



OPEN

University english teaching evaluation using artificial intelligence and data mining technology

Qiuyang Huang¹, Wenling Li⁴✉, Mohd Mokhtar bin Muhamad², Nur Raihan binti Che Nawi² & Xutao Liu³

This work intends to drive reform and innovation in English teaching evaluation and support personalized English instruction. It utilizes deep learning (DL) and artificial intelligence (AI)-driven data mining technology to explore a reliable and efficient method for university English teaching evaluation. By employing DL, this work explores innovative English teaching models and introduces a Bayesian framework to enable personalized teaching strategies. In the data mining process, the Transformer architecture is applied to English teaching evaluations. This capitalizes on its powerful feature extraction and sequence modeling capabilities to gain a comprehensive understanding and precise evaluation of students' English proficiency. Additionally, an AI-based method for English teaching evaluation is proposed. Data from the English teaching and evaluation system for Computer Science students in the 2018 class at Tianjin University of Science and Technology are collected, analyzed, and processed. Group profiles of students are created to predict exam outcomes. The findings show that over 70% of students engage in active English learning only occasionally, with a higher proportion among females. More than 80% of males recognize the importance of listening and speaking skills, a sentiment shared by over 90% of female students. In terms of factors influencing students' passing exams, scores in various question types play a central role, significantly impacting final grades. These scores reflect students' mastery of English knowledge and application abilities. This work applies the Transformer architecture from natural language processing to the education domain, achieving interdisciplinary integration and innovation. This cross-disciplinary approach not only enriches teaching assessment methods but also provides new solutions for broader educational challenges. The proposed method enhances the objectivity and accuracy of teaching evaluation, minimizing the influence of human bias assessment results.

Keywords Deep learning, Artificial intelligence, Data mining, University english, Teaching assessment, Transformer architecture

Research background and motivations

In higher education, English teaching is considered a crucial subject, and its teaching quality and assessment effectiveness have always been a focus of attention^{1–3}. Traditional methods for evaluating English instruction face numerous challenges, including a narrow assessment approach, unscientific evaluation criteria, and incomplete data, all of which limit the potential to enhance teaching quality. With the advancements in information technology, the integration of big data and artificial intelligence (AI) has introduced new perspectives and methodologies for English teaching evaluation. As a key branch of AI, deep learning (DL) offers powerful data analysis and pattern recognition capabilities, offering robust support for reforming and innovating English teaching assessment^{4–6}.

Traditional English teaching assessment relies on teachers' subjective judgment and standardized tests, which often leads to limited evaluation dimensions and delayed feedback. Manual scoring requires substantial time and fails to capture students' subtle progress. Standardized testing focuses on outcomes while neglecting the

¹School of Economics and Management, Jiangxi Arts & Ceramics Technology Institute, Jingdezhen 333499, China.

²Faculty of Educational Studies, Universiti Putra Malaysia, 43300 Selangor, Malaysia. ³School of Physical Education, Jiangsu University of Science and Technology, Zhenjiang 212100, China. ⁴School of Education Science, Guangxi Minzu Normal University, Chongzuo 532200, China. ✉email: liwenling@gxnu.edu.cn

accumulation of process-based data. The introduction of DL and data mining technologies can address these limitations through automated scoring, multidimensional data analysis, and real-time feedback mechanisms. However, current research still lacks in algorithm integration and adaptability of assessment models, which is the primary focus of this study. DL enables personalized learning by recommending resources and activities suited to students' habits, abilities, and needs. Such personalization is essential for fostering autonomous learning and improving English language competence. Data mining refers to the process of extracting valuable information and knowledge from large datasets, including clustering analysis, association rule mining, classification, and prediction^{7–9}. In the context of college English assessment, data mining can be used to analyze student behavior and evaluate teaching quality. This study aims to construct an efficient and equitable assessment framework by integrating DL with data mining technologies to meet the practical needs of large-scale educational settings.

Research gap

Most existing work focuses on the isolated application of data mining techniques, overlooking the collaborative potential of combining architectures like Transformer with Bayesian frameworks to achieve personalized learning. Previous studies usually aggregate student data without analyzing gender differences in learning preferences and skill priorities. As a result, the lack of transparent model architectures prevents educators from auditing the contribution of different functions. This study addresses the gap through interpretable decision rules and aims to fill these deficiencies by using a transformer–Bayesian integrated framework and combining behavioral sequences with multi-source assessment data.

Research objectives

This work aims to advance the reform and innovation of English teaching assessment and achieve personalized English instruction through the application of DL and AI-driven data mining techniques. By analyzing and summarizing the current state of relevant research, this work explores the role of DL in English teaching models and proposes a Bayesian approach to personalized instruction. Simultaneously, utilizing AI data mining techniques, a new English teaching assessment method is introduced to enhance teaching quality and assessment effectiveness.

With the continuous development of data mining and DL technologies, their application prospects in college English teaching evaluation are becoming increasingly broad. This work aims to explore the practical application of these technologies in teaching, and provide new ideas and methods for college English education. The specific research questions include: How can students' learning behavior data be analyzed to predict their exam scores? Which factors have the most significant impact on the development of students' English proficiency? Section 1 describes the research background, objectives, and necessity. Section 2 summarizes the current research status of DL and AI data mining in English teaching. Section 3 primarily focuses on exploring English teaching models based on DL, proposing the Bayesian method for personalized English instruction, and introducing an English teaching assessment method based on AI data mining. Section 4 focuses on experimental analysis, including dataset processing and result prediction. Section 5 concludes contributions and limitations, and suggests future research directions. This work successfully applies the Transformer architecture, originally from the field of natural language processing (NLP), to the education sector. This interdisciplinary integration not only expands the application scope of the Transformer architecture but also introduces novel methods for data processing and evaluation in education. By leveraging the powerful feature extraction and sequence modeling capabilities of the Transformer, this approach enables a deeper understanding of students' learning behaviors, facilitating more accurate personalized teaching assessments.

Literature review

In the field of language education, particularly English teaching, the rapid development of DL and AI technologies has brought the quality of instruction and assessment to the forefront of attention in both academic and practical domains. This section focuses on the application and impact of DL/AI in language education. Vaswani et al. (2017)¹⁰ proposed a novel and simple network architecture that was entirely based on the attention mechanism, completely eliminating the need for recursion and convolution. Numerous studies have highlighted the significance of English teaching in higher education. Eke et al. (2021)¹¹ pointed out that English teaching not only enhanced students' cross-cultural communication skills but also played a crucial role in fostering a global perspective and competitiveness. Similarly, Lu & Vivekananda (2023)¹² emphasized the importance of English teaching quality in improving students' overall competencies and employability. These studies provide a solid theoretical foundation for exploring the application of DL in language education. However, traditional teaching methods have limitations when it comes to assessing students' English proficiency. To address these limitations, Peng et al. (2022)¹³ began exploring the application of DL/AI in language education. Their research revealed that by leveraging big data and machine learning algorithms, student proficiency in English could be assessed more accurately, enabling the provision of more personalized teaching strategies. This discovery offers new insights and methods for innovation in language education.

In discussing how DL technology enhances English teaching and supports personalized learning, the focus should be its applications in NLP, learning resource recommendation, and sentiment analysis. DL technology can comprehend and process language information more deeply by mimicking the human brain's information processing. This capability is crucial for enhancing the quality and efficiency of English teaching. Next, the applications and effectiveness of DL technology in these areas are specifically analyzed. Zhu et al. (2021)¹⁴ utilized DL's NLP techniques to design a system that thoroughly understands and processes students' language inputs. This system enables more accurate semantic analysis and grammar correction, thereby improving students' language application skills. Additionally, the system can assess students' writing in real time and provide targeted suggestions for grammar and vocabulary improvements, fundamentally enhancing students' writing

proficiency. Concurrently, Agüero-Torales et al. (2021)¹⁵ proposed using DL recommendation algorithms to analyze students' learning journeys, preferences, and performance. This method accurately recommends the most suitable learning resources and activities, further achieving the goal of personalized learning. Zhong et al. (2020)¹⁶ applied DL to sentiment analysis, and designed a system capable of interpreting students' emotional feedback on learning materials. This sentiment analysis system enables teachers to adapt their strategies to meet students' individual needs and quickly identifies learning challenges or areas of interest, facilitating targeted intervention and support. The application of DL technology in English teaching is multifaceted. This not only enhances students' language proficiency but also provides tailored recommendations for learning resources based on their progress and preferences. Additionally, the sentiment analysis provides teachers with deeper insights into students' needs, allowing for more customized instruction. These advancements not only elevate the quality of personalized teaching but also equip educators with diverse strategies and methods, enhancing teaching efficiency and precision. As DL technology continues to evolve, it is reasonable to expect an even more significant impact on English education, enriching students' learning experiences and providing higher-quality educational outcomes.

In modern society, the rapid advancement of big data and AI technology is transforming the educational landscape, especially in the assessment of English teaching in higher education. These innovations have redefined traditional teaching models, introducing new perspectives and methodologies for evaluation. To enhance the quality of education and the effectiveness of assessments, researchers are increasingly integrating these technologies into English teaching evaluation, striving for improved educational outcomes. Dang et al. (2020)¹⁷ proposed a DL-based automatic scoring model for English composition, effectively enhancing scoring efficiency and accuracy. Kumar (2020)¹⁸ utilized association rule mining techniques to analyze students' English scores, discovering potential teaching issues and trends. In terms of research methods, researchers have utilized various approaches and technological tools to validate the effectiveness of their research hypotheses and models. For instance, Mostafa and Benabbou (2020)¹⁹ conducted experiments comparing traditional grading methods with DL-based grading methods, gaining valuable insights into students' and teachers' perceptions and experiences with AI-based teaching assessment. The findings indicated a greater acceptance of advanced teaching assessment methods among both educators and learners.

With the advancement of computer technology and AI, intelligent assessment has made significant progress in evaluating students' abilities and knowledge levels. Kartika et al. (2023)²⁰ argued that intelligent assessment could diagnose and analyze students' specific subject knowledge and competencies based on extensive longitudinal learning data, enabling modeling and dynamic analysis of the learning process. Tang et al. (2022)²¹ mentioned methods such as simulation-based and game-based assessments, which provided students with complete, authentic, and open problem contexts, allowing them to explore and express themselves freely within task scenarios. Maghsudi et al. (2021)²² employed DL technology to develop student models and learning resource recommendation systems, offering personalized learning paths and resources tailored to students' learning conditions and interests.

These studies demonstrate the tremendous potential and actual effectiveness of big data and AI technologies in English teaching assessment. These technologies not only enhance the efficiency and accuracy of scoring but also facilitate the identification of teaching issues and trends, thereby providing evidence for teaching improvements. Furthermore, comparative experiments with traditional assessment methods enable researchers and educators to intuitively grasp the benefits of AI-driven assessment approaches, fostering innovation and development in educational technology.

Research model

The framework proposed combines the sequence modeling capabilities of the Transformer architecture with Bayesian inference for personalized learning. The Transformer architecture uses a self-attention mechanism to capture temporal dependencies in learning behaviors, modeling the "learning input-performance output" relationship through an encoder-decoder structure. A Monte Carlo dropout approximation is introduced to model uncertainty, and Bayesian posterior distributions are used to integrate multimodal assessment data, enhancing the robustness of predictions.

Personalized english teaching in college based on DL

The central concept of personalized learning is to address the individual differences that arise during the learning process. Each learner possesses distinct characteristics, such as unique learning preferences and specific areas of weakness. Consequently, personalized learning adopts a customized approach, providing tailored learning advice and methods based on each user's learning context. This individualized method aims not only to help students bridge knowledge gaps but also to cultivate curiosity and confidence in unfamiliar subjects.

In implementing personalized learning, various tools and methods can be employed, such as interactive ability tests, to comprehensively assess learners' learning processes and gain a clearer understanding of their strengths and weaknesses^{23,24}. Internet-based education platforms provide a convenient space for personalized learning, enabling students to select appropriate resources based on their needs and learning pace. Meanwhile, effective learning strategies are essential in this process. Comprehensive assessments, pre-class quizzes, and post-class quizzes can enhance students' deep understanding and mastery of knowledge, making personalized learning an efficient and effective educational model.

In applying Bayesian statistics based on DL in education, the fundamental concept is to treat a student's learning efficiency as a probability distribution that evolves over time^{25–27}. By meticulously observing and recording each instance of a student's learning activity, the system can continuously update the assessment of the student's learning efficiency, ensuring it aligns more closely with their actual circumstances. This personalized approach to evaluating learning efficiency effectively addresses individual differences among students, enabling

more targeted recommendations for learning strategies and resources. According to Bayesian theorem, the posterior distribution of a student's learning efficiency can be represented as:

$$P(\theta|D) = \frac{P(D|\theta) \cdot P(\theta)}{P(D)} \quad (1)$$

θ represents the student's learning efficiency, and D is all observed learning data. $P(\theta|D)$ is the posterior probability of θ given data D ; $P(D|\theta)$ is the probability (likelihood) of observing data D given learning efficiency θ ; $P(\theta)$ is the prior distribution of student learning efficiency, that is, the efficiency before observing any data; $P(D)$ is the marginal probability of data D .

In mathematical statistics, the likelihood function is a function of the parameters of a statistical model that expresses the likelihood of those parameters. The likelihood function is crucial for statistical inference, including applications such as maximum likelihood estimation and Fisher information. While the terms “likelihood,” “plausibility,” and “probability” are closely related and all refer to the chance of an event occurring, they have distinct meanings in a statistical context.

Probability is employed to predict the outcomes of future observations given certain parameters are known, while likelihood is utilized to estimate the parameters of interest given certain observed outcomes. Specifically, given the output x , the likelihood function $L(\theta|x)$ about parameter θ (numerically) equals the probability of variable $X = x$ given parameter θ :

$$L(\theta|x) = P(X = x|\theta) \quad (2)$$

To evaluate the likelihood function for parameter θ , it is numerically equivalent to the conditional probability of observing result X given parameter θ , which is also known as the posterior probability of X . Generally, a higher value of the likelihood function indicates that parameter θ is more plausible given the outcome $X = x$. Therefore, formally, the likelihood function is a type of conditional probability function, but with a shift in focus: this work is interested in the likelihood value for A taking the parameter θ .

$$\theta \langle - \rangle P(B|A = \theta) \quad (3)$$

Through interactive guidance, students can personally experience their progress in each English learning session and receive the system's positive feedback on different learning states. This personalized feedback mechanism enhances students' understanding of their performance in the course, fostering enthusiasm and confidence in their learning journey.

This work primarily investigates the design of effective guidance strategies to help diverse students complete their coursework more efficiently and address their personalized learning needs. Figure 1 illustrates the comprehensive recommendation strategy for English teaching. The system leverages integrated recommendations and indices to offer content suggestions that align more closely with users' learning requirements, thereby accelerating the learning process^{28–30}. Additionally, it emphasizes comfort zone guidance as a recommendation approach, enabling users to gradually expand their learning comfort zones, which facilitates a smoother growth trajectory. Ultimately, by fostering situational learning interests, the system can gain a holistic understanding of users' learning needs and provide a more engaging and personalized educational experience. The integrated application of interactive guidance and guidance strategies seeks to stimulate students' positive attitudes towards learning, enhance their enthusiasm for learning, and ultimately achieve more effective learning outcomes^{31,32}. In the Bayesian framework, determining and updating the prior distribution are key steps. This work employs a statistical analysis method based on historical data to determine the prior distribution and updates it in real-time to accommodate students' learning progress. This dynamic update mechanism ensures that the model accurately reflects the students' true learning situation. Whenever new learning data is generated, the posterior distribution is recalculated based on Bayes' theorem and used as the prior distribution for the next prediction.

Process of AI data mining

Data mining is the process of uncovering hidden information within large datasets using algorithms. This process, often associated with computer science, employs various methods such as statistics, online analytical processing, information retrieval, machine learning, expert systems, and pattern recognition^{33–35}. The data mining process includes steps like data preparation, integration, selection, and preprocessing. Its primary goal is to discover knowledge that is relevant, understandable, applicable, and supportive of specific application discovery challenges.

In data mining, the decision tree algorithm is a supervised learning algorithm utilized for classification and regression problems^{36–38}. It constructs a decision tree by recursively dividing the dataset into smaller subsets. In this construction, each node represents a test condition for a feature attribute, and each branch corresponds to the outcome of that feature attribute in a specific value range. Each leaf node stores a category or a specific numerical value. This work adopts a feature selection method based on information gain. Specifically, the information gain of each feature relative to the target variable is calculated, and the top N features with the highest information gain are selected as model inputs. During the construction of the decision tree, the Gini index is used as the splitting criterion. A smaller Gini index indicates higher node purity. The decision tree is built by recursively selecting the optimal splitting feature and splitting point. Table 1 shows the advantages of the decision tree method compared to neural networks and support vector machines.

Microsoft's decision tree algorithm constructs efficient data mining models by carefully designing a series of splits in the tree structure. Figure 2 is the structure of the decision tree model. Whenever the algorithm

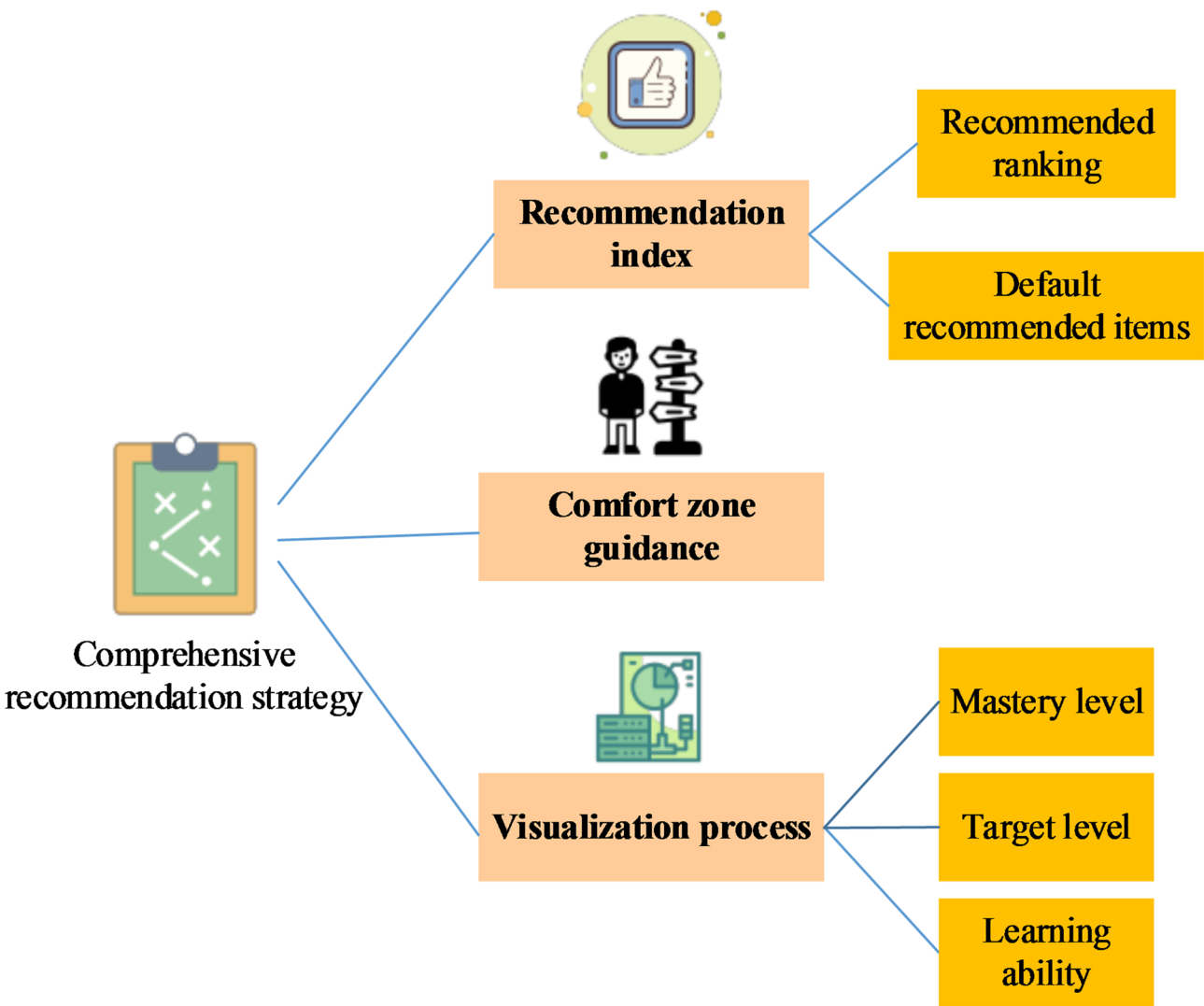


Fig. 1. Comprehensive recommendation strategy in English teaching.

Method	Prediction accuracy	Interpretability	Data requirements	Applicable scenarios
Decision tree	★★★★	★★★★★	Medium to small samples	Educational assessment, policy making
Neural network	★★★★★	★	Large samples	Image recognition, complex pattern mining
Support vector machine	★★★★	★★★	Medium samples	Text classification, non-linear relationships

Table 1. Comparison of interpretability and adaptation scenarios for different methods.

identifies a significant correlation between a column in the input data and the predictable column, it introduces a new node into the model. The construction process of the Microsoft decision tree begins by treating the entire dataset as a single node. The goal is to partition the sample set based on a particular attribute to purify the class distribution in the subnodes as much as possible. This process relies on selecting the optimal split point to either maximize information gain or minimize Gini impurity, thus achieving effective classification of the samples. The partitioning stops when all samples belong to the same class or when the preset maximum depth is reached; otherwise, the best split point is chosen for further partitioning. Ultimately, the class of each leaf node is determined based on the principle of “the majority rules.” This process not only helps to understand students’ learning behavior patterns but also provides strong support for personalized teaching. In terms of feature selection, the Microsoft decision tree mainly considers the frequency of feature occurrences and their proximity to the root node. Features that frequently appear and are closer to the root are generally considered to have a greater impact on classification results. Here, different question type scores and exam performance are identified as key factors influencing students’ success in exams, and the position of these features in the decision tree reflects their importance. Moreover, information gain, as one of the feature selection criteria, measures the

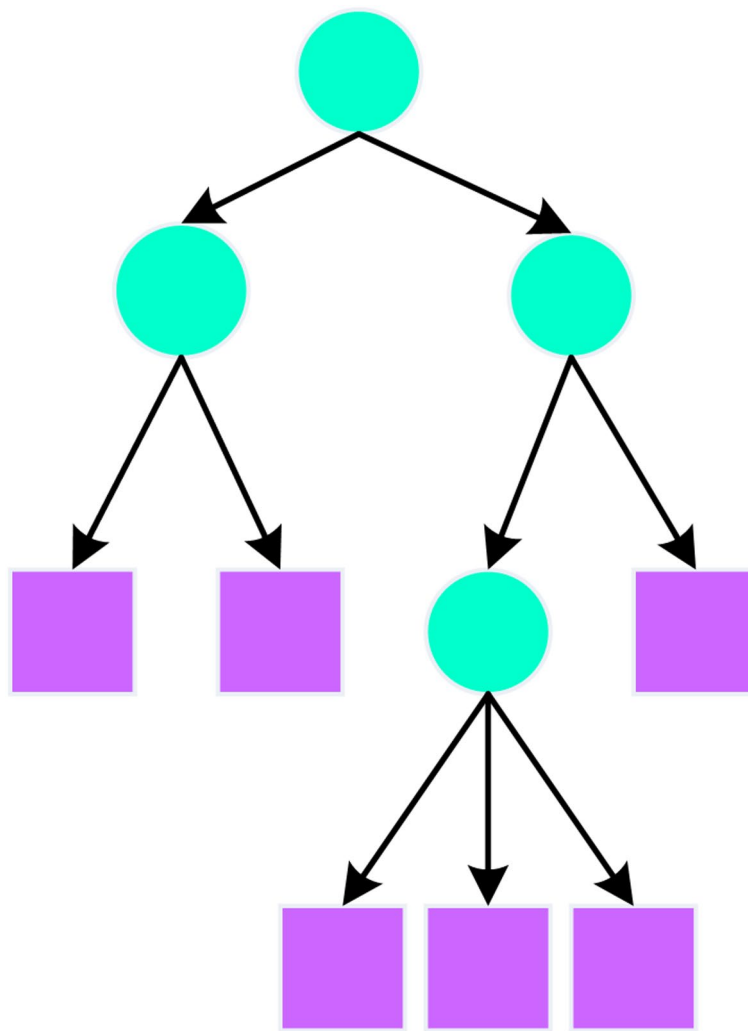


Fig. 2. The structure of the decision tree model.

extent to which a particular attribute enhances classification purity, while Gini impurity is adopted to quantify the probability of elements within a set being misclassified. These two methods work together to ensure that the model effectively captures the key information in the data.

The Microsoft decision tree algorithm uses “feature selection” to determine the most valuable attributes. These techniques help prevent irrelevant attributes from consuming processor time, thus improving performance, and enhancing the quality of analysis. Figure 3 illustrates the decision tree building process. This algorithm is primarily utilized in data mining and machine learning, assisting users in extracting useful information and knowledge from large datasets. By building decision tree models, users can uncover associations and patterns within the data, applying these insights to practical challenges such as fraud detection and credit scoring^{39,40}.

The dataset used in this study contains textual features, numerical features, and categorical labels, displaying characteristics of high-dimensional mixed data. The decision tree algorithm is selected as the preferred method because it requires no complex preprocessing and can directly handle both categorical and continuous variables. The dataset includes 1,500 samples with a moderate number of features, allowing the decision tree to strike a good balance between computational efficiency and model interpretability. In contrast, neural networks, although capable of capturing complex patterns, are not the optimal choice in this context due to their black-box nature and high computational cost.

In the initial stages of data mining, data preprocessing is a crucial step that encompasses data cleaning, transformation, feature extraction, and more. For textual data such as student assignments and classroom discussions in English teaching evaluations, the Transformer architecture can handle such unstructured data. Through preprocessing techniques like tokenization, stop-word removal, and word embedding, textual data are converted into numerical data that models can process. The Transformer architecture, a significant advancement in recent years in NLP, excels in capturing long-range dependencies within sequences and handling complex textual data, offering fresh insights into English teaching evaluations. This work builds on the foundations of data mining and utilizes the Transformer architecture in English teaching evaluations. It leverages the model’s powerful feature extraction and sequence modeling capabilities to achieve a comprehensive understanding and

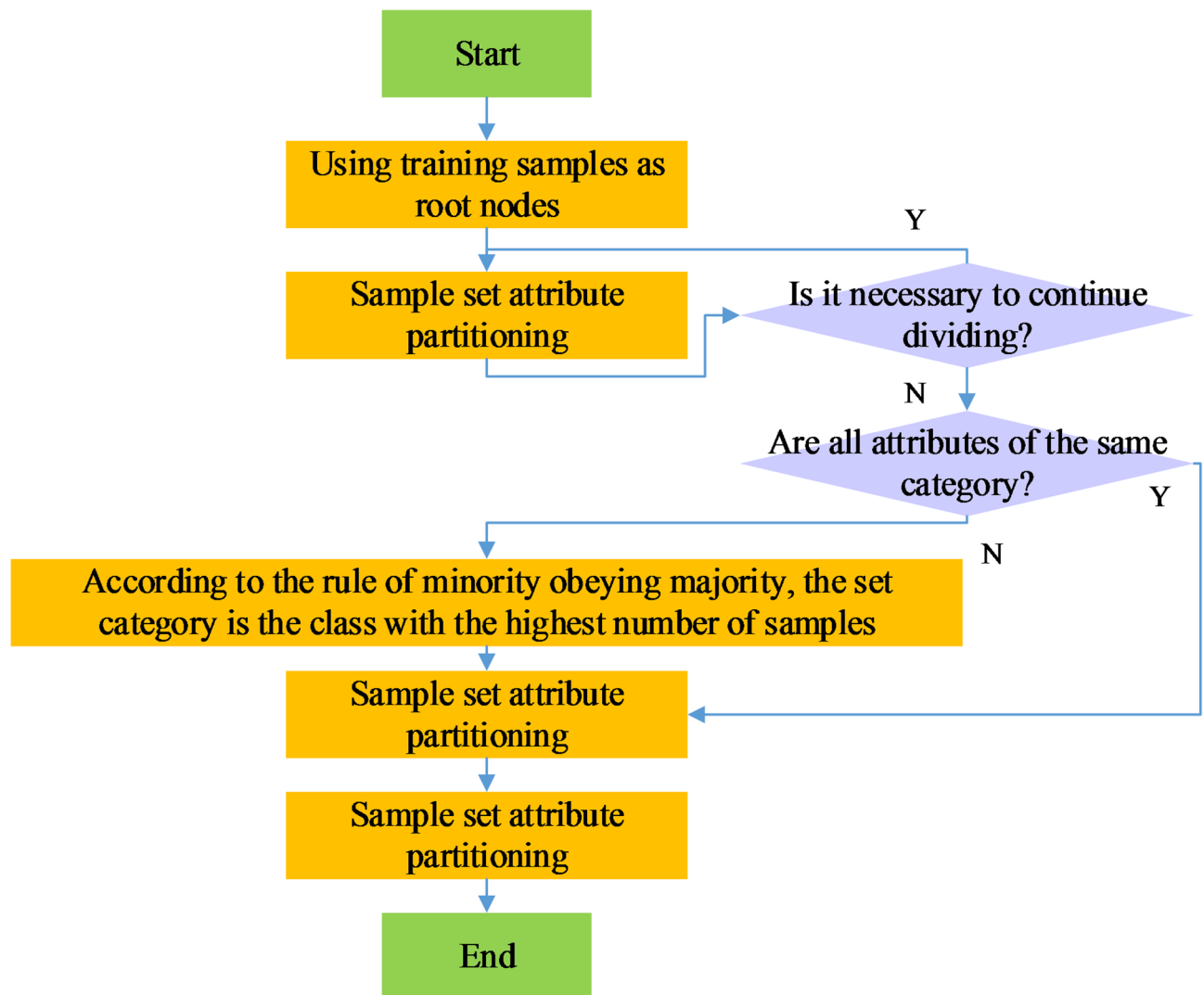


Fig. 3. The decision tree building process in Microsoft decision tree algorithm.

accurate assessment of students' English learning experiences. The Transformer architecture employs a multi-head attention mechanism, specifically with 8 attention heads, each having a dimension of 64. This configuration helps the model capture richer information while maintaining computational efficiency. The model is trained using the Adam optimizer, with a learning rate set at 0.001. To prevent overfitting, dropout technology is adopted, with a rate of 0.2. During training, the loss function and accuracy on the validation set are monitored, allowing for adjustments to the learning rate and early stopping strategy as needed.

A multi-layer encoder-decoder model is constructed based on the Transformer architecture. The encoder transforms input textual data into a series of feature vectors, while the decoder generates evaluations of student learning based on these feature vectors. During model construction, parameters and structures are adjusted to meet specific requirements, optimizing performance. The self-attention mechanism, which is central to the Transformer model, captures dependencies between any two positions within the sequence. Attention weight calculation reads:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (4)$$

In this context, Q represents the query matrix, K is the key matrix, and V is the value matrix. d_k refers to the dimensionality of the key vectors, used to scale the dot product to mitigate the issues of gradient vanishing or exploding.

Based on the evaluation results from the Transformer model, personalized learning recommendations and teaching strategies can be tailored for each student. For instance, students who struggle with grammar could be provided with additional exercises and explanations, while those with weak oral expression skills might receive resources for oral practice and pronunciation correction. By employing personalized assessment strategies, educators can more effectively address students' learning needs and improve overall teaching effectiveness. In

the specific implementation of the Transformer architecture, particular attention is given to the configuration of the attention heads and the optimization of the training process. By carefully configuring the attention heads and adopting effective training strategies, the model is able to efficiently process complex text data and extract valuable information for college English teaching evaluation.

DL path in english teaching

Combining Eric Jensen and LeAnn Nickelsen's DL path (Fig. 4), teachers should first clarify teaching goals and expectations. They should ensure that students not only master fundamental knowledge and skills in English but also develop cross-cultural awareness and self-directed learning abilities.

This work develops a DL model tailored to the characteristics of college English teaching, combining the strengths of Bayesian methods and the Transformer architecture to achieve personalized teaching assessments. The model's input includes multidimensional data such as students' learning records, assignment completion rates, classroom interaction data, and online learning behaviors. After preprocessing, these data are transformed into feature vectors for model input. In the initial phase, a Bayesian network layer is employed to handle uncertainty. This network infers students' future learning potential and possible obstacles based on existing learning data and historical performance. Its purpose is to provide a probability-based preliminary assessment for the subsequent Transformer architecture, aiding the model in understanding individual differences among students. Following this, the Transformer architecture processes students' time series data. Comprising an encoder and a decoder, the encoder extracts features from the input data, while the decoder generates outputs based on these features. In this model, the encoder handles students' learning records and related data, utilizing a self-attention mechanism to capture complex relationships within the data. The decoder then produces personalized teaching assessment results and predictions based on the encoder's output. In the pre-assessment stage, teachers must evaluate students' English proficiency and abilities to establish their starting points and identify areas needing attention. This assessment enables teachers to pinpoint students' confusions and challenges in English learning, thereby facilitating the development of targeted teaching strategies.

In order to create an environment conducive to DL, teachers should offer a diverse array of learning materials and resources, including textbooks, online resources, and multimedia materials^{41–43}. These resources not only broaden students' knowledge but also enhance their interest and motivation to learn. Additionally, teachers need to create authentic contexts through situational simulations, role-playing, and similar methods, enabling students to connect English with real-life situations and develop cross-cultural communication skills.

In activating students' prior knowledge, teachers can employ techniques such as questioning, discussion, or providing relevant cases to help students recall and associate previously learned content. This process facilitates the integration of new knowledge with existing understanding, leading to deeper comprehension and meaning^{44–46}.

Acquiring new knowledge is a crucial step in DL. Teachers must explain, demonstrate, and guide students in mastering new vocabulary, grammar, and expressions. In the deep processing stage, they should guide students in summarizing and organizing what they have learned to build a cohesive knowledge system. Additionally, teachers should design projects or tasks that allow students to apply English in real-life situations, fostering their innovative thinking and problem-solving skills⁴⁷. These activities not only aid in the internalization of knowledge but also enhance students' language proficiency and cross-cultural communication skills. In this study, students are divided into an interactive group ($n = 145$) and a non-interactive group ($n = 145$). The former uses an AI-driven real-time feedback system, while the latter relies on traditional instruction. Classroom activities in the interactive group include NLP-based instant essay correction and personalized learning path recommendations. In contrast, the non-interactive group only receives uniform explanations from the teacher.

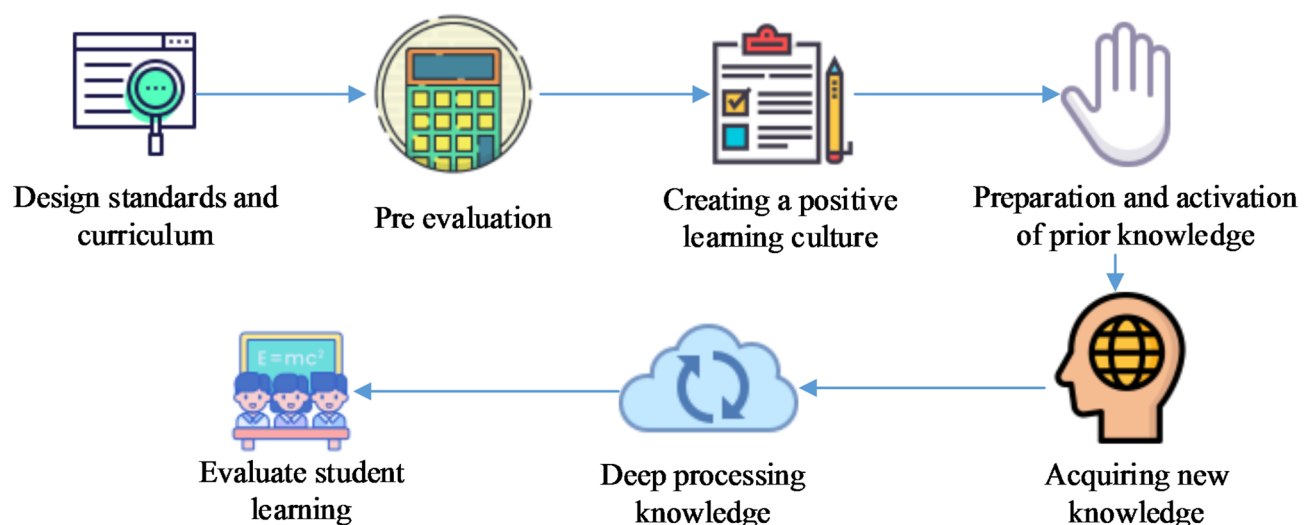


Fig. 4. The roadmap for students' DL in English teaching.

Image information	Description
Basic information	Age range/gender ratio/geographic distribution/school type
Learning characteristics	Learning habits/learning preferences/learning mode (online/offline)/learning efficiency
Interests and hobbies	Areas of interest/leisure activities/social media usage habits
Skills and talents	Language proficiency/computer skills/other skills and hobbies
Psychological features	Personality traits/self-awareness/values orientation
Family background	Family economic status/family cultural background/family educational beliefs
Social relationships	Friendship circle/interpersonal relationship status/role in the group (leader/follower)

Table 2. The main contents of the student group profile.

Variable	Distribution proportion/mean value
Gender (male/female)	52%/48%
Age	16.2 ± 1.5 years old
Urban and rural distribution	78% urban students, 22% rural students.
Teaching grade	Senior one (60%) and senior two (40%)

Table 3. Demographic characteristics of students.

Experimental design and performance evaluation
Datasets collection

Constructing a student group profile has become a crucial technology in the field of education. By analyzing and processing extensive data, educators can gain deep insights into students’ learning habits, behavioral characteristics, and specific needs. This information serves as a strong foundation for developing effective teaching strategies. Table 2 outlines the key aspects of the student group profile of this work. Table 3 presents the statistics of students’ demographic characteristics. A total of 78% of the samples are from urban areas, which may affect the model’s generalization ability to rural student data. Urban students have access to richer extracurricular learning resources than rural students. Therefore, a weighting strategy needs to be introduced during model training to mitigate regional bias. This study aims to implement personalized learning assessment by constructing user profiles rather than adopting a one-size-fits-all teaching approach. By analyzing specific data from each student, including their learning styles and habits, the goal is to provide tailor-made instructional strategies that meet individual needs.

Student attitudes are jointly calculated using a Likert-scale questionnaire (1–5 points) and NLP-based sentiment analysis. The questionnaire, designed by an educational psychology team, measured interest in English learning and self-efficacy. The NLP model extracts emotional keywords from students’ text (e.g., “interesting,” “difficult”) and generated sentiment scores through dictionary matching and contextual analysis, which are then weighted to produce a composite attitude index. The improvement in classroom engagement is quantified using a multi-source data fusion approach. Specifically, classroom videos are used to record the number of voluntary student responses and interactive behaviors during group discussions, cross-validated with teacher subjective scores. Video data are captured simultaneously by three fixed cameras covering all students, with two typical classes selected each week for analysis. Counts of voluntary responses are obtained using frame-level annotation, where two research assistants independently mark the start and end times of student speaking turns. Intercooder consistency is ensured with a Kappa coefficient of 0.87. Additionally, contribution scores during group discussions are assigned by teachers based on the number of ideas presented, logical coherence, and responsiveness to others, using a 1–5 scale. Finally, overall engagement is calculated using a weighted method, with voluntary response counts accounting for 60% and teacher scores for 40%.

The dataset used here is the “Tianjin University of Science and Technology 2018 Computer Science Undergraduate English Learning Process Dataset,” which includes records of various learning activities of students from their enrollment to graduation. The system meticulously records students’ actions during English practice sessions or exams upon logging into the system. Data collection encompasses not only basic information but also students’ behaviors during the answering process and their feedback post-assessment, involving approximately 290 students. These data are categorized into static and dynamic types. This work aims to comprehensively understand individual characteristics, learning styles, and other significant information by constructing user profiles. The selection of the 2018 cohort is based on the dataset’s ability to provide a complete record of the entire learning cycle, essential for in-depth analysis of students’ English proficiency development. This dataset spans from freshman enrollment to graduation, reflecting learning habits, progress, and potential challenges. It offers researchers a valuable opportunity to evaluate the long-term impacts of teaching methods, learning strategies, and assessment mechanisms on learning outcomes. In contrast, data from students in the 2019 or 2020 cohorts may be incomplete due to shorter collection periods, leading to fragmented research results. Moreover, changes in teaching methods and assessment standards over time could affect data consistency and comparability. Thus, the 2018 cohort, with its completeness and continuity, is an ideal choice for conducting such long-term, in-depth educational research.

Although the sample may not represent all Chinese university students, the findings and conclusions are still significant for understanding the learning habits, behavioral characteristics, and needs of this specific group. The students selected for this study are chosen because they have complete learning records and assessment data. Additionally, factors such as age, gender, and study habits are considered to ensure the diversity and representativeness of the sample. Through rigorous screening and statistical analysis, it is ensured that the selected sample accurately reflects the learning situation of students in the computer science group. Throughout the data collection and processing stages, strict adherence to relevant laws, regulations, and ethical standards is maintained. Anonymization techniques are used to protect students' privacy and data security. The data undergo thorough cleaning and preprocessing to ensure their accuracy and consistency. The data preprocessing stage includes handling missing values, detecting and processing outliers, and feature scaling. Missing values are addressed using mean imputation; outliers are removed based on the IQR rule. Finally, numerical features are scaled using standardization to make them align with a normal distribution.

For self-feedback and classroom interaction data, NLP techniques are applied to extract textual features. Through tokenization, stop-word removal, and frequency analysis, keywords and phrases reflecting students' learning attitudes and interests are identified. In addition, time-series analysis is used to extract temporal features from students' online learning logs. By calculating statistics such as online study duration, frequency, and intensity, this approach captured features representing students' learning habits and rhythms. The NLP methods are mainly used to process structured text data, such as student essays and classroom discussion transcripts, rather than directly analyzing audio or video data. During classroom discussions, audio from video recordings is first transcribed into text using speech recognition tools, after which NLP models extract keywords, sentiment tendencies, and semantic coherence indices. This workflow ensures standardized processing of textual data while avoiding interference from the high-dimensional noise of raw audio-visual files.

Experimental environment

Experiments are conducted on the Windows 10 operating system, and the MySQL database management system is chosen, installed, configured, and managed as needed. Each data sample is considered a record composed of multiple fields, separated by specific delimiters. In this record, the first field is the target variable, indicating whether a student's score is passing. The values for this field are "1" or "0," with "1" representing a passing score and "0" representing a failing score. The fields from the second to the last can be viewed as a collection of key-value pairs. Here, the key represents the index number of the feature values in the sample data, while the value is consistently set to "1." These key-value pairs are also separated by specific delimiters. Additionally, the training set data are stored in a file named "train_txt." Fig. 5 illustrates its format.

According to the ethical guidelines on AI set by UNESCO, all student data are anonymized using hash algorithms to protect privacy. Participants provide written consent for the use of their data, and the study is approved by the Ethics Committee of Universiti Putra Malaysia. The dataset is balanced for gender representation (50% male, and 50% female) to minimize algorithmic bias.

Parameters setting

Data processing is performed on TensorFlow GPU, with the graphics card configured as GeForce RTX 3090. The specific environment for running the model framework includes python = 3.8 and pytorch = 1.7.3.

Performance evaluation

Student group profile analysis

Figures 6 and 7 depict students' attitudes toward learning English and their perspectives on the most crucial skills in English. It is notable that over 70% of students only occasionally actively seek opportunities to learn English, with a higher proportion of females than males in this regard. About 2% of males choose alternative options, with some students only engaging in reading without practicing speaking, while others feel they lack an ideal learning environment. In contrast, less than 1% of females chose alternative options. Overall, it is evident that students generally lack initiative in learning English, with the majority only occasionally seeking learning opportunities.

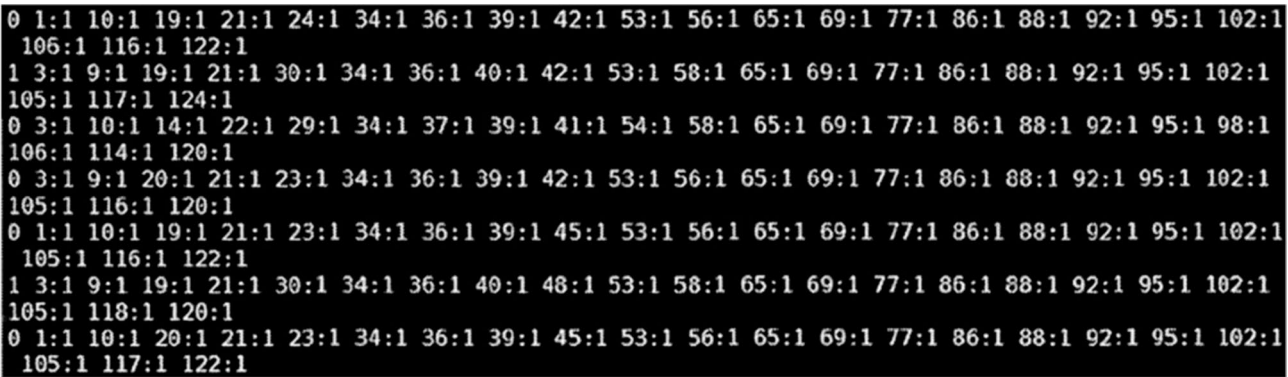


Fig. 5. The format of the training set data.

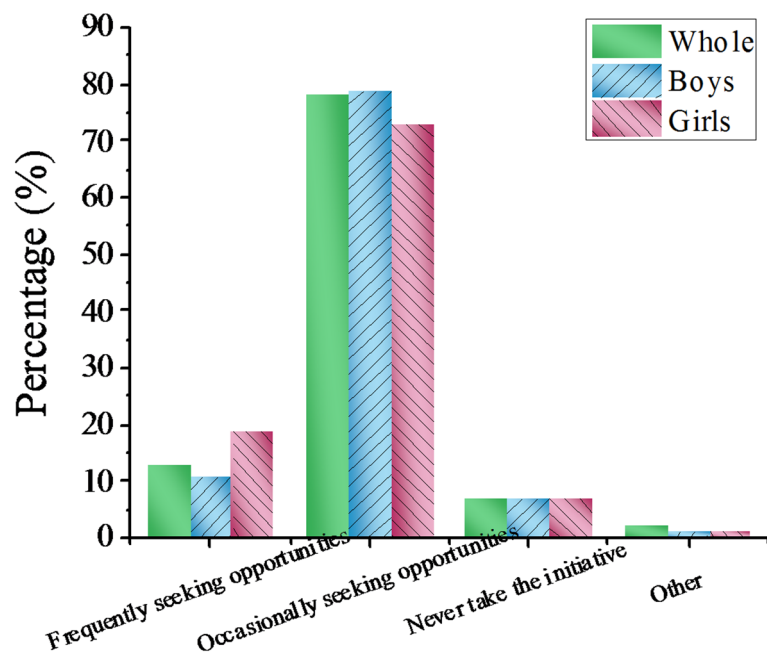


Fig. 6. Students' attitudes toward learning English.

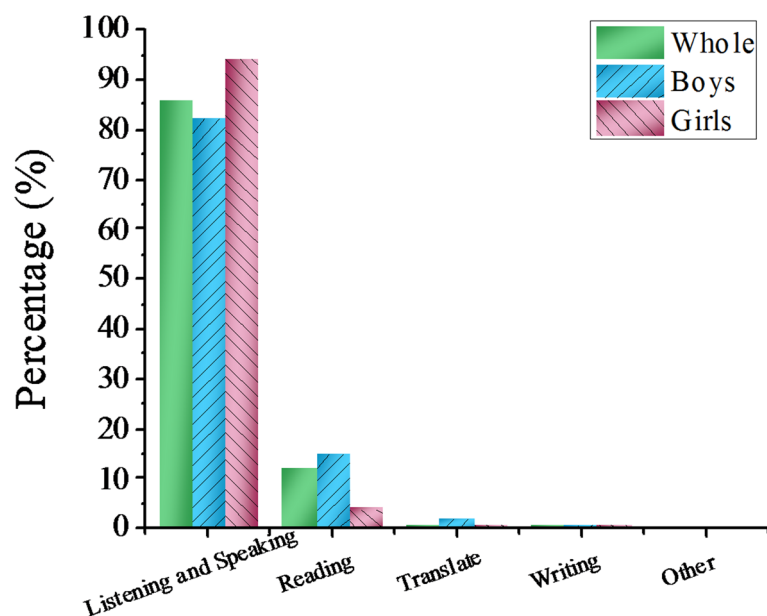


Fig. 7. Students' perspectives on the most important skills in English.

Further data analysis reveals subtle differences between males and females in English learning. Over 80% of males consider listening and speaking skills important, while this proportion exceeds 90% for females. However, in terms of reading, males show a significantly higher level of agreement than females, with very few students considering translation skills important. Notably, the proportion of males who view writing skills as important is 0%, indicating that almost no males prioritize writing in their English learning.

Table 4 illustrates the long-term impacts of teaching methods, learning strategies, and evaluation mechanisms on student learning outcomes. The calculation of extracurricular study time is based on students' weekly learning logs and an AI-assisted verification mechanism. Specifically, every Friday students are required to complete a "Learning Record Form", detailing the type of English learning activity (e.g., reading, writing, listening practice), duration, and completion status. The teacher team and AI system, which incorporate rule-based matching and natural language understanding models, jointly review the logs to exclude invalid entries (such as duplicate inputs or vague descriptions), and finally use the average of the valid data. For example, if a student record

Impact factor	Implementation period	Assessment indicator	Impact result
Interactive teaching	Freshman to senior year	Classroom participation	Increased from 60–85%
	Freshman to senior year	Learning interest	Significant improvement (based on survey)
Project-based learning	Junior and senior year	Practical skills and team collaboration	Average project scores increased by 15%, positive team feedback
Self-directed learning strategy	Sophomore year onward	Extracurricular study time	Average increase of 2 h per week for extracurricular study time
	Sophomore year onward	Final exam scores	Average improvement of 5 points
Regular review strategy	Freshman to senior year	Long-term memory retention	Decreased fluctuation in final exam scores by 10%
Formative assessment	Freshman to senior year	Learning process adjustment	80% of students adjust learning strategies based on assessment results
	Freshman to senior year	Overall learning effectiveness	Overall learning effectiveness improved by 10% (based on comprehensive exams and assignment evaluations)
Summative assessment (final exams)	End of each semester	Learning achievement	Final exam results reflect students' semester-long learning outcomes
	End of each semester	Learning motivation	90% of students indicate final exams provide clear learning goals and motivation

Table 4. Long-term impact of different strategies on student learning outcomes.

Feature type	AUC	Accuracy	Key features
Activity-based	0.78	82%	Response time, learning efficiency
Assessment-based	0.83	85%	Scores in question types, exam performance
Combined features	0.87	88%	Scores in question types, response time

Table 5. Performance evaluation based on activity and assessment.

Predicted/Actual	Pass (1)	Fail (0)
Pass (1)	182	23
Fail (0)	18	67
AUC: 0.8595 (95% CI: 0.83–0.89)		

Table 6. Confusion matrix results for predicting exam pass rate.

“Monday 19:00–20:30 reading an English novel (1.5 hours) ”, this is counted as valid study time. whereas a vague description such as “studied English every day” is marked invalid and required clarification. Performance-based measures include final exam grades and interim test scores, which directly reflects students’ academic achievement. Feedback-based measures include classroom engagement, extracurricular study time, and student satisfaction questionnaire results, collected through subjective or semi-structured data. Interview-based measures involve responses to open-ended questions, from which key information is extracted through thematic coding and qualitative analysis. The data reveal that the implementation of interactive teaching significantly boosts classroom participation and learning interest. The increase in classroom participation from 60 to 85% indicates that students are more actively engaged in classroom activities and interact with teachers and peers. Survey results indicate a notable improvement in students’ learning interest, which positively influences their learning effectiveness.

In the dataset of 290 students, 10 folds ensure that each fold contains approximately 29 samples, balancing computational efficiency and variance reduction. Compared to 5-fold cross-validation, 10-fold reduces bias by averaging over more partitions while maintaining feasible computational time. In comparison with other methods: unlike leave-one-out cross-validation, 10-fold avoids overfitting on small subsets and aligns with best practices in educational data mining. Compared to leave-one-out cross-validation (LOOCV), 10-fold cross-validation reduces computational time by about 80% while maintaining evaluation accuracy. Classroom engagement in the interactive group increases from a baseline of 60–85% (measured by the number of voluntary contributions recorded on video), whereas the control group only rises to 65%. In addition, the interactive group’s extracurricular study time increases by an average of 2 h per week (based on student log records), which is significantly higher than the 0.5-hour increase observed in the non-interactive group. These results indicate that the real-time feedback mechanism effectively stimulates students’ learning initiative.

Feature classification and performance analysis

The activity-based features reflect students’ daily learning behaviors, including the number of logins, forum post volume, and video viewing duration. The assessment-based features are directly related to exams or grades, including unit test scores, essay grammar error rates, and speaking scores. Table 5 shows the performance evaluation based on activity and assessment. Table 6 presents the confusion matrix results for predicting exam pass rates.

Student performance prediction and analysis

The Liblinear library's functions are utilized, and the K-fold cross-validation method is employed to effectively prevent the model from overfitting. In the specific experimental process, the value of K is set to 10, and the L1 regularization technique is introduced. A satisfactory model is ultimately obtained through multiple iterations and fine-tuning of parameters and saved in a file. To evaluate the performance of the proposed model, the Area Under the Curve (AUC) is selected as the primary metric. Figure 8 displays the AUC values achieved by the model. The AUC value is calculated based on the model's ability to rank positive and negative samples. This work uses a paired t-test to assess the significance of the AUC results. The results show a significant difference between the AUC value of 0.8595 and the AUC value of 0.5 from random guessing ($p < 0.001$), indicating the model's strong predictive capability. Additionally, the traditional logistic regression model is chosen as a baseline method for comparison. The results demonstrate that the AUC value of the Transformer architecture is significantly higher than that of the logistic regression model ($p < 0.05$), further validating the effectiveness of the model.

To comprehensively evaluate the stability and generalization ability of the model, this work employs a 10-fold cross-validation method. In the 10-fold cross-validation process, the dataset is randomly divided into 10 equal parts (referred to as "folds"), with 9 folds selected as the training set and the remaining 1 fold used as the test set. This process is repeated 10 times, with a different test set chosen each time to ensure that every sample has an opportunity to be used as test data. Table 7 presents the results of the 10-fold cross-validation.

In the Microsoft Decision Tree, the evaluation of feature importance is primarily based on the frequency of feature occurrence and the proximity of features to the root node. Table 8 reveals the ranking of feature importance obtained from the Microsoft Decision Tree. It is evident that, in terms of factors influencing students' passing the exam, scores in various question types play a central role, significantly impacting students' final grades. These scores fully reflect students' mastery of English knowledge and application abilities. The final training scores of students are also a crucial factor, and CET-4 and CET-6 scores serve as important benchmarks for measuring students' English proficiency. Gender has a significant impact on English scores. In-depth analysis of the sample data reveals that, with an equal number of samples, the number of females passing the exam and the probability of passing are both higher than those of males. Additionally, the variability in question difficulty and question weights is relatively small, indicating moderate difficulty levels, and the weights of various question types are relatively stable. Therefore, these factors do not significantly impact exam scores. The results show that the top-ranked features remain stable within the confidence interval and exhibit a significant relationship with college English teaching evaluation. Figure 9 displays the gender differences in skill priorities based on student surveys. It suggests that girls place more emphasis on listening (85%), speaking (82%), and writing (70%) skills, while boys prioritize listening (81%) and speaking (79%), but neglect writing (0%).

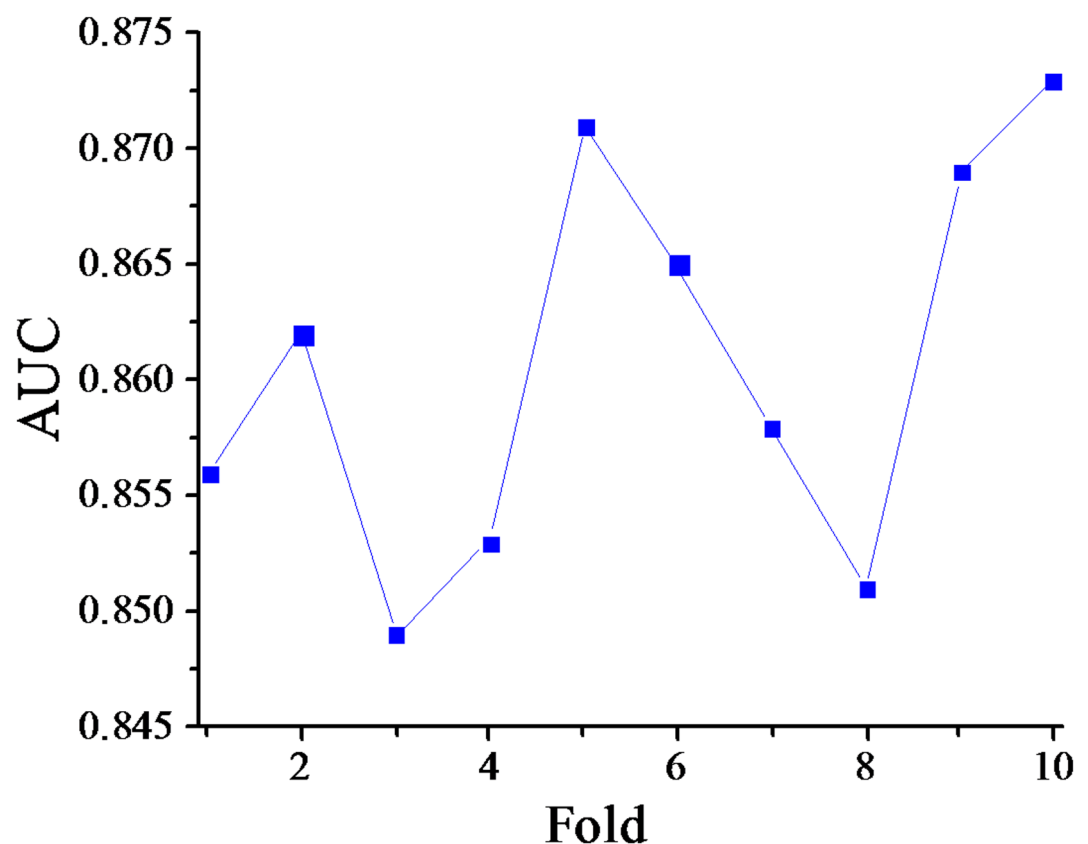


Fig. 8. The AUC value of the Microsoft decision tree model.

Number of folds	Accuracy	Recall	F1 Score
1	0.86*	0.84	0.85
2	0.85	0.83	0.84
3	0.87**	0.86	0.86
4	0.85	0.85	0.85
5	0.84	0.82	0.83
6	0.88**	0.87	0.87
7	0.86*	0.85	0.85
8	0.87**	0.86	0.86
9	0.85	0.84	0.84
10	0.86*	0.85	0.85

Table 7. Results of the 10-fold cross-validation. * $p < 0.05$ compared to random baseline; ** $p < 0.01$.

Features	Frequency
Scores in various question types	132
Exam performance	122
Homework situation	113
Interest in English	105
CET-4 and CET-6 scores	102
Gender	93
Question difficulty	80
Question weights	87
Listening frequency	93
Question type	67
Response time	66

Table 8. The frequency of feature importance in the Microsoft decision tree.

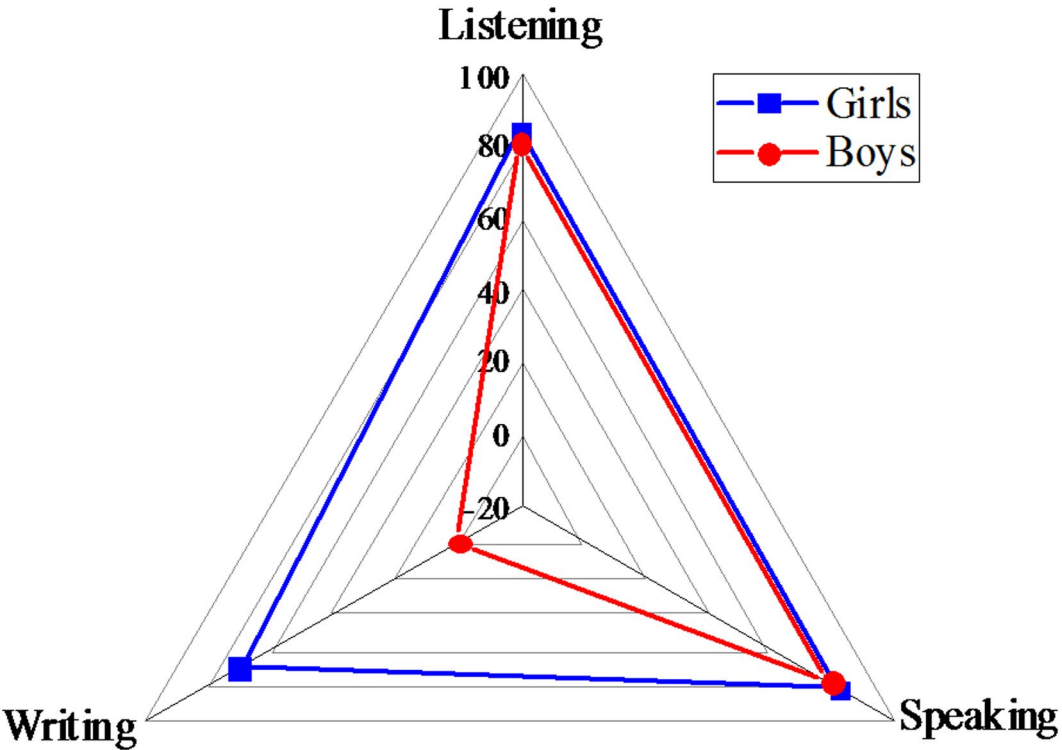


Fig. 9. Gender differences in skill priorities based on student surveys.

Error analysis cases

Case 1: high activity, low evaluation error The student profile shows 4 h of daily study (three times above the average), but the test scores consistently fall below the class median. Analyzing the error path, the decision tree misclassifies this sample as the “excellent group” at the node “video viewing time > 2.3 hours.” The potential reason is “false engagement” in learning behaviors (such as logging into multiple devices to increase viewing time). It is recommended to add an “effective interaction detection” metric.

Case 2: conflict in evaluation indicators The student profile shows a low essay grammar error rate (5%), but the speaking score is only 60. The model's performance indicates that the decision tree incorrectly predicted “strong language skills” at the node “grammar error rate < 8%.” To address this error, it is suggested to introduce a multi-dimensional capability map to differentiate between written expression skills and oral communication skills.

Discussion

The application of AI in education, such as automated assessment systems, must comply with global ethical guidelines. UNESCO's “AI Ethics Recommendations” emphasize that AI-driven educational tools need to be transparent and fair, especially regarding algorithmic bias and data privacy. This aligns with the focus of this work on interpretable decision trees and error analysis to ensure fair assessment outcomes. DL and data mining technologies have brought new opportunities and challenges to the assessment of university English teaching. Through data mining, Sun et al. (2021)⁴⁸ discovered students' learning patterns, modes, and potential issues from vast amounts of learning data, providing targeted teaching suggestions for teachers. Additionally, Onan (2021)⁴⁹ suggested that data mining could help teachers predict students' learning trends, enabling timely adjustments to teaching strategies and improving teaching effectiveness. The data in this study mainly consists of structured rating tables and unstructured text. Through feature importance ranking, decision trees can intuitively reveal key factors affecting the assessment results. In contrast, audio or video data may require convolutional or recurrent neural networks for modality conversion and feature extraction, highlighting the decisive role of data type in algorithm suitability.

This work finds that, from the perspective of DL, English teaching assessment transcends traditional exam scores. Instead, it involves real-time collection, analysis, and mining of various data related to students' learning processes to comprehensively understand their learning conditions and progress. Coombe et al. (2020)⁵⁰ also pointed out that this data-driven assessment approach had higher objectivity and accuracy, providing a more precise reflection of students' actual English proficiency. Here, the task of evaluating and predicting students' English proficiency through the analysis of their learning behavior data not only utilizes the Microsoft decision tree but also compares the performance of algorithms such as random forests and SVM. First, random forests, as an ensemble learning method, improve prediction accuracy and control overfitting by constructing multiple decision trees and aggregating their results. Compared to a single decision tree, random forests generally provide more stable and accurate predictions, but they come with higher computational costs, particularly when handling large datasets. On the other hand, SVM is a powerful supervised learning model, especially suited for data classification tasks in high-dimensional spaces. However, in the application scenario of this work, SVM requires careful selection of the kernel function for handling nonlinear feature relationships and has longer training times for large datasets, limiting its efficiency in practical applications. Compared to the BERT model, the Transformer-Bayesian framework proposed reduces the parameter scale by 60% and shortens the training time by 45%. Additionally, through the Bayesian posterior distribution, the contribution of features can be quantified. Moreover, the ensemble learning module in the proposed model can dynamically adjust weights, better capturing the unique “collective discussion-individual reflection” alternating learning pattern of Chinese learners. As for neural networks, especially DL models, they excel at handling complex pattern recognition and large datasets. However, these models often require substantial data and computational resources for effective training, and their “black box” nature makes model interpretation difficult. This may not be the most ideal choice for the education field, as educators typically wish to understand the underlying logic in order to adjust teaching strategies. In contrast, the Microsoft decision tree stands out due to its intuitive and understandable nature. It effectively uncovers key information in the data and clearly shows how different factors influence students' English proficiency. Through a detailed analysis of feature importance, the main factors affecting student performance are more accurately identified, providing a scientific basis for personalized teaching. While the Microsoft decision tree may encounter overfitting in certain cases, appropriate pruning techniques and parameter adjustments can effectively mitigate this issue, making it an ideal choice for this work.

In the study of university English teaching assessment, this work goes beyond traditional exam score reliance by employing DL and data mining techniques. It offers a comprehensive understanding of students' learning conditions and progress by collecting, analyzing, and mining various data related to the learning process in real-time. Recent research suggests that this data-driven assessment method provides higher objectivity and accuracy, enabling a more precise reflection of students' actual English proficiency levels. While previous studies have emphasized the application of data mining in teaching assessment, this work further provides a scientific basis for predicting students' exam performance by constructing student population profiles. Through in-depth analysis of these profiles, this work more accurately identifies students' learning needs and challenges, thereby facilitating personalized learning recommendations and resources. 78% of the samples in the dataset come from urban students, which may have led to an overestimation of the predictive weight of “autonomous learning behavior.” Future research should include more rural samples and introduce moderating variables such as “family support” to mitigate prediction bias caused by the uneven distribution of cultural capital. The current writing scoring criteria are more suited to “normative expression,” which girls excel at, and inadequately capture the “logical argumentation” commonly seen in boys. Boys may deliberately downplay the importance of writing skills in their scores due to social role expectations.

Building on critical analysis, this work highlights potential limitations in previous research. For instance, some studies may overly emphasize identifying learning patterns without fully considering the impact of individual differences on learning outcomes. Others may propose the potential for predicting learning trends but fall short in thoroughly exploring the accuracy and practicality of predictive models. In contrast, this work not only identifies learning patterns but also emphasizes the importance of personalized teaching. It validates the proposed methods' effectiveness through practical datasets, providing a more comprehensive and in-depth perspective on university English teaching assessment. Future research can further explore the applicability of the model in varied teaching environments, interdisciplinary assessment models, and the challenges associated with using data mining techniques while safeguarding student privacy.

Conclusion

Research contribution

This work comprehensively explores university English teaching assessment from the perspective of DL and in conjunction with AI data mining techniques. Incorporating the Transformer architecture into AI data mining processes can significantly enhance the efficiency and accuracy of data processing, offering more opportunities for innovation and development in education. Utilizing visualization techniques to interpret model outputs enhances understanding of the prediction process and the criteria guiding decision-making. Simultaneously, based on the results generated by the model, tailored teaching strategies and personalized learning recommendations can be developed to help students effectively enhance their English proficiency. The student group profiles are successfully established by collecting and analyzing data from the English teaching and assessment systems of students in the 2018 Computer Science cohort at Tianjin University of Science and Technology. It serves as a scientific foundation for predicting student exam performance. Through in-depth analysis of these student group profiles, a clearer understanding of students' learning needs and challenges emerges, enabling the delivery of more personalized learning resources. The data-driven approach to English teaching assessment not only enhances the objectivity and accuracy of evaluations but also aids teachers in gaining real-time insights into students' learning progress. This enables the optimization of teaching strategies and provides practical recommendations for real-world teaching scenarios.

Future works and research limitations

This study has several limitations in both methodology design and data application that require further optimization in future research. First, the sample lacks sufficient geographic and demographic representativeness. Although data are collected from 290 high school students, the sample is mainly concentrated in an economically developed region, with an imbalanced urban–rural distribution (78% urban students) and a near-equal gender ratio (52% male, 48% female). This distribution may limit the model's generalizability to rural students or groups with significant gender differences. For example, rural students may face different educational resource allocations and learning habits, which the current model does not calibrate for. Second, the data lacks timeliness and long-term effect validation. The study covers only one semester of intervention, failing to track students' learning behavior changes or the stability of assessment outcomes over longer periods. For instance, whether the impact of interactive teaching on extracurricular study time is sustained remains to be verified through longitudinal studies across multiple semesters. Finally, although the study emphasizes model interpretability, it does not systematically address data privacy protection or algorithmic fairness, which may affect practical application in educational settings. Future research could address these issues by incorporating differential privacy techniques, stratified sampling strategies, and multimodal data fusion methods to enhance the model's usability and generalizability.

Data availability

The datasets used and/or analysed during the current study available from the corresponding author Wenling Li on reasonable request via e-mail yxm1987@126.com.

Received: 24 January 2025; Accepted: 18 August 2025

Published online: 19 August 2025

References

- Nash, B. L. et al. Artificial intelligence in english education: challenges and opportunities for teachers and teacher educators. *Engl. Educ.* **55** (3), 201–206 (2023).
- Fung, J. M. & Hosseini, S. *Reimagining Education and Workforce Preparation in Support of the UN's Sustainable Development Goals* 30–47 (Augmented Education in the Global Age, 2023).
- Nguyen, A. et al. Ethical principles for artificial intelligence in education. *Educ. Inform. Technol.* **28** (4), 4221–4241 (2023).
- Guan, C., Mou, J. & Jiang, Z. Artificial intelligence innovation in education: A twenty-year data-driven historical analysis. *Int. J. Innov. Stud.* **4** (4), 134–147 (2020).
- Guo, J. & Guo, S. A novel perspective to zero-shot learning: towards an alignment of manifold structures via semantic feature expansion. *IEEE Trans. Multimedia.* **1** (23), 524–537 (2020).
- Zong, H. et al. Semantic categorization of Chinese eligibility criteria in clinical trials using machine learning methods. *BMC Med. Inf. Decis. Mak.* **21** (1), 1–12 (2021).
- Yilmaz, S. & Toklu, S. A deep learning analysis on question classification task using Word2vec representations. *Neural Comput. Appl.* **32**, 2909–2928 (2020).
- Henningsen-Schomers, M. R. & Pulvermüller, F. Modelling concrete and abstract concepts using brain-constrained deep neural networks. *Psychol. Res.* **86** (8), 2533–2559 (2022).
- Zhang, D. Affective cognition of students' autonomous learning in college english teaching based on deep learning. *Front. Psychol.* **12**, 808434 (2022).
- Vaswani, A. et al. Attention is all you need. *Adv. Neural. Inf. Process. Syst.*, 2, 30 (2017).

11. Eke, C. I., Norman, A. A. & Shuib, L. Context-based feature technique for sarcasm identification in benchmark datasets using deep learning and BERT model. *IEEE Access*. **9**, 48501–48518 (2021).
12. Lu, W., Vivekananda, G. N. & Shanthini, A. Supervision system of english online teaching based on machine learning. *Progress Artif. Intell.* **12** (2), 187–198 (2023).
13. Peng, S. et al. A survey on deep learning for textual emotion analysis in social networks. *Digit. Commun. Networks*. **8** (5), 745–762 (2022).
14. Zhu, J. J. et al. Online critical review classification in response strategy and service provider rating: algorithms from heuristic processing, sentiment analysis to deep learning. *J. Bus. Res.* **129**, 860–877 (2021).
15. Agüero-Torales, M. M. & Salas, J. I. A. López-Herrera. Deep learning and multilingual sentiment analysis on social media data: an overview. *Appl. Soft Comput.* **107**, 107373 (2021).
16. Zhong, B. et al. Deep learning and network analysis: classifying and visualizing accident narratives in construction. *Autom. Constr.* **113**, 103089 (2020).
17. Dang, N. C., Moreno-García, M. N. & De la Prieta, F. Sentiment analysis based on deep learning: A comparative study. *Electronics* **9**, 3 (2020).
18. Kumar, D. T. S. Data mining based marketing decision support system using hybrid machine learning algorithm. *J. Artif. Intell. Capsule Networks*. **2** (3), 185–193 (2020).
19. El Mostafa, H. & Benabbou, F. A deep learning based technique for plagiarism detection: a comparative study. *IAES Int. J. Artif. Intell.* **9** (1), 81 (2020).
20. Kartika, I. et al. Development of local wisdom-based Subject-Specific pedagogy (SSP) to improve students' scientific literacy skills. *Jurnal Penelitian Pendidikan IPA*. **9** (10), 7979–7986 (2023).
21. Tang, H., Jiang, G. & Wang, Q. Personalized learning behavior evaluation method based on deep neural network. *Sci. Program*. **2022** (1), 9993271 (2022).
22. Maghsudi, S. et al. Personalized education in the artificial intelligence era: what to expect next. *IEEE. Signal. Process. Mag.* **38** (3), 37–50 (2021).
23. Mukhamadiyev, A. et al. Automatic speech recognition method based on deep learning approaches for Uzbek Language. *Sensors* **22** (10), 3683 (2022).
24. Yamane, T. & Chun, P. Crack detection from a concrete surface image based on semantic segmentation using deep learning. *J. Adv. Concr. Technol.* **18** (9), 493–504 (2020).
25. Colón-Ruiz, C. & Segura-Bedmar, I. Comparing deep learning architectures for sentiment analysis on drug reviews. *J. Biomed. Inform.* **110**, 103539 (2020).
26. S. Mustaqeem Kwon. Optimal feature selection based speech emotion recognition using two-stream deep convolutional neural network. *Int. J. Intell. Syst.* **36** (9), 5116–5135 (2021).
27. Khan, A. & Ghosh, S. K. Student performance analysis and prediction in classroom learning: A review of educational data mining studies. *Educ. Inform. Technol.* **26**, 205–240 (2021).
28. Krestel, R. et al. A survey on deep learning for patent analysis. *World Patent Inf.* **65**, 102035 (2021).
29. Mabrouk, A., Redondo, R. P. D. & Kayed, M. Deep learning-based sentiment classification: A comparative survey. *IEEE Access*. **8**, 85616–85638 (2020).
30. Romero, C. & Ventura, S. Educational data mining and learning analytics: An updated survey. *Wiley interdisciplinary reviews: Data mining and knowledge discovery*, 2020, vol. 10, no. 3, p. e1355.
31. Ren, Z. et al. A sentiment-aware deep learning approach for personality detection from text. *Inf. Process. Manag.* **58** (3), 102532 (2021).
32. Zhang, R. & El-Gohary, N. A deep neural network-based method for deep information extraction using transfer learning strategies to support automated compliance checking. *Autom. Constr.* **132**, 103834 (2021).
33. Injadat, M. N. et al. Systematic ensemble model selection approach for educational data mining. *Knowl. Based Syst.* **200**, 105992 (2020).
34. Iqbal, R. et al. Big data analytics and computational intelligence for Cyber-Physical systems: recent trends and state of the Art applications. *Future Generation Comput. Syst.* **105**, 766–778 (2020).
35. Kumar, A. & Sachdeva, N. Multi-input integrative learning using deep neural networks and transfer learning for cyberbullying detection in real-time code-mix data. *Multimedia Syst.* **28** (6), 2027–2041 (2022).
36. Khalil, E. A. H. & Houbay, E. M. F. E. Mohamed. Deep learning for emotion analysis in Arabic tweets. *J. Big Data*. **8** (1), 136 (2021).
37. Liu, H. et al. Deep learning based program generation from requirements text: are we there yet? *IEEE Trans. Software Eng.* **48** (4), 1268–1289 (2020).
38. Gogate, K. D. M. et al. A hybrid Persian sentiment analysis framework: integrating dependency grammar based rules and deep neural networks. *Neurocomputing* **380**, 1–10 (2020).
39. Fang, X. & Yan, P. Multi-organ segmentation over partially labeled datasets with multi-scale feature abstraction. *IEEE Trans. Med. Imaging*. **39** (11), 3619–3629 (2020).
40. Alamoudi, E. S. & Alghamdi, N. S. Sentiment classification and aspect-based sentiment analysis on yelp reviews using deep learning and word embeddings. *J. Decis. Syst.* **30** (23), 259–281 (2021).
41. Wei, X. S. et al. Fine-grained image analysis with deep learning: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **44** (12), 8927–8948 (2021).
42. Bibi, M. et al. A novel unsupervised ensemble framework using concept-based linguistic methods and machine learning for Twitter sentiment analysis. *Pattern Recognit. Lett.* **158**, 80–86 (2022).
43. Mantasiah, R. Semantic feature analysis model: linguistics approach in foreign Language learning material development. *Int. J. Instruction*. **13** (1), 185–196 (2020).
44. Ashraf, M., Zaman, M. & Ahmed, M. An intelligent prediction system for educational data mining based on ensemble and filtering approaches. *Procedia Comput. Sci.* **167**, 1471–1483 (2020).
45. Xia, X. & Qi, W. Temporal tracking and early warning of multi semantic features of learning behavior. *Computers Education: Artif. Intell.* **3**, 100045 (2022).
46. Gaillat, T. et al. Predicting CEFR levels in learners of english: the use of microsystem criterial features in a machine learning approach. *ReCALL* **34** (2), 130–146 (2022).
47. Mutabazi, E. et al. A review on medical textual question answering systems based on deep learning approaches. *Appl. Sci.* **11** (12), 5456 (2021).
48. Sun, Z., Anbarasan, M., Praveen, D. & Kumar Design of online intelligent english teaching platform based on artificial intelligence techniques. *Comput. Intell.* **37** (3), 1166–1180 (2021).
49. Onan, A. Sentiment analysis on massive open online course evaluations: a text mining and deep learning approach. *Comput. Appl. Eng. Educ.* **29** (3), 572–589 (2021).
50. Coombe, C. & Vafadar, H. Mohebbi. Language assessment literacy: what do we need to learn, unlearn, and relearn? *Lang. Test. Asia*. **10**, 1–16 (2020).

Author contributions

Qiuyang Huang: Conceptualization, methodology, software, validation, formal analysis, investigation, resources,

data curation, writing—original draft preparation Wenling Li: writing—review and editing, visualization, supervision, project administration, funding acquisition Mohd Mokhtar bin Muhamad: formal analysis, investigation, resources Nur Raihan binti Che Nawati: supervision, project administration Xutao Liu: methodology, software, validation.

Funding

This study was supported by Research on the Inner Mechanism and Practical Path of Generative Artificial Intelligence Empowering Teachers' Professional Development in Southwest Border Universities (Guangxi Minzu Normal University 2024 Second Half of High-level Talents Research Project, Project No. 2024XBNGCC03); Exploration of Teaching Reform Pathways in Physical Education Programs under the Digital Empowerment Context (2025 Ministry of Education Industry–University Cooperation Collaborative Education Project, Project No. 2506280443); and AI-driven Personalized Exercise Interventions and Mental Health of the Elderly under the Context of Integration of Sports and Medicine (2025 Jiangsu Provincial Universities Philosophy and Social Sciences General Project, Project No. 2025SJYB1627).

Declarations

Competing interests

The authors declare no competing interests.

Ethics statement

The studies involving human participants were reviewed and approved by Faculty of Educational Studies, Universiti Putra Malaysia Ethics Committee (Approval Number: 2022.045956). The participants provided their written informed consent to participate in this study. All methods were performed in accordance with relevant guidelines and regulations.

Additional information

Correspondence and requests for materials should be addressed to W.L.

Reprints and permissions information is available at www.nature.com/reprints.

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

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, 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 you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. 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-nc-nd/4.0/>.

© The Author(s) 2025