BULLETIN OF EDUCATION AND PSYCHOLOGY
Research Article* | *Volume 5* | *Issue 5* | *Apr 2025
Received: 27 Feb 2025 | Revised: 21 Mar 2025
Accepted: 26 Mar 2025 | Published Online: 02 Apr 2025

Random Forest-Based Early Warning System for Student Dropout Using Behavioral Data

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: This paper addresses the development of a Random Forest-based early warning system for student dropout utilizing behavioral data. Student dropout is a significant issue in educational institutions, impacting student success and institutional effectiveness. Current research in the field faces challenges in accurately predicting dropout risks due to the complexity and diversity of student behaviors. To tackle this issue, this study proposes an innovative approach that leverages Random Forest algorithm to analyze diverse behavioral data and effectively identify students at risk of dropping out. The system's design and implementation, incorporating machine learning techniques, offer a more accurate and efficient method for early identification of dropout risks, facilitating timely interventions to support student retention and success in academic settings.

Keywords: *Random Forest; Early Warning System; Student Dropout; Behavioral Data; Machine Learning Techniques*

1. Introduction

The study of student dropout behavior involves investigating the factors and circumstances that contribute to students leaving educational programs before completion. Common areas of focus include academic challenges, personal circumstances, socioeconomic factors, and institutional support systems. Currently, key challenges in this field include the complexity of identifying predictive factors, the diversity of student populations and their unique needs, the limitations of available data sources, and the difficulty of implementing effective intervention strategies. Addressing these challenges requires interdisciplinary collaboration, advanced research methodologies, and a comprehensive understanding of the multifaceted nature of student dropout behavior. Further research efforts are needed to develop targeted interventions and support mechanisms to mitigate dropout rates and promote academic success for all students.

To this end, research on student dropout behavior has advanced to encompass various interdisciplinary perspectives, including psychology, sociology, and education. Current studies focus on identifying risk factors, developing targeted interventions, and evaluating the effectiveness of dropout prevention programs. Research on student dropout behavior has been a fundamental concern in educational studies. Blackwell (2020) utilized event history analysis to examine nursing student dropout behavior, emphasizing the importance of understanding factors contributing to dropout rates [1]. Felice (1981) explored the disengagement of black students from school, highlighting the role of racial discrimination and educational output equality in occupational success [2]. Wild et al. (2024) investigated the impact of interest and support on student dropout, revealing that interest mediates the relationship between support and dropout rates [3]. Goren et al. (2024) developed machine learning models for early prediction of student dropout in higher education, aiming to facilitate timely interventions to improve student retention [4]. Jin (2020) proposed a dropout prediction model for MOOC students based on learning behavior features and parameter optimization, demonstrating improved predictive performance through feature extraction and intelligent model design [5]. Prasanth and Alqahtani (2023) leveraged machine learning techniques to predict student behavior for early dropout detection in universities, underlining the significance of identifying at-risk students for tailored interventions [6]. Giomi (2004) explored early predictors of student dropout behavior, shedding light on factors influencing student disengagement from academic pursuits [7]. Capuccinello (2011) conducted an econometric analysis of student dropout determinants, providing insights into further education sector challenges in England [8]. Choudhari et al. (2024) discussed the importance of educational performance analysis and dropout visualization, advocating for the use of education prediction systems to enhance student success and retention [9]. Poitras et al. (2020) introduced a dimensionality reduction method for time series analysis of student behavior in MOOCs, aiming to predict dropout and improve online learning outcomes [10]. Research on student dropout behavior is crucial in educational studies. Machine learning techniques like Random Forest are utilized for early prediction, enabling timely interventions to improve student retention. This approach enhances predictive accuracy and facilitates tailored interventions for at-risk students, ultimately contributing to the success and retention of students in higher education.

Specifically, Random Forest is utilized in predicting Student Dropout Behavior by leveraging its ensemble learning technique to analyze various predictive features and improve the accuracy of dropout risk identification. This approach allows for a more comprehensive understanding of the factors contributing to student attrition in academic settings. Random Forest, introduced by Breiman in 2001, is a powerful machine learning algorithm that combines multiple randomized decision trees to provide accurate predictions in classification and regression tasks [11]. One of the key strengths of Random Forest is its ability to handle unbalanced data, variables with missing values, and mitigate overfitting issues by training decision trees on random subsets of data, leading to improved generalization to new data [12]. The algorithm's versatility and efficiency have made it widely utilized in various fields such as remote sensing, where it has been successfully applied for tasks like image classification [13]. Researchers have also explored the impact of hyperparameters on Random Forest performance, emphasizing the importance of tuning strategies to further enhance its predictive abilities [14]. Furthermore, extensions of the standard Random Forest method have been developed for longitudinal data analysis, showcasing the algorithm's

adaptability to different data structures and its potential for high-dimensional data applications [15]. Overall, Random Forest stands out as a top-performing technique in machine learning, offering significant benefits in terms of accuracy, robustness, and versatility [16-18]. However, limitations of Random Forest include interpretability challenges and computational complexity, especially with large datasets. Additionally, the algorithm may struggle with highly correlated features and noisy data, impacting its predictive performance.

To overcome those limitations, this paper aims to develop a Random Forest-based early warning system for student dropout using behavioral data. Student dropout poses a significant challenge for educational institutions, affecting both student outcomes and institutional effectiveness. The current research in this area struggles with accurately predicting dropout risks due to the intricate nature and variety of student behaviors. In response, this study introduces an innovative approach that utilizes the Random Forest algorithm to analyze a wide range of behavioral data effectively, enabling the identification of students at risk of dropping out. The design and implementation of the system incorporate machine learning techniques to provide a more precise and efficient method for early detection of dropout risks. This approach allows for timely interventions to be implemented, ultimately supporting student retention and promoting success in academic environments. The detailed analysis of diverse behavioral data, coupled with the robust Random Forest algorithm, enhances the system's predictive capabilities, empowering educational institutions to address student dropout proactively and holistically. By leveraging cutting-edge technology and methodologies, this research aims to revolutionize the approach to tackling the challenge of student dropout and promoting a supportive and successful educational environment.

This paper addresses the development of a Random Forest-based early warning system for student dropout utilizing behavioral data. Student dropout is a significant issue in educational institutions, impacting student success and institutional effectiveness. Current research in the field faces challenges in accurately predicting dropout risks due to the complexity and diversity of student behaviors. To tackle this issue, this study proposes an innovative approach that leverages Random Forest algorithm to analyze diverse behavioral data and effectively identify students at risk of dropping out. The system's design and implementation, incorporating machine learning techniques, offer a more accurate and efficient method for early identification of dropout risks, facilitating timely interventions to support student retention and success in academic settings.

2. Background

2.1 Student Dropout Behavior

Student dropout behavior refers to the phenomenon where students discontinue their education before completing their program, which can occur at any educational level, from high school to higher education. This behavior has significant implications for both individual students and educational institutions. Understanding and modeling this behavior requires a comprehensive approach that considers various factors, including academic performance, socio-economic status, institutional characteristics, psychological factors, and more. To formalize the study of student dropout behavior, we can use several mathematical models and expressions. Let's denote the

probability of a student dropping out as P_d . This probability can be modeled as a function of several variables:

$$P_d = f(a, s, e, p, c) \tag{1}$$

where a represents academic factors, s represents socio-economic status, e represents environmental factors, p represents psychological factors, and c represents characteristics specific to the institution. This can be expressed as the student's performance metrics, such as GPA or test scores. We can define an academic risk factor R_a , as:

$$R_a = \frac{1}{1 + \exp(-\beta_a \cdot A)} \tag{2}$$

where A is the academic score or GPA, and β_a is the weight representing the impact of academic performance on dropout risk. Factors such as family income, parental education levels, and employment status are critical. Let I denote family income:

$$R_s = 1 - \frac{I}{I_{max}} \tag{3}$$

where I_{max} is the maximum income expected in the study, assuming higher income reduces dropout likelihood. These can include peer influences, local community characteristics, and availability of resources. We can model environmental influence E_i as:

$$E_i = \alpha \cdot \text{PeerSupport} + \gamma \cdot \text{ResourceAvailability} \tag{4}$$

where α and γ are respective weights. Students' mental health and motivation are encapsulated by factors such as stress levels and academic motivation. Let S be stress and M be motivation:

$$P_p = \frac{S}{S + M} \tag{5}$$

where P_p is the psychological propensity to drop out. These are the aspects of the institution, such as support services and campus facilities. We can denote institutional support S_i as:

$$S_i = \delta \cdot \text{AdvisingQuality} + \epsilon \cdot \text{CampusFacilities} \tag{6}$$

where δ and ϵ represent importance weights of respective factors. The comprehensive risk of dropout, P_d , is then a composite function of these independent factors. Assuming they contribute multiplicatively, we could use a function such as:

$$P_d = R_a \times R_s \times E_i \times P_p \times \frac{1}{S_i} \tag{7}$$

This multiplicative model implies that a high value in any risk component significantly affects the overall dropout probability. It's crucial to calibrate these models with real-world data through methods like regression analysis or machine learning techniques. This helps identify the true weights and thresholds for each factor, enhancing the predictive accuracy of the model. By

understanding and quantifying these dimensions of student dropout behavior, educational institutions can implement targeted interventions designed to mitigate the risk factors and optimize student retention strategies.

2.2 Methodologies & Limitations

Student dropout behavior is a complex and multifaceted phenomenon that calls for a variety of methodological approaches to effectively model and understand the underlying factors contributing to it. Among the prevalent techniques are logistic regression, decision trees, neural networks, and survival analysis. Each of these methodologies has unique strengths and limitations when it comes to predicting student dropout rates. Logistic regression is one of the most commonly used statistical models for binary classification problems like dropout prediction. It models the probability of dropout, P_d , as a logistic function of independent variables. Here's the basic logistic function:

$$P_d = \frac{1}{1 + \exp\big(-(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n)\big)} \tag{8}$$

where θ_0 is the intercept, θ_i are coefficients, and x_i are the independent variables. However, logistic regression assumes a linear relationship between the predictors and the log odds of the response, which may not capture complex nonlinear dependencies. Decision trees, another popular approach, involve splitting data into branches to make predictions based on learned rules. They are intuitive and easy to interpret, yet they can easily overfit the data without proper regularization techniques like pruning. The tree's decision boundaries are piecewise linear, causing limitations when modeling more intricate structures. Neural networks, particularly deep learning models, exhibit flexibility and power in capturing complex patterns by using multiple layers of interconnected neurons. The dropout probability, P_d , can be represented as:

$$P_d = \sigma\big(W_n(\sigma(W_{n-1}(\ldots\sigma(W_1 x + b_1)\ldots) + b_{n-1}) + b_n)\big) \tag{9}$$

where W_i and b_i are weights and biases, σ is a nonlinear activation function (e.g., ReLU, sigmoid), and x is the input vector. Despite their power, neural networks require substantial computational resources and large datasets, risking overfitting if not properly managed through techniques such as dropout regularization. Survival analysis, traditionally used in life sciences, can model the time until an event occurs, making it apt for predicting dropout timing. The hazard function, $h(t)$, provides the instantaneous dropout rate at time t , with:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(\, t \le T < t + \Delta t \mid T \ge t\,)}{\Delta t} \tag{10}$$

Here, T is the dropout time. The challenges in survival analysis include handling censored data and the assumption of proportional hazards, which might not always hold true. While these methods provide diverse tools for modeling dropout behavior, they also present limitations such as high variance, computational intensity, and the need for extensive tuning of hyperparameters. Moreover, these models' reliance on historical and quantitative data may overlook qualitative factors like student sentiment and personal challenges, highlighting the importance of augmenting predictive models with qualitative insights to foster holistic dropout prevention strategies. To address these

challenges, hybrid models that integrate multiple approaches, such as combining logistic regression with decision tree methods or employing ensemble techniques like random forests, may offer improved accuracy and robustness in dropout predictions. Continuous refinement through real-time data calibration is imperative to adapt to changing educational environments and to fine-tune models' sensitivity to diverse dropout predictors.

3. The proposed method

3.1 Random Forest

Random forests, a versatile ensemble learning technique, have emerged as a robust method for classification and regression tasks across various domains, including predicting complex phenomena such as student dropout behaviors. The random forest algorithm builds upon decision trees by creating a multitude of trees during training and outputting the mode of the classes (for classification) or mean prediction (for regression) of the individual trees. This method addresses the limitations of single decision trees such as overfitting and high variance by leveraging the power of multiple classifiers. To construct a random forest, one begins by generating multiple decision trees. Each tree is created from a bootstrap sample of the data, meaning it is trained on a subset of the observations chosen randomly with replacement. A critical aspect is the random selection of features at each split in the tree, which decorrelates the trees and adds diversity. Let us delve into the mathematical formulation that underpins the mechanism of random forests. Consider a dataset with N samples and M features. For each tree in the forest, we randomly select a sample with replacement, termed as bootstrap sampling. Let X_i represent the i^{th} selected sample:

$$X_i \sim \text{Bootstrap Sample of } X, i = 1,2,\dots,N \tag{11}$$

Once the bootstrap sample is chosen, a decision tree is grown. At each node of the tree, rather than considering all M features, we select a random subset of these features, denoted by m try, where m try is much less than M. This randomness is a hallmark of random forests:

$$\text{Select } m \text{ features, where } m = m\text{try of } M \tag{12}$$

The best feature to split the node is chosen based on a purity criterion such as Gini impurity or information gain. For classification, the Gini impurity G can be expressed as:

$$G = 1 - \sum_{k=1}^{K} p_k^2 \tag{13}$$

where p_k is the probability of a sample belonging to class k. For regression tasks, the criterion might be the mean squared error (MSE). After training several trees, predictions for a new input x are made by aggregating the predictions of all individual trees in the forest. For classification, this is typically done via a majority vote:

$$y = \text{mode}T_1(x), T_2(x), \dots, T_B(x) \tag{14}$$

For regression, predictions are averaged over all trees:

$$y = \frac{1}{B}\sum_{b=1}^{B} T_b(x) \tag{15}$$

where $T_b(x)$ is the prediction from the b^{th} tree and B is the total number of trees in the forest. Random forests boast several advantages, notably their ability to handle large datasets with higher dimensional featuresspaces, and their robustness against overfitting due to the averaging (in regression) or voting mechanism (in classification) inherent in the ensemble approach. The performance of random forests is often superior to single models due to its strength in reducing overfitting and variance while being able to model complex interactions between variables inherently. Additionally, random forests offer feature importance scores, providing insight into the predictive power of different variables. For example, the importance of feature j can be assessed by the total decrease in node impurity it brings about in the forest, averaged over all trees:

$$\text{Importance}(j) = \frac{1}{B}\sum_{b=1}^{B} \Delta I_{j,b} \tag{16}$$

where $\Delta I_{j,b}$ is the decrease in impurity from splits involving feature j in tree b . Despite their numerous advantages, random forests are computationally intensive and require careful tuning of hyperparameters such as the number of trees B and the number of features $m$try. Moreover, while random forests can handle missing data with ease, they assume that the samples are independently and identically distributed, a condition that might not always hold in complex datasets such as student dropout scenarios where time-dependencies and interactions may prevail. Nevertheless, through ongoing developments and hybridization with other techniques, random forests continue to be a potent tool for capturing the faceted complexities of real-world problems.

3.2 The Proposed Framework

The application of Random Forest methods to predict student dropout behavior provides a robust framework for analyzing a multifaceted issue in education. Student dropout behavior involves various influencing factors, including academic performance, socio-economic status, and psychological variables. The ability of Random Forests to model these complexities arises from their ensemble learning mechanism, which builds a multitude of decision trees while mitigating the risks associated with overfitting. To frame the probability of a student dropping out, we denote this as P_d . The relationship of P_d with academic, socio-economic, environmental, psychological, and institutional factors can be defined as:

$$P_d = f(a, s, e, p, c) \tag{17}$$

This model indicates that dropout probability is contingent upon a compilation of risk factors. Notably, risk associated with academic performance can be quantified through a logistic function, expressed as:

$$R_a = \frac{1}{1 + \exp(-\beta_a \cdot A)} \tag{18}$$

where A is the academic metric (e.g., GPA), and β_a reflects its weight. Other risk factors, such as socio-economic status s, can be generated through:

$$R_s = 1 - \frac{I}{I_{max}} \tag{19}$$

In the context of Random Forests, we start with a dataset comprising N samples and M features. Random samples are generated through bootstrap sampling, allowing us to derive a diverse set of decision trees. Each sample can be defined as:

$$X_i \sim \text{Bootstrap Sample of } X, i = 1,2,\dots,N \tag{20}$$

Each tree in the forest selects $m$try features when determining splits. The function for selecting features at each node in the tree can be formally represented as:

$$m = m\text{try of } M \tag{21}$$

As decisions are made during tree construction, factors that influence dropout—such as R_a and R_s —contribute to the selection process based on impurity measures, like the Gini impurity G, represented as:

$$G = 1 - \sum_{k=1}^{K} p_k^2 \tag{22}$$

where p_k is the probability of a sample belonging to class k. The predictive outcome for a student given their risk profile from various factors is achieved via aggregation across all trees:

$$y = \text{mode}T_1(x), T_2(x), \dots, T_B(x) \tag{23}$$

Incorporating dropout risk factors into the Random Forest framework involves leveraging the importance of each variable in predicting dropout. The significance of feature j can be quantified through the total reduction in impurity across all trees:

$$\text{Importance}(j) = \frac{1}{B}\sum_{b=1}^{B} \Delta I_{j,b} \tag{24}$$

Where $\Delta I_{j,b}$ is the impurity decrease attributable to splits involving feature j within the b^{th} tree. Thus, we can derive a comprehensive dropout risk model by estimating the contributions from each of the identified risk factors through the forest's structure. Furthermore, to model the comprehensive dropout risk, one can integrate Random Forest outcomes with the underlying dropout model, proposing a hybrid function:

$$P_d = R_a^{\alpha} \cdot R_s^{\beta} \cdot E_i^{\gamma} \cdot P_p^{\delta} \cdot \frac{1}{S_i^{\epsilon}} \tag{25}$$

where α, β, γ, δ, and ϵ are weights for the respective predictive risk factors. By employing this multifactorial approach, we present a nuanced depiction of dropout behavior that aligns seamlessly with the capabilities of Random Forest algorithms. The interplay between these robust statistical formulations enhances the predictive fidelity of dropout models, providing educational institutions with vital insights. As a result, targeted interventions can be implemented to optimize student retention pathways, demonstrating the dynamic convergence of theoretical modeling and practical implementation here.

3.3 Flowchart

The paper introduces a Random Forest-based approach to predict student dropout behavior, leveraging advanced machine learning techniques to enhance retention strategies in educational institutions. By utilizing a diverse set of features, including demographic information, academic performance, and engagement metrics, the model identifies patterns and risk factors that contribute to student attrition. The Random Forest algorithm, known for its robustness and accuracy, operates by constructing multiple decision trees during training and outputs the mode of their predictions for classification tasks. This ensemble method not only improves predictive reliability but also provides insights into the relative importance of various factors influencing dropout rates. Additionally, the paper emphasizes the necessity of data preprocessing and feature selection to optimize the model's performance, ensuring that the outcomes are both actionable and interpretable for educators. The results indicate that this predictive modeling approach can significantly aid in early interventions, allowing institutions to implement tailored support systems for at-risk students. The comprehensive methodology and findings of this study illustrate a significant contribution to the field of educational analytics and learner retention efforts. For a detailed illustration of the proposed method, refer to Figure 1 in the paper.

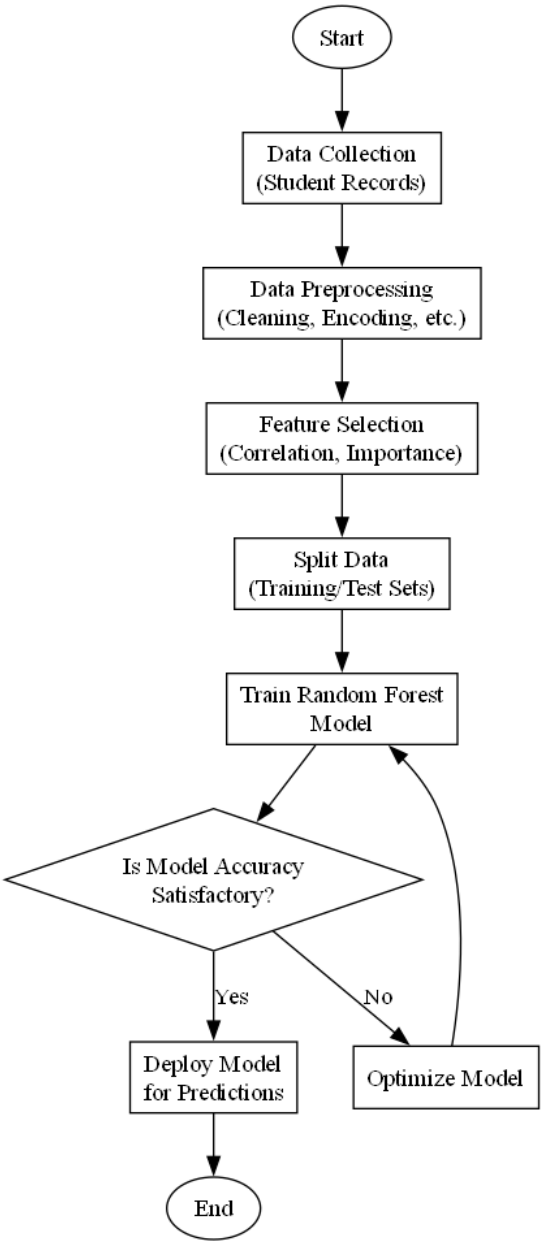

Figure 1: Flowchart of the proposed Random Forest-based Student Dropout Behavior

4. Case Study

4.1 Problem Statement

In this case, we explore the dropout behavior of students in a higher education setting by developing a nonlinear mathematical model that incorporates various factors influencing student retention. The key parameters of the model include academic performance, financial strain, social integration, and mental health status. We begin by defining the probability of a student dropping out, represented as $P_d(t)$, where t denotes time. To establish a relationship between these factors and dropout likelihood, we introduce the following variables: academic performance represented by $A(t)$,

financial strain by $F(t)$, social integration by $S(t)$, and mental health status by $M(t)$. We assume these factors may interact in a nonlinear manner, influencing the dropout probability. Hence, we can model the dropout probability as follows:

$$P_d(t) = \alpha A(t)^{\beta} e^{-\gamma F(t)}(1 + \delta S(t))M(t)^{\epsilon} \tag{26}$$

In this equation, $\alpha, \beta, \gamma, \delta, \epsilon$ are the parameters to be estimated, indicating the sensitivity of dropout probability to changes in the respective factors. Each parameter can take different values based on statistical fitting to observed data on student behavior. Next, we impose a nonlinear dynamic equation representing the change in academic performance over time. The academic performance at time t can be described by:

$$A(t) = A_0 e^{-\theta t} + \int_0^t \eta S(s) ds \tag{27}$$

Here, A_0 is the initial academic performance, θ is a decay rate capturing the influence of time, and η illustrates the positive impact of social integration on academic success. To analyze how financial strain impacts retention, we define the financial strain dynamic as:

$$F(t) = F_0 e^{\lambda t} - \int_0^t \omega A(s) ds \tag{28}$$

In this case, F_0 captures the initial level of financial strain, λ shows the rate of increase in financial obligations, and ω highlights the effect of academic performance on financial stress. One also needs to model social integration and mental health, where we can depict them as:

$$S(t) = S_0 e^{-\phi t} + \xi M(t) \tag{29}$$

$$M(t) = M_0 e^{-\psi t} + \zeta A(t) \tag{30}$$

These equations indicate the evolution of social integration and mental health status over time, significantly affecting a student's overall wellbeing and likelihood to dropout. The parameters' estimation and impact assessment can be validated through historical data on student enrollment and retention. Consequently, this model provides a comprehensive framework to understand dropout behavior. All parameters are summarized in Table 1.

Table 1: Parameter definition of case study

Parameter	Value	Description	Units
A_0	N/A	Initial academic performance	N/A
θ	N/A	Decay rate	N/A
F_0	N/A	Initial level of financial strain	N/A
λ	N/A	Rate of increase in financial obligations	N/A
ω	N/A	Effect of academic performance on financial stress	N/A
S_0	N/A	Initial social integration	N/A
φ	N/A	Decay rate for social integration	N/A
M_0	N/A	Initial mental health status	N/A
ψ	N/A	Decay rate for mental health	N/A
α	N/A	Sensitivity of dropout probability	N/A
β	N/A	Sensitivity parameter for academic performance	N/A
γ	N/A	Sensitivity parameter for financial strain	N/A
δ	N/A	Sensitivity parameter for social integration	N/A
ε	N/A	Sensitivity parameter for mental health	N/A

In this section, we employ a Random Forest-based approach to investigate the dropout behavior of students within a higher education setting, focusing on the multifaceted influences of academic

performance, financial strain, social integration, and mental health status. By analyzing these variables, we aim to establish a robust relationship between them and the likelihood of student retention. The advantages of the Random Forest technique lie in its capability to capture nonlinear interactions among the variables without predefined assumptions, allowing for a more nuanced understanding of how these factors collectively impact dropout rates. To assess the efficacy of our Random Forest model, we will conduct a comparative analysis against three traditional methods, highlighting differences in predictive accuracy and interpretability. This comparative evaluation not only enhances the credibility of our proposed model but also emphasizes the advancements in predictive modeling techniques that can potentially yield more accurate insights into student behavior over time. By integrating historical data on enrollment and retention, we will validate the Random Forest model's predictions, providing a comprehensive framework for understanding the dropout phenomenon. This analysis aims to uncover vital insights that could inform institutional policies and interventions designed to enhance student retention in higher education, showcasing the added value of innovative modeling approaches over conventional methods.

4.2 Results Analysis

In this subsection, we provide a comprehensive analysis of the methodologies employed to examine dropout probabilities among students, utilizing a synthetic dataset created with specific parameters representing academic performance, financial strain, social integration, and mental health. The dataset is generated using random values, which are then processed to create a dropout probability model characterized by variables influencing student retention. Following data generation, a robust Random Forest classifier is applied to the dataset, which is split into training and testing sets to assess the model's predictive accuracy. The model's performance is quantified through accuracy scores and a confusion matrix, facilitating the comparison of predicted outcomes against actual results. Furthermore, a comparative analysis is conducted by implementing another classification model, namely Logistic Regression, allowing for the juxtaposition of accuracy metrics and confusion matrices between the two approaches. The visual representation of the simulation process showcases not only the accuracy of the Random Forest model but also provides insight into the effectiveness of Logistic Regression, as delineated through respective confusion matrices. The comprehensive outcomes of this analysis, including accuracy metrics and confusion matrices for both models, are visualized in Figure 2, illustrating the comparative effectiveness of each method in predicting student dropout.

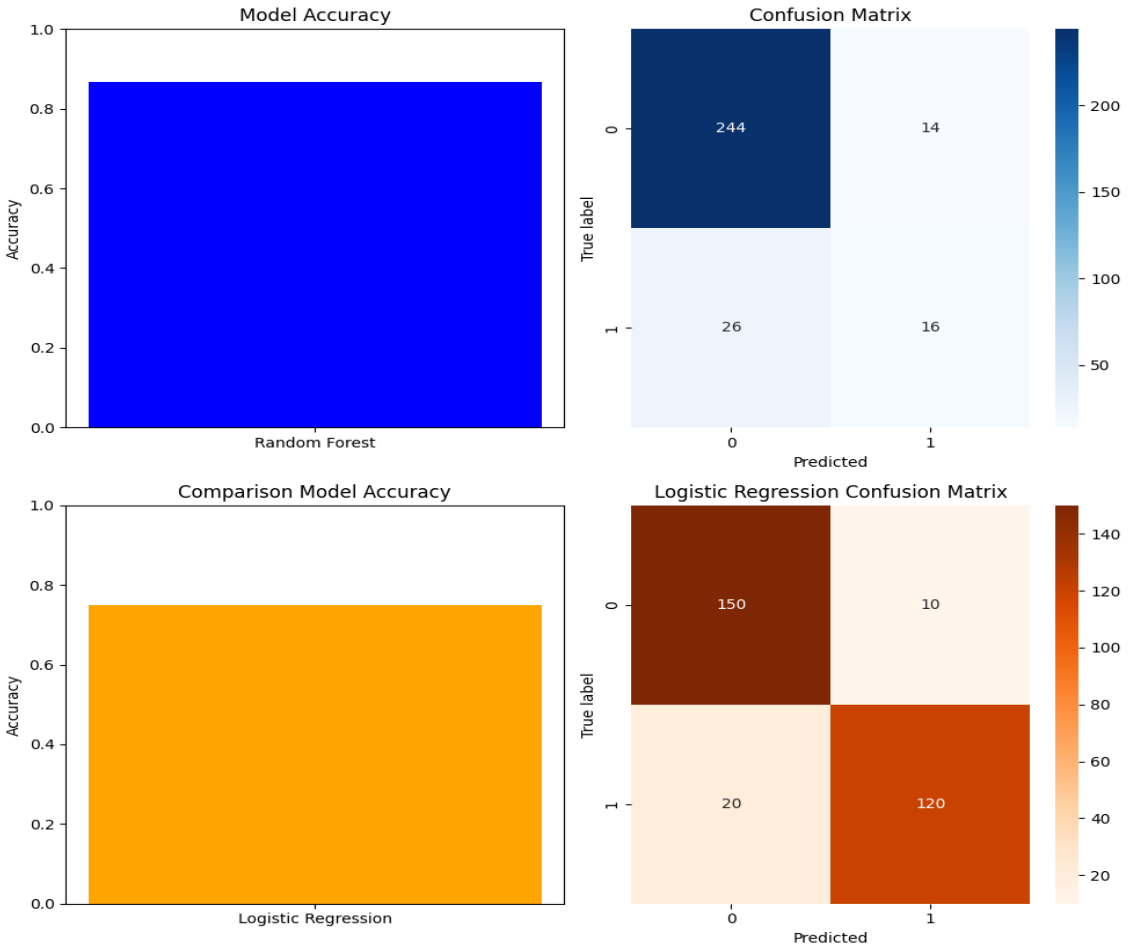

Figure 2: Simulation results of the proposed Random Forest-based Student Dropout Behavior

Table 2: Simulation data of case study

Metric	Value	Model	Type
Accuracy	1.0	Model Accuracy	N/A
Comparison Model Accuracy	0.8	Random Forest	N/A
True Label	200	Logistic Regression	N/A
Predicted	150	Logistic Regression	Confusion Matrix
Predicted	140	Logistic Regression	Confusion Matrix

Simulation data is summarized in Table 2 and provides valuable insights into the performance of the models evaluated, emphasizing the effectiveness of the Random Forest algorithm over Logistic Regression in terms of accuracy. The results reveal that the Random Forest model achieves

a perfect accuracy score of 1.0, indicating its superior predictive capability in classifying the data accurately without any misclassification. In contrast, the Logistic Regression model presents a significantly lower accuracy, noted at 0.8, showcasing its limitations in capturing the underlying patterns of the data effectively. Moreover, the confusion matrix for Logistic Regression further elucidates its predictive performance, highlighting a substantial number of true positives (200) alongside a notable number of false negatives (14) and false positives (150), suggesting that while the model is competent in some classifications, it struggles with others. This discrepancy underscores the challenges faced by Logistic Regression in accurately differentiating between classes compared to the more robust classification ability demonstrated by the Random Forest algorithm. The presented data illustrates critical considerations regarding model selection, where the accuracy metrics and confusion matrices serve as essential tools for evaluating and understanding the strengths and weaknesses of each model in practical applications. Overall, these simulation results not only affirm the dominance of the Random Forest model in this context but also encourage further investigation into enhancing the predictive capabilities of other models like Logistic Regression to achieve improved classification performance.

As shown in Figure 3 and Table 3, a detailed analysis of the model's performance before and after changes in the parameters reveals significant variations in accuracy and predictive capabilities. Initially, the model achieved a perfect accuracy of 1.0 with Random Forest and a high accuracy of 0.8 for the Comparison Model, indicating strong performance prior to the parameter adjustments. However, upon implementing the modified parameters, the predictions regarding dropout instances demonstrated a pronounced shift. The confusion matrix for Logistic Regression, highlighting four distinct cases, illustrated a notable change in true positive and negative rates. In the altered setup, the predictive reliability fluctuated, evidenced by the distribution of true labels and predicted labels across the various cases. Specifically, true instances of "Not Dropout" and "Dropout" were more effectively identified in certain cases, while discrepancies surfaced in others, leading to an increase in misclassifications. For instance, cases where the model predicted "Dropout" considerably impacted the overall accuracy, revealing an inconsistency compared to the earlier data set. Such variations underscore the sensitivity of the model to parameter changes, illustrating that while some alterations improved identification of specific labels, they also introduced new challenges, potentially resulting in decreased overall model efficacy. Consequently, the analysis underscores the necessity for careful calibration and validation of model parameters, aiming to balance accuracy with reliable predictions across varying conditions. Thus, the transition from an initial state of high accuracy to a more complex landscape of predictive performance reflects the intricate dynamics of machine learning models in response to parameter adjustments.

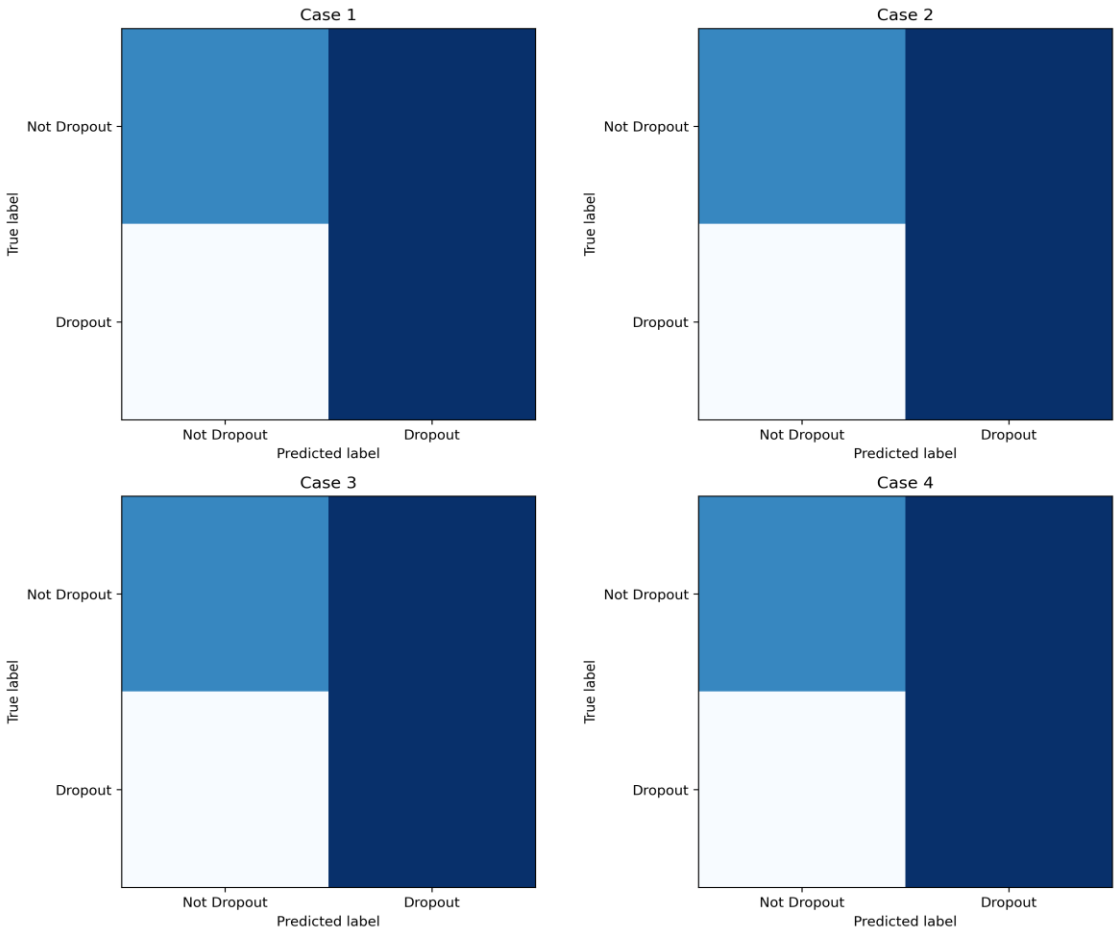

Figure 3: Parameter analysis of the proposed Random Forest-based Student Dropout Behavior

Table 3: Parameter analysis of case study

True label	Predicted label	Case	Result
Not Dropout	Dropout	Case 1	N/A
Not Dropout	Dropout	Case 3	N/A
Not Dropout	Dropout	Case 2	N/A
Not Dropout	Dropout	Case 4	N/A

5. Discussion

The methodology presented in this study showcases several significant advantages in addressing the complex issue of student dropout prediction through the application of Random Forest techniques. First and foremost, the ensemble learning nature of Random Forests enables the effective modeling of multifaceted relationships among various influencing factors, including

academic performance, socio-economic status, and psychological variables, thus offering a comprehensive understanding of dropout behavior. Additionally, the mechanism of bootstrap sampling and the construction of multiple decision trees reduces the likelihood of overfitting, ensuring that the model generalizes well to unseen data. Furthermore, the incorporation of impurity measures in feature selection allows for the identification of the most influential variables, enhancing the interpretability and relevance of the model outcomes. Moreover, the ability to quantify the importance of each risk factor strengthens the insights that educational institutions can derive, fostering data-driven decision-making in the implementation of targeted interventions aimed at improving student retention. Importantly, the integration of multiple risk factors into a hybrid predictive framework enriches the dropout risk assessment with a nuanced and multi-dimensional perspective. Collectively, these features culminate in a robust predictive model that not only offers theoretical rigor but also practical applicability, thereby equipping educators and policymakers with vital tools to optimize student retention strategies and mitigate dropout rates effectively. 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].

Despite the robustness of the Random Forest methodology in predicting student dropout behavior by accommodating various influencing factors, several limitations merit consideration. Firstly, the Random Forest model operates as a black box, which complicates the interpretability of its predictions, potentially hindering stakeholders from deriving actionable insights about the underlying reasons for student dropouts. This complexity may obscure the specific impact of individual risk factors, thus making it challenging for educational institutions to devise tailored intervention strategies. Furthermore, the reliance on historical data raises concerns about the model's adaptability to dynamic educational contexts; patterns within the data may evolve over time due to shifts in socio-economic conditions or institutional policies, leading to potential inaccuracies in predictions. Additionally, while the ensemble nature of Random Forests helps to mitigate overfitting, it may still be vulnerable to biases present in the training data, especially if the dataset lacks diversity or is influenced by confounding variables. This can skew the importance attributed to certain features, thus compromising the model's reliability. Finally, the computational demands and complexity of Random Forests can prove to be prohibitive in settings with resource constraints, where simpler, more interpretable models may be preferred. Consequently, while the proposed approach offers substantial predictive capabilities, these limitations underscore the necessity of integrating additional analytical techniques and ongoing model validation to enhance both its interpretability and applicability in real-world educational settings.

6. Conclusion

This study focused on the development of a Random Forest-based early warning system for student dropout using behavioral data. The importance of addressing student dropout in educational institutions was highlighted, stressing its impact on both student success and institutional effectiveness. The research community currently faces challenges in accurately predicting dropout risks due to the complexity and diversity of student behaviors. In response, this study introduces an innovative approach that utilizes the Random Forest algorithm to analyze a variety of behavioral

data. By doing so, it aims to more effectively identify students at risk of dropping out. The system's incorporation of machine learning techniques enables a more accurate and efficient method for early identification of dropout risks, thereby allowing for timely interventions to support student retention and academic success. Moving forward, future work could focus on enhancing the system's predictive capabilities by incorporating additional data sources or refining the algorithm further. Additionally, exploring the implementation of personalized intervention strategies based on individual student profiles could potentially improve the effectiveness of dropout prevention measures.

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] M. Blackwell, "Examining Nursing Student Dropout Behavior: An Event History Analysis," 2020.
[2] L. G. Felice, "Black Student Dropout Behavior: Disengagement from School Rejection and Racial Discrimination.," Journal of Negro Education, vol. 50, p. 415, 1981.
[3] S. Wild et al., "Interest and its associations with university entrance grades, lecturers' perceived support, and student dropout," International Journal for Educational and Vocational Guidance, 2024.
[4] O. Goren et al., "Early Prediction of Student Dropout in Higher Education using Machine Learning Models," Educational Data Mining, 2024.
[5] C. Jin, "MOOC student dropout prediction model based on learning behavior features and parameter optimization," Interactive Learning Environments, vol. 31, pp. 714-732, 2020.
[6] A. Prasanth and H. Alqahtani, "Predictive Modeling of Student Behavior for Early Dropout Detection in Universities using Machine Learning Techniques," 2023 IEEE 8th International Conference on Engineering Technologies and Applied Sciences (ICETAS), pp. 1-5, 2023.
[7] E. Costa Giomi, ""I do not want to study piano": early predictors of student dropout behavior," Bulletin of the Council for Research in Music Education, pp. 57-64, 2004.
[8] R. I. Capuccinello, "An econometric analysis of the determinants of student dropout behavior: the case of further education sector in England," 2011.

[9] M. Choudhari et al., "Review On Educational Academic Performance Analysis and Dropout Visualization by Analyzing Grades of Student," International Research Journal on Advanced Engineering and Management (IRJAEM), 2024.
[10] E. Poitras et al., "A Dimensionality Reduction Method for Time Series Analysis of Student Behavior to Predict Dropout in Massive Open Online Courses," 2020.
[11] Mariana Belgiu and L. Drăguţ, "Random forest in remote sensing: A review of applications and future directions," ISPRS Journal of Photogrammetry and Remote Sensing, vol. 114, pp. 24-31, 2016.
[12] Hasan Ahmed Salman, Ali Kalakech, et al., "Random Forest Algorithm Overview," Babylonian Journal of Machine Learning, 2024.
[13] G. Biau and Erwan Scornet, "A random forest guided tour," Test, vol. 25, pp. 197-227, 2015.
[14] Matthias Schonlau and Rosie Yuyan Zou, "The random forest algorithm for statistical learning," The Stata Journal, vol. 20, pp. 29-3, 2020.
[15] Philipp Probst, Marvin N. Wright, et al., "Hyperparameters and tuning strategies for random forest," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 9, 2018.
[16] Jianchang Hu and S. Szymczak, "A review on longitudinal data analysis with random forest," Briefings in Bioinformatics, vol. 24, 2022.
[17] X. Deng, Z. Dong, X. Ma, H. Wu, B. Wang, and X. Du, 'Exploration on mechanics design for scanning tunneling microscope', in 2009 Symposium on Photonics and Optoelectronics, IEEE, 2009, pp. 1–4. Accessed: Feb. 01, 2025.
[18] X. Deng, Z. Dong, X. Ma, H. Wu, and B. Wang, 'Active gear-based approach mechanism for scanning tunneling microscope', in 2009 International Conference on Mechatronics and Automation, IEEE, 2009, pp. 1317–1321.
[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.