti, G.R.C.: Floating photovoltaic module temperature estimation: Modeling and comparison. Renewable Energy **208**, 162–180 (2023). https://doi.org/10.1016/j.renene.2023.03.076
10. Rusin, A., Wojaczek, A.: Changes in the structure of the Polish energy mix in the transition period to ensure the safety and reliability of energy supplies. Energy **282**, 128831 (2023). https://doi.org/10.1016/j.energy.2023.128831
11. Wątróbski, J., Bączkiewicz, A., Rudawska, I.: A strong sustainability paradigm based analytical hierarchy process (SSP-AHP) method to evaluate sustainable healthcare systems. Ecological Indicators **154**, 110493 (2023). https://doi.org/10.1016/j.ecolind.2023.110493

12. Wątróbski, J., Bączkiewicz, A., Sałabun, W.: New multi-criteria method for evaluation of sustainable RES management. Applied Energy **324**, 119695 (2022). https://doi.org/10.1016/j.apenergy.2022.119695
13. Wątróbski, J., Karczmarczyk, A., Bączkiewicz, A.: Using the TOSS method in semi-autonomous passenger car selection. Sustain. Energy Technol. Assess. **58**, 103367 (2023). https://doi.org/10.1016/j.seta.2023.103367

Transforming the Tutor and Tutee Matching Process Using ICT: The Problems of Multi-Criteria Decision Analysis

Andrzej Greńczuk[1](✉), Kamila Łuczak[1], Iwona Chomiak-Orsa[1], Estera Piwoni-Krzeszowska[1], Sudha Prathyusha Jakkaladiki[2], Martina Janečková[2], and Dominik Palla[2]

[1] Wroclaw University of Economics and Business, 53-345 Wroclaw, Poland
{andrzej.grenczuk,kamila.luczak,iwona.chomiak-orsa,estera.piwoni-krzeszowska}@ue.wroc.pl

[2] Faculty of Informatics and Management, University of Hradec Králové, 500 03 Hradec Králové, Czech Republic
{sudha.jakkaladiki,martina.janeckova,dominik.palla}@uhk.cz

Abstract. Information and Communication Technologies (ICT) enables supporting decision-making processes on an increasingly broader spectrum of quality problems that have previously been difficult to quantify. One such process is academic tutoring, which aims to support student development. The quality of this process depends on the appropriate tutor selection for the tutee. The key determinant of the success of this process is the matching of the parties to this relationship. So far, all stages of the process were carried out traditionally and pairs were selected based on students' indications, which was subject to many errors. The aim is to create a concept of transforming the tutor and tutee matching process using ICT with a multi-criteria analysis. The co-participant observation conducted at the Wroclaw University of Economics and Business was used to build the concept. We believe that technologies supporting multi-criteria analysis can improve the quality of the above processes. With the right pairing, collaboration can generate much more value for both sides, and the multi-criteria analysis could be performed by artificial intelligence-based systems using artificial neural networks.

Keywords: academic tutoring · multi-criteria decision analysis (MCDA) · Information and Communication Technologies (ICT)

1 Introduction

The development of information and communication technologies (ICT) and the development needs of society have undergone spectacular changes in the last two decades. Therefore, it can be observed that education is one of the industries that has been very intensively implementing modern technologies in recent years. Moreover, new teaching methodologies are being sought because traditional methods do not meet expectations

M. Hernes et al. (Eds.): ECAI 2024, LNNS 1217, pp. 237–248, 2025.
https://doi.org/10.1007/978-3-031-78465-1_20

and their effectiveness is clearly declining [1]. Educational processes are evolving – because of equipping teaching processes with advanced technologies and using increasingly innovative teaching forms. The increase in social awareness that education should play a key role in the economic development of every country contributes to the search for development strategies and greater attention to these processes. These trends are visible primarily at the level of academic education, which has greater opportunities to make course forms more flexible. Providing high-quality education is an important task of an educational institution, and continuous improvement of the effectiveness of teaching processes is forced by the market expectations of future employers.

One of the modern forms of education is academic tutoring, defined as the essence of which is to support students' talents through individual cooperation with experienced academic tutors. Tutoring is a multi-stage process. Tutoring begins with the tutor and tutee getting introduced to each other. Then together they identify their goals, and tasks to be completed during the entire tutoring process or a selected meeting. Together they develop a plan of action and dates for future meetings. Then they check that the meaning of all the tasks and the assumptions of the work are understood in the same way. This concludes the selected meeting.

The main advantage of academic tutoring is to work individually with a student in a 1-1 relationship. Educating students in 1 and 1 is expensive for academic institutions and, what's more, it does not always yield the intended results. Some of the problems are the growing number of students or the lack of formal supervision of tutoring [2]. Switching from group educational processes to individual ones increases the standards and quality of education, but the effects are not always optimal. The main success determinant is the proper selection of individuals collaborating and cooperating in the tutor-tutee relationship which will bring a positive atmosphere to the learning process.

Currently, the processes of connecting cooperating pairs are conducted traditionally by comparing the student's choices with the recommendations of the academic teacher. Two of the authors of this article are academic tutors in the BIPS (Business Individual Study Program) at the Wroclaw University of Economics and Business (WUEB). Based on their own experience, the authors note the potential to implement multi-criteria decision analysis (MCDA) in the academic tutoring process, especially in the aspect of finding the right tutor and tutee. The complex methodology of MCDA can be facilitated to be conducted using ICT.

The current research article is destined to address potential solutions to the concerns that have been identified, based on that the authors defined the following research questions:

Question 1: In what stages of tutoring do multi-criteria problems appear?
Question 2: How can MCDA be used in the process of matching tutor and tutee during tutoring?
Question 3: What is the role of ICT in the concept of using MCDA in the matching process?

The article aims to invent a methodology for transforming the tutor and tutee matching process with a multi-criteria analysis using ICT. The use of solutions supporting multi-criteria decisions, on the one hand, may contribute to improving the efficiency of

the whole process, and that may open the doors to the prospect of scalability of this process and increase its availability to a larger group of students.

2 Theoretical Background

2.1 Matching Tutors and Tutees Process

Tutoring is one of the modern education methods. In the tutoring process, students learn and develop their competencies and knowledge under the supervision of a tutor for each student (or two or three students at a time) [3]. Face-to-face tutoring has also improved motivation-related factors such as students' interest in the tutored subject (e.g. [4]). Moreover, the tutor's task is to help the student work around difficulties. The tutor helps the student to develop a strategy for acquiring knowledge and skills while helping him understand what does not work. In addition, tutors inspire students to identify and take up challenges, invoke curiosity, and maintain their sense of control over their development process. Human tutors don't give their students clear solutions but rather tips and suggestions that motivate them to overcome challenges [5].

The relationship between the tutor and tutees is highly interactive because the tutor constantly provides feedback to tutees as they solve problems, learn, or develop their skills [5]. Some publications suggest that the interactivity facilitated by human tutoring is key to its effectiveness [6]. Optimizing the effectiveness of tutoring requires appropriate profiling of the tutoring process entities. Profiling needs the acquisition and use of data about people who want to participate in the tutoring process or are already involved in this process. This data should be used to make informed, optimal decisions about qualifications for the tutoring program and to plan the progress of the tutoring process to ensure its most excellent effectiveness [7].

Profiling in the context of tutoring involves creating detailed descriptions or models of the tutor and student based on various characteristics, including, but not limited to, general knowledge, learning or teaching preferences, behaviors, talents, goals, skills, etc. Profile modeling involves the analysis of data from multiple sources to build a comprehensive picture of the student and tutor. Tutor and student profiling aims to provide a personalized and adaptive tutoring experience based on each student's unique characteristics and needs. Artificial intelligence (AI) solutions are used in the profiling process [8].

To maximize the tutoring process's effectiveness, it is essential to match tutees and tutors. To match tutors effectively with tutees, it is vital to consider a few determinants.

Effective conduct of the tutoring process requires, first, that the tutor has appropriate knowledge of the tutoring subject. When selecting a tutor and a student, the tutor's experience, previous tutoring experience, and substantive knowledge should also be considered. It can be assumed that tutors who have successfully passed the content knowledge assessment will be more willing to provide students with accurate and comprehensive explanations during tutoring sessions [9].

Empirical research on human and computer tutoring has revealed characteristics of novice and experienced tutors [10, 11], Socratic and teaching strategies [12], patterns of collaborative dialogue in tutoring [13] and the interrelationships between tutor and tutee engagement, skills [14], motivation, and learning [15, 16]. More skillful and engaging

tutors can positively impact tutees' performance. On the other hand, a proactive and engaged tutee is more likely to lead to higher tutoring gains.

Factors such as the tutor's teaching style and ability to provide interactive feedback must also be considered to match tutors and tutees [14]. In addition, the tutee's learning style should be considered, referring to how the student prefers to perceive the learning material (e.g., graphical representation, audio materials, and text representation) [5]. Adjusting these factors can optimize a tutoring session to facilitate a successful tutoring experience for the tutee.

The interactivity and feedback provided by the tutor during the tutoring session determine the effectiveness of this process. Therefore, when matching tutors and tutees, it is essential that the tutor conducts an interactive dialogue, provides feedback, and adapts their tutoring style to the tutee's requirements so that the tutee's preferences and tutoring needs are met. However, the interactivity of both tutors and tutees is conditioned by personal traits. Personality traits can play a significant role in influencing people's behavior. Researchers use the five-factor model of personality (FFM) to describe intra-individual predispositions (i.e., temperament). The FFM dimensions represent fundamental individual differences, explain current behavior, and predict future actions [17]. Therefore, matching tutors and tutees should also consider whether they have traits like neuroticism, extroversion, openness, agreeableness, and conscientiousness because they are significantly related to tutees' behaviors. Additionally, demographic variables such as study grade, gender, and social group can be considered to understand tutees' behavior [18].

In summary, matching tutors and tutees should be based on data regarding participants' static and dynamic characteristics in the tutoring process [5]. Static characteristics include information such as gender, age, and mother tongue. Dynamic characteristics come from the features that determine tutors' and tutees' behaviors. The challenge is finding individuals' relevant dynamic attributes to optimize the matching process [19]. The dynamic characteristics include knowledge and skills, previous tutoring experience, learning and teaching styles and preferences, personal traits, and cognitive factors like the ability to learn and the ability to solve problems and make decisions.

2.2 Multi-Criteria Decision Analysis (MCDA)

MCDA contributes to identifying options and then systematically evaluating them based on a set of decision criteria MCDA is a driving force for decision making. Currently, all sub-processes of academic tutoring, which are aimed at verifying candidates applying for the program, identifying the candidates' characteristics, and, ultimately, the key stage of pairing cooperating tutor-tutees, take place in the process of identifying people with whom the program participants would like to yield optimal results. This process is conducted traditionally, and the analysis is carried out by creating simple selection matrices. Even though only thirty candidates are admitted to the program annually, this stage is one of the most labor-intensive and tedious sub-processes, especially since the analysis is performed manually by two process coordinators. At this stage, no ICT tools supporting MCDA are used. Moreover, the selection of tutoring relationship partners is currently conducted before individual people are subjected to tests such as competence tests, work style tests, or predispositions to a role in project teams. This means that this subprocess is carried out separately from the analysis of features that can significantly

increase the effectiveness of this process. Therefore, the authors of this article believe that creating a concept of using tools supporting MCDA can significantly improve not only this stage of selecting tutoring pairs but also the quality of the tutoring process.

MCDA should be used when solving decision-making problems where a large number of criteria are accumulated that may influence the outcome of the decision-making process [20]. MCDA tools are responsible for correctly identifying and defining these criteria. In the current situation of the pairing process, it is not possible to consider more variables than just the students' indications, which are often accidental and result from a misunderstanding of the purpose of the tutoring process. If this process could use ICT solutions enabling assessment resulting from various criteria in which many factors are defined, the process would be much more adequate to the expected effects.

MCDA parameters common to all the considered alternatives constitute the basis for their assessment and determination of the most optimal one in terms of the pursued goal. In the process of selecting tutor-tutee pairs, the cooperating people must have the expected characteristics so that the cooperation can bring the highest expected results.

Usually in decision-making problems, the variables that have the greatest impact on the final decision are complex. This is also the case with the qualities of both the tutor and the tutee, especially when there will be an opportunity to submit them to prior evaluations identifying such criteria as talents, work styles, and career preferences. Multi-criteria problems are based on a large amount of information of a complex and contradictory nature, often reflecting different points of view and changing over time [21]. The primary purpose of the multi-criteria analysis is to support decision-making [22]. MCDA is a certain mechanism, sequential steps that help decision-makers make choices that are based on many different variables [23].

According to the authors, multi-criteria analysis can also be used in other stages of the tutoring process, apart from the sub-process of matching pairs cooperating throughout the development process, because it can also be used to solve such decision-making problems as sorting, ranking, describing, designing, and creating a portfolio [24].

The above concept is possible to implement in the tutoring process, but only if ICT tools are used. The traditional method of assessing candidates, verifying their characteristics, and selecting pairs, i.e., the so-called matching that takes place in the current situation does not assume the possibility of using the assumptions of multi-criteria analysis - because even with such a small group size as 30 potential candidates for tutee and 30 tutors, it is too large a set of variables. Therefore, the authors, based on their experiences resulting from participation in the BIPS program and direct participation in individual tutoring sub-processes, proposed such a concept supporting decision-making processes.

3 Methodology

3.1 Research Procedure

The research for this article was carried out according to this research procedure, including both the theoretical and empirical parts (Fig. 1).

The research procedure began with the identification of the research problem based on the author's experience, as well as co-participant observation at the WUEB. The research

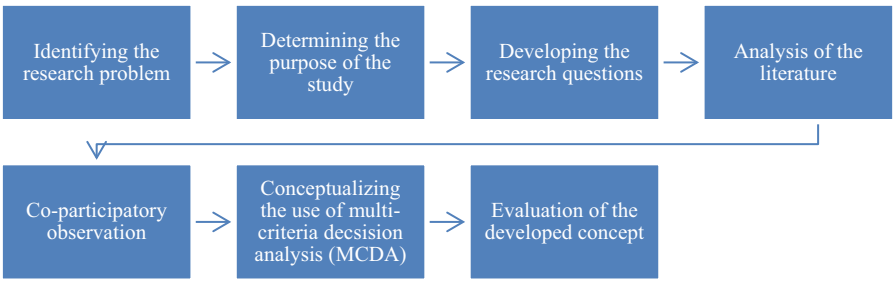


Fig. 1. Research procedure

problem was described as the necessity of transforming the process of matching tutors and tutees with the use of ICT to conduct MCDA. Therefore, the research aims to develop the concept of using multi-criteria analysis in the process of matching tutors and tutees in academic tutoring. Then the research questions presented in the introduction of the article were developed. In the next step, the authors analyzed the literature related to the described topic based on scientific papers from the Scopus database, as well as Google Scholar. Selected results were presented in the theoretical background. The authors then completed the theoretical part of the study and began to develop the concept of using MCDA in matching tutors and tutees in the academic tutoring process. The foundation for developing the concept was co-participatory observation during the BIPS program at the WUEB. The final step of the study was an evaluation of the developed concept by two authors who were tutors in the mentioned BIPS program.

3.2 The Process of Multi-Criteria Decision Analysis (MCDA)

Within the scope of this paper, the authors proposed the concept of using MCDA in the process of matching tutors and tutees in the process of academic tutoring. MCDA can be carried out within 5 stages [25] (Fig. 2).

Goals	• Develop realistic analysis objectives.
Stakeholders	• Stakeholder identification. • Identification of stakeholder needs.
Criteria	• Identify criteria for action. • Verify that the criteria meet stakeholder needs.
Weights and ratings	• Assigning weights to selected criteria. • Assigning scores to criteria.
Discussion and evaluation	• Comparison of stakeholder results to check the direction of action.

Fig. 2. Steps in multi-criteria decision analysis

It should be noted that the stages of MCDA will not be implemented in their entirety - instead, all the assumptions necessary for the entire process will be prepared. Currently, in the traditional matching process, data from the various stages are not collected. The matching process takes place before the competency tests of the tutor and tutee.

In further work, there is a possibility of extracting data from previous competency test results, entering them into the developed MCDA template, and identifying the results.

4 Results

The authors used Decision Modeling and Notation (DMN) to formulate the assumptions for conducting the multi-criteria analysis. Criteria and weights for tutor-tutee matching based on psychological tests including Gallup, FRIS, and other supplementary tests have been modeled (Fig. 3).

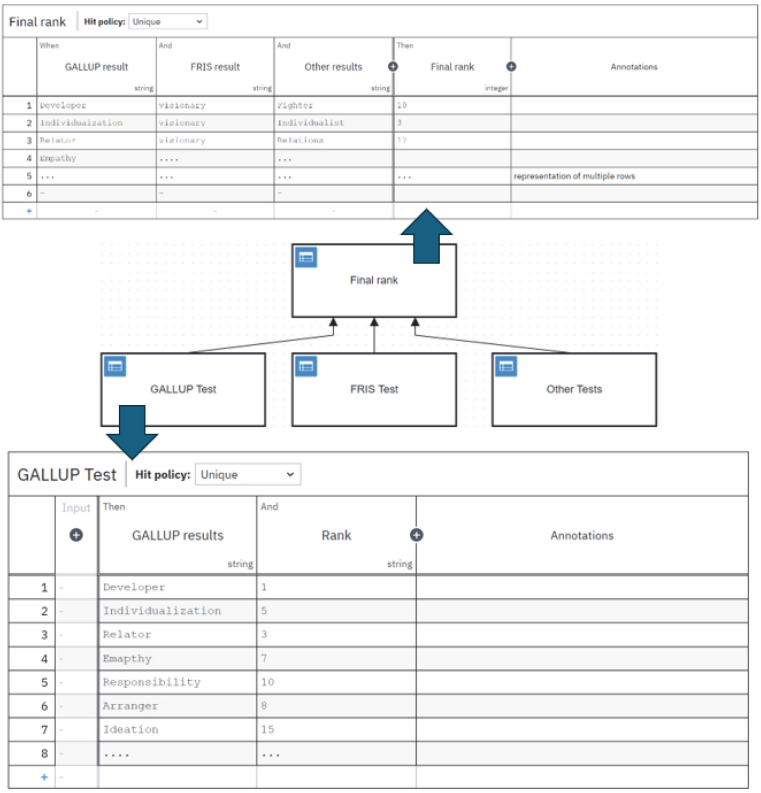


Fig. 3. Visualization and decision tables in DMN of matching assumptions

The purpose of the developed concept of using multi-criteria analysis in the tutor-tutee matching process in academic tutoring is to enhance the process and improve the accuracy of matching. Stakeholders in this analysis include students (tutees), academic teachers (tutors), and tutoring program coordinators. Among the criteria, the authors chose two tools/personality tests:

- Gallup - is a research tool created by Donald Clifton at the Gallup Institute to assess a person's so-called "talents" as well as their intensity. In the end, a participant learns about their five dominant talents, which they can use productively.
- FRIS - a tool for understanding and describing a human's style of thinking and acting based on four cognitive perspectives - facts, relations, ideas, and structures. They show the relevance of perceiving certain information.

Rules can be generated for the created decision tables (Table 1). It should be noted that only the table for the Gallup test is shown in the figure to illustrate the designed concept. If the concept is implemented, tables can be designed based on various personality tests with appropriate scales.

Table 1. Rules generated for decision tables.

Rule	Meaning
IF "GALLUP Test = = "Developer" AND "FRIS Test" = "Visionary" and "Other Tests" = "Fighter" THEN "Final rank" = 10	If the value from the GALLUP Test returns the "Developer" feature and FRIS Test returns the "Visionary" value and the Other Tests returns "Fighter" then we receive a value of 10
IF "GALLUP Test = = "Individualization" AND "FRIS Test" = "Partner" and "Other Tests" = "Individualist" THEN "Final rank" = 3	If the value from the GALLUP Test returns "Individualization" and FRIS Test returns "Partner" value and the Other Tests returns "Individualist" value, then we receive a value of 3
IF "GALLUP Test = = "Relator" AND "FRIS Test" = "Competitor" and "Other Tests" = "Relations" THEN "Final rank" = 17	If the value from the GALLUP Test returns "Relator" FRIS Test returns the "Competitor" value and Other Tests returns the "Relations" value, then we receive a value of 17

The returned "Final rank" from each rule will be used in the match-up of the tutee with a tutor. The assumption is that this matching can use more rules for a more secure match and to meet the expectations of the parties to the agreement. The values resulting from the rules prepared in this way can then serve as input values for the development and execution of a multi-criteria matching analysis. A simplified process for matching tutor and tutee is shown by the authors below (Fig. 4).

Once the tutee and tutor are accepted into the tutoring program, they complete personality tests (such as Gallup, FRIS, and others). Once filled out, an automated multi-criteria analysis is performed, which makes the connection between the tutee and the tutor. This step will be performed by AI methods using artificial neural networks, classification,

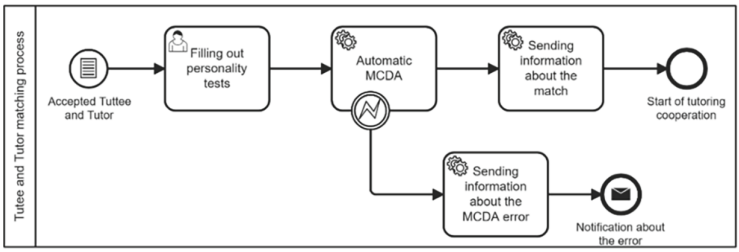


Fig. 4. A model for an improved matching process

regression, or other methods. After the pairing is done, the information about the match is sent to the interested sides. Once the information is sent, formal cooperation begins. If there are errors in the matching process, the corresponding person will be notified of the error, which will result in the application of appropriate procedures to correct the error that occurred.

5 Discussion and Conclusion

The article aimed to invent a methodology for transforming the tutor and tutee matching process with a multi-criteria analysis using ICT. Based on the study, the authors responded to the research questions.

Chi, M., Jordan, P., VanLehn, K., Litman, D. rightly noted that the key to the effectiveness of tutoring is the interactivity of its entities [7]. However, it is determined by many factors. Siler, S.A. and VanLehn, K. believe substantive knowledge is one of the most important factors [10]. Boyer, K.E., Phillips, R., Ingram, A., Ha, E.Y., Wallis, M., Vouk, M., Lester, J. believe that the teaching style determines the effectiveness of tutoring [15]. At the same time, Cohen, P.A., Kulik, J.A., Kulik, C.-L.C. rightly pointed out that the effectiveness of tutoring is also determined by the tutee's learning style [6]. Additionally, Ehrler, D.J., Evans, J.G., McGhee, R.L. emphasize that the importance of personality traits cannot be ignored [18], as they determine people's interactivity and thus influence the effectiveness of tutoring. However, the effectiveness of tutoring is not defined by the possession of a set of specific features by the tutor and the tutor but as rightly emphasized by Ehrler, D.J., Evans, J.G., McGhee, R.L. matching the features of the tutor and tutee [18]. There may be many such matching combinations, so finding the optimal pair - from the perspective of tutoring effectiveness - requires a multi-criteria analysis. It is the basis for modeling the profile of tutors and tutees (Question 1).

In the case of many tutors and tutees in a tutoring program, performing a multi-criteria analysis by a human would be time-consuming and, therefore, ineffective. As Alshaikh, F. and Hewahi, N. emphasize, AI solutions may be used in the profiling process. Thus, multi-criteria analysis could be conducted by AI-based systems using artificial neural networks (Question 3).

The DMN used to specify the rule set mainly contains three specification parts. The decision requirements graph describes how data is propagated between the various decisions made. A simple expression language is defined by how values are extracted from variables and decision tables (Fig. 5).

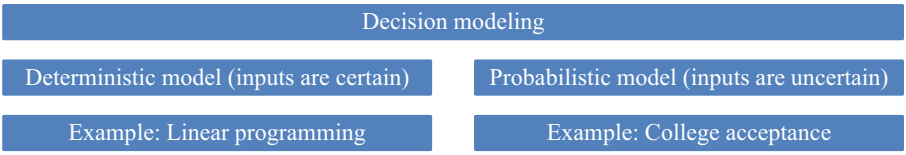


Fig. 5. Models in decision modeling

Considering their various characteristics, MCDA can be used to match tutors and tutees. Data from psychological tests, including the FFM or FRIS test, can be regarded as particularly important because, as Ehrler, D.J., Evans, J.G., McGhee, R.L write, psychological tests reveal fundamental individual differences, explain current behavior and predict future activities [18]. As Alshaikh, F., Hewahi, N. emphasize, tutor and tutee profiling aim to provide personalized and adaptive learning based on each tutee's unique characteristics and needs [9]. Thanks to the appropriate selection of pairs, cooperation can generate much greater value for both parties (Question 2).

Decision modeling notation enhances a complete understanding of the research need and improves decision-making. At the same time, it is also a scientific method of approaching problems in a replicable and repeatable way and reducing bias. When deterministic models are implemented, inputs are certain and multi-attribute decision-making and there is another stream of probabilistic models in this model there is no certainty of the data set. In our study, the primary limitation lies in the dependency on ICT systems for data accuracy and consistency. Any flaws or biases in the data input can significantly impact the outcomes of MCDA. Additionally, the scope of the study is confined to the context of the WUEB, which may limit the generalizability of the findings to other institutions or settings.

In terms of theory, our study contributes to the existing body of knowledge on the integration of ICT in educational processes, particularly in improving the accuracy of decision-making through MCDA. Practically, the implementation of our proposed ICT-assisted matching process can streamline tutor-tutee pairings, potentially reducing errors and improving the overall effectiveness of academic tutoring. This has significant implications for educational institutions aiming to leverage technology for improved student support and development. Further research can be carried forward by automating the process with portal execution semantics in which input data nodes, decision nodes, and reusable logic components. Behind each element in the decision, modeling is driven by supporting decision logic typically in the form of decision tables. In further research, we can combine BPMN (Business Process Model and Notation) and DMN processes with machine learning predictive models to attain an explainable AI solution.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Łuczak, K., Chomiak-Orsa, I., Piwoni-Krzeszowska, E.: The use of an ontology to evaluate the suitability of using Design Thinking in teaching based on the syllabus. Procedia Comput. Sci. **225**, 2714–2722 (2023). https://doi.org/10.1016/j.procs.2023.10.263
2. Luck, C.: Challenges faced by tutors in Higher Education. Psychodyn. Pract. **16**, 273–287 (2010). https://doi.org/10.1080/14753634.2010.489386
3. Bloom, B.S.: The 2 sigma problem: the search for methods of group instruction as effective as one-to-one tutoring. Educ. Res. **13**, 4–16 (1984)
4. Cohen, P.A., Kulik, J.A., Kulik, C.-L.C.: Educational outcomes of tutoring: a meta-analysis of findings. Am. Educ. Res. J. **19**, 237–248 (1982). https://doi.org/10.3102/00028312019002237
5. Alkhatlan, A., Kalita, J.: Intelligent tutoring systems: a comprehensive historical survey with recent developments. Int. J. Comput. Appl. **181**, 1–20 (2019). https://doi.org/10.5120/ijca2019918451
6. Chi, M., Jordan, P., VanLehn, K., Litman, D.: To elicit or to tell: does it matter? In: Proceedings of 14th International Conference on Artificial Intelligence in Education, pp. 197–204 (2009)
7. Bond, M., et al.: A meta systematic review of artificial intelligence in higher education: a call for increased ethics, collaboration, and rigour. Int. J. Educ. Technol. High. Educ. **21**, 4 (2024). https://doi.org/10.1186/s41239-023-00436-z
8. Alshaikh, F., Hewahi, N.: Convolutional neural network for predicting student academic performance in intelligent tutoring system. Int. J. Comput. Digit. Syst. **15**, 239–258 (2024). https://doi.org/10.12785/ijcds/150119
9. Siler, S.A., VanLehn, K.: Learning, interactional, and motivational outcomes in one-to-one synchronous computer-mediated versus face-to-face tutoring. Int. J. Artif. Intell. Educ. **19**, 73–102 (2009)
10. Evens, M., Michael, J.: One-on-One Tutoring by Humans and Computers. Psychology Press (2006). https://doi.org/10.4324/9781410617071
11. Cade, W.L., Copeland, J.L., Person, N.K., D'Mello, S.K.: Dialogue modes in expert tutoring. In: Woolf, B.P., Aïmeur, E., Nkambou, R., and Lajoie, S. (eds.) Intelligent Tutoring Systems, pp. 470–479. Springer, Berlin, Heidelberg (2008). https://doi.org/10.1007/978-3-540-69132-7_50
12. Rosé, C.P., Moore, J.D., VanLehn, K., Allbritton, D.: A Comparative evaluation of socratic versus didactic tutoring. Proc. Annu. Meet. Cogn. Sci. Soc. **23**, 869–874 (2001)
13. Graesser, A.C., Person, N.K., Magliano, J.P.: Collaborative dialogue patterns in naturalistic one-to-one tutoring. Appl. Cogn. Psychol. **9**, 495–522 (1995). https://doi.org/10.1002/acp.2350090604
14. Boyer, K.E., et al.: Characterizing the effectiveness of tutorial dialogue with hidden Markov models. In: Aleven, V., Kay, J., Mostow, J. (eds.) Intelligent Tutoring Systems, pp. 55–64. Springer, Berlin, Heidelberg (2010). https://doi.org/10.1007/978-3-642-13388-6_10
15. D'Mello, S., Taylor, R., Graesser, A.: Monitoring affective trajectories during complex learning. Proc. Annu. Meet. Cogn. Sci. Soc. **29**, 203–208 (2007)
16. Lepper, M.R., Woolverton, M., Mumme, D.L., Gurtner, J.-L.: Motivational techniques of expert human tutors: lessons for the design of computer-based tutors. In: Computers as Cognitive Tools. L. Erlbaum Associates, Hillsdale, NJ (1993)
17. Ehrler, D.J., Evans, J.G., McGhee, R.L.: Extending Big-Five theory into childhood: a preliminary investigation into the relationship between Big-Five personality traits and behavior problems in children. Psychol. Sch. **36**, 451–458 (1999). https://doi.org/10.1002/(SICI)1520-6807(199911)36:6%3c451::AID-PITS1%3e3.0.CO;2-E

18. Chen, P., Lu, Y.: An AI-powered teacher assistant for student problem behavior diagnosis. In: Niemi, H., Pea, R.D., and Lu, Y. (eds.) AI in Learning: Designing the Future. pp. 91–104. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-09687-7_6
19. Chrysafiadi, K., Virvou, M.: Student modeling approaches: a literature review for the last decade. Expert Syst. Appl. **40**, 4715–4729 (2013). https://doi.org/10.1016/j.eswa.2013.02.007
20. Trendowicz, A., Kopczyńska, S.: Adapting multi-criteria decision analysis for assessing the quality of software products. Current approaches and future perspectives. In: Advances in Computers. pp. 153–226. Elsevier (2014). https://doi.org/10.1016/B978-0-12-800162-2.00004-X
21. Belton, V., Stewart, T.J.: Multiple Criteria Decision Analysis. Springer US, Boston, MA (2002). https://doi.org/10.1007/978-1-4615-1495-4
22. Majumder, M.: Multi criteria decision making. In: Impact of Urbanization on Water Shortage in Face of Climatic Aberrations. pp. 35–47. Springer, Singapore (2015). https://doi.org/10.1007/978-981-4560-73-3_2
23. Cinelli, M., Coles, S.R., Kirwan, K.: Analysis of the potentials of multi criteria decision analysis methods to conduct sustainability assessment. Ecol. Indic. **46**, 138–148 (2014). https://doi.org/10.1016/j.ecolind.2014.06.011
24. Grigoroudis, E., Matsatsinis, N. eds: Preference Disaggregation in Multiple Criteria Decision Analysis: Essays in Honor of Yannis Siskos. Springer International Publishing : Imprint: Springer, Cham (2018). https://doi.org/10.1007/978-3-319-90599-0
25. Multi-Criteria Decision Analysis (MCDA/MCDM), https://www.1000minds.com/decision-making/what-is-mcdm-mcda#how-does-mcda-work, last accessed 2024/06/01

Sustainable Development Strategies Assessment Using the New Multi-Target TOPSIS Method

Aleksandra Bączkiewicz[1(✉)], Jarosław Wątróbski[1,2], Robert Król[1], and Rafał Pawlak[2]

[1] Institute of Management, University of Szczecin, ul. Cukrowa 8, 71-004 Szczecin, Poland
{aleksandra.baczkiewicz,jaroslaw.watrobski}@usz.edu.pl,
robert.krol@phd.usz.edu.pl
[2] National Institute of Telecommunications, ul. Szachowa 1, 04-894 Warsaw, Poland
r.pawlak@il-pib.pl

Abstract. Sustainability is a complex concept that requires consideration of multiple dimensions in assessment. To fulfill these requirements, multi-criteria decision-making methods (MCDA) are employed to aggregate multiple sustainability indicators. However, in many MCDA methods such as Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS), for example, the phenomenon of criteria compensation is observed, which is contrary to the strong sustainability paradigm. The mentioned limitation became the motivation for developing a new Multi-Target Technique for Order of Preference by Similarity to Ideal Solution (MT-TOPSIS) method based on the classic TOPSIS method, which allows multi-criteria evaluation with consideration of performances of alternatives concerning individual target values. The proposed method was used to evaluate European countries in terms of implementing the Europe 2020 strategy. The results demonstrate the complexity of evaluating sustainable development involving multiple targets and how to take them into account.

Keywords: Multi-Target TOPSIS · Sustainable development · MCDA · Sustainable Development Goals

1 Introduction

This paper presents the application of an innovative extension of the TOPSIS (Technique for Order of Preference by Similarity to Ideal Solution) method called Multi-Target Technique for Order of Preference by Similarity to Ideal Solution (MT-TOPSIS) to evaluate the European countries' implementation of the Europe 2020 program. Europe 2020 is the EU's agenda for improving strategic areas such as employment and economic growth [8]. The European Council adopted this program on 17 June 2010 as the successor to the Lisbon Strategy [2]. Its objectives include intelligent and sustainable social and economic growth in the EU for the upcoming decade [10]. The strategy's primary goals are to achieve high employment rates, productivity, and social cohesion while minimizing harmful effects on the environment [14]. To achieve its objectives, the EU adopted

M. Hernes et al. (Eds.): ECAI 2024, LNNS 1217, pp. 249–260, 2025.
https://doi.org/10.1007/978-3-031-78465-1_21

ambitious targets covering fields of employment, research and development (R&D), climate change and energy, education, and poverty reduction, which were to be achieved by the Member States by 2020. These targets were converted into individual national targets, reflecting the situation and capacity of each Member State to contribute to the common objectives of the strategy [5]. Indicators designed specially to support the Europe 2020 strategy are used to monitor progress toward the set targets and assess the distance to reach them [11]. Data on the indicators are available on the Eurostat website's Europe 2020 headline indicators section. A set of nine indicators to support the analysis of member country progress is included in Table 1. The European strategy adopted the above objectives under three key priorities: Smart growth: development of an economy based on knowledge, research, innovations, and education of the society, Sustainable growth: promoting a resource-efficient, green, competitive economy, Inclusive growth: supporting job creation and poverty reduction [9].

Targets included in the Europe 2020 strategy reflect the level of ambition of each country based on individual capabilities that they can set in their common pursuit of the Europe 2020 strategy. The indicators allow individual country performance to be assessed against the targets set. Currently, the Europe 2020 strategy has reached the end of its life cycle, making an important contribution to the socio-economic development of the EU [1]. However, the 2030 Agenda for Sustainable Development continues to implement this strategy, and Eurostat will continue to play a key role in monitoring the sustainable development of EU countries. Therefore, the strategy's conclusion is the appropriate moment to summarize and assess the degree to which individual EU countries have fulfilled their commitments under the strategy. These indicators provide a valuable statistical tool for analyzing the achievement of the EU strategy objectives and can be used at the national level and in assessing how well individual objectives have been met compared to other countries. Multiple criteria appearing in the framework of the Europe 2020 program make the Multi-Criteria Decision Analysis (MCDA) methods suitable for evaluating alternatives with regard to its criteria, allowing for the simultaneous inclusion of many criteria given in different units and with opposing objectives (maximization for profit criteria and minimization for cost criteria [6]. In previous work, researchers have attempted to evaluate European countries in the context of the implementation of the Europe 2020 strategy using multi-criteria methods such as TOPSIS and VIKOR [12]. Although these methods allow multiple evaluation criteria to be taken into account simultaneously, their basic version does not offer the possibility to evaluate alternatives in relation to individual expected target values. This is because these methods make it possible to evaluate alternatives by taking into account their performance in relation to reference points determined generally for a full set of considered alternatives [7].

The reference solutions are determined differently depending on the algorithm of the MCDA method, for example, for the TOPSIS method the reference solutions are positive ideal solution and negative ideal solution, and the alternatives are evaluated based on the Euclidean distance from them. For the VIKOR (VlseKriterijumska Optimizacija I Kompromisno Resenje) method, the reference solution is the ideal solution, similar to the SPOTIS (Stable Preference Ordering Towards Ideal Solution) method [13]. In the methods discussed above, the construction of reference solutions is performed based on the full dataset collectively for all alternatives. The alternatives are then evaluated

Table 1. Evaluation criteria of the implementation of the Europe 2020 strategy.

C_j	Criterion name	Unit	Sustainability target	Dimension (key priority)
C_1	Employment rate, age group 20–64	% of total population	Maximize	Inclusive growth
C_2	Gross domestic expenditure on R&D (GERD)	% of gross domestic product (GDP)	Maximize	Smart growth
C_3	Greenhouse gas emissions	% of amount emitted in 1990	Minimize	Sustainable growth
C_4	Share of renewable energy in gross final energy consumption	%	Maximize	Sustainable growth
C_5	Primary energy consumption	MTOE (Million tonnes of oil equivalent)	Minimize	Sustainable growth
C_6	Final energy consumption	MTOE (Million tonnes of oil equivalent)	Minimize	Sustainable growth
C_7	Greenhouse gas emissions in ESD sectors	MTOE (Million tonnes of oil equivalent)	Minimize	Sustainable growth
C_8	Early leavers from education and training	% of population aged 18-24	Minimize	Smart growth
C_9	Tertiary educational attainment	% of population aged 30-34	Maximize	Smart growth
C_{10}	People at risk of poverty or social exclusion	Thousand persons	Minimize	Inclusive growth

in relation to a single reference solution. A limitation of this approach is the lack of an individual approach for each alternative, which is required, for example, when assessing the sustainable development of countries. Different countries have different possibilities and potentials depending on their individual conditions and it may be desirable to assess their implementation of individually defined goals. An example of such a decision problem is the assessment of European countries against the implementation of goals included in the Europe 2020 strategy. In this decision problem, the aim is to assess the implementation of the goals in relation to the targeted values, which are defined individually for each country in the target matrix, instead of assessing countries concerning each other.

The lack of a customized approach to evaluating individual alternatives in relation to individually constructed reference solutions constitutes a research gap that the authors

of this paper aim to fill. Therefore, in this paper, the authors propose a novel MT-TOPSIS method that allows one to take into account the ideal reference points assigned individually for each evaluated alternative, taking into account its current objectives and the possibility of achieving them. Dataset with criteria performance and target values collected for evaluated countries from Eurostat are included in the Github repository along with the implemented MT-TOPSIS method in Python 3 at the link https://github.com/energyinpython/Multi-Target-TOPSIS. The rest of the paper is organized as follows. Section 2 provides fundamentals, assumptions, and mathematical formulas developed for MT-TOPSIS. Section 3 provides and discusses research outcomes. Finally, in Section 4, the research summary is drawn and limitations and future work directions are outlined.

2 Methodology

The proposed Multi-Target TOPSIS (MT-TOPSIS) method is the new extension of the classical TOPSIS method, allowing evaluation alternatives concerning targets set individually for each alternative. Such an extension is necessary when, due to the specificity of the problem, the individual goals for each alternative have different values, and it is recommended that the overall preference value is obtained based on them. Assume that $X = [x_{ij}]_{m\times n}$ is a decision matrix with criteria performance values for evaluated alternatives represented by Equation (1)

$$X = [x_{ij}]_{m\times n} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix} \tag{1}$$

where m denotes number of alternatives ($i = 1, 2, \ldots, m$) and n means number of criteria ($j = 1, 2, \ldots, n$), and $T = [t_{ij}]_{m\times n}$ is target matrix containing target criteria performance values individually designated for evaluated alternatives expressed by Equation (2).

$$T = [t_{ij}]_{m\times n} = \begin{bmatrix} t_{11} & t_{12} & \cdots & t_{1n} \\ t_{21} & t_{22} & \cdots & t_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ t_{m1} & t_{m2} & \cdots & t_{mn} \end{bmatrix} \tag{2}$$

Step 1. Thresholding of target matrix. This step includes thresholding of target matrix with goal criteria performance values. The target matrix contains individually set target performance values for each alternative. These values represent the goals that the alternatives individually pursue. Thus, attainment of assumed values is identified as the achievement of the targets. An alternative may achieve a better value for a criterion than the target value. For example, such a situation occurs when the alternative achieves a higher value than the target value for the profit criterion or when the alternative reaches a lower value than the target value for the cost criterion. In the TOPSIS method, the distance of the alternatives from the reference solutions is calculated using the Euclidean distance, which always has a positive value. Obtaining better-than-target results would

cause increasing distances from the reference ideal solution to be determined in the second-to-last step of the algorithm. The mentioned situation would result in a paradoxical worsening of final scores as the alternative achieves results better than the target. Therefore, a target matrix thresholding procedure is required. The target matrix thresholding procedure involves replacing the values in the target matrix with decision matrix values when the decision matrix values are higher in the case of profit criteria or lower for cost criteria. The target matrix thresholding is provided in the following Equation (3)

$$t_{ij} = \begin{cases} x_{ij} \ if \ type = 1 \ \wedge \ x_{ij} > t_{ij} \ \vee \ type = -1 \ \wedge \ x_{ij} < t_{ij} \\ t_{ij} \ if \ type = 1 \ \wedge \ x_{ij} \leq t_{ij} \ \vee \ type = -1 \ \wedge \ x_{ij} \geq t_{ij} \end{cases} \tag{3}$$

where *type* denotes criterion type and takes value 1 for profit type and -1 for cost type.

Step 2. Normalization of decision matrix and target matrix. Normalization techniques using maximum or minimum and maximum values are recommended for decision matrix and target matrix normalization in the MT-TOPSIS algorithm. These techniques include minimum-maximum normalization, linear normalization, and maximum normalization. Boundary vectors containing maximum and minimum values required for these normalization techniques are determined from the decision matrix and target matrix. A matrix combining the decision matrix and target matrix is created for this purpose by vertical concatenation of matrices X and T, as Equation (4) demonstrates.

$$A = [a_{ij}]_{2m \times n} \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \\ t_{11} & t_{12} & \cdots & t_{1n} \\ t_{21} & t_{22} & \cdots & t_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ t_{m1} & t_{m2} & \cdots & t_{mn} \end{bmatrix} \tag{4}$$

In minimum-maximum normalization normalized values of decision matrix for profit criteria r_{ij}^{+} and for cost criteria r_{ij}^{-} are obtained applying Equation (5).

$$r_{ij}^{+} = \frac{x_{ij} - min_j(a_{ij})}{max_j(a_{ij}) - min_j(a_{ij})}, \ r_{ij}^{-} = \frac{max_j(a_{ij}) - x_{ij}}{max_j(a_{ij}) - min_j(a_{ij})} \tag{5}$$

where r_{ij} denotes normalized values of the decision matrix. The normalization of the target matrix is performed analogously, for profit criteria rt_{ij}^{+} and for cost criteria rt_{ij}^{-} as shown in Equation (6).

$$rt_{ij}^{+} = \frac{t_{ij} - min_j(a_{ij})}{max_j(a_{ij}) - min_j(a_{ij})}, \ rt_{ij}^{-} = \frac{max_j(a_{ij}) - t_{ij}}{max_j(a_{ij}) - min_j(a_{ij})} \tag{6}$$

Step 3. Calculation of weighted normalized decision matrix and weighted normalized target matrix. Weighted normalized decision matrix is computed according

to Equation (7) and weighted normalized target matrix is calculated as Equation (8) demonstrates,

$$v_{ij} = w_j r_{ij} \tag{7}$$

$$vt_{ij} = w_j rt_{ij} \tag{8}$$

where w_j represents the weight of jth criterion, v_{ij} denotes weighted normalized values of the decision matrix and vt_{ij} express weighted normalized values of target matrix.

Step 4. Determination of Positive Ideal Solution (PIS) and Negative Ideal Solution (NIS). In the MT-TOPSIS algorithm, the weighted normalized target matrix vt plays the role of vectors with PIS for each alternative. NIS includes minimum values of weighted normalized decision matrix and is determined as Equation (9) shows.

$$v_j^- = \{v_1^-, v_2^-, \ldots, v_n^-\} = \{min_j(v_{ij})\} \tag{9}$$

Step 5. Calculation of Euclidean distance from PIS and NIS. Calculation of alternatives' distance from individual PIS (D_i^+) and NIS (D_i^-) included in vt is performed for each alternative according to Equation (10).

$$D_i^+ = \sqrt{\sum_{j=1}^{n}(v_{ij} - vt_{ij})^2},\ D_i^- = \sqrt{\sum_{j=1}^{n}(v_{ij} - v_j^-)^2} \tag{10}$$

Step 6. Calculation of the relative closeness to the Positive Ideal Solution. Calculation of the score for each considered alternative according to Equation (11). This step is identical to the last step of the classic TOPSIS. The C_i value is always between 0 to 1, and the alternative that has the highest C_i value is the best. It implies that for TOPSIS, the ranking is created by descending sorting of alternatives by preference value.

$$C_i = \frac{D_i^-}{D_i^- + D_i^+} \tag{11}$$

3 Results

This section presents the MT-TOPSIS ranking for countries evaluated according to the implementation of the Europe 2020 strategy in 2020, taking into account the achievement of the criteria values concerning the goals set in the strategy. The contents of this section include a dataset of performance values of evaluated countries in terms of Europe 2020 targets, target performance values determined for considered countries regarding the mentioned criteria and results of MT-TOPSIS evaluation in comparison with assessment using reference classical MCDA methods: TOPSIS, SPOTIS (Stable Preference Ordering Towards Ideal Solution) [4], and VIKOR (VlseKriterijumska Optimizacija I Kompromisno Resenje) [3], which do not consider target values individually. The selection of a set of reference MCDA methods is justified because TOPSIS is the basis of

newly developed MT-TOPSIS, and SPOTIS and VIKOR are methods considering ideal reference points in assessment. Thus, the assumptions of all methods included in the research are similar. Table 2 displays performance values of evaluated countries collected from the Eurostat database for 2020 in terms of nine indicators that are criteria of model assessment.

Table 2. Decision matrix with countries' performances collected for 2020.

Country	Symbol	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9
Belgium	BE	70	3.52	82.67	9.924	43.88	33.29	74.28	8.1	47.8
Bulgaria	BG	73.1	0.86	57.16	21.564	17.19	9.54	27.12	12.8	33.3
Czechia	CZ	79.7	1.99	64.82	16.244	37.47	24.48	67.76	7.6	35
Denmark	DK	77.8	3.03	70.69	37.204	15.32	13.15	32.49	9.3	49.8
Germany	DE	80	3.14	70.44	17.354	262.49	201.66	439.66	10.1	36.3
Estonia	EE	78.8	1.79	49.98	31.889	4.31	2.78	6.53	7.5	44.3
Ireland	IE	73.4	1.23	113.6	11.984	13.43	11.18	44.56	5	58.1
Greece	EL	61.1	1.49	90.84	19.677	19.68	14.33	44.29	3.8	43.9
Spain	ES	65.7	1.41	119.74	18.362	105.03	73.76	200.93	16	44.8
France	FR	72.1	2.35	83.1	17.216	208.36	130.23	341.13	8	48.8
Croatia	HR	66.9	1.27	75.23	28.466	7.76	6.47	16.82	2.2	34.7
Italy	IT	62.6	1.54	84.41	18.181	132.32	102.74	272.34	13.1	27.8
Cyprus	CY	74.9	0.85	153.81	13.8	2.2	1.57	4.31	11.5	59.8
Latvia	LV	76.9	0.7	45.95	40.975	4.26	3.86	9.04	7.2	49.2
Lithuania	LT	76.7	1.17	42.64	25.461	6.23	5.31	14.36	5.6	59.6
Luxembourg	LU	72.1	1.13	94.16	7.047	3.94	3.81	9.23	8.2	62.2
Hungary	HU	77.5	1.62	67.82	12.614	23.89	18.01	43.48	12.1	33.2
Malta	MT	77.3	0.66	96.14	8.488	0.74	0.54	1.38	12.6	39.8
Netherlands	NL	80	2.29	88.58	8.768	58.38	45.52	100.16	7	54
Austria	AT	74.8	3.22	102.66	33.626	29.73	26.07	50.73	8.1	41.6
Poland	PL	73.6	1.39	87.42	12.164	96.53	71	206.85	5.4	47
Portugal	PT	74.2	1.58	118.9	30.619	19.54	15.02	41.56	8.9	39.6
Romania	RO	70.8	0.47	46.84	24.29	30.92	23.53	75.46	15.6	26.4
Slovenia	SI	75.6	2.15	94.35	21.974	6.13	4.39	10.82	4.1	46.9
Slovakia	SK	72.5	0.92	59.16	16.894	15.15	10.34	21.77	7.6	39.7
Finland	FI	76.5	2.94	81.41	43.081	29.8	23.28	29.32	8.2	49.6
Sweden	SE	80.8	3.51	75.28	56.391	41.82	30.93	30.5	7.7	52.2

Table 3 provides a matrix with target values of performances assigned to particular countries in terms of nine indicators acquired from the Eurostat database. The leader of the MT-TOPSIS ranking is Croatia. Croatia achieved a leading position in the ranking of the method taking into account the Europe 2020 targets, although in the rankings of the classical methods evaluating only performances, the country ranked 14th for TOPSIS and SPOTIS and 12th for VIKOR. It means that Croatia has achieved performances realizing the goals set for the country to the greatest extent among all the countries considered. When the level of achievement of the Europe 2020 goals is considered during the evaluation, Croatia scored higher than the leaders of the classic MCDA rankings, such as Sweden, Finland, and Denmark. This result shows that the requirements for Croatia, considering the country's capabilities against the Europe 2020 criteria, are set lower than for the leaders of the classic MCDA rankings.

The second place in the MT-TOPSIS rankings went to Denmark. The country is ranked third in the TOPSIS, SPOTIS, and VIKOR rankings. This result confirms Denmark's strong position in terms of overall values and meeting the highly set Europe 2020 targets among other European countries. MT-TOPSIS ranks Sweden third. On the other hand, Sweden is the leader in the rankings of the three classic MCDA methods included in this analysis. This means that Sweden has a very high overall performance of the Europe 2020 criteria, which allows it to obtain a leading position and achieve the targets set at a good level, reaching its third place in this respect.

For comparative analysis, the results were compared with the ranking of the classical TOPSIS method, which was generated based on the criteria' values without considering the objectives' values. It can be observed that the MT-TOPSIS ranking built based on evaluation with target values differs significantly from the TOPSIS ranking generated only based on performances. In the next stage, in addition to the classic TOPSIS method, two other classic MCDA methods that are also based on reference solutions were included in the comparative analysis: SPOTIS [4] and VIKOR [3] considering only alternatives' performances. Results are displayed in Figure 1 and provided on GitHub in detail.

Among the countries that have been rated much better in the MT-TOPSIS ranking, which takes into account the level of achievement of the Europe 2020 targets individually for each country rather than generally as it happens in the classic MCDA methods can be mentioned Czech Republic, Germany, Croatia, Italy, Hungary, Netherlands, Poland, and Slovakia. On the other hand, Greece, Spain, Cyprus, Portugal, and Romania are among the weakest countries in achieving the individual Europe 2020 targets. The observed differences in results reflected in different rankings demonstrate significant differences between the MT-TOPSIS approach and classical approaches such as TOPSIS, VIKOR, and SPOTIS. These differences show that for a decision problem requiring the evaluation of alternatives against individually determined reference solutions, the use of classical MCDA approaches is not adequate, as it does not allow for the consideration of individual reference solutions for each alternative.

In the next stage of the research, Pearson's correlation coefficient was used to compare the rankings obtained by each MCDA method objectively. The correlation values displayed in Figure 2 show that the rankings of the classical MCDA methods considering only performances, i.e., TOPSIS, SPOTIS, and VIKOR, are convergent, as their correlations have high values close to 1. The correlations between the MT-TOPSIS and the

Table 3. Matrix with countries' target performances in Europe 2020 program.

Country	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9
Belgium	73.2	3	80	13	43.7	32.5	68.25	9.5	47
Bulgaria	76	1.5	80	16	16.9	8.6	26.54	11	36
Czechia	75	1	80	13	39.6	25.3	67.2	5.5	32
Denmark	80	3	80	30	17.4	14.4	32.06	10	40
Germany	77	3	80	18	276.6	194.3	410.91	10	42
Estonia	76	3	80	25	6.5	2.8	6.02	9.5	40
Ireland	69	2	80	16	13.9	11.7	37.65	8	60
Greece	70	1.2	80	18	24.7	18.4	60.05	10	32
Spain	74	2	80	20	119.8	80.1	212.39	15	44
France	75	3	80	23	219.9	131.4	342.48	9.5	50
Croatia	62.9	1.4	80	20	11.15	7	19.32	4	35
Italy	67	1.53	80	17	158	124	291.01	16	26
Cyprus	75	0.5	80	13	2.2	1.8	3.98	10	46
Latvia	73	1.5	80	40	5.4	4.5	9.99	10	34
Lithuania	72.8	1.9	80	23	6.5	4.3	15.24	9	48.7
Luxembourg	73	2.3	80	11	4.5	4.2	8.12	10	66
Hungary	75	1.8	80	13	24.1	14.4	52.83	10	34
Malta	70	2	80	10	0.7	0.5	1.17	10	33
Netherlands	80	2.5	80	14	60.7	52.2	107.36	8	40
Austria	77	3.76	80	34	31.5	25.1	47.75	9.5	38
Poland	71	1.7	80	15	96.4	71.6	205.18	4.5	45
Portugal	75	2.7	80	31	22.5	17.4	49.08	10	40
Romania	70	2	80	24	43	30.3	89.81	11.3	26.7
Slovenia	75	3	80	25	7.3	5.1	12.31	5	40
Slovakia	72	1.2	80	14	16.4	9	25.95	6	40
Finland	78	4	80	38	35.9	26.7	28.51	8	42
Sweden	80	4	80	49	43.4	30.3	36.08	7	45

TOPSIS, SPOTIS, and VIKOR rankings are lower, demonstrating their divergence. The results obtained confirm the similarity of the performance of classical MCDA approaches such as TOPSIS, VIKOR, and SPOTIS and the divergence of the results returned by MT-TOPSIS. This demonstrates that MT-TOPSIS evaluates alternatives in a different way than reference methods, taking into account individual reference solutions, which is confirmed by the low correlation with the rankings given by classical MCDA methods. Different algorithms cause the observed divergences for considering performance

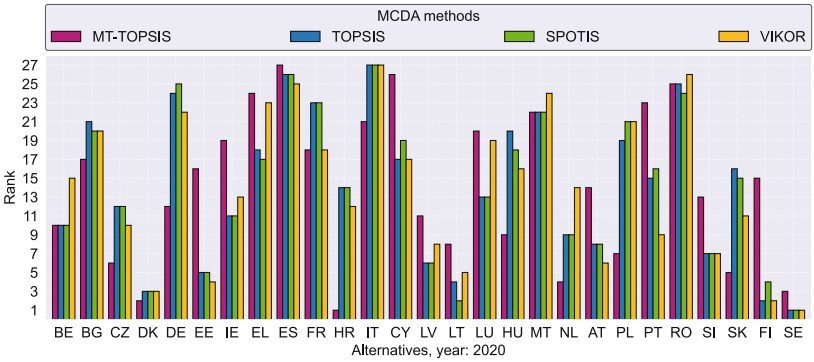


Fig. 1. MCDA rankings of evaluated countries.

data by the MT-TOPSIS method and the other classic MCDA methods included in the research.

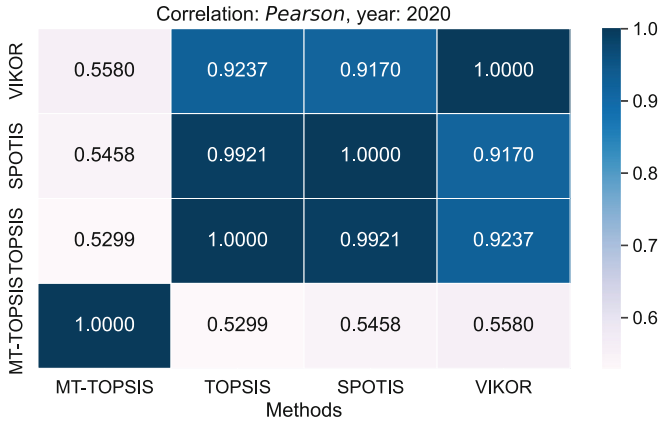


Fig. 2. Correlation between MT-TOPSIS ranking and classical MCDA rankings.

MT-TOPSIS evaluates alternatives in terms of the proximity of performances to individual targets. It means that in the chosen strategy, each country has set target values to pursue, and it is natural that they differ between countries. This is because individual countries vary in area, population, geography, climate, and government policy. Using the MT-TOPSIS method, a country closer to reaching an individual target will be evaluated better than a country with a higher performance value, but its achievement relative to the individual target is weaker.

4 Conclusions

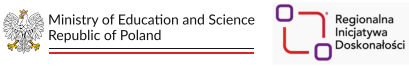
The purpose of this paper was to present a newly developed methodological framework based on the basic assumptions of the TOPSIS method, which makes it possible to evaluate sustainability goals by taking into account target values. The newly developed

MT-TOPSIS method eliminates a limitation of the classic TOPSIS algorithm, which, by aggregating all criteria into a single score, results in their full compensation, so good performances within others can compensate for poor performances of some criteria. In practical terms, the MT-TOPSIS method has been employed to perform a comprehensive assessment of the European Union countries involving a multi-criteria analysis of the level of implementation of the recommendations of the Europe 2020 strategy. The results show that MT-TOPSIS shows great usefulness when assessing the implementation of strategies for which we have target values of individual evaluation criteria.

The proposed method has demonstrated its practical usefulness in assessing European countries against the implementation of the Europe 2020 strategy. It shows potential for usefulness in other practical problems of sustainability assessment. Among them are problems of evaluation concerning the implementation of any established development programs in which specific measurable targets are set for the individual alternatives to be evaluated, such as, for example, countries, regions, cities, and companies. Such problems may include the evaluation of the renewable energy economy or the sustainable development of cities, settlements, and communities. The proposed approach can be useful in assessing and identifying potential areas of improvement when extended by sensitivity analysis.

As this is preliminary research focused on multi-criteria analysis considering individual reference solutions, a limitation of this paper is the application of only one MT-TOPSIS method considering individual reference solutions. A further limitation is the experiment with the proposed approach performed for a single dataset. Future work directions include the development of other MCDA methods such as VIKOR and SPOTIS towards the possibility of including individual reference solutions and the use of the developed approaches for other datasets.

Acknowledgments. This research was partially funded by National Science Centre, Poland 2022/45/B/HS4/02960 (J.W.), and Co-financed by the Minister of Science under the "Regional Excellence Initiative" Program RID/SP/0046/2024/01 (A.B.).



Disclosure of Interests The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Arbolino, R., Carlucci, F., De Simone, L., Ioppolo, G., Yigitcanlar, T.: The policy diffusion of environmental performance in the European countries. Ecol. Ind. **89**, 130–138 (2018). https://doi.org/10.1016/j.ecolind.2018.01.062
2. Becker, W., Norlén, H., Dijkstra, L., Athanasoglou, S.: Wrapping up the Europe 2020 strategy: a multidimensional indicator analysis. Environ. Sustain. Ind. **8**, 100075 (2020). https://doi.org/10.1016/j.indic.2020.100075

3. Bera, B., Shit, P.K., Sengupta, N., Saha, S., Bhattacharjee, S.: Susceptibility of deforestation hotspots in terai-dooars belt of himalayan foothills: a comparative analysis of vikor and topsis models. J. King Saud University-Computer Inf. Sci. **34**(10), 8794–8806 (2022). https://doi.org/10.1016/j.jksuci.2021.10.005
4. Dezert, J., Tchamova, A., Han, D., Tacnet, J.M.: The spotis rank reversal free method for multi-criteria decision-making support. In: 2020 IEEE 23rd International Conference on Information Fusion (FUSION). pp. 1–8. IEEE (2020). https://doi.org/10.23919/FUSION45008.2020.9190347
5. Eurostat: Smarter, greener, more inclusive? Indicators to support the Europe 2020 strategy (2019), https://ec.europa.eu/eurostat/web/products-statistical-books/-/ks-04-19-559
6. Fedajev, A., Stanujkic, D., Karabašević, D., Brauers, W.K., Zavadskas, E.K.: Assessment of progress towards "Europe 2020" strategy targets by using the multimoora method and the shannon entropy index. J. Clean. Prod. **244**, 118895 (2020). https://doi.org/10.1016/j.jclepro.2019.118895
7. Fura, B., Wojnar, J., Kasprzyk, B.: Ranking and classification of EU countries regarding their levels of implementation of the Europe 2020 strategy. J. Clean. Prod. **165**, 968–979 (2017). https://doi.org/10.1016/j.jclepro.2017.07.088
8. Miola, A., Schiltz, F.: Measuring sustainable development goals performance: how to monitor policy action in the 2030 agenda implementation? Ecol. Econ. **164**, 106373 (2019). https://doi.org/10.1016/j.ecolecon.2019.106373
9. Moreno, B., García-Álvarez, M.T.: Measuring the progress towards a resource-efficient European Union under the Europe 2020 strategy. J. Clean. Prod. **170**, 991–1005 (2018). https://doi.org/10.1016/j.jclepro.2017.09.149
10. Schoukens, P., De Becker, E., Smets, J.B.: Fighting social exclusion under the Europe 2020 strategy: which legal nature for social inclusion recommendations? Int. Comp. Jurisprud. **1**(1), 11–23 (2015). https://doi.org/10.1016/j.icj.2015.10.003
11. Siksnelyte, I., Zavadskas, E.K., Bausys, R., Streimikiene, D.: Implementation of eu energy policy priorities in the baltic sea region countries: sustainability assessment based on neutrosophic multimoora method. Energy Policy **125**, 90–102 (2019). https://doi.org/10.1016/j.enpol.2018.10.013
12. Ture, H., Dogan, S., Kocak, D.: Assessing Euro 2020 strategy using multi-criteria decision making methods: VIKOR and TOPSIS. Soc. Indic. Res. **142**, 645–665 (2019). https://doi.org/10.1007/s11205-018-1938-8
13. Wątróbski, J., Bączkiewicz, A., Sałabun, W.: New multi-criteria method for evaluation of sustainable RES management. Appl. Energy **324**, 119695 (2022). https://doi.org/10.1016/j.apenergy.2022.119695
14. Zorpas, A.A.: Strategy development in the framework of waste management. Sci. Total Environ. **716**, 137088 (2020). https://doi.org/10.1016/j.scitotenv.2020.137088

Author Index

M. Hernes et al. (Eds.): ECAI 2024, LNNS 1217, pp. 261–262, 2025.
https://doi.org/10.1007/978-3-031-78465-1