



Artificial intelligence applications in education: Natural language processing in detecting misconceptions

Yunus Kökver¹ · Hüseyin Miraç Pektaş² · Harun Çelik²

Received: 21 December 2023 / Accepted: 18 July 2024 / Published online: 6 August 2024
© The Author(s) 2024

Abstract

This study aims to determine the misconceptions of teacher candidates about the greenhouse effect concept by using Artificial Intelligence (AI) algorithm instead of human experts. The Knowledge Discovery from Data (KDD) process model was preferred in the study where the Analyse, Design, Develop, Implement, Evaluate (ADDIE) instructional design cycle was used. The dataset obtained from 402 teacher candidates was analysed by Natural Language Processing (NLP) methods. Data was classified using Machine Learning (ML), one of the AI tools, and supervised learning algorithms. It was concluded that 175 teacher candidates did not have sufficient knowledge about the concept of greenhouse effect. It was found that the AI algorithm with the highest accuracy rate and used to predict teacher candidates' misconceptions was Multilayer Perceptron (MLP). Furthermore, through the Enhanced Ensemble Model Architecture developed by researchers, the combination of ML algorithms has achieved the highest accuracy rate. The kappa (κ) value was examined in determining the significant difference between the AI algorithm and the human expert evaluation, and it was found that there was a significant difference, and the strength of agreement was significant according to the research findings. The findings of the current study represent a significant alternative to the prevailing pedagogical approach, which has increasingly come to rely on information technologies in the process of improving conceptual understanding through the detection of conceptual misconceptions. In addition, recommendations were made for future studies.

Keywords Misconception · Natural language processing · Machine learning · Artificial intelligence

1 Introduction

The development of AI technology has been encouraging in recent years, and activities such as scientific research, academic conferences and technical competitions in the field of AI have spread all over the world (Xue & Wang, 2022). According to Ouyang and Jiao (2021), the field of Educational Artificial Intelligence (EAI), which covers various disciplines, supports teachers in their teaching methods, enhances students' learning experiences, and facilitates the overall transformation of the education system. Initially, EAI holds the promise of enhancing instructional design and pedagogical development within teaching process. This includes capabilities such as predicting students who may be at risk (Hellings & Haelermans, 2022), automatically assessing students' performance (Zampirolli et al., 2021), tracking the progress of students' learning (Berland et al., 2015). Second, EAI presents opportunities for a profound transformation of the education system. This transformation involves highlighting the crucial role of technology (Hwang et al., 2020), enhancing environments for knowledge transfer (Yannier et al., 2020), and reshaping the dynamics of the instructor-student relationship (Xu & Ouyang, 2022). Finally, EAI proves valuable in advancing student-centered learning, encompassing features like adaptive tutoring (Kose & Arslan, 2017), recommending personalized learning resources (Zhang et al., 2020), and diagnosing students' learning deficiencies (Liu et al., 2021).

If students do not envisage the concept correctly when they first encounter it, their subsequent educational lives are negatively affected, and it causes them to learn it incorrectly (Ercan et al., 2010; Schulte, 2001). Environmental issues such as global warming, ozone layer depletion, acid rain, air-water-soil pollution and local environmental problems are discussed at all levels of formal education. A review of the literature suggests that there are misconceptions about these concepts and that it is essential to assess teachers' and teacher candidates' perceptions and cognitive structures about the environment (Ballantyne & Rain, 1995). According to Kahraman (2020), global warming is one of the important environmental problems and therefore, it was understood that students had misconceptions about the greenhouse effect. When the results and sample levels of the studies conducted in the literature regarding the concept of the greenhouse effect are examined, it is seen that the concepts of the greenhouse effect and ozone layer depletion are confused with each other, and there are not many university students and teacher candidates among the participants (Andersson & Wallin, 2000; Boyes et al., 1993; Myers et al., 2004). According to Khalid (2003), it is known that science teacher candidates have lack of knowledge and misconceptions about current environmental problems such as greenhouse effect, acid rain and ozone depletion. Determination of students' misconceptions is considered important for the correct structuring of subsequent knowledge in the process of constructing knowledge. Open-ended questions are the most widely used measurement tools in the literature to determine students' misconceptions. These questions are widely used in international tests such as "National Assessment of Educational Progress, Massachusetts Comprehensive Assessment System, Test of English as a Foreign

Language, Programme for International Student Assessment, Trends in International Mathematics and Science Study and Advanced Placement Tests”. However, these widely used tests have some limitations. Some of these limitations are the time-consuming implementation and scoring and the difficulty of ensuring content validity. In other words, to put it more clearly, according to Demirezen et al. (2023), they cannot be objectively scored accurately, and even if they are scored, detecting misconceptions is a difficult and time-consuming process for experts.

1.1 Significance of the research

Today, AI technologies such as speech recognition, facial recognition, and autonomous vehicles are increasingly changing people’s lives significantly. AI technology encompasses technical knowledge in many fields such as information processing devices, big data, and various algorithms. Generally, AI can be used to solve real-world problems by applying six main branches: computer vision, NLP, cognition and reasoning, robotics, gaming, and ethics and ML (Huang & Qiao, 2022). With the development of computer science and computing technologies, automatic, adaptive, and efficient AI technologies have widely influenced various academic fields (Xu & Ouyang, 2022). In this context, AI is widely used in various fields including healthcare, marketing, industry, and education, and many countries have invested significant resources to educate their citizens to use, adopt, and build technologies related to AI (Aung et al., 2022).

Currently, AI can be widely used in three main areas in educational processes: preparation of educational materials, design of instruction, and systematization of assessment and evaluation processes (West et al., 2023). In this process, AI plays a particularly important role in automating routine tasks. A content pool tailored to the needs of the student can be created, and instruction can be personalized through a suitable program design (Buluş & Elmas, 2024). After the instruction is completed, the student can be subjected to algorithmic assessments using AI applications to identify deficiencies. Thus, the student’s individual learning process is structured in a unique way (Wardat et al., 2023). For example, AI designed to support teachers can assess whether specific students need additional training in a particular topic by using criteria such as how much time students spend on an activity, the number of correct answers, and the number of attempts. AI can analyze whether the student needs additional training and provide a recommendation to the teacher (Cooper, 2023). Chassignol et al. (2018) have discussed the impact of AI on education under the headings of personalized educational content, innovative teaching methods, the relationship between student and instructor, and technology-supported assessment.

Research on the application of AI in various disciplines reveals the potential to transform and enhance educational experiences (Jia et al., 2024). From STEM education (Jang et al., 2022) to language learning (Pokrivcakova, 2019), mathematics education (Gadanidis et al., 2017), computer science (Di Eugenio et al., 2021), history (Bertram et al., 2021), art (Kong, 2020), and music education (Zulić, 2019), AI holds promise in providing personalized, interactive, and innovative learning opportunities (Jia et al., 2024). However, despite its increasing popularity, there is a

relatively limited number of empirical studies on the effectiveness and best practices of AI application in science education, posing significant challenges to drawing conclusive results (Jia et al., 2024).

Examining AI applications in science education, Kong and Shen (2024) developed an unsupervised machine learning-based approach to predict students' understanding of computational thinking concepts using their cognitive, affective, and demographic characteristics. Demirezen et al. (2023) utilized unsupervised machine learning-based algorithms to detect misconceptions related to atomic concepts in their study. Our study aims to develop an AI classifier based on supervised ML for detecting misconceptions among teacher candidates regarding greenhouse effect, thus distinguishing itself from the literature.

It is believed that AI systems used in education, as they analyze large amounts of data from students, can provide more detailed results than traditional assessment methods. Therefore, this notion provides innovative opportunities for qualitative analysis. Although AI has traditionally been associated with quantitative studies, the potential for developing qualitative research is increasingly being recognized. Building on this, this study aims to explore innovative ways in which AI can enrich and expand the boundaries of qualitative research. Analyzing the impact of AI on the fundamental values and principles of qualitative research will enable a discussion of the potential future of qualitative research in an AI-driven world.

Therefore, the aim of this study is to identify the misconceptions of science teacher candidates about the greenhouse effect by using an AI algorithm instead of human experts. More specifically, the following research questions (**RQs**) are addressed:

RQ#1. What are the misconceptions of teacher candidates regarding the concept of the greenhouse effect?

RQ#2. Which AI algorithm is used to predict teacher candidates' misconceptions and has the highest accuracy rate?

RQ#3. Is there a significant difference between the AI algorithm and the human expert evaluation?

2 Theoretical framework

2.1 Artificial intelligence (AI)

Presently, AI technologies, such as voice recognition, facial recognition, and autonomous vehicles, are exerting a profound and transformative impact on people's lives. The realm of AI encompasses technical expertise across diverse domains, including computing devices, big data, and various algorithms. The application of AI to address real-world challenges involves leveraging six major branches: NLP, robotics, computer vision, cognition and reasoning, gaming and ethics, and ML, as outlined by Huang and Qiao (2022). AI has gained widespread utilization across diverse fields, encompassing healthcare, marketing, industry, and education. Numerous countries have made significant investments in training their citizens to effectively utilize, adapt, and contribute to AI-related technologies, as noted by Aung

et al. (2022). Providing a precise definition of AI can be challenging, even for experts, due to several significant reasons, primarily stemming from the continually evolving nature of AI contents, as highlighted by Chen et al. (2020). Nick Bostrom also points out that numerous ground-breaking AI technologies have been integrated into universal applications without being explicitly labelled as AI (Bostrom, 2006). AI stands as an interdisciplinary field continually enriched by the contributions of researchers and experts from diverse domains, including neuroscience, psychology, and linguistics, who bring their insights, knowledge, and terminology (Chen et al., 2020). For instance, according to Russell and Norvig (2016), AI is used to characterise machines or computers that emulate “cognitive” functions associated with the human mind, such as “learning” and “problem solving.” Poole et al. (1998) conceptualized AI as the study of intelligent agents capable of perceiving their environment and achieving specific goals by maximizing probabilities. Kaplan and Haenlein (2019) defined AI as the capability of a system to accurately interpret and learn from input data, applying acquired knowledge to accomplish specific objectives. While a universally accepted definition of AI remains elusive, it is evident that AI transcends the boundaries of encompassing information science, computer science, linguistics, psychology, neuroscience, philosophy, mathematics, and various other disciplines (Chen et al., 2020). Indeed, AI emerges as an interdisciplinary field characterized by its broad scope and inclusiveness (Zhao et al., 2018). As articulated by Baker and Smith (2019), AI refers to computers engaging in cognitive tasks such as learning and problem-solving, typically associated with human mental processes. AI is defined as intelligent computer systems with the ability to memorise knowledge, perceive and manipulate their environment similar to humans, and have human-like characteristics (Huang, 2018; Lodhi et al., 2018; Welham, 2008). Also in the field of language learning and teaching tasks, as highlighted by Dodigovic (2007), AI can mimic the behaviour of both teachers and learners. The roots of AI trace back to the 1950s, marked by John McCarthy’s two-month workshop at Dartmouth College in the USA. McCarthy introduced the term “AI” in his workshop proposal in 1956, as noted by Russell and Norvig (2016).

The evolution of AI has gone through three phases (Long & Zeng, 2022). In the first phase, studies focused on creating intelligent systems by abstracting and generalising logical rules, which were then implemented in the form of computer programs. This period is referred to as the inference stage, during which scientists endeavored to apply AI through knowledge databases and reasoning in the 1970s. During this period, scientists constructed extensive and intricate expert systems to simulate the intelligence levels of human experts. A major challenge encountered with these precisely defined rules was the inability to translate many complex and abstract concepts into concrete code. For instance, the processes of recognizing images and understanding languages in human cognition cannot be effectively simulated using predefined rules. To address such challenges, a research discipline known as ML has emerged, allowing machines to learn rules directly from data. ML serves as an umbrella term encompassing various technologies and methods, including NLP, data mining, neural networks, and algorithms, as highlighted by Zawacki-Richter et al. (2019). ML, as an AI method, encompasses both supervised and unsupervised classification and profiling. This includes applications such as predicting a

student's likelihood of dropping a course, determining acceptance into a program, or identifying topics in written assignments, as outlined by Zawacki-Richter et al. (2019). According to Long and Zeng (2022), ML gained prominence as a focal point in the field of AI during the 1980s, marking the second stage. There has been a notable trend in ML towards learning complex, abstract logic through neural networks, and breakthroughs in deep neural network technology in computer vision, NLP and robotics have been evident since 2012. In fact, some software applications that point to the third stage of AI have exceeded the level of human intelligence.

2.2 AI and theoretical basis of misconceptions

The foundation of research in the field of misconceptions lies in constructivist learning. While John Dewey, Jerome Bruner, and Jean Piaget have contributed to the development of this view, its main foundations are based on Ausubel's (1968) Meaningful Learning Theory, Kelly's (1955) Personal Construct Theory, and Vygotsky's (1962) Socio-cultural Theory. Piaget's theory of cognitive development has made significant contributions to understanding students' thought processes. Teachers should know the cognitive development levels of their students for effective science education. Thus, they can make healthier decisions when determining the content of the subject, deciding on the concepts to be taught (concrete-abstract), or organizing teaching methods. According to constructivist learning, knowledge is individualized and has a hierarchically organized structure of interconnected concepts. The formation of a conceptual structure is an active process and is created by the individual over time. Continuous meaningful learning requires new knowledge to be consciously linked with existing knowledge, and the individual's existing knowledge must be tested against externally acquired knowledge (Mintzes & Wandersee, 1998).

According to Piaget, misconceptions are like a structure and are built upon one another. Misconceptions start as gaps resulting from a lack of knowledge. These gaps are randomly filled by teachers' inadequate instruction, students' existing knowledge, and encountered experiences. The knowledge constructed by the student is undoubtedly successful up to a point, but at some point, this phenomenon emerges as a misconception (Rowell et al., 1990). As many cognitive psychologists state, when people learn new concepts, they construct new information by associating it with their past knowledge and beliefs. Some definitions used instead of misconceptions in the literature are "conceptual difficulties", "mental models", "phenomenological primitives", "naive beliefs", "misunderstandings" and "alternative concepts" (Chang et al., 2007). Students form their knowledge structures in some ways by gaining their own experiences, accessing information sources, having beliefs, learning facts and interacting with the environment. Misconceptions may occur if the concepts that students form in their learning processes are not based on scientific foundations (Demirezen et al., 2023). Apart from these, reference books used by students, learning methods used by teachers and instructors in the classroom, and internet-based information can also lead to misconceptions (Tumanggor et al., 2020). According to constructivist learning, students construct their own knowledge by interpreting the information they hear, read, and see in relation to what the teacher

tells them. With the development of thinking machines, it is known that machines can now perform these skills previously only possessed by humans (Coiera, 2003; Davies, 2016). AI can play significant roles in personalizing education (Aşkar & Kızılkaya, 2006), providing the right resources at the right time (Anohina, 2007), and preventing individuals from getting lost in an overwhelming amount of information, among other technologies it can offer.

In conclusion, it is crucial to know students' existing knowledge and misconceptions to achieve meaningful learning in the teaching process. The underlying causes of students' misconceptions can be diverse and are subject to debate. As Wandersee et al. (1994) pointed out, misconceptions, especially when acquired through direct observation and perception, have hidden roots, making them difficult to identify. Misconceptions are more reliably and concretely detected when identified through various three-stage or open-ended tests. Therefore, in this study, open-ended responses were utilized to identify misconceptions related to the greenhouse effect in teacher candidates. These responses were classified using ML algorithms. In contrast to the literature, NLP has employed, thus creating feature attributes from raw data to serve as input parameters for ML algorithms.

2.3 Machine learning (ML)

Emerging from the broader field of AI, and specifically from research in pattern recognition and computational learning (Karmani et al., 2018), ML is characterised by methods that facilitate learning from data and making predictions (Kohavi, 1995). Samuel (1959) originally defined ML as the study of how learning can be achieved without explicit programming. The strength of ML lies in its ability to execute highly dynamic program instructions by modelling input data and making data-driven predictions or decisions (Chen et al., 2020). It is widely recognised that ML serves as a technique for training intelligent agents to perform specific tasks using large datasets. ML is a concept that emerges from the amalgamation of statistics and computer science, enabling machines to enhance their performance based on prior experiences, akin to human learning processes (Rasheed & Wahid, 2021). The fundamental difference lies in the source of learning, as computers learn from data, while humans learn from real-life situations and experiences (Bishop & Nasrabadi, 2006). As highlighted by Rasheed and Wahid (2021) and Long and Zeng (2022), ML comprises three types of learning algorithms: supervised, unsupervised, and reinforcement learning.

2.3.1 Supervised learning

In supervised learning, the principal aim is to discern a correspondence between a collection of input variables X and an output variable Y . This established correspondence is subsequently utilized to forecast outcomes for unobserved data, as emphasized by Cunningham et al. (2008). The algorithm is trained to comprehend the mapping relationship, expressed as $f_{\theta}:x \rightarrow y$, where f_{θ} signifies the model function, and θ denotes the parameters of the model. Throughout the training process,

the model parameters (θ) undergo optimization by minimizing the errors between the model predictions and the actual values of Y . This optimization process aims to enhance the accuracy of the model's predictions, as discussed by Long and Zeng (2022).

2.3.2 Unsupervised learning

If we contemplate a machine (or living organism) being presented with a series of inputs, denoted as $x_1, x_2, x_3, \dots$, where x_t signifies the sensory input at time t , this input- commonly denoted as data- might stem from diverse origins, including an image on the retina, pixels captured by a camera, or an audio waveform (Ghahramani, 2004). The cost of collecting labelled data is usually high. For a dataset consisting only of examples, the algorithm needs to uncover the underlying patterns in the data, leading to the classification of this algorithm as unsupervised learning as described by Long and Zeng (2022). In unsupervised learning, the system is presented with inputs $x_1, x_2, \dots$ yet lacks guided target outputs or rewards from its surroundings (Ghahramani, 2004). One particular algorithmic approach within unsupervised learning leverages the input itself as a form of supervised signal, represented as $f_\theta: x \rightarrow x$, a methodology termed self-supervised learning. Throughout the training process, the parameters are refined by minimizing the disparity between the model's anticipated output, $f_\theta(x)$, and the original input x (Long & Zeng, 2022).

2.3.3 Reinforcement learning

Reinforcement learning, as outlined by Qiang and Zhongli (2011), stands out as an online learning technology distinct from supervised and unsupervised learning. In reinforcement learning, the reinforcement signal from the environment serves as an evaluation of the quality of actions taken by the intelligent agent. However, it doesn't explicitly instruct the agent on how to produce the correct action, as noted by Harmon and Harmon (1996). Due to limited information provided by the external environment, the intelligent agent must rely on its experiences to learn. This way, the agent acquires an appropriate evaluation of the environmental state and adjusts its action strategy to adapt to the environment, as explained by Qiang and Zhongli (2011).

Supervised learning was used in this study because two experts (specialists in the field of science education) were involved in identifying misconceptions in teacher candidates. These two experts labelled the data during the data classification process. According to the literature, the supervised learning paradigm uses a teacher in the ML process (Haykin, 2009). Supervised learning provides learning by focusing on a specific target or outcome variable. This makes it possible to obtain results with high precision in certain tasks such as prediction, classification or regression. It can learn a particular task based on the interaction of the trainer, so the learned model may be highly improved on a particular task.

2.4 Supervised learning classifiers used in this study

2.4.1 Multilayer perceptron (MLP)

MLP is an artificial neural network model inspired by the biological nervous system and is known for its ability to solve complex data structures. The architecture of MLP incorporates multiple hidden layers, a design choice aimed at capturing and modeling more complex relationships within the training dataset. This structure enables the MLP to learn and represent intricate patterns and dependencies in the data, making it well-suited for handling tasks that involve non-linear and sophisticated relationships (Bisong, 2019). This model has a multi-layered structure, and each layer carries the ability to learn more abstract and higher-level representations compared to the previous one. They are a basic type of artificial neural networks and are often used in complex data analysis and classification tasks. MLP has a multi-layered structure in which many artificial nerve cells or neurons are connected. This multi-layered structure increases the ability to learn more complex relationships of data. MLP classifiers include an input layer, at least one hidden layer, and an output layer. The input layer feeds the data into the network, while the hidden layers are used to learn and create internal representations of the data. The output layer produces final predictions to assign data to specific classes. MLP updates the weight matrices and bias terms to internalize the data and learn hidden patterns in the training stage. This process is performed with the back-propagation algorithm and focuses on minimizing the error function of the network. It is evident that further investigation is warranted into enhancing the training efficiency and refining feature learning and classification capabilities within MLP (Tang et al., 2015). MLP stands as the most renowned and commonly employed architecture within neural networks (Popescu et al., 2009). Therefore, this classifier is included in our study.

2.4.2 Support vector machine (SVM)

As per Boser et al. (1992), SVM is a ML technique that acquires knowledge through examples to assign labels to objects. SVM is optimized to find the optimal separation boundary when performing the task of assigning data points to one or more classes. The main goal of SVM is to find a hyperplane that separates data points. This hyperplane tries to maximize the largest distance between two classes. The formulation of Support SVM for pattern recognition problems presents numerous benefits compared to alternative classifiers. Some of these advantages include: 1) the assurance that the obtained solution is the sole global solution once reached, 2) the solution's strong generalization properties, and 3) a robust theoretical foundation grounded in learning theory, specifically structural risk minimization (SRM), and optimization theory (Cristianini & Shawe-Taylor, 2000; Theodoridis & Koutroumbas, 2003). SVM has been used successfully in many applications due to the above features (Cortes & Vapnik, 1995; Sebald & Buklew, 2000). Therefore, SVM was used as a different classifier in our study.

2.4.3 Gradient boosting (GB)

GB stands out as an effective ML technique, especially in regression and classification problems (Kiangala & Wang, 2021). This method focuses on creating a strong estimator by combining weak decision trees. It improves the results of decision trees using the gradient descent algorithm. Each tree is added with a focus on correcting the error of previous trees, which improves the overall performance of the model. The basic principle of GB is to measure the performance of the model through the loss function and find optimal estimators by trying to minimize this loss function. This process occurs through a weighted correction of prediction errors produced by previous trees.

2.4.4 Random Forest (RF)

RF is a ML technique that generates multiple models by training individual decision trees on different subsets of observations. It excels particularly in classification and regression problems due to its ensemble approach, where predictions are aggregated across multiple decision trees. It is an ensemble method in which each tree is trained independently, and then the predictions are combined to obtain the overall result. RF ensures that trees are independent of each other by using random sampling and feature selections during the construction of each tree. The basis of the success of RF lies in the high probability that trees will make different errors and the strategy of correcting these errors with the ensemble method (Arshad et al., 2023). It generally contributes to reducing over-fitting problems and making the model more generalizable.

2.4.5 K-nearest neighbor (k-NN)

k-NN is a simple but effective ML algorithm used in classification and regression problems. Its basic principle is to predict the class or value of a data point by using the k nearest training data to that point. Although distance metrics usually use the Euclidean distance, other distance metrics may be preferred depending on the application. One of the main advantages of k-NN is that there is no complex optimization process during training the model. The model only keeps the training data in memory and uses this data to make predictions. Another important feature of this method is that it captures the local structure in the dataset. k-NN has been used successfully in various application areas, but it has been observed to be especially effective in classification problems and similarity-based pattern recognition tasks (Zhang et al., 2018).

2.4.6 Logistic regression (LR)

LR is considered a fundamental classification method in the fields of statistics and ML. LR is used to analyse the relationship between many features or variables and classify data between two or more classes based on the results of this analysis. This classifier classifies data with a logistic function. The logistic function performs a

compression operation where the result is expressed as a probability value. These probability values typically take a value between zero and one and represent the probability that the result belongs to a class. The training process of the model involves learning the weight coefficients using the data. These weight coefficients represent the contribution of data features to the class prediction. A substantial advantage of Logistic Regression is the ability to present the results of the model with statistical significance and confidence intervals. It allows researchers and decision-makers to understand how reliable the model's predictions are (Thabtah et al., 2019).

2.4.7 Enhanced ensemble model architecture

In ensemble learning, multiple models, often referred to as base models, are used, and the outputs of all models are combined to produce a single unified output (Iqbal & Wani, 2023). In other words, predictions/decisions of various models are aggregated to generate the final prediction or decision. Ensemble models have achieved quite good success in the recent past (Iqbal & Wani, 2023). Based on this, the Enhanced Ensemble Model Architecture in this study is outlined as follows: Step 1: Transformation of the greenhouse effect misconception dataset is performed. Step 2: In the second step, the transformed dataset is then fed into a community model consisting of MLP, SVM, GB, RF, k-NN, and LR classifiers. Step 3: The dataset is split into two parts: a training dataset and a test dataset. Then, the most appropriate hyper-parameters for all classifiers are selected, and the grid search parameter tuning method is used to train the community model. Step 4: In the final step, a voting mechanism is used to make predictions on the test dataset, considering the best predictions of each classifier model.

2.5 Natural language processing (NLP)

NLP is a field designed to help computers understand, interpret and produce human language. This discipline aims to enable computers to understand, translate and summarize texts, and even produce language like humans by combining the fields of AI and linguistics. One of the main goals of NLP is to give computer systems the ability to understand the complex structure of language. It includes elements such as word meaning, sentence structures, context and emotional tone of the language. NLP learns the patterns of language using large datasets and uses these learnings in future language understanding and production tasks (Khurana et al., 2023). Many applications such as text classification, speech recognition, machine translation, sentiment analysis and automatic summarization are among the areas of NLP usage (Montejo-Ráez & Jiménez-Zafra, 2022). NLP was used in this study to classify the texts written by teacher candidates regarding the concept of the greenhouse effect.

In order to detect misconceptions regarding the greenhouse effect, using AI algorithms mentioned in the theoretical framework, NLP has been used. The KDD process model was used for creating the dataset, and the ADDIE design cycle was

utilized to structure the research. Methods section has elaborated on this process in detail.

3 Methods

In the study, the ADDIE instructional design model, a widely accepted framework in the field of instructional design, was followed by using the design-based research method. There are many professional instructional designers who consider this model as a standard for developing technology-based instruction (Molenda, 2003; Morrison, 2010) and it consists of 5 stages: analysis, design, development, implementation and evaluation (Branch, 2009; Karamustafaoğlu & Pektaş, 2023). The instructional design cycle and implementation process used are given in Fig. 1.

ADDIE is commonly characterized as a linear model, yet it's crucial to acknowledge its iterative and cyclical nature. This instructional design model incorporates evaluation at each stage, emphasizing a continuous feedback loop. This iterative approach allows for ongoing assessment, refinement, and adaptation, ensuring that the instructional design process remains responsive and effective (Trust & Pektaş, 2018). Researchers often favor the ADDIE model for instructional design due to its practical applicability and smooth progression from one stage to the next (Hanafi et al., 2020). The data collection process on the basis of this research, which was created using the ADDIE model, is discussed in detail below.

3.1 Data collection process

Raw data from 451 teacher candidates enrolled in science education programmes in the education faculties of 8 different universities in Turkey, collected via a web form, were reduced to 402 data after two expert evaluations. 80% of these data were identified as training data and 20% as test data. As a result of the machine learning algorithms, an interface in .exe format was created. This interface was used to test

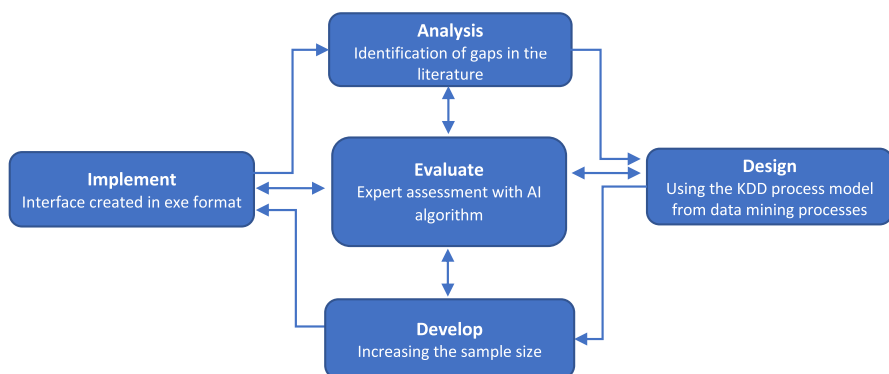


Fig. 1 ADDIE Design cycle and implementation process (Branch, 2009; Karamustafaoğlu & Pektaş, 2023)

the classification performance of the system on 59 teacher candidates. This process is discussed in detail in the following stages of the ADDIE design model.

3.1.1 Analysis

In the literature review, only one study was found to detect students' misconceptions using an AI algorithm instead of expert humans. The research conducted by Demirezen et al. (2023) aims to detect students' misconceptions about the concept of atom with a machine instead of a human. Unsupervised learning was used in this study, and two methods were proposed to classify students' answers. These methods are the transformers model and fastText algorithm, which are deep learning algorithms. As a result of the study, misconceptions were distinguished with high accuracy between 0.97 and 1.00 with the method suggested by Demirezen et al. (2023). Unlike this study in the literature, supervised learning was used in our study, which was conducted on the concept of the greenhouse effect. Two field experts did the data labelling. In this process, the data tagged by two experts were compared and the compliance rate was evaluated with the Kappa statistic. The data of 402 teacher candidates out of 451 were labelled in the same classes by two experts. Therefore, the Kappa value was calculated as 89.14. In addition, MLP, SVM, GB, RF, k-NN and LR classifiers from ML algorithms were used in our study.

3.1.2 Design

The KDD process model, a prominent data mining approach, was employed in this study, with a broader focus on KDD, CRISP-DM, and SEMMA data mining processes. Many researchers and data mining experts prefer the comprehensive and accurate nature of the KDD process model, making it a widely followed framework in the field (Shafique & Qaiser, 2014). Therefore, the KDD process model was preferred at this stage (Fig. 2).

The KDD model, as described by Brachman and Anand (1994), operates in an iterative and interactive manner, aiming to find valuable information within data. It places emphasis on the high-level application of specific data mining methods (Shafique & Qaiser, 2014). The emergence of KDD was driven by the need to analyze large volumes of data, marking a pivotal endeavor in discerning legitimate, novel, potentially valuable, and ultimately comprehensible patterns within data, as articulated by Fayyad et al. (1996).

Data mining techniques are informed by three primary academic disciplines. Initially, statistics offers rigorously defined methodologies for systematically elucidating relationships between variables. This encompasses classical approaches such as data visualization, descriptive statistics, correlation analysis, frequency distributions, multivariate exploration techniques, as well as advanced linear and generalized linear models. Secondly, modern knowledge discovery techniques stem from ML and AI. ML entails instructing computers to discern patterns within data, while artificial neural networks simulate the human brain's architecture, comprising interconnected processing elements collaborating to process information. Lastly, database systems constitute the third field, allowing the storage, access, and retrieval of

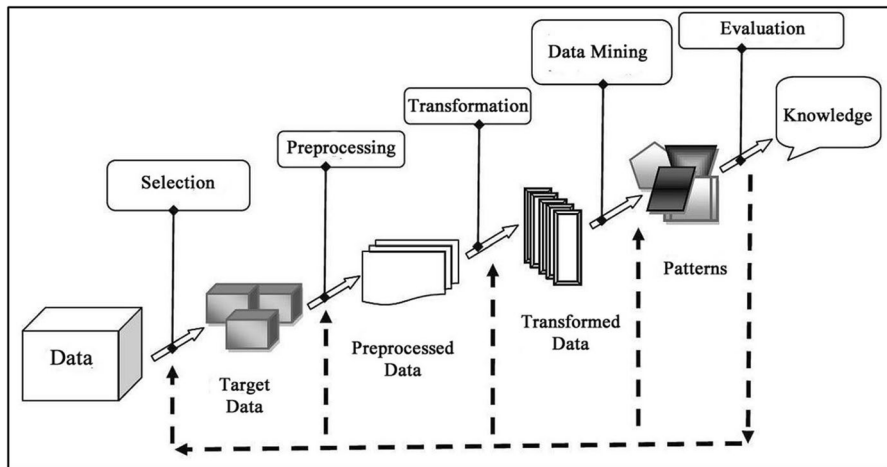


Fig. 2 KDD process model and stages

large amounts of data, thereby providing the necessary support for the computing and mining platform.

Data: The answers given to the misconception interview survey administered to teacher candidates via the web form constitute the raw data of this study.

Data Selection: The raw data was checked and meaningless data recorded in the web form (data not written for the greenhouse effect) were removed from the dataset at this stage. Additionally, data not jointly labelled by two experts were removed from this dataset.

Preprocessing: The Zemberek library, which allows morphological analysis for Turkish texts, was used as a NLP library. The following procedures were applied to provide data in plain text format as input parameters to data mining classification algorithms: Converting the entire text to lowercase, removing punctuation marks in the text, removing line spaces in the text, separating it into word markers, extracting the roots of words by removing their suffixes, removing stop-word words, separating them into sentence markers, extracting root length distributions, extracting sentence length distributions on a word basis, subtracting word richness ratios, average root length, average sentence length in words, subtracting the total number of punctuation marks, subtracting the total number of stop-words used, and subtracting the number of words written in all capital letters.

Transformation: Term Frequency-Inverse Document Frequency (TF-IDF), bag-of-word vectors, word-sentence distributions and other attribute vectors of each text, whose roots were extracted after pre-processing, were extracted, and all vectors were normalized among themselves using the min-max normalization method. After these operations are applied, the attributes obtained for each data are as follows: Average sentence length, average word length, number of punctuation marks, type-token ratio, number of stop-words, number of words written in all capital letters, word length histogram vector (as letters), sentence length histogram vector (as words), bag-of-words vector, and TF-IDF vector. Since the

TurkishMorphology module of the Zemberek library could not perform the morphological analysis of foreign words in obtaining the roots of the words in the data, the roots of these words were taken as ‘UNK’.

Data Mining: After the features were obtained, 80% of the data was divided into training data and the remaining 20% as test data. K-Folds Cross-Validation Technique was used to prevent over-fitting and selection bias problems in the selection of training and test data and to prove that the success of the model is not random depending on the selection of test data. Cross-validation is a model verification technique that tests what result a study will achieve (Kohavi, 1995). The dataset is divided into k different subsets in the K-Folds Cross-Validation technique. The (k-1) part is used for training, and the remaining is used for testing. This process is performed k times, and different test data are processed each time. In this way, the average success value is calculated by measuring the success of the model with different test data k times.

Evaluation: This study seeks to objectively analyze the outcomes by employing a range of evaluation metrics to assess the performance of the classification model. Evaluation metrics afford us insights into the classification accuracy and precision of the model. Specifically, four key metrics derived from the Confusion Matrix—Accuracy, Recall, Precision, and F1-Score—are utilized for this purpose.

Accuracy measures the model’s correct classification rate and is calculated by the following equation:

$$\text{Accuracy} = (TN + TP) / (TP + FP + TN + FN) 100$$

In binary classification, True Positives (TP) signify the count of correctly predicted positive samples, whereas True Negatives (TN) denote the count of correctly predicted negative samples. False Positives (FP) indicate the instances falsely predicted as positive, and False Negatives (FN) represent the instances falsely predicted as negative. Accuracy is used as a metric to evaluate overall model performance.

Recall measures how many true positives we detect and is calculated by the following equation:

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

Recall indicates how well the model can accurately capture positive classes and represents precision. Precision measures how many of the samples predicted as positive are actually positive and is calculated by the following equation:

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

F1-Score provides a balanced combination of Recall and Precision metrics and is calculated by the following equation:

$$F1 \text{ Score} = (2(\text{Recall} \times \text{Precision})) / (\text{Recall} + \text{Precision})$$

These metrics play a critical role in helping us evaluate the model’s performance, identifying potential weaknesses of the model and areas that need improvement.

3.1.3 Development

First, a dataset of 300 data was created in the study by selecting the data that two experts labelled in the same class. As a result of the ML applications, it was observed that the accuracy value was below 80% for each classifier. Therefore, 102 more teacher candidates participated, and a dataset of 402 data was created. It was observed that as the number of datasets increased, the accuracy value increased above 80%.

3.1.4 Implementation

At this stage, an interface in exe format was created. Using this interface, it was tested with 59 teacher candidates whether there was a significant difference between the AI algorithm and the expert assessment, which is the third sub-problem of the study.

3.1.5 Evaluation

The Kappa (κ) statistic was used at this stage to determine the reliability between the raters with 59 teacher candidates. The κ statistic, commonly employed for assessing inter-rater reliability, was proposed by Cohen in 1960. Initially designed to measure the level of agreement between two raters scoring at the classification level, the κ statistic was developed to address situations with only two raters (Cohen, 1960). Later, Fleiss extended the κ statistic in 1971 to make it applicable for assessing agreement among more than two raters (Fleiss, 1971). The κ statistic is based on some basic assumptions (Crawforth, 2001). These assumptions were expressed by Brennan and Prediger (1981) as the objects or individuals categorized in the scoring process, the raters' ratings and the categories used in scoring are independent. One notable advantage of the κ statistic is its simplicity in calculation and straightforward interpretation. Another significant and paramount advantage is that it incorporates a correction for expected agreement by chance. This correction adds a valuable aspect to the κ statistic, making it a robust measure of agreement that accounts for chance agreement in the assessments. In interpreting the κ statistics, the fit levels suggested by Landis and Koch (1977a, 1977b) in Table 1 are used.

4 Data collection tools and analysis of data

4.1 Data collection tools

The Word Association Test (WAT) was employed to elucidate the cognitive structures of science teacher candidates concerning the greenhouse effect concept. The selection of the key-concept was determined through consultation with two academic experts in the science field. During the WAT application,

Table 1 Value ranges for interpreting the Kappa statistic

Strength of Agreement (κ)
Almost Perfect (0.81–1.00)
Substantial (0.61–0.80)
Moderate (0.41–0.60)
Fair (0.21–0.40)
Slight (0.00–0.20)
Poor (<0.00)

teacher candidates were tasked with jotting down the first five words that came to mind regarding the greenhouse effect concept. The rationale behind listing the key-concept separately was to mitigate the risk of chained responses and to prevent participants from fixating on words prompted by previously written answers rather than the key-concept itself (Bahar & Özatlı, 2003). In addition, creating a sentence about the key-concept after word association makes it easier to determine the nature of the relationship established between the key-concept and the answers given to them (Gunston, 1980). Teacher candidates were given 40 seconds for the concept of the greenhouse effect. It is seen in the literature that the time given varies depending on the repetition of the key-concept. If the key-concept is repeated five times, this time is determined as 20 seconds in the studies on the subject, so students were given 20 seconds for the concept in this study (Kurt & Ekici, 2013).

4.2 Analysis of data

Both the relationship of the sentences made by the science teacher candidates with the concept of the greenhouse effect and whether they were scientifically correct were examined in grouping the sentences containing scientific information. If the sentences made by science teacher candidates were scientifically correct with all their elements, they were put in the “Sentence containing scientific information” category. If the teacher candidates made sentences that were non-scientific, used in daily life, and had content related to their past experiences and traditions, they were included in the “Unscientific and superficial sentences” category. Non-scientific sentences in this category were not considered as misconceptions because the criterion was taken into account that teacher candidates use the concept in contexts that do not have scientific meaning used in daily life and language, rather than in any different or incorrect way. In other words, even if the sentences made are correct, they may be sentences with no scientific meaning. If the science teacher candidates tried to attribute scientific meanings to the key concepts in the given sentences but confused these concepts with concepts with different and incorrect meanings, they were included in the “Sentence containing misconceptions” category. What is important here is to determine the key concepts of the teacher candidates and to reveal their misconceptions about the greenhouse subject. This category determines the importance of the study more.

First, to measure the teacher candidates’ prior knowledge on the concept of the greenhouse effect, they were asked to write Turkish words that evoke the concept of the greenhouse effect in their minds and a sentence in Turkish about the greenhouse effect and the information obtained from the students was recorded in a database. Then, each data was examined by two experts and labelled in three categories. A dataset of 402 data was created by selecting data that two experts labelled in the same category. Statistical information regarding the dataset is given in Table 2.

Miles and Huberman (1994) include some questions that need to be asked regarding internal validity, which is about the accuracy of the findings and results obtained in the study: “Are the findings compatible with the previously created conceptual

Table 2 Statistical information about the dataset

Category of sentences	Number of sentences	Percentage value (%)
Sentence containing scientific information	106	26.37
Unscientific and superficial sentences	121	30.10
Sentence containing misconception	175	43.53
Total	402	100

framework or theory?”, “Has this framework guided data collection?”. Ten AI studies conducted in the field of education in the last three years in the literature were examined based on these questions in our study (Table 3).

Table 3 Accuracy percentages and average values for AI applications in the literature

Study	Classifier	Accuracy (%)
(Liu et al., 2022)	KNN	53.9
	Naive Bayes	66
	Random Forest	72.9
	CNN	68.5
	RNN	67.9
	Att-Bi-LSTM	70.9
	BERT	77.8
	BERT-CNN	78.7
(Wambsganss et al., 2022)	BERT	54.12
(Huang et al., 2023)	BERT	95.8
	Bi-LSTM	91.5
	BERT-Bi-LSTM	94.7
(Zou et al., 2021)	BERT	83
	Random Forest	56
(Nguyen & Diederich, 2023)	Logistic Regression	59
	Multinomial Naive Bayes	52
	SVM	60
	BERT	72
(Shin et al., 2023)	Random Forest	86.7
(Demirezen et al., 2023)	fastText for q1	88
	fastText for q1	94
	fastText for q1	97
(Demirezen et al., 2023)	BERT-Base Fine Tuning (for question 1)	98
	BERT-Base Fine Tuning (for question 2)	98
	BERT-Base Fine Tuning (for question 3)	99
(Subheesh et al., 2022)	Logistic Regression	70
(Rasheed & Wahid, 2021)	Support Vector Machine	75.55
	Decision Tree	45.55
	Logistic Regression	73.33
	Random Forest	73.33
	KNN	67.77
	Linear Discriminant Analysis	69.44
	Naive Bayes	70.55

Considering the accuracy of the 33 classifiers used in these studies, the accuracy of our study is considerable.

5 Results

5.1 RQ#1. What are the misconceptions of teacher candidates regarding the concept of the greenhouse effect?

Table 4 shows the frequency and percentage distribution of teacher candidates' misconceptions about the concept of the greenhouse effect.

When Table 4 was examined, it was seen that 175 teacher candidates had misconceptions. The most common misconception sentences were “The depletion of the ozone layer causes harmful sunlight to enter our world” and “Acid rain and global warming are factors that increase the greenhouse effect.”

5.2 RQ#2. Which AI algorithm is used to predict teacher candidates' misconceptions and has the highest accuracy rate?

When Table 5 is examined, it is seen that MLP is the AI classifier with the highest validation value. It has been determined that the validation values of SVM, GB and RF classifiers are equal. In addition, thanks to the Ensemble Model

Table 4 Sentences containing misconceptions and frequency percentage values

Sentences containing misconceptions	<i>f</i>	%
The depletion of the ozone layer causes harmful sunlight to enter our world	30	17,14
Acid rain and global warming are factors that increase the greenhouse effect	30	17,14
The greenhouse effect is an agricultural greenhouse activity.	27	16,00
It causes the planet's temperature to increase and climates to change as a result of air pollution.	26	14,86
The use of chemical drugs and hormones in agriculture and GMO production are greenhouse effects.	13	6,86
The use of deodorant and spray triggers the greenhouse effect	10	5,71
Earth warming due to some irreversible rays	8	4,57
It is natural heating caused by gases in the atmosphere retaining heat from the sun.	8	4,57
Decrease in the amount of drinking water due to rapidly increasing temperature.	7	4,00
Our world gets even hotter thanks to the high heat retention properties of some greenhouse gases.	6	3,43
Heating occurs as a result of excessive increase in carbon dioxide gas.	4	2,29
Plant production using solar energy in a closed area	2	1,14
Acid rain increases the greenhouse effect and global warming	2	1,14
The greenhouse effect is the harmful gases in the atmosphere passing into the soil and killing living things.	2	1,14
Total	175	100

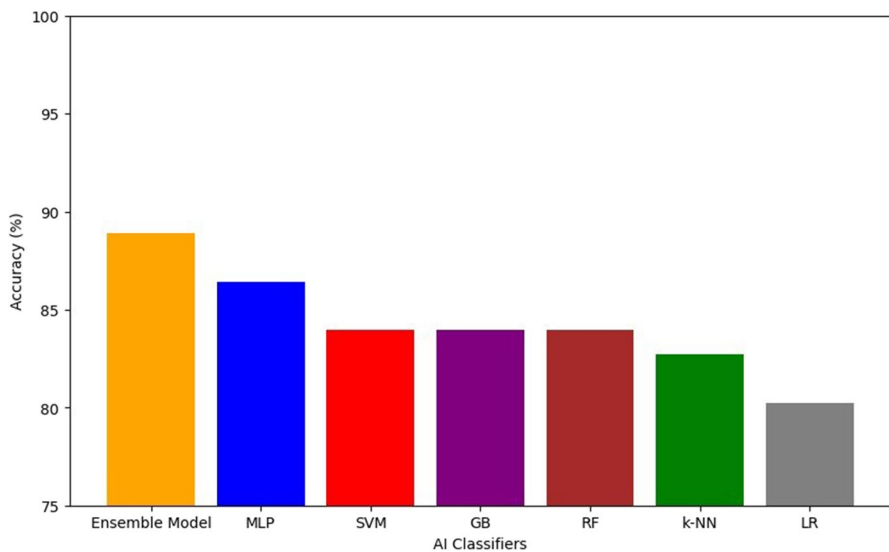
Table 5 AI classifiers and evaluation metrics

AI Classifier	Class	Recall	Precision	F1-Score	Accuracy (%)
MLP	CSI	0.72	0.9	0.8	86.42
	US	0.8421	0.667	0.7442	
	CM	0.973	0.973	0.973	
SVM	CSI	0.64	0.9412	0.7619	83.95
	US	0.7894	0.7143	0.75	
	CM	1	0.8605	0.925	
GB	CSI	0.76	0.9048	0.8261	83.95
	US	0.7368	0.6364	0.6829	
	CM	0.9459	0.9211	0.9333	
RF	CSI	0.9	0.9474	0.9231	83.95
	US	0.7097	0.88	0.7857	
	CM	0.9333	0.7568	0.8358	
k-NN	CSI	0.75	0.8824	0.8108	82.71
	US	0.8065	0.8333	0.8197	
	CM	0.9	0.7941	0.8438	
LR	CSI	0.68	0.7727	0.7234	80.24
	US	0.6316	0.6316	0.6316	
	CM	0.973	0.9	0.9351	
Enhanced Ensemble Model	CSI	0.9	0.9474	0.9231	88.90
	US	0.8065	0.926	0.8621	
	CM	0.9667	0.8286	0.8923	

CSI Sentence Containing Scientific Information, *US* Unscientific and Superficial Sentence, *CM* Sentence Containing Misconception

Architecture enhanced by the researchers, the best accuracy rate was achieved by combining the ML algorithms used (Accuracy = 88.90%). A graphical representation of this finding is given in Graph 1. Confusion Matrix information about AI classifiers is given in Appendix Figure 3.

A high accuracy value indicates that the model makes correct classifications in general. However, in unbalanced datasets, accuracy can be misleading because the performance of low-frequency classes may be underestimated. Therefore, recall and precision values provide a deeper insight. A high recall indicates that the model correctly identifies most positive examples. However, low recall means that the model misses many positive samples, which is a major problem, especially when critical positive samples are important (e.g. disease diagnosis). A high precision indicates that the majority of the model's positive predictions are correct. Low precision indicates that the model makes many false positive predictions, which can lead to false alarms. Especially in situations where precision is important (e.g. spam filtering, medical diagnostic systems, credit card fraud detection), low precision can cause annoyance to users. The F1-Score provides a balance between precision and recall, making it possible to evaluate these two



Graph 1 Graphical representation of AI classifiers and accuracy values

metrics together. The confusion matrix summarises the number of true and false classifications into four categories (True Positive, False Positive, True Negative, False Negative). Each cell in this matrix reveals in more detail how successful the model is for a particular class. For example, a high False Positive rate indicates that the model makes too many false positive predictions, which lowers the precision value. Similarly, a high False Negative rate indicates that the model missed many positive examples, which negatively affects the recall value. Considering Table 5, it is observed that all these metrics are evenly distributed in our study.

5.3 RQ#3. Is there a significant difference between the AI algorithm and the human expert evaluation?

The study used κ statistics to measure the agreement between AI and human expert classification. AI and human expert cross table is given in Table 6.

Table 6 AI and human expert cross-tabulation

Statistics Count		Expert Human			Total
		CSI	US	CM	
AI	CSI	14	2	1	17
	US	1	14	2	17
	CM	3	0	22	25
Total		18	16	25	59

Table 7 Symmetric measurement results

	Value	Asymptotic Standard Error	Approximate T	Approximate Significance (p)
Measure of Agreement Kappa	0.767	.071	8.275	0.000*
N of Valid Cases	59			

$p < 0.05^*$

Symmetrical measurement results are given in Table 7. There appears to be a significant difference between AI and human expert classification. Additionally, κ was calculated as 0.767.

6 Discussion and implications

When the findings obtained through WAT regarding the first sub-problem are examined, it is seen that the number of sentences containing misconceptions is 175. When the sentences containing misconceptions (Table 4) were examined, it was determined that the statements of the teacher candidates “The depletion of the ozone layer causes harmful sunlight to enter our world” ($f=30$) and “Acid rain and global warming are factors that increase the greenhouse effect” ($f=30$) caused the most misconceptions. Similar misconceptions have been revealed in the literature in parallel with this finding (Boyes et al., 1993; Gautier et al., 2006). Another statement that attracts attention is “The greenhouse effect is the agricultural greenhouse activity” ($f=27$). Another statement is “It causes the planet’s temperature to increase and climates to change as a result of air pollution.” ($f=26$). There are studies in the literature that support this finding (Boyes et al., 1993; Jeffries et al., 2001; Kilinc et al., 2008; Malandrakis et al., 2011; Österlind, 2005). In a limited number of studies conducted with teacher candidates, it was revealed that teacher candidates had misconceptions about the greenhouse effect (Topsakal & Altınöz, 2010), and it is emphasized that this situation arises because the greenhouse effect is a complex phenomenon (Kilinc et al., 2013; Österlind, 2005; Shepardson et al., 2009; Shepardson et al., 2011).

Regarding the second sub-problem of the study, it was found that the AI algorithm used to predict teacher candidates’ misconceptions was MLP ($Accuracy=86.42\%$), which had the highest accuracy rate. Upon reviewing the literature, it was determined that MLP represent an efficient algorithm for addressing supervised learning tasks (El Boucheffy & de Souza, 2020). MLPs enjoy widespread popularity and are utilized more frequently than other neural network architectures across a diverse range of problem domains (Park & Lek, 2016). It was found in the study that the accuracy values of other AI classification tools were SVM ($Accuracy=83.95\%$), GB ($Accuracy=83.95\%$), RF ($Accuracy=83.95\%$), k-NN ($Accuracy=82.71\%$) and LR ($Accuracy=80.24\%$). Finally, it was observed that the accuracy rate increased

to 88.90% with the Enhanced Ensemble Model Architecture, which incorporates all ML algorithms. According to this model, since the results are obtained through the voting method of all ML algorithms, the prediction results provided by the proposed model can be more confidently validated than those from individually applied algorithms. This finding is consistent with the findings of Alam et al. (2021).

According to the findings regarding the third sub-problem, the kappa (κ) value was examined to determine the significant difference between the AI algorithm and the expert evaluation. According to Landis and Koch (1977a, 1977b), the strength of fit ($0.61 < \kappa = 0.767 < 0.80$) was found to be significant. Therefore, it was concluded that there was a significant difference between the AI classifiers and human expert classification ($0.05 > p = 0.00$) and a significant level of agreement was observed. A similar finding was found by Göktepe Körpeoğlu and Göktepe Yıldız (2023). As a result, AI-supported tools and applications in education are used to improve the quality of services provided to students and teachers (Suh & Ahn, 2022). It was seen that the predictions of the developed AI model and human expert evaluations were similar.

In consequence, the findings of the study have emerged as an important alternative in today's pedagogical approach where the support of information technologies in the process of developing conceptual understanding through the detection of conceptual misconceptions is becoming increasingly important. This is because research on the application of artificial intelligence in various disciplines shows promise in transforming and enhancing educational experiences, providing personalised, interactive and innovative learning opportunities (Jia et al., 2024). In the process of detecting and eliminating misconceptions in educational research, it has been clearly seen that the classification of the data set analysed by NLP with ML algorithms will provide a significant advantage in terms of time management. In this process, artificial intelligence plays a particularly important role in automating routine tasks (West et al., 2023). According to Buluş and Elmas (2024), after the teaching is completed, the student can be subjected to algorithmic evaluations using artificial intelligence applications to identify deficiencies. Thus, the individual learning process of the student is structured in a unique way (Wardat et al., 2023). Artificial intelligence can analyse whether the student needs additional education and make suggestions to the teacher (Cooper, 2023).

6.1 Conclusion

In this study, misconceptions among teacher candidates regarding the greenhouse effect have been identified. These findings were determined through qualitative analyses conducted by both human experts and AI. Moreover, it was concluded that there is a promising level of agreement between the classifications made by AI classifiers and human experts, indicating significant accuracy/consistency. Therefore, our study has reached the conclusion that qualitative analyses can also be performed with AI classifiers. These results point to a promising future for NLP applications in education. This study has demonstrated the feasibility of using supervised AI to classify conceptual change. Evaluating qualitative data obtained from open-ended

questions through AI offers significant advantages in terms of quality and time. Thus, this notion provides innovative opportunities for qualitative analyses.

This study presents new findings in Turkey by using AI methods to analyze teacher candidates' misconceptions about greenhouse effect. It is anticipated that the ability to predict misconceptions using AI algorithms will bring a different dimension to both the research and education communities. The limited existing studies in the literature and the findings of this research indicate that the NLP method provides successful and reliable results. This study serves as an example demonstrating that it is possible to predict teacher candidates' conceptual understandings using AI.

6.2 Limitations and future work

In the context of AI applications in education, the potential bias is associated with the reliability of the AI results. This study can address this bias through a comparison between human experts and AI, along with kappa (κ) value results. Additionally, supervised learning, which involves two domain experts labelling the data, enhances the reliability of the results obtained from AI. Another approach to mitigate potential bias in AI applications is the implementation of Ensemble models, which increase the reliability of the study. This study focuses on AI applications in education. In this context, the applicability of the study is limited to the concept of greenhouse effect and science teacher candidates. Furthermore, when examining the research findings, the main reason for the accuracy values detected by AI algorithms being lower than the predicted value is thought to be due to the small number of words in the sentences written by teacher candidates, which prevents the AI models from fully learning from the dataset. This result constitutes a limitation of the study. However, this insight may provide an advantage for researchers preparing to work on similar topics. In order to achieve a more consistent result in the dataset to be created, it would be advisable to increase the number of words in each line and to include longer sentences in the dataset.

This study, limited to qualitative data related to the greenhouse effect concept, aims to detect misconceptions using AI. Therefore, AI applications can be preferred for research topics such as conceptual change, misconceptions, perceptions of a situation, and awareness in various disciplines. In this regard, the study will lead other researchers in creating innovative opportunities for qualitative data analysis. Additionally, this study may shed light on future studies aimed at identifying different misconceptions about science subjects. Designed to support teachers, AI can assess whether particular students need additional instruction on a particular subject using criteria such as how much time students spend on an activity, the number of correct answers, and the number of attempts.

Appendix

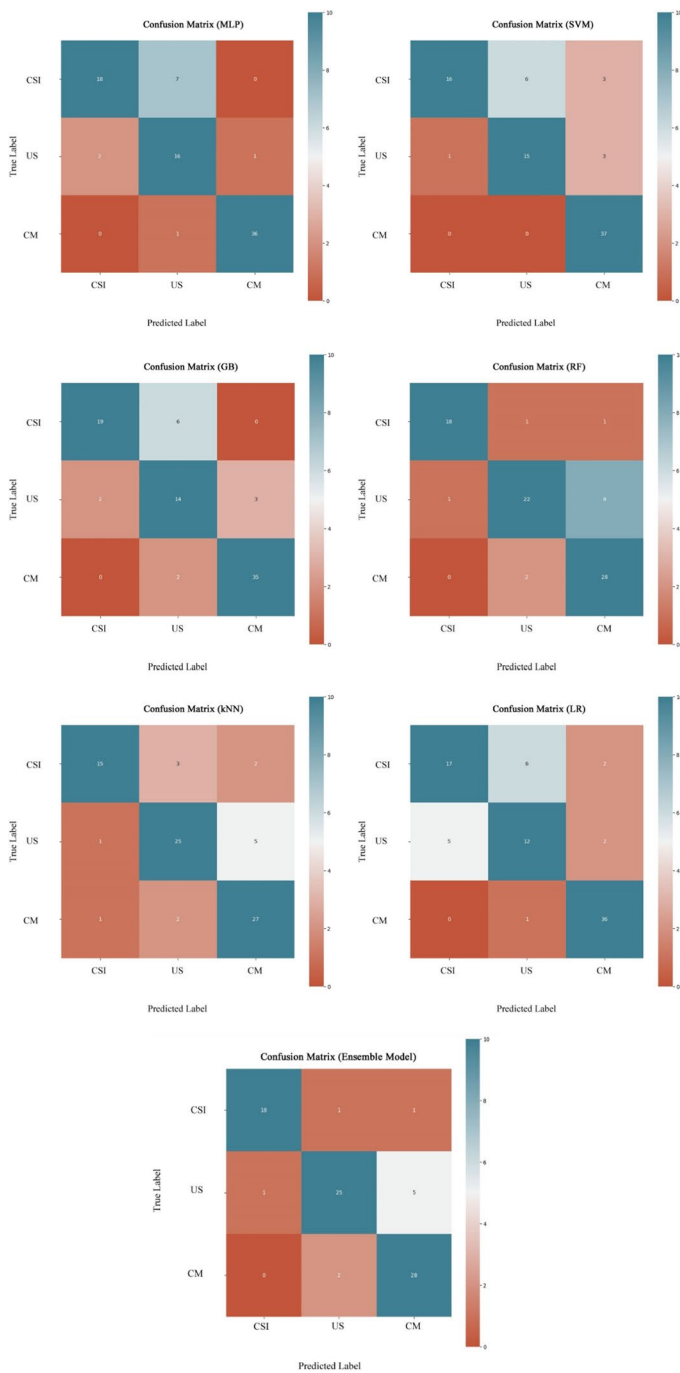


Fig. 3 Confusion matrix information of the AI classifiers

Funding Open access funding provided by the Scientific and Technological Research Council of Türkiye (TÜBİTAK). The authors declare that they have received no funding, grants or other support in the preparation of this article.

Data availability Available from the corresponding author upon reasonable request.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Alam, K. S., Bhowmik, S., & Prosun, P. R. K. (2021). Cyberbullying detection: an ensemble based machine learning approach. In 2021 third international conference on intelligent communication technologies and virtual mobile networks (ICICV) (pp. 710–715). IEEE. DOI: <https://doi.org/10.1109/ICICV50876.2021.9388499>.
- Andersson, B., & Wallin, A. (2000). Students' understanding of the greenhouse effect, the societal consequences of reducing CO2 emissions and the problem of ozone layer depletion. *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching*, 37(10), 1096–1111.
- Anohina, A. (2007). Advances in intelligent tutoring systems: Problem-solving modes and model of hints. *International Journal of Computers Communications & Control*, 2(1), 48–55. Retrieved October 23, 2023, from <https://univagora.ro/jour/index.php/ijccc/article/view/2336/805>
- Arshad, A., Jabeen, M., Ubaid, S., Raza, A., Abualigah, L., Aldiabat, K., & Jia, H. (2023). A novel ensemble method for enhancing internet of things device security against botnet attacks. *Decision Analytics Journal*, 8, 100307. <https://doi.org/10.1016/J.DAJOUR.2023.100307>
- Aşkar, P., & Kızılkaya, G. (2006). Eğitim yazılımlarında eğitsel yardımcı kullanımı: Eğitsel ajan [Using pedagogical assistants in educational software: A pedagogical agent]. *Hacettepe University Journal of Faculty of Education*, 31(31), 25–31.
- Aung, Z. H., Sanium, S., Songsaksupachok, C., Kusakunniran, W., Precharattana, M., Chuechote, S., et al. (2022). Designing a novel teaching platform for AI: A case study in a Thai school context. *Journal of Computer Assisted Learning*, 38(6), 1714–1729. <https://doi.org/10.1111/jcal.12706>
- Ausubel, D. P. (1968). *Educational Psychology: A cognitive view*. Holt, Rinehart, and Winston.
- Bahar, M., & Özathı, N. S. (2003). Kelime iletişim test yöntemi ile lise 1. sınıf öğrencilerinin canlıların temel bileşenleri konusundaki bilişsel yapılarının araştırılması [investigating the cognitive structures of high school 1st grade students about the basic components of living things with word communication test method]. *Journal of Balıkesir University Institute of Science and Technology*, 5(2), 75–85.
- Baker, T., & Smith, L. (2019). Educ-AI-tion rebooted? Exploring the future of artificial intelligence in schools and colleges. Retrieved October 10, 2023, from https://media.nesta.org.uk/documents/Future_of_AI_and_education_v5_WEB.Pdf
- Ballantyne, R., & Rain, J. (1995). Enhancing environmental conceptions: An evaluation of cognitive conflict and structured controversy learning units. *Studies in Higher Education*, 20(3), 293–303. <https://doi.org/10.1080/03075079512331381565>

- Berland, M., Davis, D., & Smith, C. P. (2015). Amoeba: Designing for collaboration in computer science classrooms through live learning analytics. *International Journal of Computer-Supported Collaborative Learning*, 10, 425–447. <https://doi.org/10.1007/s11412-015-9217-z>
- Bertram, C., Weiss, Z., Zachrich, L., & Ziai, R. (2021). Artificial intelligence in history education. Linguistic content and complexity analyses of student writings in the CAHisT project (Computational assessment of historical thinking). *Computers and Education: Artificial Intelligence* (p. 100038). <https://doi.org/10.1016/j.caeai.2021.100038>
- Bishop, C. M., & Nasrabadi, N. M. (2006). Pattern recognition and machine learning (Vol. 4, No. 4, p. 738). : Springer.
- Bisong, E. (2019). The multilayer perceptron (MLP). In *Building machine learning and deep learning models on Google cloud platform*. Apress. https://doi.org/10.1007/978-1-4842-4470-8_31
- Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In D. Haussler (Ed.), *5th annual ACM workshop on COLT* (pp. 144–152). ACM Press.
- Bostrom, N. (2006). Quantity of experience: Brain-duplication and degrees of consciousness. *Minds and Machines*, 16, 185–200. <https://doi.org/10.1007/s11023-006-9036-0>
- Boyes, E., Chuckran, D., & Stanisstreet, M. (1993). How do high school students perceive global climatic change: What are its manifestations? What are its origins? What corrective action can be taken? *Journal of Science Education and Technology*, 2, 541–557. <https://doi.org/10.1007/BF00695323>
- Brachman, R. J., & Anand, T. (1994, July). The process of knowledge discovery in databases: A first sketch. In *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining* (pp. 1–11).
- Branch, R. M. (2009). Instructional design: The ADDIE approach. *Springer Science & Business Media*. <https://doi.org/10.1007/978-0-387-09506-6>
- Brennen, R. L., & Prediger, D. J. (1981). Coefficient kappa: Some uses, misuses, and alternatives. *Educational and Psychological Measurement*, 41(1981), 687–699.
- Buluş, B., & Elmas, R. (2024). Yapay zeka uygulamalarının kimya eğitiminde kullanımı alternatif araçlar [the use of artificial intelligence applications as alternative tools in chemistry education]. *Journal of Turkish Chemical Society Section C: Chemistry Education (JOTCSC)*, 9(1), 1–28. <https://doi.org/10.37995/jotcsc.1366999>
- Chang, H. P., Chen, J. Y., Guo, C. J., Chen, C. C., Chang, C. Y., Lin, S. H., Su, W. J., Lain, K. D., Hsu, S. Y., Lin, J. L., Chen, C. C., Cheng, Y. T., Wang, L. S., & Tseng, Y. T. (2007). Investigating primary and secondary students' learning of physics concepts in Taiwan. *International Journal of Science Education*, 29(4), 465–482. <https://doi.org/10.1080/095006090601073210>
- Chassignol, M., Khoroshavin, A., Klimova, A., & Bilyatdinova, A. (2018). Artificial intelligence trends in education: A narrative overview. *Procedia Computer Science*, 136, 16–24. <https://doi.org/10.1016/j.procs.2018.08.233>
- Chen, X., Xie, H., Zou, D., & Hwang, G. J. (2020). Application and theory gaps during the rise of artificial intelligence in education. *Computers and Education: Artificial Intelligence*, 1, 100002. <https://doi.org/10.1016/j.caeai.2020.100002>
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37–46.
- Coiera, E. (2003). Clinical decision support systems. *Guide to health informatics*, 2(1), 1–12. Retrieved November 2, 2023, from <http://www.coiera.com/aimd.htm>
- Cooper, G. (2023). Examining science education in ChatGPT: An exploratory study of generative artificial intelligence. *Journal of Science Education and Technology*, 32(3), 444–452. <https://doi.org/10.1007/s10956-023-10039-y>
- Cortes, C., & Vapnik, V. N. (1995). Support vector networks. *Machine Learning*, 20(3), 273–297.
- Crawforth, K. (2001). Measuring the interrater reliability of a data collection instrument developed to evaluate anesthetic outcomes (Doctoral Dissertation). Available from Proquest Dissertations and Theses database. (UMI No. 3037063).
- Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge Univ. Press.
- Cunningham, P., Cord, M., & Delany, S. J. (2008). Supervised learning. In M. Cord & P. Cunningham (Eds.), *Machine learning techniques for multimedia. Cognitive technologies*. Springer. https://doi.org/10.1007/978-3-540-75171-7_2
- Davies, N. (2016). Can robots handle your healthcare? *Journal of Engineering Technology*, 11(9), 58–61. <https://doi.org/10.1049/et.2016.0907>

- Demirezen, M. U., Yilmaz, O., & Ince, E. (2023). New models developed for detection of misconceptions in physics with artificial intelligence. *Neural Computing and Applications*, 35(12), 9225–9251. <https://doi.org/10.1007/S00521-023-08414-2/FIGURES/13>
- Di Eugenio, B., Fossati, D., & Green, N. (2021). *Intelligent support for computer science education: Pedagogy enhanced by artificial intelligence*. CRC Press.
- Dodigovic, M. (2007). Artificial intelligence and second language learning: An efficient approach to error remediation. *Language Awareness*, 16(2), 99–113. <https://doi.org/10.2167/la416.0>
- El Boucheffy, K., & de Souza, R. S. (2020). Learning in Big Data: Introduction to Machine Learning. Knowledge Discovery in Big Data from Astronomy and Earth Observation: Astrogeoinformatics, 225–249. DOI: <https://doi.org/10.1016/B978-0-12-819154-5.00023-0>.
- Ercan, F., Taşdere, A., & Ercan, N. (2010). Kelime ilişkilendirme testi aracılığıyla bilişsel yapının ve kavramsal değişimin gözlenmesi. *Journal of Turkish Science Education*, 7, 136–154.
- Fayyad, U. M., Piatetsky-Shapiro, G., & Smyth, P. (1996). Knowledge Discovery and Data Mining: Towards a Unifying Framework. In KDD (Vol. 96, pp. 82–88).
- Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5), 378382.
- Gadanidis, G., Hughes, J. M., Minniti, L., & White, B. J. (2017). Computational thinking, grade 1 students and the binomial theorem. *Digital Experiences in Mathematics Education*, 3, 77–96. <https://doi.org/10.1007/s40751-016-0019-3>
- Gautier, C., Deutsch, K., & Rebich, S. (2006). Misconceptions about the greenhouse effect. *Journal of Geoscience Education*, 54(3), 386–395. <https://doi.org/10.5408/1089-9995-54.3.386>
- Ghahramani, Z. (2004). Unsupervised learning. In O. Bousquet, U. von Luxburg, & G. Rätsch (Eds.), *Advanced lectures on machine learning* (ML 2003. Lecture notes in computer science) (Vol. 3176). Springer. https://doi.org/10.1007/978-3-540-28650-9_5
- Göktepe Körpeoğlu, S., & Göktepe Yıldız, S. (2023). Comparative analysis of algorithms with data mining methods for examining attitudes towards STEM fields. *Education and Information Technologies*, 28(3), 2791–2826. <https://doi.org/10.1007/s10639-022-11216-z>
- Gunston, R. F. (1980). Word association and the description of cognitive structure. *Research in Science Education*, 10, 45–53.
- Hanafi, Y., Murtadho, N., & Ikhsan, M. A. (2020). Reinforcing public university student's worship education by developing and implementing mobile-learning management system in the ADDIE instructional design model. *International Journal of Interactive Mobile Technologies*, 14(2). <https://doi.org/10.3991/ijim.v14i02.11380>
- Harmon, M. E., & Harmon, S. S. (1996). Reinforcement learning: A tutorial. WL/AAFC, WPAFB Ohio, 45433, 237–285.
- Haykin, S. (2009). *Neural networks and learning machines*, 3/E. Pearson Education India.
- Hellings, J., & Haelermans, C. (2022). The effect of providing learning analytics on student behaviour and performance in programming: A randomised controlled experiment. *Higher Education*, 83, 1–18. <https://doi.org/10.1007/s10734-020-00560-z>
- Huang, C., Tu, Y., Han, Z., Jiang, F., Wu, F., & Jiang, Y. (2023). Examining the relationship between peer feedback classified by deep learning and online learning burnout. *Computers & Education*, 207, 104910. <https://doi.org/10.1016/J.COMPEDU.2023.104910>
- Huang, S. P. (2018). Effects of using artificial intelligence teaching system for environmental education on environmental knowledge and attitude. *Eurasia Journal of Mathematics, Science and Technology Education*, 14(7), 3277–3284. <https://doi.org/10.29333/ejmste/91248>
- Huang, X., & Qiao, C. (2022). Enhancing computational thinking skills through artificial intelligence education at a STEAM high school. *Science & Education*, 1–21. <https://doi.org/10.1007/s11191-022-00392-6>
- Hwang, G. J., Xie, H., Wah, B. W., & Gašević, D. (2020). Vision, challenges, roles and research issues of artificial intelligence in education. *Computers and Education: Artificial Intelligence*, 1, 100001. <https://doi.org/10.1016/j.caeai.2020.100001>
- Iqbal, T., & Wani, M. A. (2023). Weighted ensemble model for image classification. *International Journal of Information Technology*, 15(2), 557–564. <https://doi.org/10.1007/s41870-022-01149-8>
- Jang, J., Jeon, J., & Jung, S. K. (2022). Development of STEM-based AI education program for sustainable improvement of elementary learners. *Sustainability*, 14(22), 15178. <https://doi.org/10.3390/su142215178>

- Jeffries, H., Stanisstreet, M., & Boyes, E. (2001). Knowledge about the 'Greenhouse Effect': Have college students improved? *Research in Science & Technological Education*, 19(2), 205–221. <https://doi.org/10.1080/02635140120087731>
- Jia, F., Sun, D., & Looi, C. K. (2024). Artificial intelligence in science education (2013–2023): Research trends in ten years. *Journal of Science Education and Technology*, 33(1), 94–117. <https://doi.org/10.1007/s10956-023-10077-6>
- Kahraman, S. (2020). Fen bilgisi öğretmen adaylarının sera etkisi kavramı ile ilgili bilişsel yapıları [pre-service science teachers' cognitive structure on the concept of greenhouse effect]. *Inonu University Journal of the Graduate School of Education*, 7(14), 42–55.
- Kaplan, A., & Haenlein, M. (2019). Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. *Business Horizons*, 62(1), 15–25. <https://doi.org/10.1016/j.bushor.2018.08.004>
- Karamustafaoğlu, O., & Pektaş, H. M. (2023). Developing students' creative problem solving skills with inquiry-based STEM activity in an out-of-school learning environment. *Education and Information Technologies*, 28(6), 7651–7669. <https://doi.org/10.1007/s10639-022-11496-5>
- Karmani, P., Chandio, A. A., Korejo, I. A., & Chandio, M. S. (2018). A review of machine learning for healthcare informatics specifically tuberculosis disease diagnostics. In *Intelligent Technologies and Applications: First International Conference, INTAP 2018, Bahawalpur, Pakistan, October 23–25, 2018, Revised Selected Papers 1* (pp. 50–61).
- Kelly, G. A. (1955). *The psychology of personal constructs* (p. 1955). W.W. Norton & Company.
- Khalid, T. (2003). Pre-service high school teachers' perceptions of three environmental phenomena. *Environmental Education Research*, 9(1), 35–50. <https://doi.org/10.1080/13504620303466>
- Khurana, D., Koli, A., Khatter, K., & Singh, S. (2023). Natural language processing: State of the art, current trends and challenges. *Multimedia Tools and Applications*, 82(3), 3713–3744. <https://doi.org/10.1007/S11042-022-13428-4/FIGURES/3>
- Kiangala, S. K., & Wang, Z. (2021). An effective adaptive customization framework for small manufacturing plants using extreme gradient boosting-XGBoost and random forest ensemble learning algorithms in an industry 4.0 environment. *Machine Learning with Applications*, 4, 100024. <https://doi.org/10.1016/J.MLWA.2021.100024>
- Kilinc, A., Eroglu, B., Boyes, E., & Stanisstreet, M. (2013). Could organisms and ecosystems be used as motivators for behaviour to reduce global warming? The views of school students. *International Research in Geographical and Environmental Education*, 22(3), 191–208. <https://doi.org/10.1080/10382046.2013.817663>
- Kilinc, A., Stanisstreet, M., & Boyes, E. (2008). Turkish students' ideas about global warming. *International Journal of Environmental and Science Education*, 3(2), 89–98.
- Kohavi, R. (1995). A Study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, Vol. 2, Montreal, 1137–1145.
- Kong, F. (2020). Application of artificial intelligence in modern art teaching. *International Journal of Emerging Technologies in Learning (iJET)*, 15(13), 238–251. Retrieved October 19, 2023, from <https://www.learntechlib.org/p/217610/>
- Kong, S. C., & Shen, W. (2024). Using students' cognitive, affective, and demographic characteristics to predict their understanding of computational thinking concepts: A machine learning-based approach. *Interactive Learning Environments*, 1–14. <https://doi.org/10.1080/10494820.2024.2331148>
- Kose, U., & Arslan, A. (2017). Optimization of self-learning in computer engineering courses: An intelligent software system supported by artificial neural network and vortex optimization algorithm. *Computer Applications in Engineering Education*, 25(1), 142–156. <https://doi.org/10.1002/cae.21787>
- Kurt, H., & Ekici, G. (2013). Biyoloji öğretmen adaylarının bağımsız kelime ilişkilendirme testi ve çizme-yazma tekniğiyle "Osmoz" kavramı konusundaki bilişsel yapılarının belirlenmesi [Determining biology student teachers' cognitive structure on the concept of "osmosis" through the free word-association test and the drawingwriting technique]. *Turkish Studies (Elektronik)*, 8(12), 809–829.
- Landis, J. R., & Koch, G. (1977a). The measurement of observer agreement for categorical data. *Biometrics*, 33, 159–174. <https://doi.org/10.2307/2529310>

- Landis, J. R., & Koch, G. G. (1977b). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, 363–374. <https://doi.org/10.2307/2529786>
- Liu, C., Hou, J., Tu, Y. F., Wang, Y., & Hwang, G. J. (2021). Incorporating a reflective thinking promoting mechanism into artificial intelligence-supported English writing environments. *Interactive Learning Environments*, 1–19. <https://doi.org/10.1080/10494820.2021.2012812>
- Liu, S., Liu, S., Liu, Z., Peng, X., & Yang, Z. (2022). Automated detection of emotional and cognitive engagement in MOOC discussions to predict learning achievement. *Computers & Education*, 181, 104461. <https://doi.org/10.1016/J.COMPEDU.2022.104461>
- Lodhi, P., Mishra, O., Jain, S., & Bajaj, V. (2018). StuA: An intelligent student assistant. *International Journal of Interactive Multimedia and Artificial Intelligence*, 5(2), 17–25. <https://doi.org/10.9781/ijimai.2018.02.008>
- Long, L., & Zeng, X. (2022). Beginning deep learning with TensorFlow: Work with Keras. *MNIST Data Sets, and Advanced Neural Networks*. <https://doi.org/10.1007/978-1-4842-7915-1>
- Malandrakis, G., Boyes, E., & Stanisstreet, M. (2011). Global warming: Greek students' belief in the usefulness of pro-environmental actions and their intention to take action. *International Journal of Environmental Studies*, 68(6), 947–963. <https://doi.org/10.1080/00207233.2011.590720>
- Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis*. SAGE.
- Mintzes, J. J., & Wandersee, J. H. (1998). Reform and innovation in science teaching: A human constructivist view. In J. J. Mintzes, J. H. Wandersee, & J. D. Novak (Eds.), *Teaching science for understanding. A human constructivist view*. Academic Press.
- Molenda, M. (2003). In search of the elusive ADDIE model. *Performance Improvement*, 42(5), 34–36. <https://doi.org/10.1002/pf.4930420508>
- Montejo-Ráez, A., & Jiménez-Zafra, S. M. (2022). Current approaches and applications in natural language processing. *Applied Sciences*, 12(4859), 1–6. <https://doi.org/10.3390/APP12104859>
- Morrison, G. R. (2010). *Designing effective instruction* (6th ed.). John Wiley & Sons.
- Myers, G., Boyes, E., & Stanisstreet, M. (2004). School students' ideas about air pollution: Knowledge and attitudes. *Research in Science & Technological Education*, 22(2), 133–152. <https://doi.org/10.1080/0263514042000290868>
- Nguyen, H., & Diederich, M. (2023). Facilitating knowledge construction in informal learning: A study of TikTok scientific, educational videos. *Computers & Education*, 205, 104896. <https://doi.org/10.1016/J.COMPEDU.2023.104896>
- Österlind, K. (2005). Concept formation in environmental education: 14-year olds' work on the intensified greenhouse effect and the depletion of the ozone layer. *International Journal of Science Education*, 27(8), 891–908. <https://doi.org/10.1080/09500690500038264>
- Ouyang, F., & Jiao, P. (2021). Artificial intelligence in education: The three paradigms. *Computers and Education: Artificial Intelligence*, 2, 1–6. <https://doi.org/10.1016/j.caeai.2021.100020>
- Park, Y. S., & Lek, S. (2016). Artificial neural networks: Multilayer perceptron for ecological modeling. In *Developments in environmental modelling* (Vol. 28, pp. 123–140). Elsevier. <https://doi.org/10.1016/B978-0-444-63623-2.00007-4>
- Pokrivcakova, S. (2019). Preparing teachers for the application of AI-powered technologies in foreign language education. *Journal of Language and Cultural Education*, 7(3), 135–153. <https://doi.org/10.2478/jolace-2019-0025>
- Poole, D. I., Goebel, R. G., & Mackworth, A. K. (1998). *Computational intelligence* (Vol. 1). Oxford University Press <https://www.cs.ubc.ca/~poole/ci/front.pdf>
- Popescu, M. C., Balas, V. E., Perescu-Popescu, L., & Mastorakis, N. (2009). Multilayer perceptron and neural networks. *WSEAS Transactions on Circuits and Systems*, 8(7), 579–588.
- Qiang, W., & Zhongli, Z. (2011). Reinforcement learning model, algorithms and its application. In 2011 International Conference on Mechatronic Science, Electric Engineering and Computer (MEC) (pp. 1143–1146). IEEE.
- Rasheed, F., & Wahid, A. (2021). Learning style detection in E-learning systems using machine learning techniques. *Expert Systems with Applications*, 174, 114774. <https://doi.org/10.1016/J.ESWA.2021.114774>
- Rowell, J. A., Dawson, C. J., & Lyndon, H. (1990). Changing misconceptions: A challenge to science educators. *International Journal of Science Education*, 12(2), 167–175. <https://doi.org/10.1080/0950069900120205>
- Russell, S. J., & Norvig, P. (2016). *Artificial intelligence: A modern approach*. Pearson Education Limited.

- Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, 3(3), 210–229. <https://doi.org/10.1147/rd.33.0210>
- Schulte, P. L. (2001). *Preservice primary teacher alternative conceptions in science and attitudes toward teaching science (Doktora Tezi)*. New Orleans Üniversitesi.
- Sebald, D. J., & Buklew, J. A. (2000). Support vector machine techniques for nonlinear equalization. *IEEE Transactions on Signal Processing*, 48(11), 3217–3227. <https://doi.org/10.1109/78.875477>
- Shafique, U., & Qaiser, H. (2014). A comparative study of data mining process models (KDD, CRISP-DM and SEMMA). *International Journal of Innovation and Scientific Research*, 12(1), 217–222. Retrieved October 16, 2023, from <http://www.ijisr.issr-journals.org/>
- Shepardson, D. P., Niyogi, D., Choi, S., & Charusombat, U. (2009). Seventh grade students' conceptions of global warming and climate change. *Environmental Education Research*, 15(5), 549–570. <https://doi.org/10.1080/13504620903114592>
- Shepardson, D. P., Niyogi, D., Choi, S., & Charusombat, U. (2011). Students' conceptions about the greenhouse effect, global warming, and climate change. *Climatic Change*, 104(3–4), 481–507. <https://doi.org/10.1007/s10584-009-9786-9>
- Shin, J., Balyan, R., Banawan, M. P., Arner, T., Leite, W. L., & McNamara, D. S. (2023). Pedagogical discourse markers in online algebra learning: Unraveling instructor's communication using natural language processing. *Computers & Education*, 205, 104897. <https://doi.org/10.1016/J.COMPEDU.2023.104897>
- Subheesh, N. P., Sobin, C. C., Ali, J., & Varsha, M. (2022). Classification of Students' Misconceptions in Individualised Learning Environments (C-SMILE): An Innovative Assessment Tool for Engineering Education Settings. IEEE Global Engineering Education Conference, EDUCON, 2022-March, 795–800. DOI: <https://doi.org/10.1109/EDUCON52537.2022.9766572>.
- Suh, W., & Ahn, S. (2022). Utilizing the metaverse for learner-centered constructivist education in the post-pandemic era: An analysis of elementary school students. *Journal of Intelligence*, 10(1), 17. <https://doi.org/10.3390/jintelligence10010017>
- Tang, J., Deng, C., & Huang, G. B. (2015). Extreme learning machine for multilayer perceptron. *IEEE Transactions on Neural Networks and Learning Systems*, 27(4), 809–821.
- Thabtah, F., Abdelhamid, N., & Peebles, D. (2019). A machine learning autism classification based on logistic regression analysis. *Health Information Science and Systems*, 7(1), 1–11. <https://doi.org/10.1007/S13755-019-0073-5/TABLES/7>
- Theodoridis, S., & Koutroumbas, K. (2003). *Recognition* (2nd ed.). Academic Press.
- Topsakal, Ü. U., & Altınöz, N. (2010). İlköğretim öğretmen adaylarının sera etkisi ile ilgili kavramları algılama düzeyleri [the perception levels of elementary school teacher candidates about greenhouse effect]. *Ercincan University Journal of Education Faculty*, 12(1), 147–163.
- Trust, T., & Pektas, E. (2018). Using the ADDIE model and universal design for learning principles to develop an open online course for teacher professional development. *Journal of Digital Learning in Teacher Education*, 34(4), 219–233. <https://doi.org/10.1080/21532974.2018.1494521>
- Tumanggor, A. M. R., Supahar, K. H., & Ringo, E. S. (2020). Using four-tier diagnostic test instruments to detect physics teacher candidates' misconceptions: Case of mechanical wave concepts. *Journal of Physics: Conference Series*, 1440, 1–8. <https://doi.org/10.1088/1742-6596/1440/1/012059>
- Vygotsky, L. S. (1962). *Thought and language*. MIT Press.
- Wambsganss, T., Janson, A., & Leimeister, J. M. (2022). Enhancing argumentative writing with automated feedback and social comparison nudging. *Computers & Education*, 191, 104644. <https://doi.org/10.1016/J.COMPEDU.2022.104644>
- Wandersee, J. H., Mintzes, J. J., & Novak, J. (1994). Research on alternative conceptions in science. In D. L. Gabel (Eds.), *Handbook of research on science teaching and learning*. Macmillan.
- Wardat, Y., Tashtoush, M. A., AlAli, R., & Jarrah, A. M. (2023). ChatGPT: A revolutionary tool for teaching and learning mathematics. *Eurasia Journal of Mathematics, Science and Technology Education*, 19(7), 1–18. <https://doi.org/10.29333/ejmste/13272>
- Welham, D. (2008). AI in training (1980–2000): Foundation for the future or misplaced optimism? *British Journal of Educational Technology*, 39(2), 287–303. <https://doi.org/10.1111/j.1467-8535.2008.00818.x>
- West, J. K., Franz, J. L., Hein, S. M., Leverentz-Culp, H. R., Mauser, J. F., Ruff, E. F., & Zemke, J. M. (2023). An analysis of AI-generated laboratory reports across the chemistry curriculum and student perceptions of ChatGPT. *Journal of Chemical Education*, 100(11), 4351–4359. <https://doi.org/10.1021/acs.jchemed.3c00581>

- Xu, W., & Ouyang, F. (2022). The application of AI technologies in STEM education: A systematic review from 2011 to 2021. *International Journal of STEM Education*, 9(1), 1–20. <https://doi.org/10.1186/s40594-022-00377-5>
- Xue, Y., & Wang, Y. (2022). Artificial intelligence for education and teaching. *Wireless Communications and Mobile Computing*, 2022, 1–10. <https://doi.org/10.1155/2022/4750018>
- Yannier, N., Hudson, S. E., & Koedinger, K. R. (2020). Active learning is about more than hands-on: A mixed-reality AI system to support STEM education. *International Journal of Artificial Intelligence in Education*, 30, 74–96. <https://doi.org/10.1007/s40593-020-00194-3>
- Zampirolli, F. A., Borovina Josko, J. M., Venero, M. L., Kobayashi, G., Fraga, F. J., Goya, D., & Savegnago, H. R. (2021). An experience of automated assessment in a large-scale introduction programming course. *Computer Applications in Engineering Education*, 29(5), 1284–1299. <https://doi.org/10.1002/cae.22385>
- Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education—where are the educators? *International Journal of Educational Technology in Higher Education*, 16(1), 1–27. <https://doi.org/10.1186/s41239-019-0171-0>
- Zhang, J., Oh, Y. J., Lange, P., Yu, Z., & Fukuoka, Y. (2020). Artificial intelligence chatbot behavior change model for designing artificial intelligence chatbots to promote physical activity and a healthy diet. *Journal of Medical Internet Research*, 22(9), e22845. <https://doi.org/10.2196/22845>
- Zhang, S., Li, X., Zong, M., Zhu, X., & Wang, R. (2018). Efficient kNN classification with different numbers of nearest neighbors. *IEEE Transactions on Neural Networks and Learning Systems*, 29(5), 1774–1785. <https://doi.org/10.1109/TNNLS.2017.2673241>
- Zhao, H., Li, G., & Feng, W. (2018). Research on application of artificial intelligence in medical education. In 2018 International Conference on Engineering Simulation and Intelligent Control (ESAIC) (pp. 340–342). IEEE. <https://doi.org/10.1109/ESAIC.2018.00085>
- Zou, W., Hu, X., Pan, Z., Li, C., Cai, Y., & Liu, M. (2021). Exploring the relationship between social presence and learners' prestige in MOOC discussion forums using automated content analysis and social network analysis. *Computers in Human Behavior*, 115, 106582. <https://doi.org/10.1016/j.chb.2020.106582>
- Zulić, H. (2019). How AI can change/improve/influence music composition, performance and education: Three case studies. *INSAM Journal of Contemporary Music, Art and Technology*, 2, 100–114 <https://insamjournal.com/index.php/ij/issue/view/issue-2/11>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Yunus Kökver¹  · Hüseyin Miraç Pektaş² · Harun Çelik² 

✉ Yunus Kökver
ykokver@ankara.edu.tr

Harun Çelik
haruncelik@kku.edu.tr

¹ Computer Technologies Department, Ankara University Elmadag Vocational School, Ankara, Turkey

² Kırıkkale University Faculty of Education, Mathematics and Science Education, Kırıkkale, Turkey