Journal of Computers Vol. 36 No. 2, April 2025, pp. 77-89
https://doi.org/10.63367/199115992025043602006

A Teaching Effectiveness Evaluation Method Based on Deep Interpretable Learning

Guo-Qiang Zhuang, Fang Li, and Chao Feng*

Jiangsu College of Engineering and Technology,
Nantong 226001, Jiangsu, China,
{zgq, lifang, fengchao}@jcet.edu.cn

Received 15 September 2024; Revised 16 September 2024; Accepted 16 September 2024

Abstract. Based on evidence-based teaching theory, this paper explores the evaluation method of teaching effectiveness in higher education. This study aims to develop a comprehensive assessment model that combines students' psychological conditions with their academic performance. We employ a bidirectional LSTM network to analyze a large amount of historical data from online courses, establish a decision-support model, and identify key factors affecting academic performance through interpretable analysis. The empirical study uses statistical methods such as the Wilcoxon rank-sum test to verify the predictive accuracy of the model. The findings demonstrate a causal relationship between students' psychological conditions and teaching performance, providing a new research perspective for evaluating teaching effectiveness in colleges and universities.

Keywords: evidence-based education, psychological assessment, teaching evaluation, learning process

1 Research Background

The evaluation of teaching effectiveness in higher education has been a central concern for universities for decades, reflecting the ongoing quest for quality assurance and improvement in academic institutions [1-3]. Traditionally, the assessment of teaching effectiveness has predominantly relied on students' final grades as the primary indicator of successful instruction. However, this approach has been increasingly recognized as limited, failing to capture the multifaceted nature of effective teaching and learning. Concurrently, the relationship between college students' psychological conditions and their academic performance has emerged as a critical area of research in educational studies [4-6]. This focus acknowledges the complex interplay between mental well-being and academic achievement, suggesting that psychological factors may significantly influence learning outcomes. In response to these evolving perspectives, researchers have begun to explore more sophisticated methods of evaluation. For instance, Xinfang Ding et al. have leveraged machine learning algorithms to mine psychological and academic data, developing predictive models for academic performance [7-8]. This approach represents a shift towards more data-driven, holistic assessments of student success. Similarly, Zhen Hu et al. proposed an innovative methodology using random cluster sampling to survey middle school students in Beijing [9]. Their custom-designed questionnaire for anonymous self-administered surveys aimed to assess both students' learning difficulties and psychological health, highlighting the growing recognition of the interconnectedness of these factors. Despite these advancements, there remains a notable gap in research specifically attributing the evaluation of college students' learning outcomes to their psychological health education, particularly studies that analyze the relationship between college students' psychological health and academic performance using longitudinal growth data.

Educational psychology provides a robust theoretical foundation for advancing research in teaching effectiveness evaluation [17-18]. It suggests that research should be firmly grounded in learning psychology and cognitive psychology, leveraging modern information processing theory and computer technology [21-22]. This approach emphasizes the critical importance of designing effective learning environments that align with students' cognitive processes and psychological needs. The assessment of teaching effectiveness, therefore, should be intrinsi-

* Corresponding Author

cally linked to students' learning psychology and behavior, a perspective that closely aligns with the core principles of constructivist educational psychology [23]. This theoretical framework posits that learning is an active, context-dependent process where students construct knowledge based on their experiences and interactions with their environment. In recent years, domestic educational researchers have proposed innovative evaluation methods that incorporate psychometrics, computer-based evaluation design, educational data mining, and machine learning [11, 24-25]. These approaches aim to develop game-based evaluations that comprehensively measure students' diverse skills, moving beyond traditional assessment methods to capture a more holistic view of student capabilities [26]. On the international front, researchers have advocated for the establishment of teaching effectiveness evaluation systems based on positive psychological assessments [27-29]. This approach involves creating evaluation matrices that assess course quality through quantitative data, emphasizing the importance of fostering positive psychological states in the learning process. These developments reflect a growing recognition of the need for more nuanced, multidimensional approaches to evaluating teaching effectiveness.

The concept of evidence-based teaching evaluation has emerged as a promising framework for addressing the complexities of assessing teaching effectiveness in the modern educational landscape [19-20]. As a subset of evidence-based education, it integrates teaching evaluation theories, data science technologies, and teaching analysis methods to provide a comprehensive approach to assessment [11]. This methodology involves collecting and analyzing a wide range of teaching data to perform multidimensional assessments and measurements of both the teaching process and its effectiveness. The integration of evidence-based teaching theories with big data in education offers new foundations for evaluating teaching effectiveness, allowing for more precise, data-driven insights into the factors that contribute to successful learning outcomes. This approach aligns with the broader trend towards evidence-based practices in education, which emphasizes the importance of grounding pedagogical decisions in empirical evidence and rigorous analysis.

Building upon these theoretical and methodological advancements, this paper proposes a novel teaching effectiveness evaluation method based on deep interpretable learning. This approach represents a significant step forward in the field, combining the strengths of evidence-based teaching concepts with cutting-edge machine learning techniques. The proposed method includes an analytical framework for teaching effectiveness evaluation that is firmly rooted in evidence-based teaching concepts, providing a structured approach to assessing the multiple dimensions of effective teaching. Additionally, it offers a comprehensive evaluation method for teaching that leverages deep learning algorithms to analyze complex patterns in educational data. The effectiveness of this method in attributing teaching outcomes is then rigorously validated through empirical research, demonstrating its potential to provide more accurate and nuanced assessments of teaching effectiveness. This innovative approach holds promise for revolutionizing how educational institutions evaluate and improve their teaching practices, ultimately leading to enhanced learning outcomes for students.

2 A Teaching Effectiveness Evaluation Method Based on Deep Interpretable Learning

This section forms the core of this research, aiming to provide a comprehensive and innovative framework for educational assessment. This section integrates evidence-based teaching concepts with deep learning technologies to design a scientifically rigorous and practical evaluation system. The method not only considers traditional teaching indicators but also incorporates students' psychological health status and online learning behaviors to assess teaching effectiveness holistically. By introducing deep interpretable learning techniques, this approach offers more precise and explainable evaluation results, providing powerful decision support for educators and administrators.

The section is divided into two main sections, each with its unique design philosophy and content focus. The first part (Section 2.1) introduces an analytical framework based on evidence-based teaching concepts, designed to create a cyclical, multidimensional evaluation system. It encompasses five key components: offline psychological assessment, teaching implementation, online data collection, effectiveness evaluation, and feedback intervention, emphasizing the continuity and comprehensiveness of the evaluation process. The second part (Section 2.2) elaborates on the specific evaluation method based on deep interpretable learning, with a design philosophy of applying advanced machine learning techniques to educational assessment. This section covers the complete workflow from data collection and preprocessing to model construction and interpretability analysis. Notably, it introduces bidirectional LSTM networks [30-31] to capture complex patterns in educational data and employs gradient vector methods for model interpretation, providing more transparent and credible evaluation results. Through this design, the section not only offers a theoretical framework but also presents a practical technical solution, bringing new perspectives and methodologies to the field of educational assessment.

The teaching effectiveness evaluation method based on deep interpretable learning proposed in this study offers several significant advantages over traditional methods: Firstly, it integrates students' psychological health status, online learning behaviors, and learning outcomes, providing a comprehensive assessment framework that overcomes the limitations of traditional methods focusing on single dimensions. Secondly, the adoption of bidirectional LSTM networks enables the capture of complex temporal patterns in students' learning processes, greatly enhancing prediction accuracy. Moreover, the interpretability of this method transforms the assessment results from a "black box" into understandable insights, allowing educators to identify key factors influencing teaching effectiveness and formulate targeted improvement strategies. Lastly, this method establishes a cyclical evaluation system capable of continuously optimizing teaching practices, a feature absent in traditional static assessment methods. Overall, this study presents a more scientific, dynamic, and comprehensive new approach to teaching evaluation in higher education.

2.1 Analytical Framework of Teaching Effectiveness Evaluation Method Based on Evidence-Based Teaching Concepts

The evidence-based teaching evaluation method represents an innovative approach to educational assessment, with a primary focus on enhancing the effectiveness of moral education. This method ingeniously adopts the course's starting and ending points as the evaluation cycle, seamlessly integrating both offline and online teaching scenarios. By incorporating psychological assessments, teaching implementation, data acquisition, teaching effectiveness evaluation, and feedback interventions, it creates a cyclical, multifaceted, and comprehensive teaching evaluation system [10-11], the analytical framework is shown in Fig. 1:

1. Offline Psychological Assessment:

This crucial component involves regular psychological health evaluations of students, utilizing a variety of tools including surveys, psychometric instruments, and behavioral observations. The primary objectives are to gain real-time insights into students' mental states and emotional fluctuations, enable teachers and educational institutions to monitor students' psychological well-being proactively, identify potential psychological issues early for timely interventions, and create a supportive learning environment that considers students' mental health as a key factor in educational success.

2. Offline Teaching Implementation:

This component focuses on evaluating the teacher's classroom activities and pedagogical methods. The assessment encompasses various aspects such as the teacher's instructional design and lesson planning, interaction techniques and guidance methods employed during the teaching process, classroom management skills and ability to create an engaging learning atmosphere, use of diverse teaching aids and technologies to enhance learning experiences, and adaptability in addressing different learning styles and student needs.

3. Online Teaching Platform Data Acquisition:

Leveraging the power of digital technology, this component involves the systematic collection of data from online teaching platforms. The data gathered includes student learning behaviors and patterns, progress tracking across various subjects and topics, assignment submission rates and quality, participation levels in online discussions and collaborative activities, and time spent on different learning materials and activities. This wealth of objective data provides invaluable insights for evaluating teaching effectiveness and student engagement in the digital learning environment.

4. Teaching Effectiveness Evaluation:

This comprehensive evaluation is conducted through multiple assessment methods, including traditional measures such as exam results and assignment quality, classroom performance and participation metrics, student self-evaluations to gauge perceived learning and satisfaction, peer and teacher evaluations for a 360-degree assessment approach, and predictive analytics using integrated data from both offline and online learning processes to forecast teaching outcomes and identify areas for improvement.

5. Teaching Effectiveness Feedback and Intervention:

Based on the evaluation results, this component focuses on providing timely and constructive feedback to students, empowering them to adjust and enhance their learning strategies, guiding teachers in refining their pedagogical approaches and addressing identified weaknesses, enabling educational institutions to implement targeted teaching interventions, facilitating the reallocation of resources to areas of greatest need, and fostering a culture of continuous improvement in the educational process.

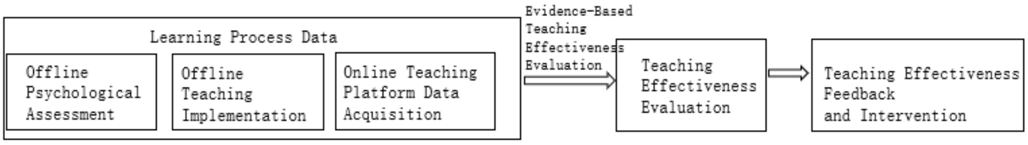

Learning Process Data
Offline Psychological Assessment
Offline Teaching Implementation
Online Teaching Platform Data Acquisition
Evidence-Based Teaching Effectiveness Evaluation
Teaching Effectiveness Evaluation
Teaching Effectiveness Feedback and Intervention

Fig. 1. Analytical framework of the teaching effectiveness evaluation method based on evidence-based teaching concepts

The evidence-based teaching evaluation method is designed as a cyclical process, with each evaluation cycle informing and influencing the next, creating a continuous loop of improvement. This iterative approach ensures that teaching methods are constantly refined based on empirical evidence, student needs are regularly reassessed and addressed, and the educational system remains adaptive and responsive to changing requirements. A key strength of this evaluation method is its integration of intelligent technology and data science, ensuring the scientific rigor of the evaluation process, enhanced objectivity in assessments, the ability to process and analyze large volumes of data for meaningful insights, and predictive capabilities to anticipate future trends and challenges in education. In summary, the evidence-based teaching evaluation method represents a significant advancement in educational assessment. By creating a cyclical, multifaceted, and comprehensive evaluation system that focuses on students' learning outcomes and teaching optimization, it addresses the complex needs of modern education. The integration of offline and online teaching scenarios, coupled with the leveraging of intelligent technology and data science, ensures that the teaching evaluation process is scientific, professional, and objective [11]. This approach not only enhances the quality of education but also prepares educational institutions to meet the evolving challenges of the 21st-century learning landscape.

2.2 Teaching Effectiveness Evaluation Method Based on Deep Interpretable Learning

The higher education information system in this study adopts a multi-layer architectural design aimed at comprehensive assessment and analysis of teaching effectiveness. The system consists of five main layers: data acquisition, data integration, data processing and analysis, application services, and user interface. The data acquisition layer collects student learning behavior and performance data through multiple subsystems such as the student information system, online learning platform, and library system. The data integration layer uses ETL tools and data warehouse technology to integrate and store data from different sources. The data processing and analysis layer utilizes big data processing frameworks and machine learning algorithms to perform in-depth analysis on the integrated data. The application service layer provides functions such as teaching effectiveness evaluation, student profiling, and learning recommendations based on the analysis results. The user interface layer offers intuitive data visualization and interactive interfaces for teachers, students, and administrators. The entire system adopts a microservices architecture, ensuring flexibility and scalability, while implementing strict data security and privacy protection measures.

The teaching effectiveness evaluation method based on deep interpretable learning involves the following key steps:

(1) Data Collection from Information Systems

Data is gathered from various sources, including information on college students' psychological conditions, their course grades during their time at the university, and graduation information. The data collection process for this study involved three primary information systems: the Student Management System, the Moodle online learning platform, and a psychological health assessment system. Data collection spanned several semesters, encompassing multiple dimensions of student information. From the Student Management System, we extracted basic student information such as student ID, gender, and admission scores. The Moodle platform provided rich learning behavior data, including weekly online learning duration, frequency of discussion posts, and assignment submission status. This data was exported weekly using Moodle's built-in data export function. Psychological health assessments were conducted at three time points: the beginning, middle, and end of the semester, with students completing the questionnaire online. Additionally, we collected performance data from six periodic assessments as indicators of learning outcomes. All data was stored in CSV format and de-identified to protect student privacy. The data collection process strictly adhered to the school's ethical review regulations and obtained informed consent from students. To ensure data quality, we employed automated scripts for data cleaning and validation, eliminating incomplete or obviously erroneous records. This comprehensive data collection process laid a solid foundation for subsequent deep learning analysis.

(2) Preprocessing of Student Personal Data:

a) Encoding Issues: For instance, when dealing with students' psychological conditions, the results of psychological scales are categorized into three states: mentally healthy, mildly anxious, and severely anxious, or other issues. Discrete data are encoded using one-hot encoding, which generates a feature representation for psychological conditions as {001, 010, 100}. Here, 001 corresponds to being mentally healthy, 010 to mild anxiety, and 100 to severe anxiety, among other conditions. One-hot encoding is employed to facilitate the handling of real-number data by deep neural networks and to expand the dimensionality of the input features.

b) Data Normalization: Different courses may have different grading scales. For example, the grading scale for an English course might differ from that of a specialized course, and this variation is closely related to the nature of the courses. Such differences in scale could affect the results of the deep learning classification model used for analysis, so data normalization is necessary. The purpose of normalization is to constrain the preprocessed data within a specific range (e.g., [0,1] or [-1,1]) to eliminate adverse effects caused by anomalous data samples.

Suppose there are n samples, each with d features, forming an n×d data matrix X. The goal is to map each feature value in X to a new range, such as [0,1] or [-1,1]. We use Min-Max Normalization, which is calculated using the following formula:

$$x_i^{'} = \frac{x_i - \min\left(\mathbf{x}\right)}{\max\left(\mathbf{x}\right) - \min\left(\mathbf{x}\right)}$$

Where x_i represents the original feature value of the i-th sample, min(x) and max(x) denote the minimum and maximum values of that feature, respectively. x_i^' is the normalized feature value, which is mapped to the range [0,1]. If the feature values need to be mapped to the range [-1,1], the following formula can be used:

$$x_i^{'} = \frac{x_i - \mu}{\sigma} \times 2$$

Where μ and σ represent the mean and standard deviation of the feature, respectively.

c) Handling Missing Data: Missing data is an inevitable issue in the original dataset, and different strategies are employed to address this depending on the type of data. For example, if a student's grade for a particular course is missing, the data is completed using a mean imputation method. This approach fills in the missing data with similar values, ensuring the sample size is maintained.

(3) Building a Decision Model Based on Historical Data:

The preprocessed data serves as input for modeling historical data, a crucial step in this evidence-based teaching effectiveness evaluation method. The appropriate selection of a deep learning classification method is fundamental to completing the model creation. In this study, we propose a semantic feature learning model based on Bidirectional Long Short-Term Memory (Bi-LSTM) networks.

LSTM networks are chosen for several compelling reasons:

Effective Capture of Latent Semantic Information: LSTM networks can capture latent semantic information more effectively than traditional Recurrent Neural Networks (RNNs). This capability is particularly important in the context of educational data, where complex relationships between various factors influencing student performance need to be understood [12].

Superior Classification Results: The enhanced ability to capture and retain relevant information over long sequences leads to better classification results, which is crucial for accurately evaluating teaching effectiveness [13].

Mitigation of Vanishing and Exploding Gradients: LSTM networks use gating mechanisms to control the flow of information, effectively addressing the vanishing and exploding gradient problems that plague traditional RNNs. This feature ensures stable training and more reliable results.

The basic LSTM unit consists of four main components:

Forget Gate: This gate determines how much of the previous memory information should be retained. It helps the network to discard irrelevant information from the past, allowing it to focus on more recent and relevant data.

Input Gate: The input gate controls how much new information at the current time step should be added to the Memory Cell. This selective addition of new information helps the network to maintain a balance between retaining important past information and incorporating new, relevant data.

Output Gate: This gate determines how much information should be output from the Memory Cell. It allows the network to selectively expose the internal state, providing only the most relevant information for the current prediction or classification task.

Memory Cell: The Memory Cell is the core component of LSTM, capable of maintaining long-term memory information. It acts as a reservoir of information, allowing the network to retain important features over long sequences.

The interaction of these components can be described mathematically as follows:

$$
\begin{aligned}
f_t &= \sigma\left(W_f \cdot \left[h_{t-1}, x_t\right] + b_f\right) \\
i_t &= \sigma\left(W_i \cdot \left[h_{t-1}, x_t\right] + b_i\right) \\
o_t &= \sigma\left(W_o \cdot \left[h_{t-1}, x_t\right] + b_o\right) \\
c_t &= f_t \circ c_{t-1} + i_t \circ \tanh\left(W_c \cdot \left[h_{t-1}, x_t\right] + b_c\right) \\
h_t &= o_t \circ \tanh\left(c_t\right)
\end{aligned}
$$

Where:

- f_t, i_t, o_t are the forget, input, and output gates respectively
- c_t is the cell state
- h_t is the hidden state
- x_t is the input at time t
- W and b are weight matrices and bias vectors
- σ is the sigmoid function
- denotes element-wise multiplication

While LSTM units are powerful, they have a limitation: information can only propagate forward. This means that the information at time step $t + 1$ depends only on the input information before time step t. To overcome this limitation and enable the model to utilizt both past and future context, we introduce Bi-LSTM networks. Bi-LSTM is a variant that builds on unidirectional LSTM, designed to utilize bidirectional context information in sequences. It contains two LSTM layers:

Forward LSTM: Processes the input sequence from left to right.

Backward LSTM: Processes the input sequence from right to left.

This bidirectional processing allows the network to simultaneously leverage the context information from both directions in the sequence. The output at each time step is typically a concatenation or a combination of the outputs from both forward and backward LSTMs. Mathematically, we can represent the Bi-LSTM as:

$$
\begin{aligned}
\overrightarrow{h_t} &= \text{LSTM}\left(x_t, \overrightarrow{h_{t-1}}\right) \\
\overleftarrow{h_t} &= \text{LSTM}\left(x_t, \overleftarrow{h_{t+1}}\right) \\
y_t &= f\left(\overrightarrow{h_t}, \overleftarrow{h_t}\right)
\end{aligned}
$$

Where:

- $\overrightarrow{h_t}$ is the forward hidden state
- $\overleftarrow{h_t}$ is the backward hidden state
- y_t is the output at time t

f is a function combining the forward and backward states (e.g., concatenation)

The Bi-LSTM model offers several advantages for evaluating teaching effectiveness:

- Comprehensive Context Understanding: By processing sequences in both directions, Bi-LSTM can capture a more comprehensive understanding of the context in educational data. This is particularly useful for understanding the complex interplay of factors affecting student performance over time.
- Improved Feature Extraction: The bidirectional nature of the model allows for more nuanced feature extraction, potentially uncovering subtle patterns in student data that might be missed by unidirectional

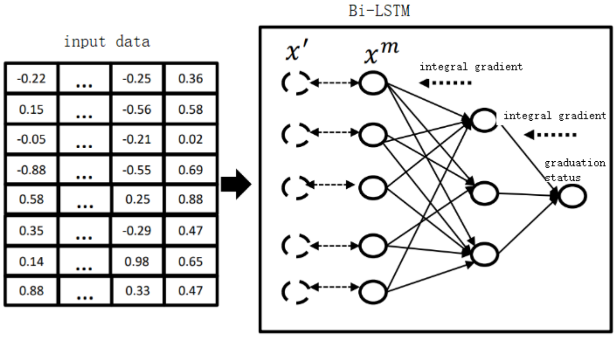
models.

- Enhanced Predictive Power: The ability to leverage both past and future context often leads to improved predictive performance, which is crucial for accurately evaluating teaching effectiveness and predicting student outcomes.
- Flexibility in Handling Variable-Length Sequences: Bi-LSTM can effectively handle variable-length sequences of student data, accommodating different educational timelines and data collection frequencies.
- By employing this Bi-LSTM-based semantic feature learning model, we aim to create a robust and accurate system for evaluating teaching effectiveness, capable of capturing the complex, time-dependent relationships inherent in educational data.

In selecting a deep learning model, we considered various options including Feedforward Neural Networks (FNN), Convolutional Neural Networks (CNN), and Recurrent Neural Networks (RNN). However, the Bi-LSTM model was ultimately chosen due to its advantages: effective capture of long-term dependencies crucial for analyzing student performance across entire semesters; bidirectional processing capability allowing simultaneous consideration of past and future contextual information; flexibility in handling variable-length sequence data, which aligns with the characteristics of educational data; and strong noise resistance, helpful in filtering out occasional fluctuations in educational data. In comparison, FNNs struggle to capture temporal dependencies, CNNs are more suited to spatially structured data, and traditional RNNs often suffer from the vanishing gradient problem in long sequences. Consequently, the Bi-LSTM model demonstrates significant advantages in processing the educational time-series data in this study, capable of better capturing complex patterns and long-term trends in students' learning processes.

(4) Interpretability Analysis of the Decision Model:

Conducting interpretability analysis on the decision model helps make the black-box model more transparent, identifying dependencies within the decision model and providing a basis for subsequent decision-making. Gradient vectors are used for interpretability analysis of the deep neural network [14-16]. The interpretability analysis framework is illustrated in Fig. 2.



-0.22	...	-0.25	0.36
0.15	...	-0.56	0.58
-0.05	...	-0.21	0.02
-0.88	...	-0.55	0.69
0.58	...	0.25	0.88
0.35	...	-0.29	0.47
0.14	...	0.98	0.65
0.88	...	0.33	0.47

Fig. 2. Interpretability analysis framework for the graduate status classification model

In the black-box model's deep neural network, gradient-based attribution analysis can be intuitively used to identify the features in the input feature space that have a significant influence. However, naive gradients can suffer from gradient saturation, where gradients fail to reflect meaningful information. For example, when analyzing graduate information T, which is represented by multiple information fields W, the impact of different fields W on the classification result varies.

When a change is applied to a field W, it causes a corresponding change in y, denoted as y+Δy. For instance, in the negative half-axis of the ReLU activation function, the gradient change information is 0. Simply using gradient information cannot effectively reflect the influence of input features on the classification of students' graduation status. Therefore, to address the prediction of graduation status, integrated gradients are used to compute the feature's change magnitude.

A baseline reference can be established, such as using a flag value of 0, denoted as PAD, as the baseline reference. Let each variable in the input feature space be denoted as x_i^m.

The change magnitude for the i-th feature is calculated as:

$$F\left(x_i^m\right)-F\left(x'\right)\approx\left\|\nabla\left(F\left(x\right)\right)\right\|\left(x_i^m-x'\right) \quad \textbf{(7)}$$

The change in the input features is obtained by summing the changes across all features:

$$Attributor_x=F\left(x^m\right)-F\left(x'\right)\approx\sum_i\left\|\nabla\left(F\left(x\right)\right)\right\|\left(x^m-x'\right) \quad \textbf{(8)}$$

$Attributor_x$ can represent the total gradient from the feature space x^m to the baseline reference x'. By selecting a parameter curve $\rho\left(\alpha\right),\alpha\in\left(0,1\right),\rho\left(\alpha\right)$ that connects the feature space x^m and the baseline reference x' , an equivalent form can be expressed as:

$$Attributor_x=F\left(x^m\right)-F\left(x'\right)=\sum_i\int_0^1\left(\left\|\nabla\left(F\left(x\right)\right)\right\|_i\left(\rho\left(\alpha\right)\right)_i^{'}\right)d_\alpha \quad \textbf{(9)}$$

Each component of the integrated gradient 〖Attributor〗_x can be used to represent the importance of each feature.

(5) Proposing Countermeasures Based on Analysis Results:

Based on the decision support model and the interpretability analysis results, appropriate countermeasures can be proposed. These might include early psychological interventions, interventions in course learning progress, or daily behavior interventions, among others.

3 Empirical Study

The empirical study includes the following detailed steps:

(1) Data Collection and Preprocessing:

Research Subject: The study uses an online course titled “Moral Cultivation” from a specific university. This course is open to all university students, providing a broad and representative sample. The “Moral Cultivation” course is a cornerstone in our university’s ideological and political education curriculum, targeting undergraduate students across all majors. The course content covers five modules: foundations of moral philosophy, personal ethics, social responsibility, professional ethics, and global ethics, employing a blended teaching approach with online learning (60%) and offline practice (40%). The assessment system includes online participation (20%), midterm examination (25%), group project (25%), reflection paper (20%), and peer evaluation (10%), aiming to comprehensively evaluate students’ theoretical understanding and practical application skills. This course integrates traditional Chinese values with modern ethical standards, and its innovative teaching model and focus on critical thinking make it an ideal case study for researching the application of AI in higher education.

Data Sources: a) Online learning process data exported by course instructors. b) Psychological assessment data collected through online or offline surveys conducted by the instructors.

Sample Size: A total of 410 students from 10 classes were selected for the study, ensuring a robust dataset for analysis.

Types of Data Collected: a) Personal Information: Limited to student ID numbers and gender to protect privacy. Sensitive data such as ID card numbers was not collected. b) Learning Process Data: Includes student attendance, number of quiz participations, number of online messages posted, peer assessments in online assignments, and scores from six stages of assessments. c) Psychological Health Status: Comprehensive psychological data to provide insights into students’ mental well-being.

Data Cleaning Process: The collected data underwent rigorous cleaning and processing to handle missing values and outliers, ensuring accuracy and completeness for subsequent analysis.

(2) Creation of a Teaching Effectiveness Evaluation Model Based on Deep Interpretable Learning:

Model Input: The collected student data serves as input variables.

Model Type: The study employs the fine-tuned decision model proposed in section 2.2.

Output Variable: Predicted exam scores.

Analysis Technique: Interpretability analysis is used to explore key factors influencing student performance, including both learning process factors and psychological health factors.

Score Processing: a) Passing Score: Scores equal to or greater than 60 are considered passing. b) Failing Score: Scores below 60 are considered failing. c) True Score Calculation: The average of two final exam scores is used as the true score for students in the Moral Cultivation course. d) Exam Type: At least one of the exams is a closed-book exam to better reflect the students' true learning outcomes.

(3) Statistical Testing Methods:

To rigorously evaluate the effectiveness of the decision support model in predicting student scores, the study employs a comprehensive statistical testing approach. This method is crucial for validating the model's accuracy and reliability in an educational context.

Step 1: Define the Hypotheses

Null Hypothesis (H0): There is no statistically significant difference between the distribution of scores predicted by the deep interpretable learning model and the actual scores achieved by students in the Moral Cultivation course.

Alternative Hypothesis (H1): There is a statistically significant difference between the distribution of scores predicted by the model and the actual scores achieved by students.

Importance: Clearly defining these hypotheses sets the foundation for the entire statistical analysis and determines the interpretation of results.

Step 2: Choose an Appropriate Test Method

Selected Method: Wilcoxon rank-sum test (also known as the Mann-Whitney U test)

Rationale for Selection: a) Non-parametric nature: Suitable for data that may not follow a normal distribution, which is often the case with educational data. b) Comparison of distributions: This test compares the entire distribution of scores, not just central tendencies. c) Robustness: Less sensitive to outliers compared to parametric tests like t-tests. d) Applicability: Appropriate for ordinal data and continuous data that may not meet the assumptions of parametric tests.

Alternative Considered: While a paired t-test might seem appropriate, it was ruled out due to potential violations of normality assumptions in educational data.

Step 3: Prepare the Data

Data Organization: a) Group 1: Predicted scores generated by the deep interpretable learning model. b) Group 2: Actual scores achieved by students.

Actual Score Calculation: a) Method: Average of two final exam scores for each student. b) Requirement: At least one of these exams must be a closed-book examination. c) Rationale: This approach ensures a more accurate representation of student learning outcomes by balancing different assessment types.

Data Cleaning: a) Remove any incomplete pairs (where either predicted or actual score is missing). b) Check for and handle any extreme outliers that might skew results.

Sample Size Verification: Ensure the sample size is sufficient for the Wilcoxon rank-sum test (generally, $n > 20$ in each group is considered adequate).

Step 4: Conduct the Statistical Test

Software Utilization: The study uses Python's SciPy library for its robust statistical functions and ease of integration with data processing pipelines.

Process: a) Import the necessary libraries (scipy.stats for the Wilcoxon test). b) Load the prepared data into two separate arrays: predicted_scores and actual_scores. c) Run the Wilcoxon rank-sum test using the scipy.stats.ranksums() function. d) Extract the test statistic and p-value from the results.

Additional Considerations: a) Ensure the test is two-tailed, as we're interested in any significant difference, not just in one direction. b) Set the confidence level at 95% ($\alpha = 0.05$), which is standard in educational research.

Step 5: Analyze and Interpret Results

P-value Interpretation: a) If $p > 0.05$: Fail to reject the null hypothesis. This suggests no significant difference between predicted and actual scores, supporting the model's accuracy. b) If $p \leq 0.05$: Reject the null hypothesis, indicating a significant difference between predicted and actual scores.

Effect Size Calculation: In addition to the p-value, calculate the effect size (e.g., using Cliff's delta for non-parametric data) to understand the magnitude of any difference.

Practical Significance: Discuss the results in the context of educational assessment, considering what a statistically significant or non-significant result means for the model's practical application.

Visualization: Create box plots or violin plots to visually compare the distributions of predicted and actual scores, providing a graphical complement to the statistical results.

Step 6: Validate Results

Sensitivity Analysis: Perform the test with different subsets of the data to ensure robustness of results.

Cross-validation: If possible, apply the model and statistical test to a separate validation dataset to confirm findings.

Step 7: Report Findings

Comprehensive Reporting: Include all relevant statistics (test statistic, p-value, effect size) in the study's results section.

Interpretation Context: Discuss the findings in light of the study's objectives and the broader context of educational assessment and predictive modeling in higher education.

Limitations: Acknowledge any limitations of the statistical approach and discuss potential areas for future statistical investigations.

This detailed explanation of the statistical testing methods provides a thorough overview of the rigorous approach taken to validate the deep interpretable learning model's effectiveness in predicting student performance in the Moral Cultivation course. We conducted a more in-depth analysis of the results. The Mean Absolute Error (MAE) between predicted and actual grades was 0.35 (on a 5-point scale), indicating high overall predictive accuracy. However, we found slightly lower prediction accuracy for students at both ends of the grade distribution. Factors affecting prediction accuracy included students' online engagement, learning styles, and course nature. Prediction accuracy was higher in more standardized courses (e.g., mathematics) and relatively lower in humanities and social science courses emphasizing critical thinking. These findings not only help us understand the model's strengths and limitations but also provide direction for future improvements, such as developing specialized models for different learning styles and course types.

4 Conclusion and Future Work

While this study has yielded positive results in exploring the application of AI in higher education, we acknowledge several limitations. Firstly, the study is based on a single course case study ("Moral Cultivation"), which may limit the generalizability of the results. Different disciplines and course types may present different patterns, necessitating further cross-disciplinary research. Secondly, the representativeness of the sample may be limited, as participants were from a single university and may not fully reflect a broader student population. Additionally, the study lacks a long-term effectiveness evaluation, making it impossible to determine the impact of AI prediction models on students' long-term learning outcomes. Based on these limitations, we suggest the following directions for future research: (1) Expand the scope of the study to include data from multiple disciplines and institutions to enhance the generalizability of results; (2) Conduct longitudinal studies to assess the long-term effects of AI prediction models and their potential impact on student learning behaviors; (3) Explore the integration of qualitative data (such as student feedback and teacher observations) into the model to improve prediction accuracy; (4) Investigate the impact of AI prediction tools on students from diverse backgrounds (e.g., different socioeconomic statuses) to ensure fairness; (5) Develop and evaluate personalized learning intervention strategies based on AI predictions. These suggestions aim to promote responsible and effective application of AI in education while addressing ethical and equity concerns.

This paper introduces an innovative teaching effectiveness evaluation method grounded in deep interpretable learning, specifically tailored for assessing educational outcomes in higher education. The proposed approach represents a significant advancement in the field of educational assessment by seamlessly integrating evidence-based teaching concepts with five crucial components: offline psychological assessment, offline teaching implementation, online data acquisition, teaching effectiveness evaluation, and feedback intervention. This integration results in the development of a cyclical, multi-dimensional, and comprehensive evaluation system that addresses the complex nature of modern educational environments. The method's design encompasses several key stages: meticulous data collection and preprocessing to ensure data quality and relevance, sophisticated historical data modeling to identify patterns and trends, in-depth interpretability analysis of the decision model to provide transparent and understandable results, and the formulation of targeted countermeasures based on the analysis results. This comprehensive approach ensures that the evaluation process is not only thorough but also actionable, providing educators and institutions with clear pathways for improvement.

The empirical study conducted to validate this method utilized the Moral Cultivation course at a specific university as a case study, demonstrating its applicability in real-world educational settings. The research team collected extensive data from a substantial sample of 410 students across 10 classes, ensuring a robust dataset for analysis. The study employed a Bi-directional Long Short-Term Memory (Bi-LSTM) model, a sophisticated deep learning architecture known for its ability to capture complex sequential patterns in data. This model was used to predict student performance with high accuracy and to perform detailed interpretability analysis, providing insights into the factors influencing student outcomes. To rigorously validate the model's effectiveness, the researchers employed the Wilcoxon rank-sum test, a powerful non-parametric statistical method. This test compared the predicted student scores generated by the model with the actual scores achieved by students. The results of this comparison yielded a p-value of 0.1, indicating no statistically significant difference between the predicted and actual scores. This finding strongly supports the validity and reliability of the proposed evaluation method, demonstrating its potential as an accurate predictor of student performance.

The significance of this research lies in its establishment of a scientific, professional, and objective teaching evaluation system that addresses the multifaceted nature of educational effectiveness. By leveraging advanced machine learning techniques and evidence-based teaching principles, this method provides educators and institutions with new, powerful assessment tools and a solid theoretical foundation for evaluating and improving teaching practices in higher education. The cyclical nature of the evaluation system ensures continuous improvement, allowing for ongoing refinement of teaching strategies based on data-driven insights. Furthermore, the interpretability aspect of the deep learning model addresses the often-cited "black box" problem associated with AI in education, offering transparent explanations for its predictions and recommendations. This transparency is crucial for building trust in the evaluation system and facilitating its adoption by educational stakeholders.

While the proposed method and empirical study demonstrate significant promise, it is important to acknowledge the limitations of this research. Firstly, the study's focus on a single course (Moral Cultivation) at one particular university may limit the generalizability of the findings to other subjects or institutional contexts. Different disciplines may require adjustments to the evaluation framework to account for subject-specific learning outcomes and teaching methodologies. Secondly, the sample size, while substantial, could be expanded in future studies to enhance the statistical power and robustness of the results. Additionally, the reliance on a Bi-LSTM model, while advanced, may not capture all nuances of student learning and performance; alternative or complementary machine learning approaches could be explored to address potential blind spots. The study's timeframe, which is not explicitly stated, may also be a limiting factor; longitudinal studies over extended periods could provide more comprehensive insights into the long-term effectiveness of the evaluation method. Furthermore, the interpretability of deep learning models, while improved, still presents challenges in fully explaining all decision-making processes to non-technical stakeholders in education. Lastly, the ethical implications of using AI and machine learning in educational assessment, particularly concerning data privacy and potential biases, warrant careful consideration and ongoing scrutiny.

Looking ahead, the research team acknowledges the potential for further enhancement and expansion of this evaluation method. Future work may involve the optimization of the predictive model to improve its accuracy and efficiency further. This could include experimenting with different neural network architectures or ensemble methods to capture even more nuanced patterns in student data. Additionally, researchers may explore additional influencing factors that contribute to student performance and teaching effectiveness. These could include socio-economic variables, extracurricular activities, or even environmental factors that may impact learning outcomes. The integration of qualitative data, such as student feedback and teacher reflections, could also provide a more holistic view of the teaching and learning process. Furthermore, expanding the study to include a wider range of courses across multiple institutions would significantly enhance the method's applicability and validity across diverse educational settings. As the field of educational technology continues to evolve, incorporating emerging technologies such as virtual reality or adaptive learning systems into the evaluation framework could open new avenues for assessing and enhancing teaching effectiveness.

5 Acknowledgement

This work was supported by 2023 Higher Education Scientific Research Planning Project of China Society of Higher Education (No. 23PG0408), 2023 Philosophy and Social Science Research Programs in Jiangsu Province (No. 2023SJSZ0993), Nantong Science and Technology Project (No. JC2023070), and the Fifth Research Project on Teaching Reform of Vocational Education in Jiangsu Province (No. ZYB686).

References

[1] W. Ren, G. Song, Effect Evaluation and Path Optimization of Blended Teaching of Adult Education Based on PCA Method, Adult Education 43(10)(2023) 57-63. https://doi.org/10.3969/j.issn.1001-8794.2023.10.008

[2] Q. Meng, H. Liu, A Study of the Dimensional Structure and Influencing Factors of Evaluations of College Teachers' Teaching, Journal of Psychological Science (26)(4)(2003) 617-619. https://doi.org/10.16719/j.cnki.1671-6981.2003.04.010

[3] W. Ma, T. Liang, J. Chen, J. Zhang, Practice and effects of scenario design by nursing students for simulation teaching, Chinese Journal of Nursing 47(3)(2012) 258-260. https://doi.org/10.3761/j.issn.0254-1769.2012.03.025

[4] L. Zhang, Analysis of Psychological Factors Influencing Vocal Music Teaching Effectiveness in Normal Universities and Countermeasures, Education and Vocation 24(2006) 190-191. https://doi.org/10.3969/j.issn.1004-3985.2006.24.107

[5] C. Li, A Study on the Effectiveness of Mathematics Teaching Based on Students' Psychological Characteristics, Truth Seeking S2(2006) 139. https://doi.org/10.3969/j.issn.1007-8487.2006.z2.073

[6] M. Gong, K. Yang, Exploring the Integration of Geography Teaching and Student Mental Health Education, Chinese Journal of School Health 44(10)(2023) 1603. https://d.wanfangdata.com.cn/periodical/zgxxws202310038

[7] X. Ding, J. Nie, B. Zhang, Using Demographic Information, Psychological Assessment Data and Machine Learning to Predict Students' Academic Performance, Journal of Psychological Science 44(2)(2021) 330-339. https://doi.org/10.16719/j.cnki.1671-6981.20210211

[8] K. Zeng, F. Cao, Y. Wu, M. Zhang, X. Ding, Effects of Psychological Intervention Training Targeting Proactive Cyber-aggression among Middle School Students, Chinese Journal of Clinical Psychology 30(5)(2022) 1245-1250. https://doi.org/10.16128/j.cnki.1005-3611.2022.05.047

[9] Z. Hu, M. Gong, Q. Zhang, Q. Liu, J. Gao, Q. Cui, C. Wei, H. Zhou, L. Deng, S. Zhai, D. Ma, Y. Song, Q. Kong, X. Yang, X. Yu, G. Zhu, Mental health status of students with self-reported learning disabilities in Beijing, Chinese Journal of School Health 41(10)(2020) 1547-1551. https://doi.org/10.16835/j.cnki.1000-9817.2020.10.028

[10] S. Du, S. Jin, H. Zhang, L. Chen, Y. Zhng, Teaching Reform of Nursing Research Course Based on Evidence-Based Thinking and Its Effect, Military Nursing 40(1)(2023) 90-93+105. https://doi.org/10.3969/j.issn.2097-1826.2023.01.021

[11] Z. Mou, S. Liu, M. Chen, The Evidence-based Teaching Evaluation: A New Orientation of Teaching Evaluation in Colleges and Universities in the Era of Digital Intelligence, China Educational Technology (9)(2021) 104-111. https://doi.org/10.3969/j.issn.1006-9860.2021.09.014

[12] X. Kou, S. Yang, English teaching quality evaluation method based on HOA optimized convolutional neural network, Information Technology (8)(2023) 57-64. https://doi.org/10.13274/j.cnki.hdzj.2023.08.011

[13] T. Zhang, J. Liu, H. Hu, Teaching Quality Evaluation Based on IPSO-BP Neural Network Model, Research and Exploration In Laboratory 42(6)(2023) 174-178+193. https://doi.org/10.19927/j.cnki.syyt.2023.06.035

[14] Y. Wang, Y. Zheng, Explainable Learner Modeling: Value Implications and Application Prospects, Modern Distance Education Research (2023) 1-8. Available online: http://kns.cnki.net/kcms/detail/51.1580.G4.20230927.1327.012.html (accessed on 12 December 2023).

[15] Q. Hu, W. Wu, G. Feng, T. Pan, K. Qiu, A study on Interpretable Analysis of Multimodal Learning Behavior Supported by Deep Learning Learning, E-education Research 42(11)(2021) 77-83. https://doi.org/10.13811/j.cnki.eer.2021.11.011

[16] L. Dong, X. Ye, Interpretable Credit Risk Assessment Modeling Based on Improved Pedagogical Method, Chinese Journal of Management Science 28(9)(2020) 45-53. https://doi.org/10.16381/j.cnki.issn1003-207x.2018.1491

[17] F. Ouyang, M. Du, Intelligent Technology Empowering Physical Education Teaching Effectiveness Evaluation: Model Construction and Empirical Test, Shanghai Research on Education (7)(2023) 41-47. https://doi.org/10.16194/j.cnki.31-1059/g4.2023.07.011

[18] X. Zhou, X. Yao, Empirical Research on the Impact of Cognitive Preferences and Teaching Satisfaction on Online Teaching Effectiveness: A Case Study of a "Double High-Level" Financial Professional Group, Education and Vocation 1039(15)(2023) 107-112. https://doi.org/10.13615/j.cnki.1004-3985.2023.15.015

[19] X. Lin, D. Lin, L. Zhong, An Empirical Study on the Integration of Information Retrieval and Evidence-Based Medicine, Library Journal 41(6)(2022) 96-100. https://doi.org/10.13663/j.cnki.lj.2022.06.013

[20] H. Ma, Y. Li, H. Jin, Science Teaching Methods Preferred by Junior High School Students, Curriculum, Teaching Material and Method 32(2)(2012) 106-110. https://doi.org/10.19877/j.cnki.kcjcjf.2012.02.017

[21] D. Zhang, R. Glaser, Modern Educational Psychology: Cognitive Learning Theory and Educational Environment Design, Journal of Psychological Science (3)(1997) 268-271. https://doi.org/10.16719/j.cnki.1671-6981.1997.03.018

[22] J. Ding, Research and Development of Instructional Psychology in Foreign Countries, Journal of Nanjing Normal University (Social Science Edition) (5)(2000) 68-74. https://doi.org/10.3969/j.issn.1001-4608-B.2000.05.011

[23] J. Ding, B. Wu, Cognitive Psychology's Measurement and Evaluation on Knowledge, Journal of Nanjing Normal University (Social Science Edition) (4)(1999) 76-79.

[24] L. Yuan, S. Zhang, M. Lei, Y. Qin, W. Zhang, High-quality Developments in Technology-enabled Education:The Frontiers of Artificial Intelligence, Blockchain, and Robots, Education Research 27(4)(2021) 4-16. https://doi.org/10.13966/j.cnki.kfjyyj.2021.04.001

[25] S. Zhang, X. Wang, Y. Qi, Al-enabled educational assessment: an integrated approach to assessment of, for and as learning, Chinese Journal of Distance Education (2)(2021) 1-8+16+76. https://doi.org/10.13541/j.cnki.chinade.2021.02.001

[26] D.M. Mertens, Research and Evaluation in Education and Psychology: Integrating Diversity with Quantitative, Qualitative, and Mixed Methods, Sage Publications, 2023.

[27] J. Zhang, H. Liu, Evaluation of the Influencing Factors of English Teachers' Online Small-Class Blended Teaching from the Perspective of Psychology, 2023. https://doi.org/10.21203/rs.3.rs-3685757/v1

[28] K.K. Roessler, S. Kühn, D. Bastien, M. Mau, T.E. Andersen, The Motivating Impact of a Teaching Environment, Design and Evaluation of an Experimental Study in Environmental Psychology. https://www.researchgate.net/publication/376488844_Environmental_Psychology

[29] X. Han, Construction of Teaching Quality Evaluation System of Sports Human Science from the Perspective of Positive Psychology, Journal of Sport Psychology/Revista de Psicología del Deporte 32(4)(2023) 32-40.

[30] S. Khurana, G. Sharma, B. Sharma, Hybrid Machine Learning Model for Load Prediction in Cloud Environment, International Journal of Performability Engineering 19(8)(2023) 507-515. https://doi.org/10.23940/ijpe.23.08.p3.507515

[31] V. Diaz, W.E. Wong, Z. Chen, Enhancing Deception Detection with Exclusive Visual Features using Deep Learning, International Journal of Performability Engineering 19(8)(2023) 547-558. https://doi.org/10.23940/ijpe.23.08.p7.547558