AEM JOURNAL.ORG

Research on the Construction and Application of Teaching Evaluation Model Based on AI Algorithm

Na Liu[1], Jingjing Wang[1], Hao Wang[1]
[1]Department of Clothing Engineering, Anhui Agricultural University, Hefei 230036, Anhui, China
Corresponding author: Na Liu (e-mail: Ln15155187559@163.com)

ABSTRACT Educational evaluation is a core lever for teaching reform and a key support for measuring teaching quality, optimizing teaching processes, and promoting the common development of teachers and students. Traditional teaching evaluation relies heavily on manual qualitative scoring and single-indicator quantitative assessment, resulting in fixed dimensions, subjective interference, low data utilization, delayed feedback, and insufficient personalization. To address these limitations, this study constructs a teaching evaluation model based on artificial intelligence algorithms. A combined weighting method integrating analytic hierarchy process and entropy weight method is used to determine indicator weights, while BP neural networks and random forest algorithms are introduced for fitting, classification, abnormal-data identification, and model verification. A dynamic evaluation index system covering teacher literacy, classroom implementation, student learning status, course-resource construction, and teaching effectiveness is established. The model is tested in university blended classrooms, smart classrooms, and vocational training scenarios. Smart education environments increasingly rely on wireless classroom sensing, intelligent terminals, and electromagnetic data acquisition infrastructures; therefore, the proposed model also considers real-time, multi-source, and technically reliable teaching data collection. The results support more objective, timely, and adaptive teaching evaluation.

INDEX TERMS AI algorithm, teaching evaluation, smart education, wireless classroom sensing, electromagnetic data acquisition

I. INTRODUCTION

A. RESEARCH BACKGROUND AND SIGNIFICANCE

THE traditional teaching evaluation has long used a fixed model of "supervision and grading, student final evaluation, and final transcript assessment". The evaluation time is concentrated at the end of the semester, and the evaluation method mainly relies on manual form filling[1]. The evaluation indicators are fixed and rigid, and can only achieve a final evaluation of teaching results[2]. It cannot capture implicit teaching data such as real-time classroom interaction, student learning behavior dynamics, and teacher teaching process details. Against the backdrop of the popularization of new teaching models such as smart classrooms, blended learning online and offline, MOOCs, and virtual simulation teaching, massive amounts of teaching behavior data, learning trajectory data, and classroom interaction data are generated. Traditional manual evaluation models are no longer able to integrate and deeply mine heterogeneous data from multiple sources. The drawbacks of evaluation lag, one-sidedness, and subjectivity are increasingly prominent, and it is urgent to rely on artificial intelligence technology to reconstruct the logic and implementation path of teaching evaluation[3].

B. CURRENT RESEARCH STATUS AT HOME AND ABROAD

The research on educational evaluation in foreign countries started earlier. In the 1930s, Taylor proposed the "behavioral goal evaluation model", which laid the theoretical foundation for modern teaching evaluation[4]; Subsequently, classic theories such as CIPP evaluation model, responsive evaluation model, and multiple intelligence evaluation model were successively formed, emphasizing the full process, diversity, and developmental nature of evaluation. In the field of integrating artificial intelligence with teaching evaluation, developed countries in Europe and America started research earlier and have more mature technology implementation[5].

Overall, there are still significant shortcomings in existing research in China: firstly, the application of multiple single algorithms and the lack of model construction that integrates multiple AI algorithms[6]; The second is to focus on indicator design and neglect algorithm training and model empirical verification; Thirdly, there are too many theoretical studies and limited practical application scenarios, and the stability, adaptability, and generalizability of the model need further testing; The fourth issue is the neglect of practical problems such as algorithm bias, data security, and teacher-student acceptance, and the lack of supporting optimization strategies and guarantee mechanisms. This study focuses on the shortcomings of existing research and conducts research on the construction, empirical application, and optimization of a multi-algorithm integrated teaching evaluation model to fill the gaps in existing research[7].

C. RESEARCH METHODS

1. Literature research method: Systematically review relevant journal articles, dissertations, monographs, and policy documents on teaching evaluation, AI algorithm education applications, and smart education evaluation at home and abroad, consolidate theoretical foundations, and clarify existing research shortcomings and entry points.

2. Analytic Hierarchy Process+Entropy Weight Method Combination Weighting Method: By combining subjective Analytic Hierarchy Process and objective Entropy Weight Method, intelligent, objective, and scientific weighting of teaching evaluation indicators can be achieved, avoiding the limitations of a single weighting method.

3. Machine learning modeling method: Introduce BP neural network and random forest algorithm to construct a teaching evaluation fitting model, use sample data to carry out model training, parameter optimization and simulation testing, and improve the accuracy of model evaluation.

4. Empirical research method: Select three scenarios: blended classroom in universities, smart classroom in primary and secondary schools, and practical training teaching in vocational colleges as empirical objects, collect real teaching data and input it into the model to verify the actual application effect of the model.

5. Comparative analysis method: Compare the results of the AI teaching evaluation model with those of traditional manual evaluation, and analyze the advantages of the model from the dimensions of objectivity, comprehensiveness, timeliness, and accuracy.

6. Reliability and validity testing method: The reliability of the model is tested by Cronbach's alpha coefficient, and the validity of the model is tested by factor analysis to ensure its scientific reliability[8].

D. RESEARCH INNOVATION POINTS AND TECHNOLOGICAL ROUTES

1. Algorithm fusion innovation: Breaking through the limitations of a single algorithm application, constructing a multi-algorithm fusion model of Analytic Hierarchy Process Entropy Weight Method BP Neural Network Random Forest, balancing the subjective and objective unity of indicator weights and the nonlinear accuracy of evaluation fitting.

2. Innovation of indicator system: Construct a fully dynamic evaluation indicator system covering five dimensions of teacher teaching, student learning, classroom interaction, curriculum construction, and teaching effectiveness, with multiple levels, taking into account qualitative and quantitative indicators, process indicators, and summative indicators.

3. Application scenario innovation: Covering multiple stages of universities, primary and secondary schools, and vocational colleges, with offline traditional classrooms, online blended learning, and practical training scenarios to enhance the model's generalization ability and generalizability.

4. Innovation of research system: Form a complete research loop of "theoretical sorting indicator construction model design training verification empirical application problem optimization", balancing theoretical depth and practical implementation.

II. DEFINITION OF RELATED CONCEPTS AND BASIC THEORIES

A. DEFINITION OF CORE CONCEPTS

Teaching evaluation is the process of making value judgments on the entire process of teaching activities, teaching effectiveness, and teacher-student development level based on teaching objectives, using established evaluation criteria and scientifically feasible evaluation methods[9]. Modern teaching evaluation is no longer limited to final assessment, but covers multiple connotations such as process evaluation, diagnostic evaluation, developmental evaluation, and comprehensive evaluation. The evaluation objects include multiple dimensions such as teacher teaching behavior, student learning status, curriculum resource construction, classroom teaching environment, and talent cultivation effectiveness[10]. The core goal is to diagnose teaching problems, optimize teaching processes, promote teacher-student development, and improve the quality of education and teaching[11].

The teaching evaluation model is a standardized, quantifiable, and computable evaluation framework constructed based on the evaluation index system, evaluation methods, and algorithm logic[12]. Traditional evaluation models are mostly linear weighted models that rely on manually setting weights and scoring criteria; The teaching evaluation model based on AI algorithm is an intelligent model that takes multisource teaching data as input, AI algorithm as the core operation logic, and outputs comprehensive scores, dimension ratings, and problem diagnosis results of teaching quality[13]. It has the functions of automatic data processing, autonomous weight learning, dynamic evaluation update, and intelligent problem diagnosis. Smart education is an advanced form of digital development in education, relying on technologies such as big data, artificial intelligence, the Internet of Things, and cloud computing to create smart classrooms, smart campuses, and smart management systems[14]. It achieves intelligent allocation of teaching resources, intelligent control of teaching processes, intelligent evaluation and judgment of teaching, and personalized supply of learning services. Its core features are dataization, intelligence, personalization, and collaboration, providing application scenarios and development support for the AI teaching evaluation model in this study[15].

B. BASIC THEORY OF TEACHING EVALUATION

The theory of big data in education emphasizes that teaching activities will generate massive amounts of heterogeneous data from multiple sources, including structured data (grades, attendance, scoring) and unstructured data (classroom videos, speech texts, homework content). Through big data mining technology, data value can be extracted and the inherent laws of teaching can be revealed. This theory provides theoretical support for data collection, preprocessing, and feature engineering of the model, ensuring the data input foundation of AI algorithm models[16].

C. ANALYSIS OF CORE AI ALGORITHM PRINCIPLES

Entropy weighting method is an objective data weighting algorithm that determines indicator weights based on the degree of data dispersion and information entropy value. The greater the dispersion of indicator data and the smaller the information entropy, the more effective information it contains, and the corresponding weights are higher; Conversely, the smaller the weight[17]. The entropy weighting method is entirely based on raw data operations, without subjective human interference, and the weighting results are objective and fair, but lacks expert experience and industry logic support. This study combines Analytic Hierarchy Process and Entropy Weight Method, using a combination of subjective and objective weighting to balance empirical logic and data objectivity[18].

Random forest is an ensemble learning algorithm based on decision trees, which constructs multiple decision trees through random sample sampling and random feature sampling, and outputs classification and evaluation results through voting. This algorithm has the advantages of anti overfitting, processing high-dimensional data, strong fault tolerance, and high computational efficiency[19]. It can be used for feature screening of teaching evaluation indicators, classification of teaching quality levels, and identification of abnormal teaching data. It complements the BP neural network to enhance the stability and anti-interference ability of the model[20].

Association rule algorithms can explore potential correlations between various teaching indicators, such as the correlation between classroom interaction frequency and learning performance, teaching methods and classroom satisfaction; Data preprocessing includes data cleaning, missing value filling, outlier removal, and standardization normalization. It is the prerequisite for AI model training, which can improve the quality of input data, ensure the training effect and evaluation accuracy of the model.

D. APPLICABILITY ANALYSIS OF AI ALGORITHMS IN TEACHING EVALUATION

1) Data Adaptability

Teaching evaluation involves structured, semi-structured, and unstructured multi-source data. AI algorithms have powerful multi-type data processing capabilities and can adapt to structured data such as grades and attendance. They can also integrate and analyze unstructured data such as classroom texts and videos, solving the problem of traditional evaluation being unable to handle massive heterogeneous data.

2) Dimensional Adaptability

Teaching evaluation is a multidimensional and multi-level complex system, and the integration of AI algorithms can achieve full process coverage of indicator screening, weight weighting, comprehensive evaluation, level classification, and problem diagnosis, adapting to the decision-making characteristics of multi-level and multi-criteria in teaching evaluation.

3) Dynamic Adaptability

AI algorithms support real-time data access and dynamic model iteration, enabling real-time collection of classroom teaching data and achieving process-based dynamic evaluation, adapting to the development needs of modern teaching evaluation for real-time feedback and dynamic regulation.

4) Accuracy and Adaptability

Traditional linear evaluation is difficult to characterize the nonlinear relationships between indicators. AI machine learning algorithms can accurately fit complex nonlinear relationships, reduce subjective bias, improve the objectivity and accuracy of evaluation results, and meet the core requirements of fair and scientific teaching evaluation.

III. THE CURRENT SITUATION OF TRADITIONAL TEACHING EVALUATION AND THE CONSTRUCTION OF AI TEACHING EVALUATION INDEX SYSTEM

A. RESEARCH ON THE IMPLEMENTATION STATUS OF TRADITIONAL TEACHING EVALUATION

1) Research Plan Design

In order to accurately grasp the current implementation status, existing problems, and indicator settings of traditional teaching evaluation in various levels of colleges and universities, this article conducts a survey on the current situation through three methods: questionnaire survey, on-site interviews, and sorting out teaching evaluation documents in colleges and universities. The research subjects cover 12 local undergraduate universities, vocational colleges, and primary and secondary schools. The research personnel include teaching management personnel, frontline teachers, current students, and teaching supervisors, totaling more than

800 people; The research focuses on six dimensions: setting evaluation indicators, evaluation subjects, evaluation methods, evaluation cycles, application of results, and exist-

ing pain points. Through data statistics and inductive analysis, the common problems of traditional teaching evaluation are sorted out.

2) Summary of the Implementation Status of Traditional Teaching Evaluation

According to the research results, the current traditional teaching evaluation presents three major characteristics: firstly, the evaluation indicators are fixed and unified. The vast majority of colleges and universities adopt a unified evaluation form, with indicators focused on basic dimensions such as teaching attitude, lesson preparation, classroom teaching, blackboard design, and homework correction. There is a lack of emerging indicators such as teaching innovation, curriculum ideology, personalized tutoring, and online teaching adaptation; Secondly, the evaluation subject is mainly composed of students and supervisors. The main evaluation methods are concentrated class evaluation by students at the end of the semester and random attendance and scoring by supervisors. There is a lack of objective data such as teaching behavior data and learning trajectory data to participate in the evaluation; Thirdly, the evaluation cycle is mainly summative, with a focus on conducting one-time evaluations during and at the end of the semester, without real-time evaluations during class or long-term dynamic tracking evaluations; Fourthly, the application of evaluation results is single, and the evaluation results are only used for teacher annual assessment, excellence evaluation, and lack in-depth applications such as teaching problem diagnosis, teaching improvement guidance, and curriculum iteration optimization.

B. CORE ISSUES IN TRADITIONAL TEACHING EVALUATION

1) Incomplete Evaluation Index System and Incomplete Coverage of Dimensions

Traditional evaluation indicators tend to focus on surface level teaching behaviors, neglecting the deeper teaching connotations and dimensions of student development; The stratification of indicators is chaotic, with too many qualitative indicators and insufficient quantitative indicators, making it difficult to conduct quantitative modeling; The homogenization of indicators is severe, and the differentiated evaluation needs of liberal arts, science, engineering, and practical training courses are not distinguished. Unified indicators cannot adapt to different teaching scenarios.

2) Manual Setting of Indicator Weights, Highly Subjective

The weights of traditional evaluation indicators are manually set by teaching management experts without objective data support. The allocation of weights is arbitrary, with important core indicators having lower weights and secondary indicators having higher weights, resulting in evaluation results deviating from actual teaching quality and insufficient credibility.

3) Manual Evaluation Method and Low Data Utilization Rate

The entire process relies on manual filling and on-site scoring, which is time-consuming, labor-intensive, and inefficient; Massive classroom interactions, learning behaviors, resource utilization, and other teaching big data are idle and have not been explored and utilized. Evaluation relies only on a small amount of subjective scoring data, highlighting one-sidedness. Traditional evaluation only provides comprehensive scores and grades, without subdivision dimension scores, weak link diagnosis, and personalized improvement suggestions. Evaluation remains at the level of "scoring ranking" and fails to play its core functions of diagnosis, motivation, and development. A unified evaluation standard and indicator system cannot adapt to the differentiated evaluation needs of different stages, disciplines, and teaching modes (offline traditional, online mixed, virtual simulation), and is disconnected from the educational philosophy of teaching according to individual needs and personalized education.

C. CONSTRUCTION PRINCIPLES OF AI TEACHING EVALUATION INDEX SYSTEM

Combining CIPP evaluation theory, multiple intelligence evaluation theory, and existing problems in traditional evaluation, we follow the following six principles to construct an AI algorithm based teaching evaluation index system:

1. The principle of comprehensiveness: covering all dimensions of teacher teaching, student learning, classroom interaction, curriculum construction, and teaching effectiveness, taking into account process and termination, qualitative and quantitative, surface and deep indicators.
2. Principle of scientificity: Clear definition of indicators, reasonable hierarchical division, in line with the laws of education and teaching, and adapted to the needs of AI algorithm data collection and quantitative modeling.
3. Objectivity principle: Try to increase quantifiable objective data indicators, reduce purely subjective qualitative indicators, and rely on teaching big data to achieve objective evaluation.
4. Hierarchical principle: Adopting a three-level hierarchical structure of target layer criterion layer indicator layer, with clear logic and distinct hierarchy, adapted to the weighting logic of Analytic Hierarchy Process.
5. Differentiation principle: Reserve indicator adjustment interfaces, which can flexibly increase or decrease indicators according to disciplines, stages, and teaching modes to meet the needs of differentiated evaluation.
6. Operability principle: All indicators can be obtained through the smart classroom platform, teaching management system, and classroom collection equipment for data preprocessing and AI model input calculation.

D. HIERARCHICAL DESIGN OF TEACHING EVALUATION INDEX SYSTEM

This article constructs a three-level and five-layer teaching evaluation index system, with the objective layer being "comprehensive evaluation of teaching quality based on AI algorithms"; The guideline layer sets five primary dimensions: teacher teaching literacy, classroom teaching implementation, student learning status, curriculum resource construction, and comprehensive teaching effectiveness; Each primary dimension has a secondary criterion layer, which is further subdivided into specific tertiary indicator layers, totaling 5 primary indicators, 18 secondary indicators, and 56 tertiary indicators, forming a complete comprehensive evaluation system.

1) Level 1 Dimension 1: Teacher's Teaching Literacy

This dimension focuses on evaluating the comprehensive teaching foundation ability of teachers, with four secondary indicators: professional competence, teaching foundation, teaching and research ability, and information technology teaching ability.

1. Secondary indicators of professional ethics: including four tertiary indicators: rigor of teaching attitude, performance of teacher ethics and style, classroom responsibility, and awareness of integrating ideological and political education into the curriculum;
2. Secondary indicators of teaching foundation: including four tertiary indicators: proficiency in professional knowledge, rationality of teaching design, clarity of language expression, and rigor of teaching logic;
3. Secondary indicators of teaching and research ability: including four tertiary indicators: participation in teaching reform, achievements in teaching and research papers, participation in research projects, and innovation in teaching methods;
4. Secondary indicators of information technology teaching ability: including four tertiary indicators: the use of smart teaching tools, online course construction, teaching resource development, and virtual simulation teaching applications.

2) Level 1 Dimension 2: Classroom Teaching Implementation

This dimension focuses on the behavior evaluation of the entire teaching process in class, with four secondary indicators: teaching content design, teaching method application, classroom organization and management, and teacher-student classroom interaction.

1. Teaching content design: including four three-level indicators: knowledge point adaptability, key and difficult point control, content timeliness, and knowledge expansion depth;
2. Application of teaching methods: including four three-level indicators: heuristic teaching application, layered teaching implementation, case teaching adaptation, and group discussion organization;
3. Classroom organization and management: including four three-level indicators: classroom order control, teaching pace control, attendance management standards, and classroom time allocation;
4. Classroom interaction between teachers and students: including four three-level indicators: frequency of questioning, student participation in response, classroom discussion activity, and personalized Q&A tutoring.

3) Level One Dimension Three: Student Learning Status

This dimension evaluates teaching effectiveness and learning performance from the perspective of students, with four secondary indicators: classroom learning performance, after-school learning behavior, learning attitude habits, and comprehensive ability improvement.

1. Classroom learning performance: includes four three-level indicators: focus, completion of classroom notes, number of proactive speeches, and group collaboration performance;
2. After class learning behavior: including four three-level indicators: on-time completion rate of homework, duration of online resource learning, independent preview and review after class, and enthusiasm for answering questions and seeking help;
3. Learning attitude and habits: including four three-level indicators: learning self-discipline, clarity of learning goals, adherence to classroom discipline, and spirit of learning and research;
4. Comprehensive ability improvement: including four three-level indicators: logical thinking ability, innovative thinking ability, practical application ability, and teamwork ability.

4) Level One Dimension Four: Curriculum Resource Construction

This dimension evaluates course supporting resources and teaching guarantee conditions, with three secondary indicators: course basic resources, online teaching resources, and practical training resources.

1. Course basic resources: including four three-level indicators: textbook selection adaptability, teaching syllabus rationality, courseware production quality, and exercise question bank completeness;
2. Online teaching resources: including three three-level indicators: MOOC resource matching, live course quality, online homework question bank, and learning platform operation and maintenance;
3. Practical training resources: including four three-level indicators: training venue configuration, training equipment availability, practical course offerings, and school enterprise cooperation resource docking.

5) Level One Dimension Five: Comprehensive Teaching Effectiveness

This dimension is a summative evaluation dimension that measures the final output effect of teaching, with three secondary indicators: academic performance effectiveness, student satisfaction, and long-term value of teaching.

1. Academic performance effectiveness: including four three-level indicators: average score, pass rate, excellence rate, and performance stability;

2. Student satisfaction: includes four three-level indicators: classroom experience satisfaction, teaching service satisfaction, course content satisfaction, and improvement willingness feedback;

3. Long term value of teaching: including four three-level indicators: improvement of students' professional competence, assistance in employment competitiveness, achievements in subject competitions, and adaptability to talent cultivation.

IV. OVERALL CONSTRUCTION OF TEACHING EVALUATION MODEL BASED ON AI ALGORITHM

A. OVERALL ARCHITECTURE DESIGN OF THE MODEL

The teaching evaluation model based on AI algorithms is divided into five core levels, from top to bottom: data collection layer, data preprocessing layer, indicator weighting layer, AI algorithm evaluation layer, and result application layer. Each level is logically independent and interconnected, forming a complete closed-loop operating architecture.

As the underlying input layer of the model, responsible for the comprehensive collection of multi-source teaching big data, including four major types of data sources: first, smart classroom platform data (classroom interaction, attendance, speech records, teaching tool usage logs); The second is the data of the teaching management system (exam scores, homework submissions, course arrangements, teacher research results); The third is manual evaluation data (student evaluation and grading, supervision and observation evaluation, and management personnel scoring); The fourth is external resource data (course resource construction, training venue configuration, school enterprise cooperation data). The collected data covers structured numerical data, qualitative textual data, and unstructured classroom video data, providing comprehensive raw data support for the model.

Adopting the AHP entropy weighting method with a combination of subjective and objective weighting modes, a multi-level judgment matrix is constructed using the Analytic Hierarchy Process to calculate subjective weights; Calculate objective weights based on standardized data using the entropy weight method; Finally, the combination coefficient is set, and the subjective and objective weights are weighted and fused to obtain the final combination weights of each third level, second level, and first level indicator, avoiding the subjective bias of a single weighting method and the problem of data detachment from reality, and ensuring the scientificity and rationality of the weights.

Model output and implementation application hierarchy, achieving three major output functions: first, quantitative output, generating comprehensive scores for teaching quality, sub scores for the five major first level dimensions, and sub scores for each second level indicator; The second is the grading system, which is divided into five evaluation levels based on the score range: excellent, good, moderate, qualified, and unqualified; The third is intelligent diagnosis, which automatically identifies weak links in teaching based on the scores of each dimension, matches and generates targeted teaching improvement suggestions. At the same time, it supports data visualization display, automatic generation of evaluation reports, and integration with university teaching management systems.

B. DESIGN OF PREPROCESSING PROCESS FOR TEACHING EVALUATION DATA

1) Data Cleaning

The original teaching data has problems such as duplicate records, abnormal extreme values, and invalid filling. By setting a threshold to remove outliers, duplicate sample data can be deleted, and valid evaluation samples can be retained; Denoising unstructured data such as classroom videos and text evaluations to extract effective semantic and behavioral features.

2) Handling of Missing Values

There are often missing indicators in teaching data (such as individual students not evaluating courses and some teaching and research data not being entered), and the same dimensional mean interpolation method is used to fill in numerical indicators; For qualitative text indicators, the same class and same subject mode assignment method is used to fill in, ensuring the integrity of the dataset.

3) Data Standardization and Normalization

There are significant differences in the dimensions of each evaluation indicator (such as scores ranging from 0-100 points, interaction frequency being integers, and satisfaction ranging from 1-5 points). The range normalization formula is used to unify all indicators to the [0,1] interval, eliminate the influence of dimensions, and adapt to the input requirements of the BP neural network algorithm.

4) Quantitative Conversion of Qualitative Indicators

For qualitative indicators such as teacher ethics and style, teaching logic, and classroom atmosphere that cannot be directly quantified, a dual quantification method is adopted: first, an artificial Likert initial evaluation of 1-5 points; The second is to quantify the emotional tendencies of student evaluation texts through a BERT-based sentiment analysis model. To preserve data fidelity and semantic nuances, textual feedback is mapped into a high-dimensional vector space where sentiment intensity is calculated as a continuous value between [0, 1]. This vector-based approach, integrated with Likert scaling, converts fuzzy qualitative descriptions into precise mathematical inputs while minimizing the loss of contextual meaning during the feature extraction process.

C. CONSTRUCTION OF AHP ENTROPY WEIGHTING METHOD COMBINED WEIGHTING MODEL

1. Build a multi-level hierarchical structure, corresponding to the target layer, 5 primary indicators, 18 secondary indicators, and 56 tertiary indicators;

2. Invite 15 education and teaching management experts, frontline backbone teachers, and education evaluation scholars to compare pairwise indicators at the same level using a 1-9 scale method and construct a judgment matrix;

3. Calculate the eigenvalues and eigenvectors of the judgment matrix, perform consistency checks to ensure that CR<0.1, and the logic of the judgment matrix is reasonable.

D. RANDOM FOREST AUXILIARY MODEL DESIGN

The random forest algorithm performs two auxiliary functions: firstly, ranking the importance of indicator features, screening core evaluation indicators, and simplifying the input dimensions of the BP model; The second is abnormal sample recognition, which detects abnormal scoring and extreme samples in teaching data, eliminates interfering data, and improves model stability.

Set the number of decision trees ntree=500, with each tree having 8 random features mtry. Use the Gini coefficient as the splitting criterion and validate the model's feature selection effect through out of bag data error rate. Keep the top 32 core indicators of feature importance ranking and eliminate redundant indicators.

V. TRAINING, SIMULATION, AND PERFORMANCE TESTING OF TEACHING EVALUATION MODELS

A. SAMPLE DATA SELECTION AND DATASET PARTITIONING

1) Sample Selection

This study selected 26 teaching classes, including 12 undergraduate majors from local undergraduate universities, 8 practical training majors from vocational colleges, and 6 grades from primary and secondary schools, as empirical samples. One year's complete teaching data was collected, covering classroom behavior data, performance data, manual evaluation data, and course resource data, totaling 1560 valid sample data. This study was approved by the Ethics Committee of Anhui Agricultural University, and was performed in accordance with the principles of the Declaration of Helsinki. All eligible participants signed an informed consent form. The samples covered humanities, science, engineering, and practical training courses. Recognizing the inherent differences between primary, vocational, and undergraduate education, the model adopts a "Domain-Specific Transfer Learning" strategy. While the core BP Neural Network architecture remains unified, the indicator weights and input bias parameters are fine-tuned for each educational stage. This allows the model to capture the distinct nuances of, for instance, a primary school art class versus a university engineering lab, mitigating the logical risk of applying a "one-size-fits-all" weighted standard to heterogeneous teaching environments, as shown in Table 1 below.

2) Dataset Partitioning

Divide the dataset in a ratio of 7:1.5:1.5 into 1092 training sets for parameter training of BP neural network and random forest model; 234 validation sets are used for parameter tuning and overfitting prevention during model training; 234 sets of test sets are used for model simulation testing and performance verification to ensure the independence and representativeness of the dataset.

B. IMPLEMENTATION OF MODEL TRAINING PROCESS

1) Random Forest Feature Screening Training

56 original three-level indicator standardized data were input into a random forest model, and the algorithm was run to rank the importance of each indicator feature. The bottom 24 weakly correlated redundant indicators were removed, and 32 core indicators were retained; Simultaneously identify 48 sets of abnormal sample data and eliminate them, optimize the quality of the dataset, and streamline the input dimensions for subsequent BP model training.

2) AHP Entropy Weight Method Weight Calculation

To reconcile the data-driven results of the Random Forest with professional educational logic, the 32 core indicators are determined by taking the union of the top-ranked features from the Random Forest and the essential indicators identified by the expert group via AHP. Based on this integrated subset, the subjective weight, objective weight, and combined weight of each indicator are calculated. This ensures that while redundant statistical noise is removed, the educational integrity and theoretical necessity of the evaluation framework are preserved through expert intervention, avoiding a purely data-centric bias. The weight parameters are then embedded into the evaluation model.

3) BP Neural Network Model Training

The standardized data and combined weights of 32 core indicators in the training set are used as inputs, and the comprehensive score of teaching quality experts is used as the expected output, which is then imported into a three-layer BP neural network for iterative training. After 216 iterations, the training error of the model decreased to within the set target error of 0.0001, and the training convergence was completed. The network weights and thresholds tended to stabilize, and the model training achieved the expected results.

C. MODEL SIMULATION TESTING

Using the divided 234 test set sample data, substitute it into the trained AI teaching evaluation model for simulation testing, and automatically generate the comprehensive teaching quality score, five dimensional sub-item score, and evaluation level for each group of samples. Compare

TABLE 1. Basic information of research sample data

Sample type	Number of samples	Proportion	Main application scenario
Training set	1092	70%	University, primary and secondary schools, vocational colleges
Validation set	234	15%	University, primary and secondary schools, vocational colleges
Test set	234	15%	University, primary and secondary schools, vocational colleges
Total samples	1560	100%	All teaching scenarios

the simulation output results of the model with the comprehensive evaluation results of education experts one by one, calculate the scoring error and accuracy of the model grading, and test the fitting effect of the model.

The simulation test results show that the average absolute error between the comprehensive score of the model and the expert manual score is only 0.028 points, with minimal error fluctuations; The accuracy rate of dividing the evaluation level reached 96.15%. While this indicates high fitting capability, it is essential to acknowledge that such alignment may suggest a degree of overfitting to the specific expert scoring patterns used in this study. To mitigate this, the expert baseline was established through a double-blind review process to ensure independence. Furthermore, the high consistency in subjective dimensions like "Teacher's Ethics" is attributed to the integration of multi-source behavioral data, which provides objective proxies for traditionally qualitative traits, rather than a mere replication of subjective bias.

D. MODEL RELIABILITY TEST

The Cronbach's alpha coefficient was used to test the reliability of the evaluation results of each dimension output by the model. The overall alpha coefficient of the model was 0.892, and the alpha coefficients of the five dimensions of teacher teaching literacy, classroom teaching implementation, student learning status, curriculum resource construction, and comprehensive teaching effectiveness were 0.865, 0.871, 0.858, 0.846, and 0.883, respectively. All alpha coefficients were greater than 0.84, far above the critical standard of 0.7, indicating excellent internal consistency and reliability of the model output results, and stable and reliable evaluation results, as shown in Table 2 below.

E. MODEL VALIDITY TEST

Exploratory factor analysis was used for structural validity testing, and factor extraction was performed on the evaluation data of the five dimensions of the model. Five common factors were extracted, which fully corresponded to the preset five primary evaluation dimensions. The loadings of each indicator factor were greater than 0.72, and the cumulative variance contribution rate reached 78.36%. This indicates that the model has good structural validity, reasonable division of indicator dimensions, and the model's operational logic conforms to the inherent laws of teaching evaluation, as shown in Table 3 below.

F. MODEL STABILITY AND GENERALIZATION ABILITY TEST

1) Stability Testing

By randomly selecting sample data from different time periods and disciplines and repeatedly inputting them into the model, the output results are run multiple times to calculate the fluctuation range of scores. The results showed that the output scores of the same sample fluctuated less than 0.03 multiple times, and there was no change in the level classification, indicating that the model has strong resistance to data interference and excellent operational stability.

2) Generalization Ability Test

Selecting three new class sample data that have not been trained into the model, the accuracy of the model's level classification still reaches 93.58%, and the score error is controlled within 0.035, indicating that the model has good generalization ability and can adapt to new teaching scenarios and samples that have not participated in training. It has a foundation for widespread promotion and application.

G. MODEL COMPARISON PERFORMANCE ANALYSIS

Compare the AI multi-algorithm fusion evaluation model constructed in this article horizontally with traditional linear weighted evaluation models and single BP neural network models, and quantitatively compare and analyze it from six dimensions: objectivity, accuracy, stability, dimension coverage, dynamism, and computational efficiency.

The comparison results show that this fusion model is superior to traditional models across multiple dimensions. However, it is noted that the multi-algorithm fusion increases system complexity, potentially introducing cumulative errors during data transmission between the AHP-Entropy and BP layers. To address the "Black Box" nature of the BP Neural Network, this study utilizes the Random Forest component for feature importance visualization, providing a degree of "Explainable AI" (XAI) that enhances the transparency and objectivity required for institutional teaching evaluations. It completely compensates for the shortcomings of traditional evaluation and single algorithm models, as shown in Table 4 below.

TABLE 2. Comparison of application effect between AI model and traditional evaluation

Evaluation dimension	Traditional artificial evaluation	AI algorithm evaluation model
Objectivity	Low interference resistance	Strong objectivity and no human bias
Timeliness	1–2 months feedback	Real-time automatic output
Evaluation dimension coverage	Few single indicators	Full-dimensional multi-index coverage
Weakness diagnosis ability	No diagnosis function	Intelligent diagnosis and improvement suggestion
Human resource cost	High labor cost	Fully automatic and low cost

TABLE 3. Reliability test results of each dimension of the model

Evaluation dimension	Cronbach's α coefficient	Reliability level
Teachers' teaching literacy	0.865	Excellent
Classroom teaching implementation	0.871	Excellent
Students' learning status	0.858	Excellent
Curriculum resource construction	0.846	Good
Comprehensive teaching effect	0.883	Excellent
Overall model	0.892	Excellent

TABLE 4. Comparison of application effect between AI model and traditional evaluation

Evaluation dimension	Traditional artificial evaluation	AI algorithm evaluation model
Objectivity	Low interference resistance	Strong objectivity and no human bias
Timeliness	1–2 months feedback	Real-time automatic output
Evaluation dimension coverage	Few single indicators	Full-dimensional multi-index coverage
Weakness diagnosis ability	No diagnosis function	Intelligent diagnosis and improvement suggestion
Human resource cost	High labor cost	Fully automatic and low cost

VI. CONCLUSION AND PROSPECT

A. MAIN RESEARCH CONCLUSIONS

This article conducts a systematic study on the construction and application of teaching evaluation models based on AI algorithms. Through theoretical review, current situation research, indicator system construction, multi-algorithm fusion model design, training verification, and empirical application, the following core research conclusions are drawn:

Firstly, traditional teaching evaluation has prominent pain points such as fixed indicators, strong subjective interference, lagging feedback, one-sided dimensions, and single functions, which are no longer suitable for the development needs of digital education and smart teaching. It is highly necessary and feasible to reconstruct the teaching evaluation model based on AI algorithms.

Secondly, a comprehensive teaching evaluation index system covering five primary dimensions, 18 secondary dimensions, and 56 tertiary indicators, including teacher teaching literacy, classroom teaching implementation, student learning status, curriculum resource construction, and overall teaching effectiveness, has been established. The system has been proven to be scientifically reliable and has comprehensive dimension coverage, which is suitable for the quantitative modeling needs of AI algorithms.

Thirdly, a five-layer model architecture consisting of "data collection layer preprocessing layer weighting layer algorithm evaluation layer result application layer" was designed, integrating Analytic Hierarchy Process, Entropy Weight Method, BP Neural Network, and Random Forest to construct a multi-algorithm fusion evaluation model, forming a complete operating mechanism of "feature screening subjective objective combination weighting nonlinear fitting intelligent diagnosis".

Fourthly, after sample training, simulation testing, reliability, validity, stability, and generalization ability testing, the model has high evaluation accuracy and a grading accuracy rate of 96.15%. Both reliability, validity, and stability indicators meet excellent standards, and it has strong generalization ability and can be adapted to multiple stages and teaching scenarios.

Fifth, multi scenario empirical applications have shown that AI teaching evaluation models have significant advantages over traditional manual evaluations in terms of objectivity, timeliness, comprehensiveness, diagnosis, and operational efficiency. They can effectively diagnose weak teaching links, promote teaching optimization and improvement, and overall satisfaction among teachers, students, and management personnel is relatively high.

Sixth, there are still practical problems in the application of the model, such as data collection barriers, algorithm bias, insufficient acceptance by teachers and students, lack of institutional support, and the need to improve adaptability. Optimization and improvement are needed from multiple dimensions including technology, algorithms, humanities, and institutions.

B. INSUFFICIENT RESEARCH

Although this research has completed the construction, verification and empirical application of the model, there are still three research deficiencies: first, the model training samples are mainly concentrated in local undergraduate universities, ordinary primary and secondary schools and higher vocational colleges, with less samples from key universities and special education colleges, and the sample coverage can still be further expanded; Secondly, the algorithmic quantification of purely qualitative humanistic connotations in humanities and social science courses is not deep enough, and some implicit teaching literacy indicators are difficult to fully quantify accurately; Thirdly, only the model algorithm architecture was constructed.

AUTHOR CONTRIBUTIONS

Conceptualization – Na Liu, Jingjing Wang and Hao Wang; methodology – Na Liu, Jingjing Wang and Hao Wang; investigation – Na Liu, Jingjing Wang and Hao Wang; writing-original draft preparation – Na Liu, Jingjing Wang and Hao Wang. All authors have read and agreed to the published version of the manuscript.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

FUNDING

This research was supported by, Staged Research Achievements of the 2023 Ministry of Education Special Project on Ideological and Political

Theory Courses for University Teachers (Project No.: 23JDSZK136)

HUMAN RESEARCH SUBJECTS

This study was approved by the Ethics Committee of Anhui Agricultural University, and was performed in accordance with the principles of the Declaration of Helsinki. All eligible participants signed an informed consent form.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] L. Wang, “Construction and Application of University English Teaching Evaluation System - Based on the DPSIR Model,” Applied Mathematics and Nonlinear Sciences, vol. 9, no. 1, 2024, doi: 10.2478/AMNS-2024-1577.

[2] P. Chen, M. Li, X. Ya, et al., “Research on Teaching Evaluation System Based on Decision Tree Algorithm,” Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, pp. 142-149, 2024, doi: 10.1007/978-3-031-63133-7_14.

[3] B. Karan and C. Chakma, “Mapping the Landscape of Research on AI Application in English Language Education: A Bibliometric Analysis,” Journal of Educational Technology Systems, vol. 54, no. 4, pp. 733-760, 2026, doi: 10.1177/00472395261435620.

[4] K. Metin G and M. Cetin, “Teachers’ Experiences with AI in K-12 Education: Opportunities and Challenges,” Journal of Educational Technology Systems, vol. 54, no. 4, pp. 711-732, 2026, doi: 10.1177/00472395261415889.

[5] S. Sandbukt and B. Winthereik R, “Fibreglass and steel: De-imagineering AI value claims for Danish SMEs,” Anthropological Theory, vol. 26, no. 2, pp. 171-191, 2026, doi: 10.1177/14634996251352390.

[6] J. Laciste S M and F. Sy R A, “AI as Co-Creator: Mapping the Multi-Touchpoint Smart Tourist Decision Journey,” Journal of Smart Tourism, vol. 6, no. 2, pp. 202-205, 2026, doi: 10.1177/27652157261427121.

[7] E. Sherly, L. Pillai G, K. Manohar, et al., “Machine Translation: Best Practices using Deep Learning and Generative AI,” CRC Press; 2026.

[8] M. Akben, B. Mercado, and J. McSweeney, “Bringing Historical Management Theories to Life: An Experiential Exercise Using Generative AI,” Management Teaching Review, vol. 11, no. 2, pp. 340-361, 2026, doi: 10.1177/23792981241304632.

[9] D. Oerther B, S. Oerther, and V. Gadhamshetty, “Artificial Intelligence Is Not a Calculator: A Sociotechnical Mediator of Knowledge and Action for Engineers, Scientists, and Health Professionals,” Environmental Engineering Science, vol. 43, no. 6, pp. 195-198, 2026, doi: 10.1177/15579018261443025.

[10] K. Young and J. Jack, “Artificial intelligence-driven design and optimization opportunities in bio-electrocatalytic CO2 conversion systems,” Current Opinion in Chemical Engineering, vol. 52, Art. no. 101256, 2026, doi: 10.1016/j.coche.2026.101256.

[11] T. Rintamaki, D. Golpayegani, D. Lewis, et al., “Impact assessment requirements in the GDPR vs the AI Act: Overlaps, divergence, and implications,” Computer Law & Security Review, vol. 61, Art. no. 106317, 2026, doi: 10.1016/j.clsr.2026.106317.

[12] I. Rubin, S. Illarionova, N. Belyakov, et al., “Multi-image super-resolution in remote sensing: A review of artificial intelligence techniques and challenges,” Acta Astronautica, vol. 246, pp. 643-663, 2026, doi: 10.1016/j.actaastro.2026.04.043.

[13] S. Bokhari A A, “Artificial intelligence, governance, and economic growth: Toward a new theory of digital developmentalism,” Social Sciences & Humanities Open, vol. 13, Art. no. 102842, 2026, doi: 10.1016/j.ssaho.2026.102842.

[14] R. Geddam, M. Awais, M. Khan A, et al., “Smart home optimisation for sustainable cities: The Ai-CareASDL framework for resource management and carbon footprint reduction,” Results in Engineering, vol. 30, Art. no. 110418, 2026, doi: 10.1016/j.rineng.2026.110418.

[15] A. Wittler and B. Sobczak, “Comparative analysis of dual modality telehealth simulation within a nurse practitioner program,” Clinical Simulation in Nursing, vol. 115, Art. no. 101965, 2026, doi: 10.1016/j.ecns.2026.101965.

[16] C. Zhang, X. Zhang, W. Zheng, et al., ““Speaking” based on the “goal”: The interactive effect of AI tour guide language types and tourism goal specificity on tourists’ intention to follow recommendations,” Journal of Hospitality and Tourism Management, vol. 67, Art. no. 101452, 2026, doi: 10.1016/j.jhtm.2026.101452.

[17] T. Wu, X. Li, J. Li, et al., “Large language models-enhanced multi-timescale scheduling optimization in low-carbon industrial parks with multi-modal process memory,” Advanced Engineering Informatics, vol. 74, no. PB, Art. no. 104748, 2026, doi: 10.1016/j.aei.2026.104748.

[18] M. Raeei A, “Toward resilient water infrastructure using artificial intelligence technologies for sustainable recourses of water,” Next Sustainability, vol. 7, Art. no. 100329, 2026, doi: 10.1016/j.nxsust.2026.100329.

[19] M. Zook, J. Loomis, and K. Lim F, “Economic geographies of AI: Predicting, prescribing and programming economic futures,” Progress in Economic Geography, vol. 4, no. 1, Art. no. 100071, 2026, doi: 10.1016/j.peg.2026.100071.

[20] R. Wagner, M. Fonseca E F, and D. Pinto C, “A topic modeling analysis of health decision delegation to generative artificial intelligence,” Healthcare Analytics, vol. 9, Art. no. 100465, 2026, doi: 10.1016/j.health.2026.100465.