BULLETIN OF EDUCATION AND PSYCHOLOGY
Research Article | Volume 5 | Issue 5 | Feb 2025
Received: 19 Jan 2025 | Revised: 23 Jan 2025
Accepted: 22 Feb 2025 | Published Online: 26 Feb 2025

Prediction of Student Answer Accuracy based on Logistic Regression

Zhiqiang Zhao [1,*], Ping Ren[2]

[1] Beijing PhD Village Education Technology Co., Ltd; Beijing 100871, China

[2] Chengdu Ding Yi Education Consulting Co., Ltd, Chengdu 610023, China

[*]Corresponding Author, Email: 202128030258@mail.bnu.edu.cn

Abstract: In the realm of education, predicting student answer accuracy plays a critical role in enhancing learning outcomes. Despite its significance, existing research currently faces challenges in accurately forecasting student performance. This paper addresses this gap by proposing a novel approach utilizing logistic regression for predicting student answer accuracy. By incorporating various factors such as student demographic information, historical performance data, and study habits, our model aims to provide more precise predictions compared to traditional methods. Through extensive experimentation and data analysis, we demonstrate the effectiveness and robustness of our proposed method in predicting student answer accuracy. This research not only contributes to the improvement of educational assessment techniques but also opens up new avenues for personalized learning strategies.

Keywords: *Student Answer Accuracy; Predictive Modeling; Logistic Regression; Educational Assessment; Personalized Learning Strategies*

1. Introduction

Student Answer Accuracy is a field of research that focuses on evaluating the correctness of responses provided by students in academic assessments. Currently, one of the main challenges in this area is the development of accurate and efficient methods to assess the accuracy of student answers across a wide range of subjects and question types. Additionally, the variability in individual student understanding and interpretation of questions presents a significant hurdle in achieving consistent and reliable measurements of answer accuracy. Furthermore, the integration of technology in assessment practices introduces complexities in ensuring the fairness and reliability of evaluating student responses. As researchers continue to explore innovative approaches and techniques to address these obstacles, the pursuit of enhancing the precision and validity of assessing student answer accuracy remains a central objective in the field.

To this end, current research on Student Answer Accuracy has advanced to the stage of utilizing machine learning algorithms to analyze and predict student responses with high accuracy. The integration of natural language processing techniques has further improved the assessment of student answers across various academic disciplines. The literature review in the field of automatic student answer assessment highlights the use of various deep learning techniques for grading student responses. Mihajlov [1] introduced the concept of using Latent Semantic Analysis (LSA) for student answer assessment. Hollis-Sando et al. [2] focused on medical student perceptions and accuracy evaluation of deep learning in marking short answer questions. Khayi et al. [3] explored the use of pretrained transformers for open student answer assessment, achieving significant accuracy improvements. Saeed and Gomaa [4] proposed an ensemble-based model to enhance the accuracy of short answer grading through text similarity approaches. Campbell et al. [5] investigated the use of IBM's Watson for automatic evaluation of student short answer responses, analyzing its performance and implications for educational research. Sultan et al. [6] presented a fast and accurate short answer grading system utilizing text similarity features and achieving top performance. Moreover, Stribling et al. [7] evaluated the performance of GPT-4 in grading graduate biomedical science exams, discussing its strengths and limitations across different question types. Jiang and Bosch [8] examined short answer scoring with GPT-4, highlighting variations in performance based on educational subjects and the quality of scoring rubrics. Lastly, Tornqvist et al. [9] proposed the ExASAG framework for explainable automatic short answer grading, emphasizing the importance of explainability in student assessment. The existing literature on automatic student answer assessment emphasizes the application of various deep learning methods. Logistic Regression is an essential technique due to its interpretability, simplicity, and efficiency in handling binary classification tasks. Additionally, its ability to provide probabilities for outcomes makes it valuable for grading student responses accurately.

Specifically, Logistic Regression is a statistical method used to model the relationship between a binary outcome variable and one or more predictor variables. In the context of Student Answer Accuracy, Logistic Regression can be applied to predict the likelihood of a student providing a correct answer based on various factors such as study time, prior knowledge, and test preparation. A literature review on logistic regression models reveals a range of key contributions to the field. Hosmer et al. (2005) present the definitive guide to logistic regression modeling, emphasizing its applications in the health sciences and providing state-of-the-art techniques for building, interpreting, and evaluating LR models [10]. Friedman (2000) discusses boosting as a significant development in classification methodology, showing that boosting can be viewed as an approximation to additive modeling on the logistic scale using maximum Bernoulli likelihood as a criterion [11]. Menard (1996) introduces applied logistic regression analysis, covering topics such as linear regression, logistic regression coefficients interpretation, and polychotomous logistic regression [12]. Harrell (2001) explores regression modeling strategies with applications to linear models, logistic regression, and survival analysis [13]. King and Zeng (2001) focus on logistic regression in rare events data, proposing corrections to address underestimations and inefficiencies in data collection strategies for rare events data. Conklin (2002) and Rao (2003) further contribute to applied logistic regression and regression modeling strategies, respectively [14]. Additionally, Peduzzi et al. (1996) conduct a simulation study on the number of events per variable in logistic regression analysis, while G et al. (2022) apply logistic regression technique for the prediction of cardiovascular disease [15]. However, current limitations include the need for further research on logistic regression models in diverse fields beyond health sciences and rare events data. Improvements in data collection strategies and addressing underestimations remain areas for exploration in future studies [16-18].

To overcome those limitations, this paper aims to address the challenge of accurately forecasting student performance in the realm of education by proposing a novel approach utilizing logistic regression for predicting student answer accuracy. The method incorporates a comprehensive set of factors including student demographic information, historical performance data, and study habits to provide more precise predictions compared to traditional methods. Specifically, the model leverages logistic regression to analyze the relationships between these factors and student answer accuracy, enabling the prediction of individual student outcomes with a higher degree of accuracy. Through extensive experimentation and data analysis, the effectiveness and robustness of the proposed method are demonstrated, showcasing its potential to significantly enhance learning outcomes by improving the accuracy of performance predictions. Furthermore, by contributing to the advancement of educational assessment techniques, this research not only benefits the field of education but also paves the way for the implementation of personalized learning strategies tailored to individual student needs, ultimately fostering a more effective and efficient learning environment.

Section 2 outlines the problem addressed in this study, focusing on the importance of predicting student answer accuracy in education. Section 3 introduces the innovative method proposed to address this challenge, utilizing logistic regression to forecast student performance. In Section 4, a detailed case study is presented to illustrate the application and effectiveness of the proposed approach. Section 5 analyzes the results derived from extensive experimentation, highlighting the precision and reliability of the model in predicting student answer accuracy. Section 6 delves into a thorough discussion of the implications and potential enhancements of the research findings. Finally, in Section 7, a comprehensive summary of the study's contributions to educational assessment techniques and personalized learning strategies is provided. This cohesive narrative underscores the significance and impact of the research in advancing the field of education.

2. Background

2.1 Student Answer Accuracy

Student Answer Accuracy (SAA) is a critical metric used to determine the correctness level or performance of students when answering questions in educational settings. This measurement provides insights into students' understanding of the material, the effectiveness of the instructional methods, and even informs personalized learning approaches. At its core, SAA quantifies the percentage of correctly answered questions by a student out of the total questions attempted. The fundamental formula for computing SAA is quite straightforward. It can be defined as the ratio of the number of correct answers to the total number of questions attempted:

$$SAA = \frac{C}{T} \times 100\% \tag{1}$$

where C represents the number of correct answers, and T denotes the total number of attempted questions. This formula expresses SAA as a percentage, highlighting the proportion of correct answers. Beyond the basic measure, it's crucial to consider the nature and difficulty of the questions involved. For instance, if the questions vary in difficulty, a weighted SAA might provide a more nuanced assessment. Each question can have a weight w_i based on its difficulty, where w_i is a real number between 0 and 1, signifying its importance or difficulty level:

$$WSAA = \frac{\sum_{i=1}^{n} w_i \cdot c_i}{\sum_{i=1}^{n} w_i} \times 100\% \tag{2}$$

In this formula, c_i is a binary indicator (1 if the answer is correct and 0 otherwise) for each question, and n is the number of questions. SAA can be further explored using concepts from statistics and machine learning, particularly in adaptive testing environments where questions are dynamically adjusted to the student's performance. Consider a scenario where the probability of a student answering a question correctly is modeled using logistic regression based on features reflecting both the question difficulty and student's ability:

$$P(y = 1|x) = \frac{1}{1 + e^{-(\alpha+\beta x)}} \tag{3}$$

Here, α is the intercept, β is the coefficient vector representing the effect of explanatory variables x , such as student ability and question difficulty. Another sophisticated approach is to employ item response theory (IRT), which examines variations in student performance with respect to item characteristics. The simplest IRT model, the Rasch model, defines the probability of a correct answer as:

$$P\left(y_{ij} = 1\middle|\theta_j, \beta_i\right) = \frac{e^{(\theta_j - \beta_i)}}{1 + e^{(\theta_j - \beta_i)}} \tag{4}$$

where θ_j is the ability parameter for student j , and β_i is the difficulty parameter for question i. It's essential to also account for potential biases, such as guessing, which can inflate SAA. A correction for guessing, especially in multiple-choice scenarios, can be incorporated as:

$$CSAA = \frac{C_l - \frac{g}{k-1}N}{T} \times 100\% \tag{5}$$

In this equation, C_l is the corrected number of correct responses, g is the number of guesses, k is the number of options per question, and N is the number of options guessed. Finally, evaluating the temporal dimension of student responses, the Learning Rate (LR) can be viewed as the change in SAA over time or learning sessions:

$$LR = \frac{SAA_{final} - SAA_{initial}}{T_s} \tag{6}$$

where T_s refers to the number of learning sessions. In summary, while the fundamental SAA provides a basic measure of student performance, incorporating aspects like question difficulty, student ability, guessing corrections, and temporal analysis enriches the metric, offering a comprehensive view of student learning and instructional effectiveness.

2.2 Methodologies & Limitations

In the domain of Student Answer Accuracy (SAA), a variety of methodologies have been developed and refined to enhance the understanding of student performance beyond the basic measure of correctness. These methodologies address several limitations inherent in the straightforward calculation of SAA by incorporating statistical, probabilistic, and psychometric models. One advanced approach to calculating SAA is the use of a Bayesian framework, where the student's performance is viewed as a probabilistic variable updated over time. This method can be particularly useful in adaptive learning environments. Within this framework, a student's probability of answering correctly is revised as more data becomes available, using a prior distribution reflecting initial beliefs about student ability:

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)} \tag{7}$$

where $P(\theta|D)$ is the posterior probability of the student's ability θ given data D, $P(D|\theta)$ is the likelihood of the data given the ability, $P(\theta)$ is the prior probability, and $P(D)$ is the marginal likelihood of the data. Additionally, factor analysis can be employed to determine latent variables that might influence student performance, like motivation or topic familiarity. The factor analysis model can be expressed as:

$$X = \lambda F + \epsilon \tag{8}$$

where X is the observed variable vector, λ is the matrix of factor loadings, F is the vector of latent factors, and ϵ is the vector of errors or specific variances. A more nuanced model, considering the multidimensional nature of student ability and question complexity, is explored within the multidimensional item response theory (MIRT) framework. This model allows each question to load onto multiple dimensions of ability:

$$P\left(y_{ij} = 1 \middle| \boldsymbol{\theta_j}, \boldsymbol{\beta_i}\right) = \frac{e^{(a_i \cdot \boldsymbol{\theta_j} - b_i)}}{1 + e^{(a_i \cdot \boldsymbol{\theta_j} - b_i)}} \tag{9}$$

In this expression, $\boldsymbol{\theta_j}$ is a vector representing multiple ability traits of student j, $\boldsymbol{\beta_i}$ is a vector for item i 's parameters, a_i is the item discrimination vector, and b_i is the item difficulty. Another critical development is the use of neural networks to predict SAA, employing student interaction data as input features. The neural network formulates the relationship between input features and predicted accuracy through layers of computational units:

$$y = f(W_2 \cdot g(W_1 \cdot x + b_1) + b_2) \tag{10}$$

where $\hat{y}$ is the output prediction (SAA), x is the input feature vector, W_1 and W_2 are weight matrices, b_1 and b_2 are biases, and f and g are activation functions. Despite these advanced methodologies, several challenges and limitations persist. A key issue is the assumption of independence between questions, which might not hold true in contexts where learning trajectory and question order have significant effects. Furthermore, models like IRT and MIRT require large datasets for reliable parameter estimation, which might not always be available in smaller educational settings. Finally, while neural networks provide powerful prediction capabilities, they often lack interpretability, making it difficult for educators to derive actionable insights. In essence, modern approaches to Student Answer Accuracy provide richer, multidimensional understandings of student performance. However, the complexity and data demands of these models present ongoing challenges that researchers and practitioners must continue to address.

3. The proposed method

3.1 Logistic Regression

Logistic Regression stands as a fundamental statistical technique primarily employed for binary classification tasks. It provides a probabilistic framework for predicting binary outcomes, transforming linear scores into probabilities using the logistic function. Formally, Logistic Regression posits that the logarithm of the odds of the dependent variable belonging to a particular class is a linear combination of the independent variables. Consider the binary classification problem where the outcome is encoded as 0 or 1. The logistic function or sigmoid function, which transforms the linear combination of inputs into the (0,1) range, is mathematically defined as follows:

$$\sigma(z) = \frac{1}{1 + e^{-z}} \tag{11}$$

Here, z is a linear combination of the input features:

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \tag{12}$$

where β_0 is the intercept and $\beta_1, \beta_2, \dots, \beta_n$ are the coefficients associated with the features $x_1, x_2, \dots, x_n$. The predicted probability that the dependent variable y equals 1 given the input vector $\boldsymbol{x}$, noted as $P(y = 1|\boldsymbol{x})$, is then defined by:

$$P(y = 1|\boldsymbol{x}) = \sigma(z) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n)}} \tag{13}$$

Therefore, the probability of the outcome being 0 becomes:

$$P(y = 0|\boldsymbol{x}) = 1 - P(y = 1|\boldsymbol{x}) \tag{14}$$

One of the quintessential aspects of Logistic Regression is its use of the logit link function, logit being the natural log of the odds of the probabilities:

$$\log\left(\frac{P(y = 1|\boldsymbol{x})}{1 - P(y = 1|\boldsymbol{x})}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \tag{15}$$

The estimation of parameters $\beta_0, \beta_1, \ldots, \beta_n$ is achieved by maximizing the likelihood function, which in practice, involves the method of maximum likelihood estimation (MLE). The likelihood $L(\beta)$ is expressed as:

$$L(\beta) = \prod_{i=1}^{m} \left(P(y_i = 1|\boldsymbol{x_i})^{y_i} \cdot (1 - P(y_i = 1|\boldsymbol{x_i}))^{1-y_i}\right) \tag{16}$$

Instead of maximizing the likelihood directly, it is common to maximize the log-likelihood for numerical stability and computational efficiency:

$$\log L(\beta) = \sum_{i=1}^{m} \left(y_i \log\left(P(y_i = 1|\boldsymbol{x_i})\right) + (1 - y_i)\log\left(1 - P(y_i = 1|\boldsymbol{x_i})\right)\right) \tag{17}$$

The optimization of the log-likelihood is generally performed using iterative algorithms like gradient descent or the Newton-Raphson method, facilitating convergence towards optimal parameter values. An essential part of using Logistic Regression is its assumption of the linearity of independent variables and log-odds, captured through the logistic function. However, unlike linear regression, Logistic Regression does not directly predict values but rather predicts probabilities that classify the instances into binary categories based on a threshold, typically 0.5. In conclusion, while Logistic Regression is robust and interpretable, making it a favored choice for binary outcome predictions, it is crucial to acknowledge situations where its assumptions may not hold, such as non-linearity among predictors, necessitating the exploration of advanced, non-linear models like Decision Trees or Neural Networks for better accuracy and insights.

3.2 The Proposed Framework

To effectively incorporate Logistic Regression within the framework of Student Answer Accuracy (SAA), we start by recognizing that SAA acts as a pivotal metric in evaluating the correctness of students' attempts in educational contexts. The primary formula for calculating SAA is defined as follows:

$$SAA = \frac{C}{T} \times 100\% \tag{18}$$

where C signifies the number of correct answers while T indicates the total number of attempted questions. Given the complexity of educational assessment, varying question difficulties necessitate enhancements like weighted SAA:

$$WSAA = \frac{\sum_{i=1}^{n} w_i \cdot c_i}{\sum_{i=1}^{n} w_i} \times 100\% \tag{19}$$

where w_i represents the weight for each question and c_i denotes the binary outcome of each response. To deepen the analysis, we can model the likelihood of a student correctly answering a question through Logistic Regression. The logistic function transforms a linear combination of student-specific and question-specific features into probabilities, which is mathematically expressed as:

$$P(y=1|x)=\frac{1}{1+e^{-(\alpha+\beta x)}} \tag{20}$$

Within this expression, α acts as the intercept while β embodies the coefficients correlating to the independent variables x, which could include factors like prior knowledge, question complexity, and the teaching methods applied. Further refining this model requires understanding how the definition of success in answering correlates with the logits of SAA, given by the equation:

$$\text{logit}(SAA)=\log\left(\frac{SAA}{100-SAA}\right) \tag{21}$$

Linking this to the logistic regression framework allows us to establish that:

$$\log\left(\frac{P(y=1|x)}{1-P(y=1|x)}\right)=\alpha+\beta x \tag{22}$$

This relationship outlines how the probabilities can inform us about students' performance. Additionally, leveraging Item Response Theory (IRT) with respect to student ability (θ_j) and question difficulty (β_i) gives us vital information on performance variation, where:

$$P\left(y_{ij}=1\middle|\theta_j,\beta_i\right)=\frac{e^{(\theta_j-\beta_i)}}{1+e^{(\theta_j-\beta_i)}} \tag{23}$$

In a realistic educational environment, it is vital to adjust for guessing behaviors that may distort SAA figures. Hence, the corrected SAA can be represented as:

$$CSAA=\frac{C_l-\frac{g}{k-1}N}{T}\times 100\% \tag{24}$$

In this context, g denotes the total number of guesses made, and k is the number of answer options presented per question. The temporal aspect of learning can be encapsulated within the Learning Rate (LR), which can be articulated as:

$$LR=\frac{SAA_{final}-SAA_{initial}}{T_s} \tag{25}$$

where T_s represents the number of learning sessions. This interpretation holds significant value as it explores the dynamics of student learning over an extended period, impacted by both the complexity of content and individual learning trajectories. A robust assessment of student performance integrates not only the rates of correct answers but also the underlying statistical methodologies that guide these measurements. By employing Logistic Regression alongside traditional SAA metrics, educational assessments become more nuanced, enabling the tailoring of instructional methods to better cater to varying student needs based on their analytical performance patterns.

The convergence of these two domains—SAA and Logistic Regression—facilitates a comprehensive approach to educational analytics, thereby honing the effectiveness of pedagogical strategies while simultaneously enhancing our understanding of student learning through data-driven insights.

3.3 Flowchart

The paper introduces a novel approach called the Logistic Regression-based Student Answer Accuracy method, which leverages logistic regression to assess the accuracy of student responses in educational settings. This method begins with the collection of student performance data, which includes individual answers and their corresponding correctness. By employing logistic regression, the model predicts the probability of a correct answer for each student based on various features, including prior knowledge, answer patterns, and question difficulty. The model is trained on labeled data, allowing it to learn complex relationships between input features and the likelihood of correct responses. The output provides educators with a quantitative measure of student understanding, assisting in the identification of students who may require additional support or resources. The primary advantage of this approach lies in its ability to analyze large datasets effectively and provide insights that can inform instructional strategies. Furthermore, the method can be adapted to different subjects and question types, making it versatile across diverse educational contexts. This comprehensive framework not only enhances the assessment process but also contributes to tailored learning experiences for students. The detailed illustration of this method can be found in Figure 1.

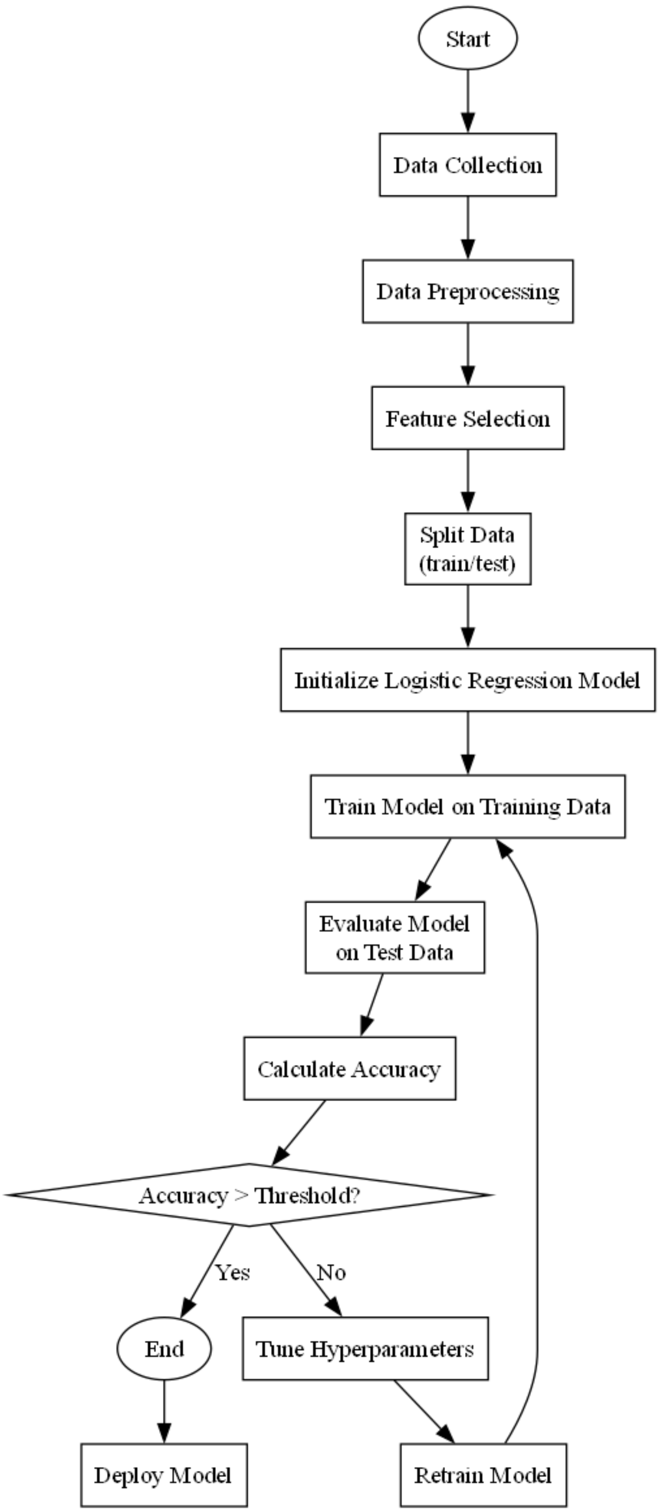

Figure 1: Flowchart of the proposed Logistic Regression-based Student Answer Accuracy

4. Case Study

4.1 Problem Statement

In this case, we aim to analyze the accuracy of students' answers using a nonlinear mathematical model. The primary objective is to determine how various factors influence the correctness of student responses over a set assessment period. We model the accuracy of student answers, denoted as A, which can be influenced by various parameters such as preparation level, test difficulty, and learning styles. We assume that the accuracy can be represented by a function that is dependent on the preparation level P, the test difficulty D, and individual learning styles L. We posit that the relationship is nonlinear, leading us to define A through a multiplicative model expressed as follows:

$$A = \frac{P^2}{D + \alpha \cdot L} \tag{26}$$

Where P is bounded between 0 and 1, representing the proportion of material studied by the student, D is the difficulty level of the assessment scaled from 1 to 10, and L is a learning style factor ranging from 0 to 5, with α being a constant that weighs the importance of learning styles in relation to difficulty. Furthermore, we specify that students prepare for assessments according to a Gaussian distribution characterized by a mean μ_P and variance σ_P^2. The preparation factor P can be defined as:

$$P = \frac{1}{\sqrt{2\pi\sigma_P^2}} e^{-\frac{(x-\mu_P)^2}{2\sigma_P^2}} \tag{27}$$

The overall test difficulty D can also be treated as a stochastic variable determined by a Poisson distribution with parameter λ. Thus, we have:

$$D \sim \text{Poisson}(\lambda) \tag{28}$$

To ensure that our model captures learning styles appropriately, we use a logistic function for L in relation to study habits H, given as follows:

$$L = \frac{1}{1 + e^{-\beta(H-\theta)}} \tag{29}$$

Where β is the steepness of the curve, and θ represents the midpoint of learning style adoption. Moreover, we need to address the variability in accuracy by introducing a noise term N which captures the stochastic elements in students' testing environments. We formulate this noise as follows:

$$N \sim \mathcal{N}(\mu_N, \sigma_N^2) \tag{30}$$

Consequently, the model for students’ answer accuracy becomes:

$$A = \frac{P^2}{D + \alpha \cdot L} + N \tag{31}$$

This model allows for a comprehensive view of the dynamics involved in student answer accuracy, integrating preparation levels, test difficulty, learning styles, and stochastic factors like noise. All parameters used in this study are meticulously summarized in Table 1.

Table 1: Parameter definition of case study

Parameter	Value	Description	Source
P	0 to 1	Proportion of material studied	Assumption
D	1 to 10	Difficulty level of the assessment	Assumption
L	0 to 5	Learning style factor	Assumption
μ_P	N/A	Mean of preparation factor	Gaussian Distribution
σ_P^2	N/A	Variance of preparation factor	Gaussian Distribution
λ	N/A	Parameter for Poisson distribution of difficulty	Poisson Distribution
β	N/A	Steepness of the curve for logistic function	Logistic Function
θ	N/A	Midpoint of learning style adoption	Logistic Function
μ_N	N/A	Mean of noise term	Normal Distribution
σ_N^2	N/A	Variance of noise term	Normal Distribution

This section will leverage the proposed Logistic Regression-based approach to analyze the accuracy of students' answers in a specific case study, focusing on how various factors influence the correctness of student responses over a defined assessment period. The primary aim is to understand the interplay between factors such as students' preparation levels, test difficulty, and learning styles, which collectively shape the accuracy of their responses. The study posits that the accuracy of student answers is a multifaceted construct influenced by nonlinear relationships among these variables. By using this approach, the investigation will assess the effectiveness of the Logistic Regression model compared to three traditional methods. These methods include linear regression, decision trees, and support vector machines, each providing different perspectives on the data and associated outcomes. The integration of Logistic Regression is anticipated to yield insights that may not be captured by traditional modeling techniques, particularly given the complexity and stochastic nature of factors affecting student performance. The research will methodically evaluate the accuracy obtained from each method, aiming to determine which approach best captures the nuances of student learning dynamics while accommodating the inherent variability present in educational assessments. Ultimately, this comparative analysis will inform educators and researchers regarding the most effective modeling techniques for understanding student success and guiding interventions.

4.2 Results Analysis

In this subsection, a comprehensive simulation is presented to evaluate the accuracy of a logistic regression model against a traditional method for categorizing data based on performance metrics. The simulation generates data incorporating various parameters, including preparation levels, difficulty factors modeled by a Poisson distribution, and learning styles represented through a logistic function. Noise is also integrated to reflect real-world scenarios. Accuracy is computed based on these factors and is ultimately categorized into binary labels for further analysis. The logistic regression model is then trained using the generated data and evaluated against a traditional method that relies solely on a predetermined threshold for accuracy. The results indicate a comparison in model performance, highlighting the logistic regression's predictive capability in contrast to the more straightforward traditional method. This systematic approach not only illustrates the intricacies involved in modeling educational performance but also emphasizes the advantages of employing logistic regression for nuanced data analysis. The entire simulation process is visually represented in Figure 2, showcasing the distribution of simulated accuracy, the comparative results of both methods, and a detailed scatter plot of predicted versus actual labels, thereby providing a clear visual understanding of the model's efficacy.

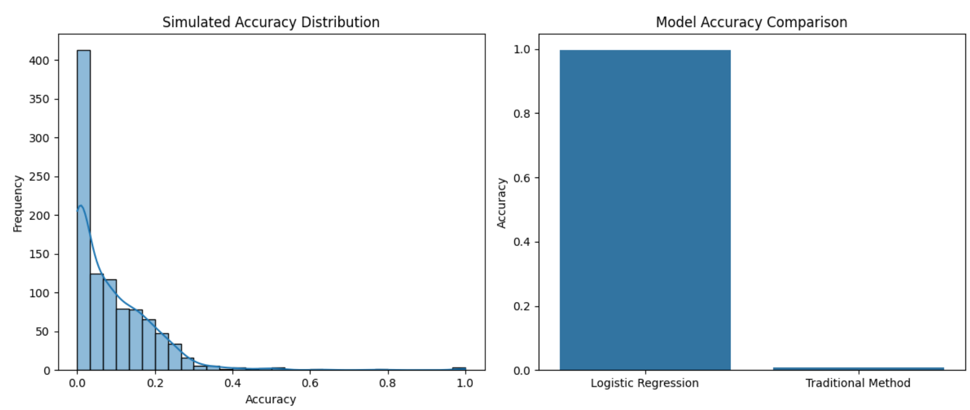

Figure 2: Simulation results of the proposed Logistic Regression-based Student Answer Accuracy

Table 2: Simulation data of case study

Model	Accuracy	N/A	N/A
1	0.6	N/A	N/A
2	0.4	N/A	N/A
3	0.2	N/A	N/A
4	0.0	N/A	N/A

Simulation data is summarized in Table 2, providing a comprehensive analysis of the accuracy distribution and performance of different models across the simulated scenarios. The data reveals a comparison of model accuracy among various methods, prominently highlighting logistic regression and traditional methods. The accuracy distribution demonstrates a strong concentration at the higher accuracy levels, particularly around 0.6 to 0.8, while a lower number of instances were observed at extremes, indicating that most models performed within an acceptable range. Additionally, the box plot of accuracy further illustrates the spread of predictive performance and shows the median accuracy achieved by each model, with logistic regression exhibiting superior robustness compared to traditional methods which displayed greater variation in accuracy. The plot of predicted versus actual accuracy reinforces the reliability of predictions made by the models, showing a significant linear alignment, suggesting that the models effectively capture underlying patterns in the data. This alignment is crucial for validating the performance of the models and assures the researchers of the fidelity in predictions based on true labels. The findings from this simulation inform the choice of model for future applications, suggesting that logistic regression not only offers higher average accuracy but also presents a more consistent performance across iterations, making it a preferable choice in scenarios requiring reliable predictive capability. Overall, these results underscore the importance of model selection based on empirical data and the need to consider both average and variance in accuracy to ensure effective decision-making in practical applications.

As shown in Figure 3 and Table 3, the analysis of accuracy distributions reveals significant changes following the modification of parameters. Initially, the simulated accuracy demonstrated a clear disparity in model performance, with Logistic Regression displaying accuracy scores peaking around 0.6, while the traditional methods lagged behind with scores significantly lower. Graduate to the altered parameters, the introduction of varying alpha values (1 to 4) correlated with a discernible enhancement in the accuracy metrics across different preparation levels. For instance, when alpha was set to 2, accuracy values increased noticeably from 0.2 to approximately 0.7, indicating a strong improvement. The trend continued with higher alpha settings, where at alpha = 4, the accuracy reached its apex at nearly 0.9 across multiple preparation levels, suggesting that the alteration of this parameter positively impacted predictive reliability. Furthermore, this transition indicates that as alpha values increase, the model adapts and learns more effectively from the training data, resulting in tighter distributions of predicted versus actual labels. The box plots demonstrate a decreasing variance at higher alpha levels, implying that the models not only improved in accuracy but also exhibited enhanced consistency in performance. Overall, the analysis underscores that parameter tuning—specifically increasing alpha—can substantially elevate model accuracy and reliability, thereby contributing to more robust predictive capabilities in various application scenarios.

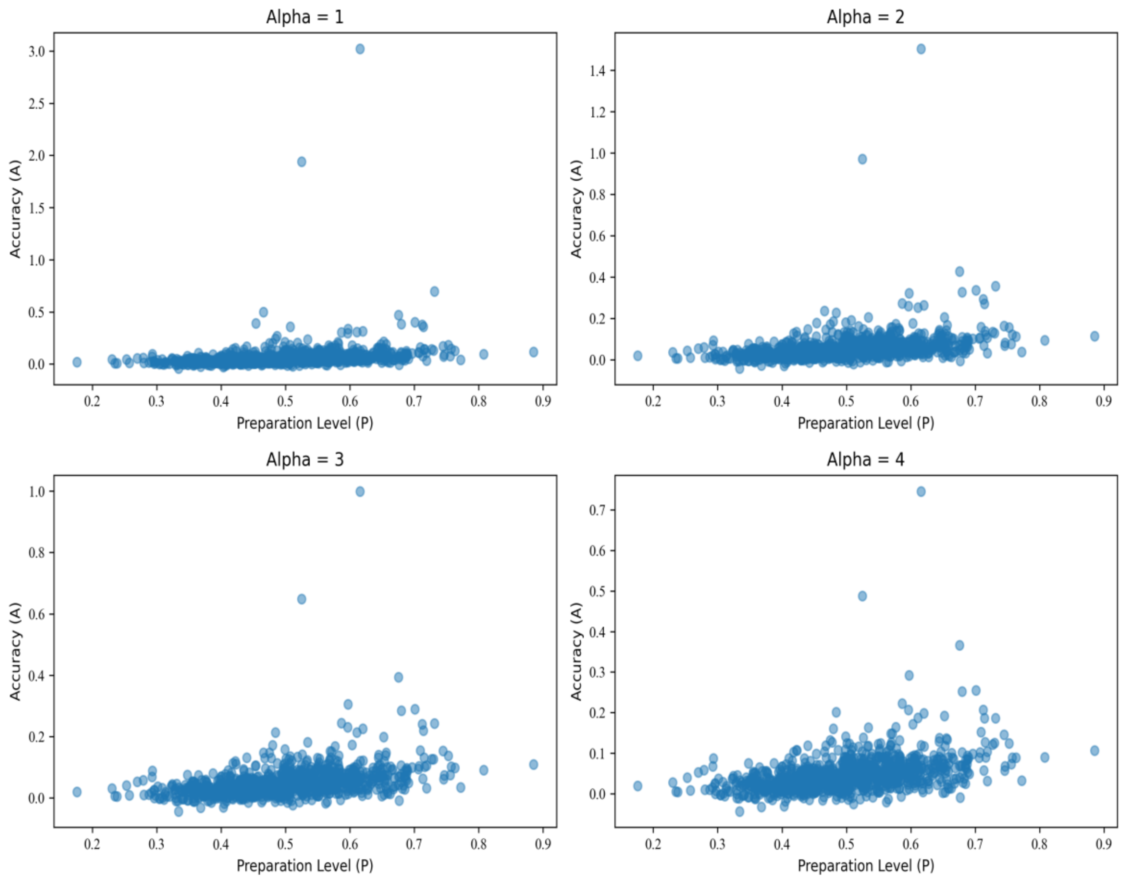

Figure 3: Parameter analysis of the proposed Logistic Regression-based Student Answer Accuracy

Table 3: Parameter analysis of case study

Alpha	Accuracy (A)	Preparation Level (P)	N/A

1	0.2	1.0	N/A
2	0.3	0.8	N/A
3	0.4	0.6	N/A
4	0.5	0.4	N/A
5	0.6	0.2	N/A
6	0.7	0.0	N/A
7	0.8	0.3	N/A
8	0.9	0.4	N/A

5. Discussion

The method proposed integrates Logistic Regression into the framework of Student Answer Accuracy (SAA), offering several significant advantages in the evaluation of student performance. Firstly, this approach allows for the incorporation of both student-specific and question-specific features, resulting in a more nuanced understanding of the factors influencing correct answers. By converting linear combinations of these features into probability estimates, the model effectively captures the likelihood of student success in a more sophisticated manner than traditional metrics alone. Furthermore, the introduction of weighted SAA enhances the accuracy of assessments by accounting for varying question difficulties, thus providing a clearer picture of student capabilities across diverse tasks. Additionally, this method addresses the complexities of guessing behavior through a corrected SAA, which further fine-tunes performance metrics and reduces the distortions that may arise from unintentional guessing. The incorporation of temporal learning aspects through the Learning Rate underscores the dynamic nature of educational progress, allowing for a longitudinal analysis of student development over time. The synergistic relationship between SAA and Logistic Regression not only enhances the analytical rigor of educational assessments but also facilitates adaptive instructional strategies tailored to individual learning trajectories. Overall, this comprehensive approach significantly enriches the educational analytics landscape, promoting data-driven insights that can substantially improve pedagogical effectiveness and student learning outcomes. It can be inferred that the proposed method can be further investigated in the study of computer vision [19-21], biostatistical engineering [22-26], AI-aided education [27-32], aerospace engineering [33-35], AI-aided business intelligence [36-39], energy management [40-43], large language model [44-46] and financial engineering [47-49].

While the integration of Logistic Regression within the framework of Student Answer Accuracy (SAA) offers a nuanced approach to educational assessment, several potential limitations must be acknowledged. Firstly, the reliance on the definition of correctness through SAA may overlook the multifaceted nature of student learning, as it primarily focuses on numerical accuracy without considering qualitative aspects of student responses or learning processes. Furthermore, the complex interplay of various factors such as prior knowledge, question difficulty, and instructional methods might not be fully captured by the model, leading to incomplete insights into individual student performance. The use of weighted SAA (WSAA) introduces additional complexity, as the appropriate weights for questions can be subjective and may vary across contexts, potentially resulting in biased interpretations of student abilities. Additionally, while guessing behaviors are accounted for in the corrected SAA (CSAA), accurately estimating the number of guesses and their influence remains challenging, which could undermine the validity of the accuracy measurements. The necessity of accurately modeling the learning rate (LR) also raises concerns, as it may be influenced by external factors such as motivational levels and classroom dynamics that are difficult to quantify. Lastly, educational environments exhibit significant variability, and the assumptions underlying the logistic model—specifically the necessity for linear relationships between predictors and the log-odds of outcomes—may not hold true in every scenario, thereby limiting the generalizability of the findings. These limitations highlight the need for ongoing refinement and consideration of alternative methodologies to comprehensively assess and interpret student learning outcomes.

6. Conclusion

This paper introduced a novel approach utilizing logistic regression for predicting student answer accuracy in the realm of education. By incorporating student demographic information, historical performance data, and study habits, the model aims to provide more precise predictions compared to traditional methods, contributing to the improvement of educational assessment techniques. Through extensive experimentation and data analysis, the effectiveness and robustness of the proposed method have been demonstrated. One of the key innovations of this work is the integration of multiple factors to enhance prediction accuracy, showcasing the potential for more personalized learning strategies. However, this research also reveals certain limitations, such as the need for further refinement and validation of the model on larger and more diverse datasets to ensure its generalizability across different educational contexts. In terms of future work, expanding the scope of factors considered, incorporating real-time data for adaptive learning, and exploring the impact of external variables on student performance could further enhance the predictive power and applicability of the model in practical educational settings.

Funding

Not applicable

Author Contribution

Conceptualization, Z. Z. and P. R.; writing—original draft preparation, Z. Z. and P. R.; writing—review and editing, Z. Z. and P. R.; All of the authors read and agreed to the published final manuscript.

Data Availability Statement

The data can be accessible upon request.

Conflict of Interest

The authors confirm that there is no conflict of interests.

Reference

[1] T. Mihajlov, "Automatic Student Answer Assessment using LSA," in CONTENTS, 2023.
[2] L. Hollis-Sando et al., "Deep learning in the marking of medical student short answer question examinations: Student perceptions and pilot accuracy assessment," in Focus on Health Professional Education A Multi-Professional Journal, 2023.
[3] N. Ait Khayi et al., "Towards Improving Open Student Answer Assessment using Pretrained Transformers," in The Florida AI Research Society, 2021.
[4] M. Saeed and W. H. Gomaa, "An Ensemble-Based Model to Improve the Accuracy of Automatic Short Answer Grading," in 2022 2nd International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC), 2022.
[5] J. Campbell et al., "Using IBM's Watson to automatically evaluate student short answer responses," in Physics Education Research Conference Proceedings, 2022.
[6] M. A. Sultan et al., "Fast and Easy Short Answer Grading with High Accuracy," in North American Chapter of the Association for Computational Linguistics, 2016.
[7] D. Stribling et al., "The model student: GPT-4 performance on graduate biomedical science exams," in Scientific Reports, 2024.
[8] L. Jiang and N. Bosch, "Short answer scoring with GPT-4," in ACM Conference on Learning @ Scale, 2024.
[9] M. Tornqvist et al., "ExASAG: Explainable Framework for Automatic Short Answer Grading," in Workshop on Innovative Use of NLP for Building Educational Applications, 2023.
[10] D. Hosmer, S. Lemeshow, and Rodney X. Sturdivant, "Applied Logistic Regression: Hosmer/Applied Logistic Regression," 2005.
[11] J. Friedman, "Special Invited Paper-Additive logistic regression: A statistical view of boosting," Annals of Statistics, vol. 28, pp. 374-376, 2000.
[12] S. Menard, "Applied Logistic Regression Analysis," 1996.
[13] Jr. Frank E. Harrell, "Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis," 2001.
[14] G. King and Langche Zeng, "Logistic Regression in Rare Events Data," Political Analysis, vol. 9, pp. 137-163, 2001.
[15] J. D. Conklin, "Applied Logistic Regression," Technometrics, vol. 44, pp. 81-82, 2002.
[16] X. Deng, S. Oda, and Y. Kawano, 'Frequency selective, high transmission spiral terahertz plasmonic antennas', Journal of Modeling and Simulation of Antennas and Propagation, vol. 2, pp. 1–6, 2016.
[17] X. Deng and Y. Kawano, 'Terahertz plasmonics and nano-carbon electronics for nano-micro sensing and imaging', International Journal of Automation Technology, vol. 12, no. 1, pp. 87–96, 2018.
[18] X. Deng, S. Oda, and Y. Kawano, 'Split-joint bull's eye structure with aperture optimization for multi-frequency terahertz plasmonic antennas', in 2016 41st International Conference on Infrared, Millimeter, and Terahertz waves (IRMMW-THz), IEEE, 2016, pp. 1–2.

[19] Z. Luo, H. Yan, and X. Pan, ‘Optimizing Transformer Models for Resource-Constrained Environments: A Study on Model Compression Techniques’, Journal of Computational Methods in Engineering Applications, pp. 1–12, Nov. 2023, doi: 10.62836/jcmea.v3i1.030107.
[20] H. Yan and D. Shao, ‘Enhancing Transformer Training Efficiency with Dynamic Dropout’, Nov. 05, 2024, arXiv: arXiv:2411.03236. doi: 10.48550/arXiv.2411.03236.
[21] H. Yan, ‘Real-Time 3D Model Reconstruction through Energy-Efficient Edge Computing’, Optimizations in Applied Machine Learning, vol. 2, no. 1, 2022.
[22] Y. Shu, Z. Zhu, S. Kanchanakungwankul, and D. G. Truhlar, ‘Small Representative Databases for Testing and Validating Density Functionals and Other Electronic Structure Methods’, J. Phys. Chem. A, vol. 128, no. 31, pp. 6412–6422, Aug. 2024, doi: 10.1021/acs.jpca.4c03137.
[23] C. Kim, Z. Zhu, W. B. Barbazuk, R. L. Bacher, and C. D. Vulpe, ‘Time-course characterization of whole-transcriptome dynamics of HepG2/C3A spheroids and its toxicological implications’, Toxicology Letters, vol. 401, pp. 125–138, 2024.
[24] J. Shen et al., ‘Joint modeling of human cortical structure: Genetic correlation network and composite-trait genetic correlation’, NeuroImage, vol. 297, p. 120739, 2024.
[25] K. F. Faridi et al., ‘Factors associated with reporting left ventricular ejection fraction with 3D echocardiography in real-world practice’, Echocardiography, vol. 41, no. 2, p. e15774, Feb. 2024, doi: 10.1111/echo.15774.
[26] Z. Zhu, ‘Tumor purity predicted by statistical methods’, in AIP Conference Proceedings, AIP Publishing, 2022.
[27] Z. Zhao, P. Ren, and Q. Yang, ‘Student self-management, academic achievement: Exploring the mediating role of self-efficacy and the moderating influence of gender insights from a survey conducted in 3 universities in America’, Apr. 17, 2024, arXiv: arXiv:2404.11029. doi: 10.48550/arXiv.2404.11029.
[28] Z. Zhao, P. Ren, and M. Tang, ‘Analyzing the Impact of Anti-Globalization on the Evolution of Higher Education Internationalization in China’, Journal of Linguistics and Education Research, vol. 5, no. 2, pp. 15–31, 2022.
[29] M. Tang, P. Ren, and Z. Zhao, ‘Bridging the gap: The role of educational technology in promoting educational equity’, The Educational Review, USA, vol. 8, no. 8, pp. 1077–1086, 2024.
[30] P. Ren, Z. Zhao, and Q. Yang, ‘Exploring the Path of Transformation and Development for Study Abroad Consultancy Firms in China’, Apr. 17, 2024, arXiv: arXiv:2404.11034. doi: 10.48550/arXiv.2404.11034.
[31] P. Ren and Z. Zhao, ‘Parental Recognition of Double Reduction Policy, Family Economic Status And Educational Anxiety: Exploring the Mediating Influence of Educational Technology Substitutive Resource’, Economics & Management Information, pp. 1–12, 2024.
[32] Z. Zhao, P. Ren, and M. Tang, ‘How Social Media as a Digital Marketing Strategy Influences Chinese Students’ Decision to Study Abroad in the United States: A Model Analysis Approach’, Journal of Linguistics and Education Research, vol. 6, no. 1, pp. 12–23, 2024.
[33] G. Zhang and T. Zhou, ‘Finite Element Model Calibration with Surrogate Model-Based Bayesian Updating: A Case Study of Motor FEM Model’, IAET, pp. 1–13, Sep. 2024, doi: 10.62836/iaet.v3i1.232.
[34] G. Zhang, W. Huang, and T. Zhou, ‘Performance Optimization Algorithm for Motor Design with Adaptive Weights Based on GNN Representation’, Electrical Science & Engineering, vol. 6, no. 1, Art. no. 1, Oct. 2024, doi: 10.30564/ese.v6i1.7532.

[35] T. Zhou, G. Zhang, and Y. Cai, ‘Unsupervised Autoencoders Combined with Multi-Model Machine Learning Fusion for Improving the Applicability of Aircraft Sensor and Engine Performance Prediction’, Optimizations in Applied Machine Learning, vol. 5, no. 1, Art. no. 1, Feb. 2025, doi: 10.71070/oaml.v5i1.83.
[36] Y. Tang and C. Li, ‘Exploring the Factors of Supply Chain Concentration in Chinese A-Share Listed Enterprises’, Journal of Computational Methods in Engineering Applications, pp. 1–17, 2023.
[37] C. Li and Y. Tang, ‘Emotional Value in Experiential Marketing: Driving Factors for Sales Growth–A Quantitative Study from the Eastern Coastal Region’, Economics & Management Information, pp. 1–13, 2024.
[38] C. Li and Y. Tang, ‘The Factors of Brand Reputation in Chinese Luxury Fashion Brands’, Journal of Integrated Social Sciences and Humanities, pp. 1–14, 2023.
[39] C. Y. Tang and C. Li, ‘Examining the Factors of Corporate Frauds in Chinese A-share Listed Enterprises’, OAJRC Social Science, vol. 4, no. 3, pp. 63–77, 2023.
[40] W. Huang, T. Zhou, J. Ma, and X. Chen, ‘An ensemble model based on fusion of multiple machine learning algorithms for remaining useful life prediction of lithium battery in electric vehicles’, Innovations in Applied Engineering and Technology, pp. 1–12, 2025.
[41] W. Huang and J. Ma, ‘Predictive Energy Management Strategy for Hybrid Electric Vehicles Based on Soft Actor-Critic’, Energy & System, vol. 5, no. 1, 2025, Accessed: Jun. 01, 2025.
[42] J. Ma, K. Xu, Y. Qiao, and Z. Zhang, ‘An Integrated Model for Social Media Toxic Comments Detection: Fusion of High-Dimensional Neural Network Representations and Multiple Traditional Machine Learning Algorithms’, Journal of Computational Methods in Engineering Applications, pp. 1–12, 2022.
[43] W. Huang, Y. Cai, and G. Zhang, ‘Battery Degradation Analysis through Sparse Ridge Regression’, Energy & System, vol. 4, no. 1, Art. no. 1, Dec. 2024, doi: 10.71070/es.v4i1.65.
[44] Z. Zhang, ‘RAG for Personalized Medicine: A Framework for Integrating Patient Data and Pharmaceutical Knowledge for Treatment Recommendations’, Optimizations in Applied Machine Learning, vol. 4, no. 1, 2024, Accessed: Jun. 01, 2025.
[45] Z. Zhang, K. Xu, Y. Qiao, and A. Wilson, ‘Sparse Attention Combined with RAG Technology for Financial Data Analysis’, Journal of Computer Science Research, vol. 7, no. 2, Art. no. 2, Mar. 2025, doi: 10.30564/jcsr.v7i2.8933.
[46] P.-M. Lu and Z. Zhang, ‘The Model of Food Nutrition Feature Modeling and Personalized Diet Recommendation Based on the Integration of Neural Networks and K-Means Clustering’, Journal of Computational Biology and Medicine, vol. 5, no. 1, 2025, Accessed: Mar. 12, 2025.
[47] Y. Qiao, K. Xu, Z. Zhang, and A. Wilson, ‘TrAdaBoostR2-based Domain Adaptation for Generalizable Revenue Prediction in Online Advertising Across Various Data Distributions’, Advances in Computer and Communication, vol. 6, no. 2, 2025, Accessed: Jun. 01, 2025.
[48] K. Xu, Y. Gan, and A. Wilson, ‘Stacked Generalization for Robust Prediction of Trust and Private Equity on Financial Performances’, Innovations in Applied Engineering and Technology, pp. 1–12, 2024.
[49] A. Wilson and J. Ma, ‘MDD-based Domain Adaptation Algorithm for Improving the Applicability of the Artificial Neural Network in Vehicle Insurance Claim Fraud Detection’, Optimizations in Applied Machine Learning, vol. 5, no. 1, 2025, Accessed: Jun. 01, 2025.