# Analysis of English Teaching Achievement Prediction Using Big Data

Yanan Bai[1], Junfang Guo [2, *]
[1]Inner Mongolia Medical University, Inner Mongolia, China, 010110
[2]Faculty of Foreign Languages, Anyang University, Anyang 455000, Henan, China
E-mail: JunfangGuo202306@outlook.com, victoriadada@163.com
[*]Corresponding author

**Keywords:** deep intervention, big data analysis, English teaching, achievement prediction

**Recieved:** June 27, 2023

*An effective data analysis tools can encourage students to enroll in online degree programs, provide access to the data and information stored in the LMS, and provide a solid foundation for decisions that will improve the administration and instruction of online degree programs. Using data analysis tools, English teachers can have a better understanding of their students' learning situations, track their development, and make better management decisions. Decision tree algorithms and methods for data analysis are first examined. Data mining has also made use of the C4.5 decision tree technique to build a prediction model for English scores. The prediction of English teaching performance is examined from the standpoint of teachers' in-depth intervention through the study of data linked to English learning, such as surveys and gathered student test scores. The findings of the survey demonstrate that the model has been tested and validated. The model outperforms the standard accuracy of 97.5% with forecast accuracy levels of 98.20%, 99.10%, 99.40%, 98.70%, and 98.90%. More practice questions and more instructor involvement in English lessons should be the focus of future studies. As a point of reference for prediction management, we provide a method for predicting students' success in English classes using big data analysis.*

*Povzetek: Predstavljena je uporaba algoritmov odločitvenih dreves in analize velikih podatkov za napoved dosežkov pri poučevanju angleščine.*

## 1 Introduction

Teaching intervention usually refers to the teaching operation mode in which teachers of various disciplines use the classroom teaching platform of their courses and use psychological knowledge to help students "understand and change themselves" [1]. Teaching intervention is conducive to students' self-adjustment at procedure of learning, improving the level of knowledge and skills, and achieving better teaching effects [2]. In the fields of information technology and educational technology, data mining, and statistical prediction models predict whether students can complete or pass courses based on variables such as effort level and grade point average [3]. Teachers provide students with effective feedback information through the acquired data, guide students to use appropriate resources to complete teaching interventions, and improve students' performance. The intensity of teacher intervention is used as a standard, divided into shallow and deep teaching interventions [4]. In-depth teaching intervention means that teachers comment on the task works submitted by students and give detailed suggestions [5].

Big data is an evolving concept in recent years. Big data provides possibilities to innovation and related advancements in technology. It is employed to describe and analyze the enormous quantities of data created in the information explosion period [6]. Extracting information that individuals are uncertain at prior is known as data mining. It remains effective for substantial volumes of unreliable, insufficient and inconsistencies [7]. Inversely processed data, determine possible patterns, forecast future developments and support decision-making are the deep data analysis approach, mainly utilizes artificial intelligence (AI), machines learning (ML) and statistical techniques [8]. Data mining and learning analytics technology are used to establish relevant systems to extract and standardize it and offer an appropriate framework for making decisions for the most effective possible implementation of the networked learning and management procedures [9]. Additionally, Zhu (2022) considered that the data-driven schools, assessment and educational improvement, prompted by arrived big data. The field of education gave rise to data mining techniques. It presented training techniques and tactics for English listening prediction using data mining technologies. This enhances listening comprehension skills and teaches pupils to employ predictive techniques. We mined and examined the listening data provided by an English skill development system. Features are selected based on data on children's listening. To predict children's listening, listening assessments from prior learners are utilized and presented. Analysis and evaluation demonstrate that data mining methods can effectively predict students' listening skills [10]. The K-means, support vector machine (SVM), decision tree, and Bayesian models are the four

primary methods used in data mining. An approach to estimating a discrete function's value is the decision tree algorithm. It is a common approach to categorization. Initially processing occurs on the data. Available standards and decision trees are produced employing induction algorithms. Then, decisions are taken to analyze the updated information. Using an ensemble of standards, decision trees are used to classify data. ID3, C4.5, The classification and Regressive trees (CART) and other algorithms are prevalent methods for decision trees. High classification accuracy, easy pattern generation and significant ability to noisy data are the three benefits of the decision tree technique. This approach for inductive reasoning has become extremely prevalent.

From the perspective of extensive intervention on big data analysis, many learning data are generated in learning process of learners are collected and analyzed. The learning characteristics and problems of learners are to enable teachers to predict the students' English achievement to judge the learning effect. Mixed data from brick-and-mortar classrooms and questionnaires are used as the basis. A DTA is used for classification. Deep intervention examines the big data analysis-based forecasting of English teaching effectiveness. Table 1 display the related work.

Table 1: Summary of related works

| References | Key features | Methodologies | Outcomes |
|---|---|---|---|
| [11] | Customized guidance for students. Enhanced student and teacher performance. | Analytic Hierarchy Process (AHP) approach to examine the variables that influence student success. K-means clustering for data processing. A genetic algorithm that optimizes initial standards and measures. For predicting scores, Back Propagation (BP) neural network was utilized. | Conducted among English major students. Outperformed standard approaches for forecast accuracy. The AHP, genetic algorithm, and BP neural network accurately forecast student scores. |
| [12] | Concentrate on Hungarian native speakers in Romania. English Learning (L3) and Romanian (L2) vocabularies. | Applied meaning in form depending on assessment of receptive vocabulary. Regression analyses were utilized. | Significant positive connections among lexical knowledge in English and Romanian, and accentedness (AA). Scores of English vocabulary emerged as greatest forecaster, providing over 30% of the variation. Indicates the lexical testing in typologically similar languages assesses the same underlying characteristic. Lexical knowledge is established to be a key explanatory aspect of accentedness, especially in multilingual. |
| [13] | Examining at the correlation between personality characteristics, global competency, and English success. | A web-based survey was conducted on 555 Chinese university students. | Extraversion and conscientiousness are favorable predictors of English success. Extraversion and openness have favorably forecast all characteristics of global capability (global perspectives, abilities and capabilities). Global perspectives are positively predicted by accepting. Global knowledge and integrity have positive associations and adverse associations with global attitudes. Neuroticism predicts global abilities and perspectives significantly. Motivation to learn English moderates' association between extraversion, English success, and global knowledge. |

| | | | |
|---|---|---|---|
| | | | - Positive correlations are greater with higher degrees of motivation. |
| [14] | Seven datasets were obtained from three institutions in Taiwan and Japan. Performance metrics for a risk detection system utilizing eight categorization methods. | Collection of data at U1, U2, and U3 institutions. Performance is evaluated using eight different categorization methods. | Factors impacting predictive performance were identified, including the values of the Spearman correlation coefficient, the number of crucial features and the number of significant categories. Higher performance of Cases 1.3 and 2.2 from the U1 and U2 datasets were attributed to a large number of significant characteristics, categories, and Spearman correlation coefficient values for significant characteristics. |
| [15] | Deep artificial neural network (DANN) utilization. | Artistic qualities taken from real learning environments Selecting "Stream Data". A comparison of logistic regression (LR) and SVM models. The inclusion of data related to evaluations and impact. | Obtains classification accuracy (84%-93%). Performance of SVM (79.95%-89.14%) and LR (79.82%-85.60%). Observe increased performance among students interested in accessing prior lecture content. Assists in developing pedagogical support frameworks for institutions. Enables higher education decision-making for sustainable education. |
| [16] | The analysis of an ordinary Chinese college student's academic achievement fluctuations. Implementation of methods to forecast performance. | Designing and implementing a questionnaire for data collecting. Chi-square test for assessing questionnaire contents and recognizing important aspects. Creation of four categorization forecasting algorithms utilizing machine learning. | Identification of major factors impacting academic success. Identifying characteristics of students failing exams. Support Vector Machine Classifier (SVC) method was discovered to provide the most reliable and optimal result. The average rate of recall obtains 82.83%, the Precision rate is 86.18 %, and the Accuracy rate is 80.96%. |
| [17] | Supervised learning methods (SVM, LR). | Several experiments employing various technologies. | The Sequential Minimal Optimization (SMO) technique outperforms LR. Enhanced accuracy achieved with SMO. Forecasting insights for future student behaviours. Classification of student achievement as excellent or poor. Identifying important factors (e.g., teacher effectiveness, student motivation) to lower dropout rates. |

# 2. Research methods and models

## 2.1 Deep intervention analysis

Intensive intervention is a teaching tool. Teacher interventions typically include all interventions that impact learning [18]. The data analysis results of the learning platform show that teachers can improve learners' intelligence levels and emotional cognition by adjusting learners' self-efficacy [19]. Teachers' personalized interventions positively affect students' academic performance and emotional cognition. According to the different intensities of teacher intervention, the concept of intervention is divided into shallow and deep teaching interventions [20]. The shallow intervention in classroom teaching is mainly direct. In-depth intervention is mainly based on indirect intervention. In-depth teacher intervention means teachers implement clear, differentiated, and targeted interventions on learners. Typical deep intervention

induction models are divided into primary, secondary, and tertiary interventions [21]. Figure 1 illustrates the analysis of the workflow of student behaviour.

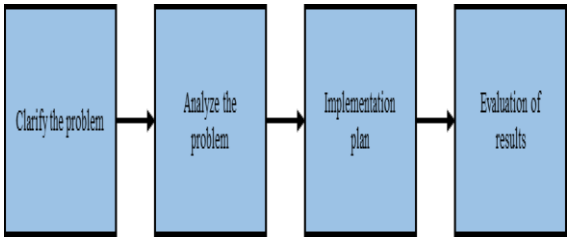



Figure 1: Process of student behavior analysis

The four phases of behavioral assessment are depicted in Figure 1 as: (1) identify the problem, define the behavioral problem in observable terms, and make a reliable record of its frequency, intensity, and duration; (2) analyze the problem, confirm the existence of the problem, identify the student variables and educational variables that help to solve the problem and formulate an appropriate plan; (3) implement the plan, execute the plan and provide corrective feedback to ensure that it is executed according to the predetermined plan; (4) problem assessment to evaluate the effect of the intervention [22]. Figure 2 depicts the flow of the student deep intervention model.

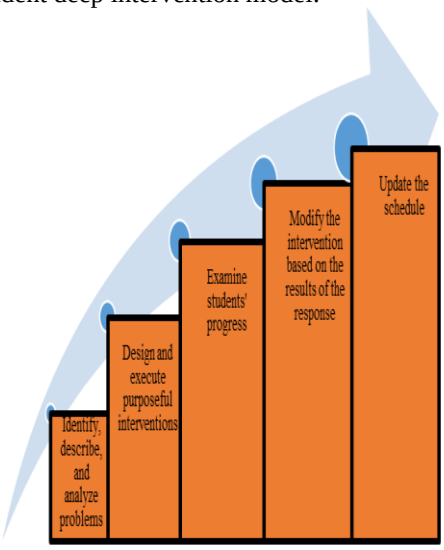



Figure 2: Flow of deep intervention

In Figure 2, the deep intervention includes four steps :(1) identify, describe, and analyze the problem; (2) design and implement targeted interventions; (3) observe the progress of students and modify the intervention measures according to the results of student's response to the intervention; (4) plan and arrange the following measures in the process of problem-solving [23]. Since the deep intervention response model is implemented in the framework of multi-level intervention, these four steps need to be performed at each level of intervention [24].

## 2.2 Big data analysis technology

The data mining method is the theoretical foundation of large data analysis. Extraction of implicitly unknown but possibly important knowledge and information from a huge quantity of noisy, ambiguous, incomplete and random data from applications is known as data mining [25]. This kind of technology is an emerging discipline formed by the intersection and integration of multiple disciplines, integrating mature tools and technologies in many disciplines, including database technology, statistics, ML, pattern recognition, AI, neural networks (NN), etc. [26]. Data mining techniques are classified from a theoretical point of view represented in Figure 3.

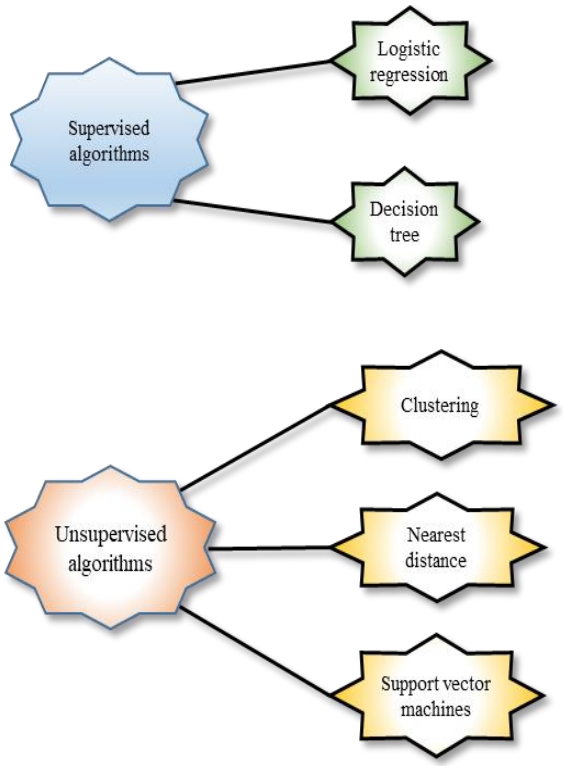



Figure 3: An empirical approach to classifying data mining techniques

In Figure 3, data mining techniques can be divided as unsupervised and supervised techniques. The LR, decision trees, etc are the supervised algorithms. Clustering, nearest neighbor distance, SVM and other techniques are mainly employed as unsupervised learning [27]. The classification of data mining in terms of application is shown in Figure 4: Figure 4 shows data mining, separated into classification, regression, cluster analysis, association rules, time series, and deviation-checking algorithms [28].

Data mining is mainly divided into four stages. Every step has specific requirements. To perform modifications and start again, it needs to stop the process and return into the previous step, when particular stage fails to produce the desired results. The process of data mining involves cyclic and interrelated steps [29]. Four stages of data mining are shown in Figure 5:

Data mining could be fragmented into four steps in Figure 5: assessment, representation, issue formulation and data preparation [30]. The majority of the times, data mining experts have to communicate extensively with the demander to obtain basic information about the data earlier. Data mining technique and employment are combined based on the type of target data [31] and pre-processing occurs on the data extracted from the collection. There are four further sub-steps in this process. Figure 6 illustrates the data preparation processes.

In Figure 6, data preparation includes four steps: 1. data selection. A mining task selects a dataset from a data source. (2) Data pre-processing. Since the data to be analyzed is generally disorganized and contains noise, the target data is subjected to some simple processing. (3) Data conversion. The pre-processed data is formatted as needed, such as discretization, normalization, etc. (4) Data loading. The processed data is loaded into a database, which has certain specifications to facilitate manipulation [32].

Analysis of the processed data is done at the data mining process. For analysis, an appropriate mining technique is selected based on the data mining requirements and goals established within the issue description process [33]. The core technology of data mining technology is classification and clustering technology. The algorithms included in data classification and clustering techniques are shown in Figure 7.

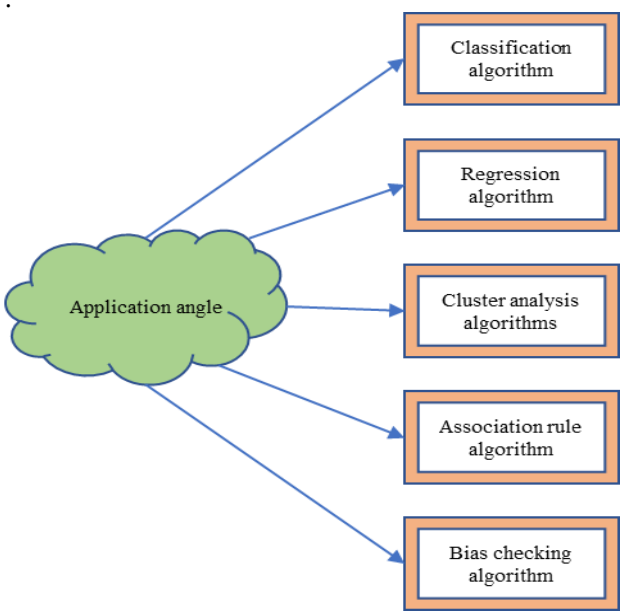



Figure 4: Classification of data mining techniques from an application perspective

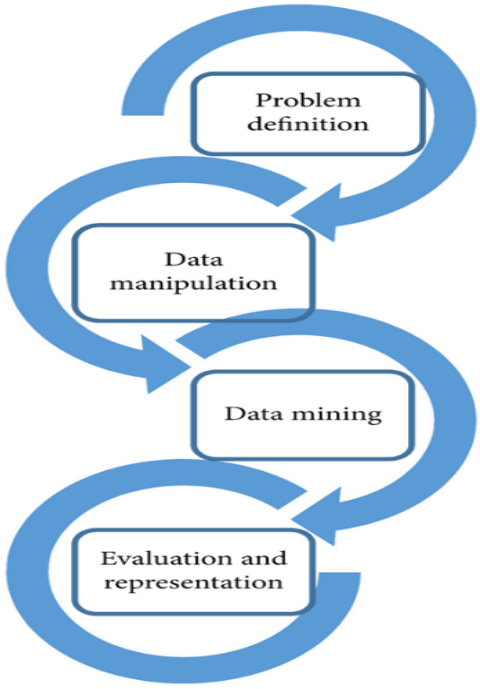



Figure 5: Stages of data mining techniques

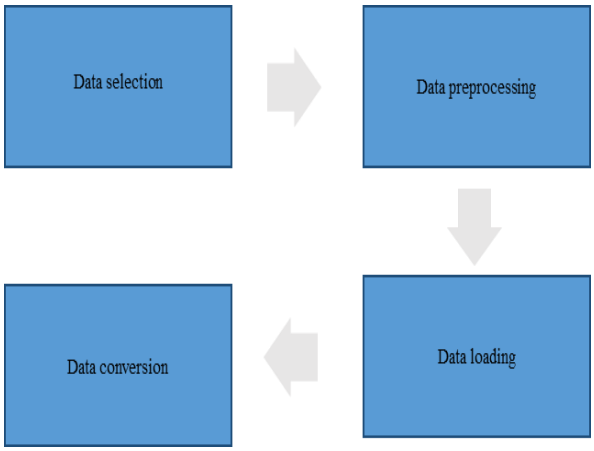



Figure 6: Steps of data preparation

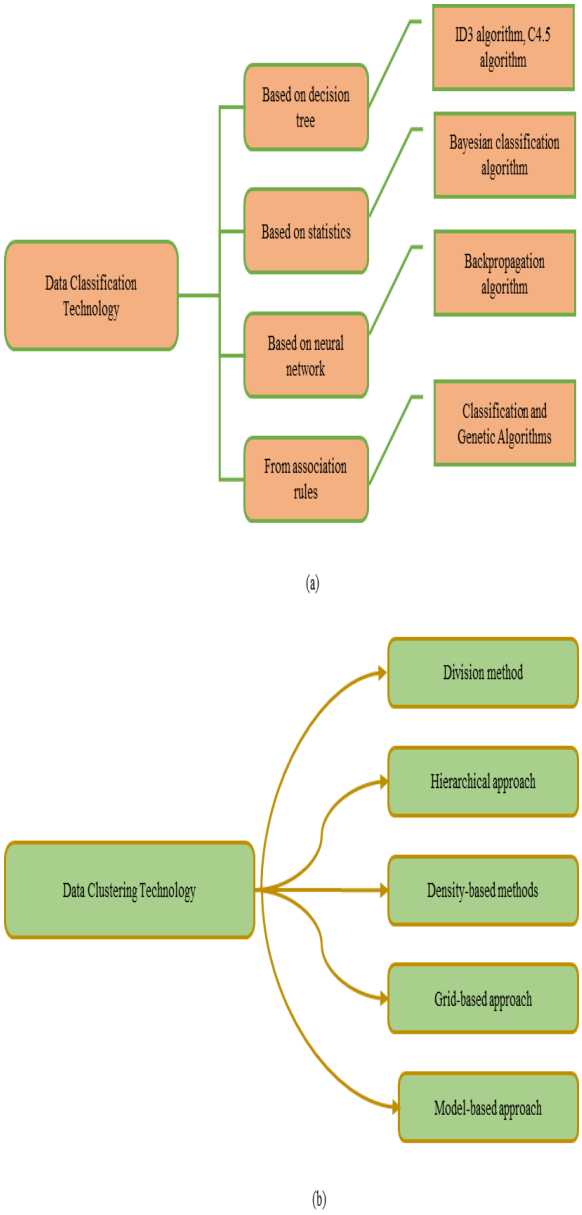



Figure 7: Data Classification and Clustering Algorithms (a) Data Classification Algorithm; (b) Data Clustering Algorithm

Data mining approaches for classification and clustering are depicted in Figure 7. Analyzing a group of objects in the database to identify their shared features can be described as data categorization in Figure 7(a). They are categorized by established standards into several groups. Classification techniques now include statistical techniques such as the Bayesian classification algorithm, neural network techniques including the backpropagation algorithm and decision tree techniques consisting of Iterative Dichotomiser 3 (ID3) and C4.5 algorithm. Figure 7(b) illustrates that the objects inside the group are connected to the data clustering attempt and the objects related to other groups are disconnected. The greater data clustering effect, more similarity exists within a class and greater variation within groups.

## 2.3 Decision tree technical analysis

The DTA is a commonly used classification algorithm in data mining. The node at the top level of the decision tree is the root node, which contains the collection of all data in the dataset. Each internal node represents a test on a feature, is a judgment condition, and contains a data set that satisfies all conditions from the root node to the node in the data set. The data set corresponding to the internal node is divided into two or more child nodes. The number of branches is determined by the characteristics of the features on the internal nodes. In the decision tree construction process, choosing a split node is the most important. The important attributes are selected to judge the internal nodes analyzing the data set with class tags. The process is iterated until a complete tree structure is generated or a specified threshold is reached to end the iteration. The structure of the decision tree is shown in Figure 8:

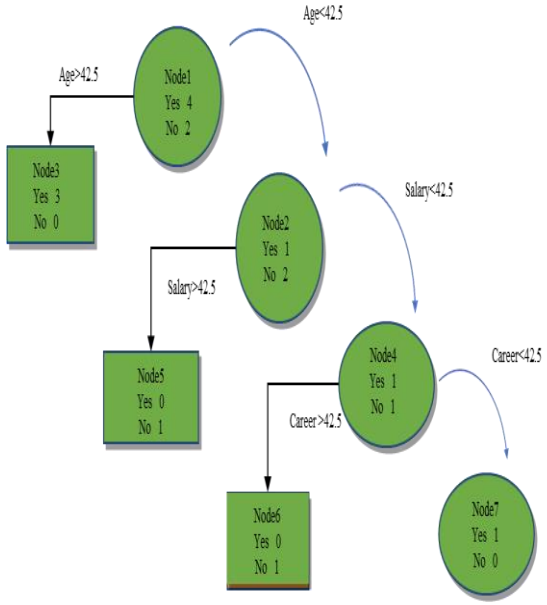



Figure 8: Structure of a decision tree

The decision tree process in Figure 8 is mainly split into two parts. First, by categorizing and evaluating the data, an initial training data set is established. The associated decision tree is established using the regression technique based on the training data set. The decision tree's inaccuracy is subsequently tested in the second phase.

During the building process, the decision tree identifies the node with the greatest data gain to be the current split node. Making a decision can decrease the quantity of data needed to split the training samples into the final tree structure. The

classification information entropy of the training data set S is determined using Eq. (1) if it is segmented based on the category attribute C.

$$H(S,C) = -\sum_{i=l}^{m} p_i log_2(p_i) \quad (1)$$

A total of *m* categories appear and $p_i$ is probable that the classification $i$ exhibits throughout the training pair. Eq. (2) indicates the categorization entropy of information for the conditioned attribute *A* used to split the training set of data *S* in relation to *C* if it is divided based on that attribute.
:

$$H(C) = -\sum_{j=i}^{v} \frac{|S_j|}{|S|} H(S_j, C) \quad (2)$$

$v$ is the quantity of condition attribute values $A$. The attribute's information gain $A$ splitting dataset $S$ showed in Eq. (3):

$$gain(C) = H(S,C) - H(S,A|C) \quad (3)$$

Equation (4) illustrates the data gained rate associated with attribute $A$ in the C4.5 method that splits the dataset $S$:

$$gain\_ratio(S,A|C) = \frac{gain(S,A|C)}{H(S,A)} \quad (4)$$

The probability distribution's Gini index in the classification tree is shown in Eq. (5):

$$Gini(p) = \sum_{k=1}^{k} (1 - p_k) = 1 - \sum_{k=1}^{k} p_k^2 \quad (5)$$

*K* is the number of categories; $p_k$ is the possibility that the data points depends with the K-th category.
If the sample point in the two-class issue has the possibility p for falling into the first group. The probability distribution's Gini index is shown in Eq. (6):

$$Gini(p) = 2p(1-p) \quad (6)$$

The Gini index of sample set D, as declared is shown in Eq. (7):

$$Gini(D) = 1 - \sum_{k=1}^{k} (\frac{|C_k|}{|D|})^2 \quad (7)$$

$C_k$ is the sample subset of *k*-th group at sample *D* and *K* is the sample groups amount.

If the feature A assumes a certain potential value a, than the dataset D is split into two portions, D1 and D2 and the Gini index at this time is shown in Eq. (8):

$$Gini(D,A) = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2) \quad (8)$$

The set's uncertainty has been represented by the Gini index. The calculation of the decision tree gain error is shown in Eq. (9):

$$\alpha = \frac{R(t) - R(T_t)}{\left|N_{T_i}\right| - 1} \quad (9)$$

$R(T_t)$ Indicates the total error costs of each leaf in the subtree $T_t$ except node *t*. $\left|N_{T_i}\right|$ is the subtree's total amount of leaf nodes *t*. *R(t)* represents the error cost of a leafless node *t*, as shown in Eq. (10):

$$R(t) = r(t) * p(t) \quad (10)$$

*r(t)* is the rate of error for node t. *p(t)* represents proportion of subtree $T_t$ to the total data.
Mainly, CART, C4.5, and ID3 algorithms are utilized in decision trees. The algorithm for the decision tree model, one of the most prominent in data mining, can enhance English teaching model classification and provide a solid foundation in data analysis. The decision tree method can improve the processing of the varied range of data used in English language instruction by performing measures such as data pruning. For the intent of operating the model, the decision tree method is utilized as the basic approach in the research.

# 3 Model for predicting english teaching outcomes using decision tree algorithm

The C4.5 on the decision tree technique serves as the foundation for the outcome prediction model for English education. The C4.5 method selects features based on information gain rate that assists in reducing the issue of excessive data gain resulting from an excessive number of eigenvalues. The algorithm performs superior on the basis of classification. Figure 9 displays the particular model:

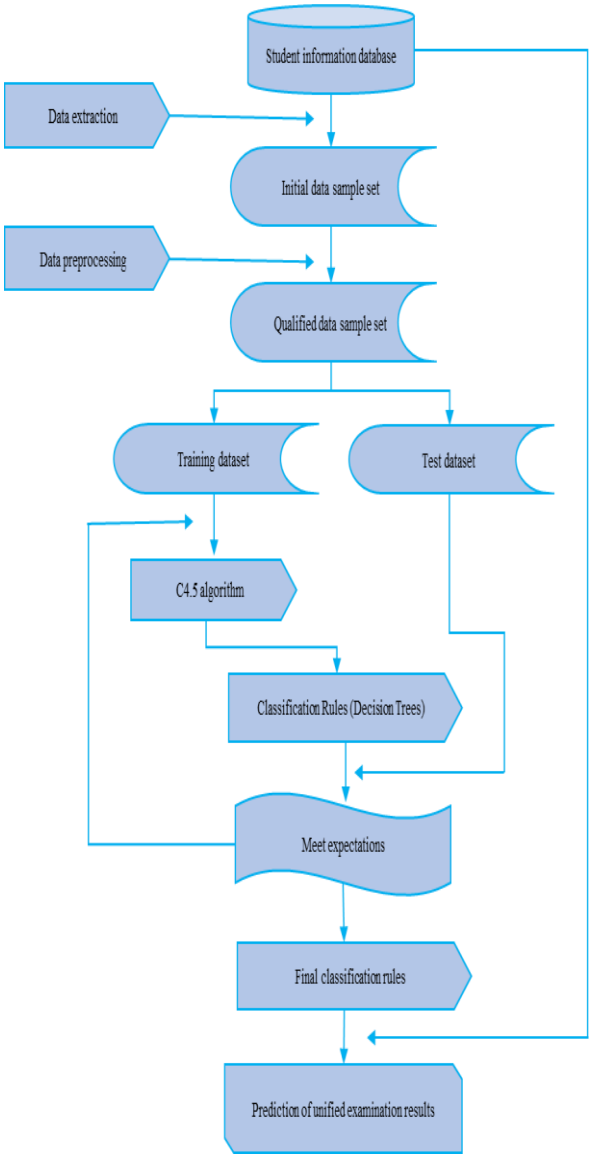



Figure 9: English teaching performance prediction model using the decision tree method

In Figure 9, firstly, certain sample information is extracted from the student information database to form an initial sample set. The initial sample set is subjected to pre-processing to remove data attributes and contents irrelevant to the mining target to form a qualified data set. Data mining provides clean, accurate, and more targeted data. It can enhance the effectiveness and precision of outcomes while reducing the load on algorithms. Testing and training datasets are separate categories of relevant datasets. The outcomes of the decision tree method can be fitted with the test dataset when it has been trained using the training data set. To output the result if the algorithm reaches the desired outcome, perform the training again.

# 4 Experimental data collection

## 4.1 Data collection process

The data collecting procedure entails obtaining information on the English scores of students from different grades at a primary school in Anyang City. This information is used as a sample data set for research on variables impacting English teaching performance.

The system that manages student registration gathers fundamental data on students, including their full names, genders, grades and other details.

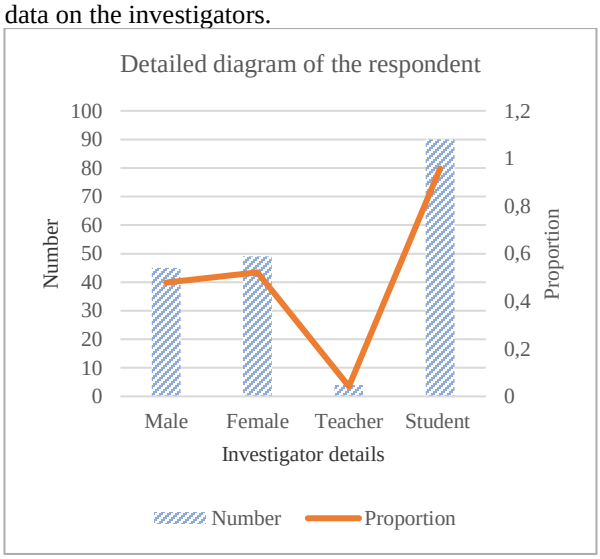
The examination scores information is gathered from students. The data table contains student numbers, names, and test-related information. The information on student achievement at school is also collected.

## 4.2 Selection criteria

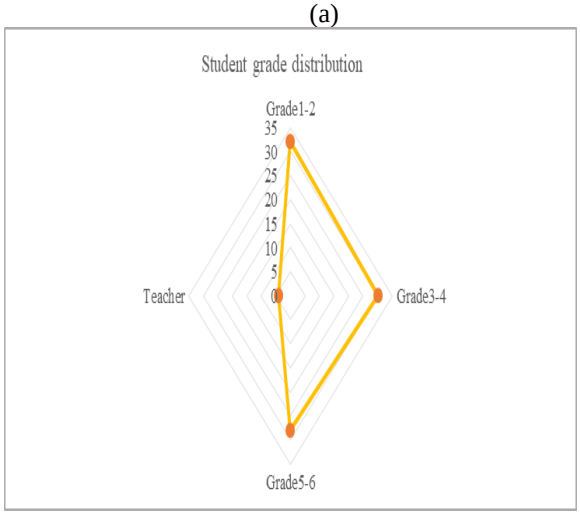
English scores of students from different grades at a primary school in Anyang City have been chosen as the sample information.

The variables influencing English teaching effectiveness are investigated using the obtained data which includes basic student information, test scores, and academic achievement. To assess the forecasting model method's operational impact, a subset of sample values is randomly chosen from the original sample set to serve as training samples.

Through a random survey of the teaching performance of students in various grades in a primary school in Anyang, and the opinions of teachers and staff on students' performance, following distribution of 100 questionnaires, 94 genuine questionnaires are found providing an efficient recovery rate of 94%. Figure 10 displays the fundamental data on the investigators.



(a)




(b)

Figure 10: The respondent information include: (a) respondent identity; (b) respondent age distribution

In Figure 10, there are 45 males, 49 females, four teachers, 90 students, 32 students from grade one to grade two, 30 students in grades three to four, and 28 students in grades five and six.

The statistical software from Statistical Product Service Solutions (SPSS) is utilized for data analysis. After computation, the Cronbach's alpha coefficient for the questionnaire's data is 0.797. The questionnaire's results are reliable as extended as they fell within the parameters of its dependability.

# 5 Experimental results

## 5.1 Model effect analysis

The experimental sample is a portion of the original sample set's data that was determined at random. Figure 11 displays the obtained values for the sample set's accuracy for prediction and processing time efficiency.

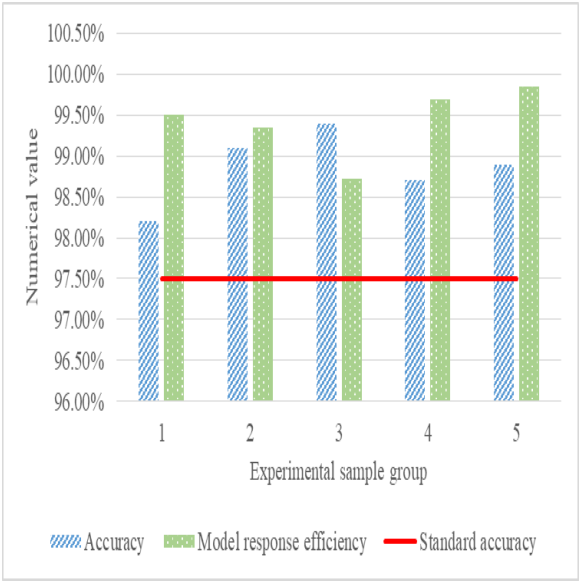
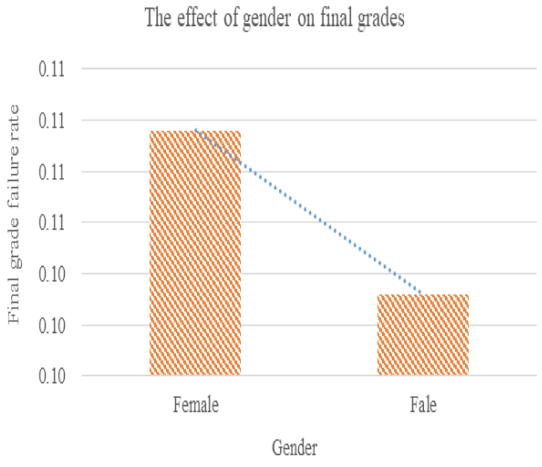



Figure 11: The model operating effect's outcomes

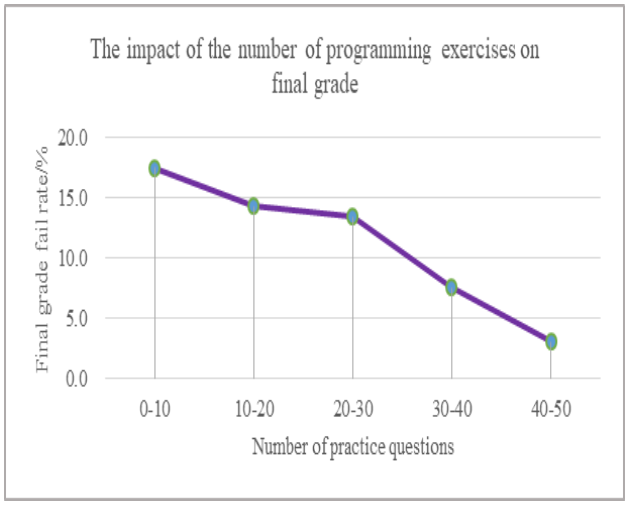
In Figure 11, the running effects of the five sample groups set to test the model are used separately. The model's prediction accuracy is 98.20%, 99.10%, 99.40%, 98.70%, and 98.90%, higher than the standard accuracy of 97.5%. Additionally, the model has a high response efficiency and an average responding efficiency of 99.42%. Thus, it is feasible to employ and effectively utilize the decision tree algorithm-based English teaching performance prediction model.

An examination of the variables influencing the effectiveness of English teaching

The effects of gender and the number of practice questions on grades are analyzed separately, illustrated in Figure 12.




(a)



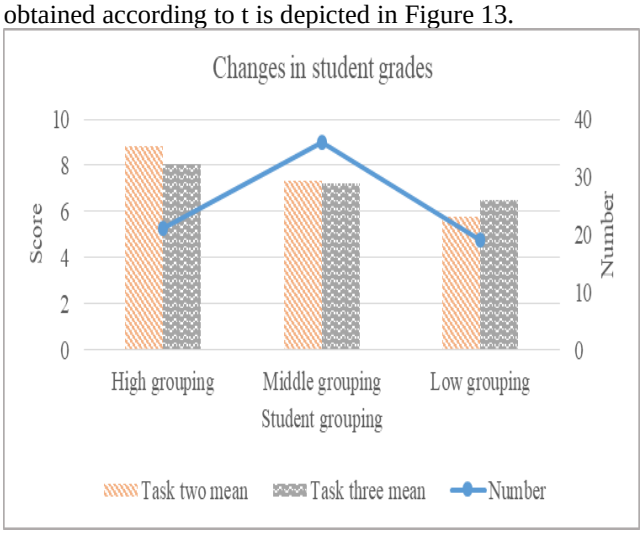

(b)

Figure 12: Effects of gender and number of practice questions on grades (a) effect of gender on final grade; (b) effect of the number of practice questions on grade

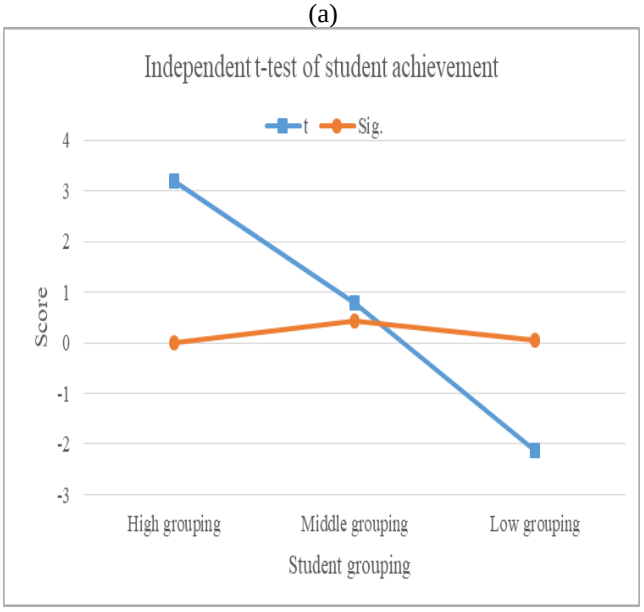
In Figure 12, there are 49 girls. There are forty-five guys and five students that failed to earn the final grade among them. Five pupils received a failing mark in their final exam. The boys' last-year failure rate, which stands at 11%, is calculated as the percentage of the number of boys that failed the midterm to the overall number of boys. The proportion of female students that failed their midterm final grade to all female students is 10%. This represents the percentage of female students that failed their final grade. For this dataset, there is a 1% difference in final grade failure rates between boys and girls. Therefore, gender has a less obvious effect on teaching achievement. With the increase in the number of practice questions, the proportion of failing grades at the end of the term decreases successively, which are 17.49%, 14.29%, 13.46%, 7.59%, and 3.05%. The percentage of final grade failure decreases as more code questions are performed. Therefore, the data suggest that the number of practice questions influences teaching performance.

Predicting English learning outcomes with deep intervention and large data analysis

The effect of a deep teaching intervention on teaching performance is further analyzed. After the course is launched, students' grades are predicted by setting three tasks. From the third prediction task, after each prediction, 76 students are selected for analysis and divided into high grouping, middle grouping, and low grouping, respectively. Students in the middle and low groupings are given teaching intervention, while those in the high grouping are not given teaching intervention. After obtaining the actual grades of all students for Task 3, the mean scores for Tasks 2 and 3 are calculated. Paired samples t-tests are used to obtain instructional intervention results in SPSS software. Among them, t is the significance test statistic, and the probability Sig. of the corresponding significance test statistic is obtained according to t is depicted in Figure 13.




(a)




(b)

Figure 13: Graph of survey results (a) changes in grades; (b) independent t-test analysis

In Figure 13, (1) high grouping: no intervention, Tasks 3 and 2 show a significant decrease trend; (2) middle grouping: The mean result of the trial group students in Tasks 3 and 2 shows a trend of improvement. There was an apparent decrease in the control group's student's scores, and the decrease was observed consistently. (3) low grouping: there is a significant trend of improvement compared with the average scores of students in task three and task two. The control group's pupils' test results increased, but not significantly, showing an overall trend of improvement. These data show that teaching intervention improves student

performance and the greater the investment in teaching intervention, the better the effect.

## 5.2 Comparative analysis

Analyzing the overall performance of this research, we utilize existing models are random forest (RF) [34], NN [34], SVM [34], LR [34], naïve Bayes (NB) [34], and k-nearest neighbors (KNN) [34]. The parameters employed for both cutting-edge and proposed models are precision, recall, accuracy and F1-score. Figure 14 and Table 2 represent the comparative outcomes of accuracy and precision. Compared to the existing models, our proposed decision tree model attains superior results in precision (87.9 %) and overall accuracy (98.86 %).

Table 2: Outcomes values of accuracy and precision

| Model | Accuracy (%) | Precision (%) |
|---|---|---|
| RF [34] | 74.6 | 75.2 |
| NN [34] | 74.6 | 74.8 |
| SVM [34] | 73.5 | 73.5 |
| LR [34] | 71.7 | 70 |
| NB [34] | 71.3 | 70.6 |
| KNN [34] | 69.9 | 69.1 |
| Decision Tree [Proposed] | 98.86 | 87.9 |

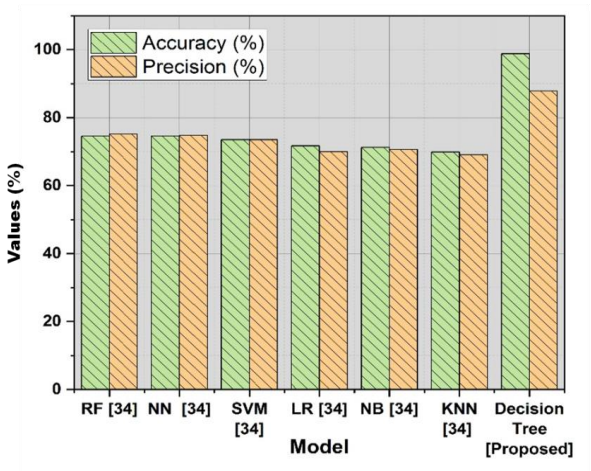



Figure 14: Outcomes of accuracy and precision

Figure 15 and Table 3 show the comparative outcomes of recall and F1-score. In contrast to the proposed and existing models, our proposed decision tree model obtains a better recall (90.2%), and f1-score (88.5%).

Table 3: Outcomes values of recall and F1-score

| Model | Recall (%) | F1-Score (%) |
|---|---|---|
| RF [34] | 74.6 | 72.1 |
| NN [34] | 74.6 | 72.3 |
| SVM [34] | 73.5 | 70.4 |
| LR [34] | 71.7 | 68.5 |
| NB [34] | 71.3 | 69.2 |
| KNN [34] | 69.9 | 69.4 |
| Decision Tree [Proposed] | 90.2 | 88.5 |

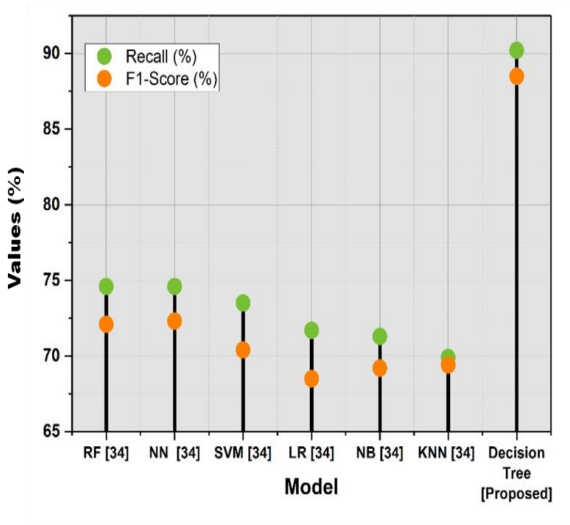



Figure 15: Outcomes of recall and F1-score

# 6 Discussion

The drawbacks of RF [34] include over fitting, longer training on big datasets, difficult ensemble interpretation, and memory-intensive computing. Over fitting, inadequate interpretability, computational complexities, data hunger, and sensitivity to adversarial assaults are among the disadvantages of NN [34]. SVM [34] may struggle with huge datasets, need precise parameter tuning, and are susceptible to noisy or overlapping data. LR [34] potential limitations involve over fitting, the inability to capture complicated correlations, and the necessity for linear decision limits. NB [34] presupposes independence between characteristics, which might result in poor efficiency when characteristics are associated or interacting in a nonlinear

manner. KNN [34] are sensitive to irrelevant characteristics, require a significant amount of computing resources, biased to majority classes, and are difficult to comprehend in high-dimensional environments. By analyzing the drawbacks of existing models, our proposed decision tree model achieved better performance.

# 7 Conclusion

A model for the prediction of English teaching results is developed, utilizing the decision tree algorithm. Under intensive intervention, the model is studied. The survey results show that the model prediction accuracy is 97.5% higher than the standard. Additionally, the average response efficiency of the model is 99.42%; Male students have an 11% failing rate; 10% of female students fail their classes. Therefore, the effect of gender on teaching performance is less obvious. As the number of practice questions increases, the failure rate gradually decreases. Thus, the data suggest that the number of practice questions affects instructional performance. The number of practice questions has a significant impact on English teaching performance. Teachers' intervention can improve students' English performance, and intervention intensity will also improve students' performance. Therefore, follow-up research should increase the number of practice questions and teacher intervention in English teaching. Due to the short time and limited sample size, there are certain deficiencies in the scope and depth of teaching intervention. In the future, this work will expand the scope of teaching interventions and take deeper teaching interventions. In addition, big data analysis technology keeps pace with the times and will update and utilize new technologies in the follow-up, deeply integrate theory and practice, and design an English teaching achievement prediction model with more teaching theory characteristics. The innovation lies in using computer and using data mining techniques to examine the traditional teaching results, making the results more credible. Additionally, teacher intervention factors are introduced to ensure the integrity of teaching performance predictors.

# Limitations and future work

Limitations include inaccuracies caused by varied student backgrounds, linguistic complexity, and incapacity to properly reflect individual learning methods and socio-cultural effects. Improve prediction models through continual data refinement, incorporate adaptive learning approaches, and investigate socio-cultural aspects to ensure reliable English teaching performance forecasts.

# Data availability statement

The data used to support the findings are available from the corresponding author upon request.

# Conflicts if interest

The authors declare that they have no conflicts of interest.

# References

[1] R. C. Robey, A. Danson, J. Evans et al., “50 using a targeted teaching intervention to drive up the quality of discharge summaries,” *shop 75+. Age and Ageing*, vol. 50, no. Supplement_1, pp: i12-i42, 2021.

[2] X. Liu, F. Gao, Q. Jiao, “Massive open online course fast adaptable computer engineering education model,” *Complexity*, vol. 2021, no. 1, pp: 1-11, 2021.

[3] J. B. Bush, “Software-based intervention with digital manipulatives to support student conceptual understandings of fractions,” *British Journal of Educational Technology*, vol. 52, no. 6, pp: 2299-2318, 2021.

[4] P. Kumar, R. Bhatnagar, K. Gaur et al., “Classification of imbalanced data:review of methods and applications,” *IOP Conference Series: Materials Science and Engineering*, vol. 1099, no. 1, pp: 012077 (8pp), 2021.

[5] J. Pellet, M. Weiss, F. Zúiga et al., “Implementation and preliminary testing of a theory-guided nursing discharge teaching intervention for adult inpatients aged 50 and over with multimorbidity: a pragmatic feasibility study protocol,” *Pilot and Feasibility Studies*, vol. 7, no. 1, pp: 1-13, 2021.

[6] M. Acharya, K. P. Acharya, K. Gyawali et al., “Discussing professor yin kejing's drug use law for mammary hyperplasia based on data mining technology,” I*nternational Journal of Clinical and Experimental Medicine*, vol. 5, no. 3, pp: 403-407, 2021.

[7] H. Huo, Y. Chang, and Y. Tang, “Analysis of treatment effect of acupuncture on cervical spondylosis and neck pain with the data mining technology under deep learning,” *The Journal of Supercomputing*, vol. 78, no. 4, pp: 5547-5564, 2021.

[8] Q. Zhou, M. Zhang, and B. Ki-Hyung, “Retracted article: edge computing and financial service industry financing risk innovation based on data mining technology,” *Personal and Ubiquitous Computing*, vol. 25, no. Suppl 1, pp: 19-19, 2021.

[9] L. Lu, Q. Wen, X. Hao et al., “Acupoints for tension-type headache: a literature study based on data mining

technology," *Evidence-based Complementary and Alternative Medicine*, vol. 2021, no. 3, pp: 1-10, 2021.

[10] Y. Zhu. "Data Mining Technology-Based English Listening Prediction Strategy and Its Training Approach", *Security and Communication Networks*, vol. 2022, 2022.

[11] G. Li, and W., Gao, "Achievement Prediction of English Majors Based on Analytic Hierarchy Process and Genetic Algorithm," *Mobile Information Systems, 2022*, 2022.

[12] C.Z. Szabo, U. Stickler, and L. Adinolfi, "Predicting the academic achievement of multilingual students of English through vocabulary testing," *International Journal of Bilingual Education and Bilingualism*, *24*(10), pp.1531-1542, 2021.

[13] C. Cao, and Q. Meng, "Exploring personality traits as predictors of English achievement and global competence among Chinese university students: English learning motivation as the moderator," *Learning and Individual Differences*, *77*, p.101814, 2020.

[14] A.Y. Huang, O.H. Lu, J.C. Huang, C.J. Yin, and S.J. Yang, "Predicting students' academic performance by using educational big data and learning analytics: evaluation of classification methods and learning logs," *Interactive Learning Environments*, *28*(2), pp.206-230, 2020.

[15] H. Waheed, S.U. Hassan, N.R. Aljohani, J. Hardman, S. Alelyani, and R. Nawaz, "Predicting academic performance of students from VLE big data using deep learning models," *Computers in Human behavior*, *104*, p.106189, 2020.

[16] D. Wang, D. Lian, Y. Xing, S. Dong, X. Sun, and J. Yu, "Analysis and prediction of influencing factors of college student achievement based on machine learning," *Frontiers in Psychology*, *13*, p.881859, 2022.

[17] E.S. Bhutto, I.F. Siddiqui, Q.A. Arain, and M. Anwar, "Predicting students' academic performance through supervised machine learning," In *2020 International Conference on Information Science and Communication Technology (ICISCT)* (pp. 1-6). IEEE, 2020.

[18] M. H. Hefter, and K. Berthold, "Active ingredients and factors for deep processing during an example-based training intervention," *Learning Environments Research*, vol. 25, no. 1, pp: 17-39, 2022.

[19] M. M. Beudels, A. Preisfeld, K. Damerau, "Impact of an Experiment-Based Intervention on Pre-Service Primary School Teachers' Experiment-Related and Science Teaching-Related Self-Concepts," *Interdisciplinary Journal of Environmental and Science Education*, vol. 18, no. 1, pp: e2258, 2022.

[20] P. A. Burns, A. A. Omondi, M. Monger et al., "Meet me where I am: An evaluation of an HIV patient navigation intervention to increase uptake of PrEP among black men who have sex with men in the deep south," *Journal of Racial and Ethnic Health Disparities*, vol. 9, no. 1, pp: 103-116, 2022.

[21] M. Balalavi, H. C. Huang, T. F. Tsai et al., "Applying Taiwanese indigenous health literacy for designing an elders' prevention fall course: a statistical analysis and deep learning approach," *The Journal of Supercomputing*, vol. 77, no. 3, pp: 2355-2382, 2021.

[22] O. Hannula, R. Vanninen, S. Rautiainen et al., "Teaching limited compression ultrasound to general practitioners reduces referrals of suspected DVT to a hospital: a retrospective cross-sectional study," *The Ultrasound Journal*, vol. 13, no. 1, pp: 1-7, 2021

[23] É. Fülöp, "Developing problem-solving abilities by learning problem-solving strategies: an exploration of teaching intervention in authentic mathematics classes," *Scandinavian Journal of Educational Research*, vol. 65, no. 7, pp: 1309-1326, 2021.

[24] F. Yao, and A. Zhang, "Integration of education management and mental health in psychological crisis intervention in colleges and universities." *ASP Transactions on Psychology and Education*, vol. 2, no. 1, pp: 31-38, 2021.

A. Dogan, and D. Birant, "ML and data mining in manufacturing," *Expert Systems with Applications*, vol. 166, pp: 114060, 2021.

[25] H. Thakkar, V. Shah, H. Yagnik et al., "Comparative anatomization of data mining and fuzzy logic techniques used in diabetes prognosis," *Clinical eHealth*, vol. 4, pp:12-23, 2021.

[26] H. X. Wang, and S. Smys, "Big Data Analysis and Perturbation using Data Mining Algorithm," *Journal of Soft Computing Paradigm (JSCP)*, vol. 3, no. 1, pp: 19-28, 2021.

[27] M. A. Schorn, S. Verhoeven, L. Ridder et al., "A community resource for paired genomic and metabolomic data mining," *Nature Chemical Biology*, vol. 17, no. 4, pp: 363-368, 2021.

[28] C. Liu, Y. Cui, X. Li et al., "microeco: an R package for data mining in microbial community ecology," *FEMS microbiology ecology*, vol. 97, no. 2, pp: fiaa255, 2021.

[29] Z. S. Ageed, S. R. Zeebaree, M. M. Sadeeq et al., "Comprehensive survey of big data mining approaches in cloud systems," *Qubahan Academic Journal*, vol. 1, no. 2, pp: 29-38, 2021.

[30] Y. Pan, and L. Zhang, "A BIM-data mining integrated digital twin framework for advanced project management," *Automation in Construction*, vol. 124, pp: 103564, 2021.

[31] P. Liu, Q. Q. Wang, W. Liu, "Enterprise human resource management platform based on FPGA and data mining," *Microprocessors and Microsystems*, vol. 80, pp: 103330.

A. Assad, and A. Bouferguene, "Data Mining Algorithms for Water Main Condition Prediction—Comparative Analysis," *Journal of Water Resources Planning and Management*, vol. 148, no. 2, pp: 04021101, 2022.

[32] M. Yağcı, "Educational data mining: prediction of students' academic performance using machine learning algorithms," *Smart Learning Environments*, *9*(1), p.11, 2022.