





Research Article

ElimRidge-HFM: A Mathematical Integration of Feature Selection, Boosting, and Hyperparameter Search for Enhanced Predictive Performance in E-Learning

Naga Satya Koti Mani Kumar Tirumanadham^{1,2*}, Thaiyalnayaki Sekhar¹, Suresh Babu Chandolu³, J. Hymavathi⁴, K Varada Rajkumar⁵, Punugupati Chiranjeevi⁶

¹Department of Computer Science and Engineering, Bharath Institute of Higher Education and Research, Selaiyur, Tamil Nadu, India

²Department of Computer Science and Engineering, Sir C R Reddy College of Engineering, Eluru, India

³Department of Computer Science and Engineering, Dhanekula Institute of Engineering and Technology, Gangur, Vijayawada, Andhra Pradesh, India

⁴Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Guntur, Andhra Pradesh, India

⁵Department of Computer Science and Engineering (AIML), MLR Institute of Technology, Hyderabad, Telangana, India

⁶Department of Mathematics and Humanities, R. V. R. & J. C. College of Engineering, Guntur, Andhra Pradesh, India
E-mail: manikumar1248@gmail.com

Received: 22 December 2024; **Revised:** 20 May 2025; **Accepted:** 27 May 2025

Abstract: The rising demand for e-learning platforms requires reliable predictive models to improve learning results. This study establishes ElimRidge-Hybrid Fusion Model (HFM) as a new predictive framework for e-learning data sets that solves the problems of feature selection, model overfitting, and potential dataset complexity. Thus, the objective of the study is as follows: Enhance the predictive precision and interpretability of the model by applying feature selection techniques using Ridge Regularization (L2) in conjugation with Recursive Feature Elimination (RFE) and using a Hybrid Fusion Model (HFM), which is the combination of AdaBoost and CatBoost. For the hyperparameter tuning, Gaussian Process-Enhanced Random Search (GP-RS) is used to explore the hyperparameters of the space efficiently with the least number of random trials. For data clearing this framework uses Synthetic Minority Over-sampling Technique (SMOTE) for class balancing, Interquartile Range (IQR) for outlier removal and, Z-score normalization for normalization. Experimental results demonstrate that ElimRidge-HFM significantly improves predictive performance, achieving 97% accuracy, 97% precision, and 97% F1-score. Statistical validation ($p < 0.05$) confirms its superiority over traditional models, effectively handling imbalanced and noisy e-learning datasets. These findings highlight its potential for enhancing personalized learning strategies and optimizing student engagement. These results further strengthen the claims that ElimRidge-HFM has great ability to handle imbalanced and noisy data sets due to its scalability. This research contributes to the development of e-learning techniques by providing opportunities to introduce decisions and/or strategies when the students' performance is low or during cases of inequality in students' learning opportunities.

Keywords: e-learning, ridge regularization, recursive feature elimination, hybrid fusion model, Gaussian process-enhanced random search

MSC: 68T05, 62H30

1. Introduction

E-learning, is a new era in education since learning is facilitated through the use of technology enhanced learning environments. With technology gradually transferring to people's lives, e-learning is a convenient and flexible approach to learning [1]. The current generation of learners can learn from computers, tablets and even from smart phones at any time of their convenience. This carries a lot of benefits especially when students have to combine their studies with work or other endeavors that come their way. This is a versatile field that is open to a wide range of formats as a form of learning starting from MOOC to corporate training. Perhaps, one of the most important benefits of e-learning is that it will afford everyone the chance to study [2]. The main advantages of e-learning include the opportunity for students without regard to their geographical location, socio-economic status to gain access to quality education. It also favours Individualized education whereby technologies may enable a teacher or system to provide the learner with an education program that will meet the learner's learning schedule, preferred method of learning, and his/her level of learning. Also, e-learning has proven to be very effective for the creation of class community via classroom communications, discussions and group work, as well as assignments even in physically wrong. Thus, e-learning also has several issues that can be encountered during the process of it [3].

The first of them is on equal distribution of the required technology as many learners lack stable internet connection or digital devices. It also important to enhance the players skills on online learning since a section of the learners might not understand technicalities involved in the use of the online learning platforms. However, it is really tough to keep students interested in the virtual space, which has opportunities for distractions, and does not have the enthusiasm which is achieved through attending the classroom [4]. To make the e-learning efficient, we need to have interactivity, use of multimedia, and good interface design in our instruction [5]. Therefore, through such approaches, e-learning is competent to provide substantial and appealing learning experiences which can facilitate further development in educational experiences. The current study focuses on data preprocessing, feature selection, and model construction with regard to e-learning data analytics using predictive modelling techniques. The specific objectives of this research are to improve the accuracy and objectivity of the predictive models and also to address issues such as a class imbalance ratio, how to handle outliers, and how to standardize features. Considering that class balancing methods include SMOTE, research ensures that the model treats all the students equally without discrimination. After removing all the irrelevant features for acquiring final dataset only important and useful feature are obtained subsequently, Analysis of Variance (ANOVA) F-statistics, Recursive feature elimination and Lasso methods are used for providing final dataset. At model building phase, combination of the multiple classifiers including Logistic Regression, Decision Tree, and K-Nearest Neighbour (KNN) are included. Thus, the paper presents a detailed plan of the approach that proves useful for constructing stable and accurate prediction models in e-learning with a focus on the increase of students' success rates individual approach.

Despite advancements in machine learning for e-learning, existing predictive models face challenges in handling feature selection inefficiencies, imbalanced datasets, and computationally expensive hyperparameter tuning. Many traditional approaches suffer from overfitting, inadequate categorical data handling, and exhaustive hyperparameter search techniques that limit scalability. This study introduces ElimRidge-HFM, a novel hybrid framework that integrates Recursive Feature Elimination (RFE) with Ridge Regularization (L2) for optimal feature selection, Hybrid Fusion Model (HFM) combining AdaBoost and CatBoost for robust classification, and Gaussian Process-Enhanced Random Search (GP-RS) for efficient hyperparameter tuning. By addressing these key challenges, ElimRidge-HFM enhances generalization, reduces computational cost, and improves classification accuracy, making it a scalable and practical solution for predictive analytics in e-learning environments.

1.1 Research gap

A key focus of EDM research has been this type of scenario, in which available datasets are sparse and contain only a limited number of labelled data and ordinal classes. While current models can optimize fully labelled datasets

consisting of nominal classification, such models are very rigid in educational environments when labelled instances are scarce and the class labels are ordered naturally. The application of semi-supervised ordinal classification seems a promising approach; however, immense research is required to validate proposed methods in the various fields of education including the implementation of deep learning in combination with the semi-supervised techniques. Furthermore, the much more sophisticated preprocessing methods and adjustment of hyperparameters that were often useful in other fields have also been very sparsely used on the educational data with ordinal dimensions. This gap had already identified the possibility to integrate semi-supervised learning with deep learning and special preprocessing for ordinal data in education. They could enhance the efficiency of accuracy of classification and models as well as contribute to the creation of individualized educational knowledge big data. Filling out these gaps would allow for better aimed and more evidence-based decision-making and improve how predictive models are employed across a wide swath of educational contexts to advance individualized learning supports and successes.

1.2 Research questions

1. How can semi-supervised ordinal classification methods be effectively used on different educational datasets with limited labelled data?
2. Can combining deep learning with semi-supervised learning improve prediction accuracy for educational datasets with ordered class labels?
3. What preprocessing and hyperparameter tuning methods work best to optimize models for educational data with ordered classes?
4. How can hybrid models combining semi-supervised and deep learning approaches help create personalized insights in educational data?

The proposed work primarily contributes to the preceding steps:

1. Proposed the ElimRidge-HFM model for accurate prediction with interpretable models for e-learning.
2. Applied SMOTE preprocessing techniques, IQR and Z-score to ensure better quality of the data.
3. Interpreted numerical and categorical data using Hybrid Fusion Model (HFM).
4. Used Gaussian Process augmented Random Search (GP-RS) for optimizing hyperparameters.
5. Obtained accuracy of 97% of the target values, which is higher than with conventional models.
6. Given a framework for working with the imbalanced and noisy data on a large scale.
7. Supported the interventions for target students and enhanced the results of learning.

The paper is structured into six sections. Section 2, Literature Review, discusses existing research and identifies gaps. Section 3, Proposed Methodology, explains the techniques used, including feature selection and hybrid modelling. Sections 4, 5, and 6 cover Experimental Results, Discussion, and Conclusion and Future Scope, presenting findings, analysing results, and suggesting future research directions.

2. Literature review

In 2019, Al Fanah et al. [6], focuses on comprehension of e-learners' behaviour by applying models of data mining, association rules, and classification into learning behaviour analysis of the online students. It used data collected from Higher Education Institutions by applying Random Forests, Logistic Regression, and Bayesian Networks for grouping into three categories: high, medium, and low. Results showed that Bayesian Networks had the highest score at 80%, followed by the Random Forest at 63% and the Logistic Regression model at 58%. It focuses on data analytics for improving the e-learning environment and guides decision-making about supporting educators in targeting their interventions.

In 2020, Enoughwure et al. [7], addressed the challenge of predicting students' performance in engineering drawing courses, which form a very central part of any engineering learning process. The paper adopted SMOTE-augmented Machine Learning Algorithm to apply both Logistic Regression and Decision Trees towards predicting student outcomes, thereby resulting in predictive accuracies between 67% and 78%, where logistic regression had the highest value. The use

of SMOTE in balancing the dataset proved very crucial in improving model performance, thus showing that it is possible to apply machine learning techniques to predict students' performance and help strategize interventions.

In 2021, Unal et al. [8], introduced the Semi-Supervised Ordinal Classification (SSOC) approach for addressing the challenges of educational data classification when labelled data are limited and ordinal class labels are used. Combining semi-supervised learning with ordinal classification achieved high accuracy even when only a small percentage of the data was labelled. It was specifically useful in education-related applications in which ordered class labels, for example, students' performance levels, mattered. The approach brought up the worth of SSOC in terms of optimization of classification performance in support of data-driven decision-making within educational settings.

In 2022, Liu et al. [9], introduced the Predictive Analytics and Intelligent Modeling (PAIM) model as a means of improving prediction models using evolutionary Spiking Neural Networks. The model was more potent than traditional algorithms as it integrated data pre-processing and feature extraction techniques. Such an approach allows the earlier identification of at-risk students and gives institutions an input to offer targeted academic support. The study focused on how personalized teaching methods and early interventions can improve academic performance and the potential of SNN in educational data mining.

In 2023, Gupta et al. [10], applied the hyperparameter-tuned machine learning models in predicting diabetes using the PIMA Indian Diabetes dataset. The comparative study of the classifiers performed is K-Nearest Neighbors, Decision Trees, and Random Forests with the highest accuracy achieved as 88.61%. The study clearly highlighted that hyperparameter tuning and preprocessing of data is necessary to develop reliable predictive models that underscore the significance of optimization of models in disease prediction in healthcare.

In 2024, Farhood et al. [11], gave a deep understanding of artificial intelligence techniques, which predict learning outcomes from machine learning models like Random Forest and XGBoost, and deep learning models like Gradient-Boosted Neural Networks (GBNN). In the course of this study, it was observed that GBNN performed best among the deep learning models while Random Forest and XGBoost dominated the machine learning model. This study showed how feature selection and hyperparameter tuning significantly improve the quality of prediction in prediction accuracy due to the development of a personalized learning plan and greater educational success with AI solutions.

Various machine learning techniques have been noted to be applied for the prediction of performance in an e-learning setting in a review of existing literature. The studies range from data mining and classification models [6] to the application of SMOTE for balancing the dataset [7], semi-supervised ordinal classification [8], and Spiking Neural Networks [9]. Research is also focused on hyperparameter tuning and feature selection to improve predictive accuracy [10, 11]. These works form the basis of increased importance of advanced models to improve educational outcomes and decision-making. While previous studies have explored various ensemble learning, feature selection, and hyperparameter tuning techniques, they often exhibit limited generalization, high computational costs, or inadequate handling of imbalanced data. For instance, Al Fanah et al. [6] applied classification models but lacked sophisticated feature selection, leading to overfitting and reduced interpretability. Gupta et al. [10] demonstrated the benefits of hyperparameter tuning but relied on computationally expensive grid search techniques. Compared to these approaches, ElimRidge-HFM uniquely integrates feature selection, hybrid boosting models, and probabilistic optimization, leading to superior predictive performance. The integration of Recursive Feature Elimination (RFE) with Ridge Regularization (L2) enhances feature selection, while HFM's fusion of AdaBoost and CatBoost ensures strong generalization in both categorical and numerical data. These advantages make ElimRidge-HFM an efficient and scalable alternative to conventional methodologies.

3. Proposed methodology

The proposed methodology consists of a few advanced techniques to optimize the performance of predictive modelling. This method is based on Data Cleaning, so that the dataset is assured to be clean and without inconsistency and missing values; so, this is a good foundation for developing a model. SMOTE [12] handles class imbalance because it generates synthetic samples for underrepresented classes to ensure that the model is unbiased. IQR [13] is used to detect outliers that may influence the result of the model. The next step is the application of the ElimRidge technique,

which combines RFE [14] and Ridge regularization [15] for feature selection. It reduces overfitting while keeping the majority features to enhance the efficiency of the model. Lastly, the application of the technique called ElimRidge-HFM, which is the efficient integration of AdaBoost [16] and CatBoost [17], focuses on boosting up the adaptive boosting in order to increase accuracy by focusing on misclassified data and meanwhile, it focuses on excelling in categorical features; thus, introducing HFM that boosts both classification accuracy and generalization capacity. Hyperparameter Tuning improves the model performance by using GP-RS in which GP-RS utilizes probabilistic modelling with the help of Gaussian Process [18] incorporating randomness into the search. The past hyper-parameter space is searched intelligently using the previously estimated hyper-parameter settings focusing on more promising regions to improve its effectiveness as much as the search results and more efficiency than traditional grid searching. The combination of ElimRidge-HFM and GP-RS leads to a strong, high-performance model that can work well with both numerical and categorical data, enhance the accuracy of prediction, and generalize with lower computational costs (see in Figure 1).

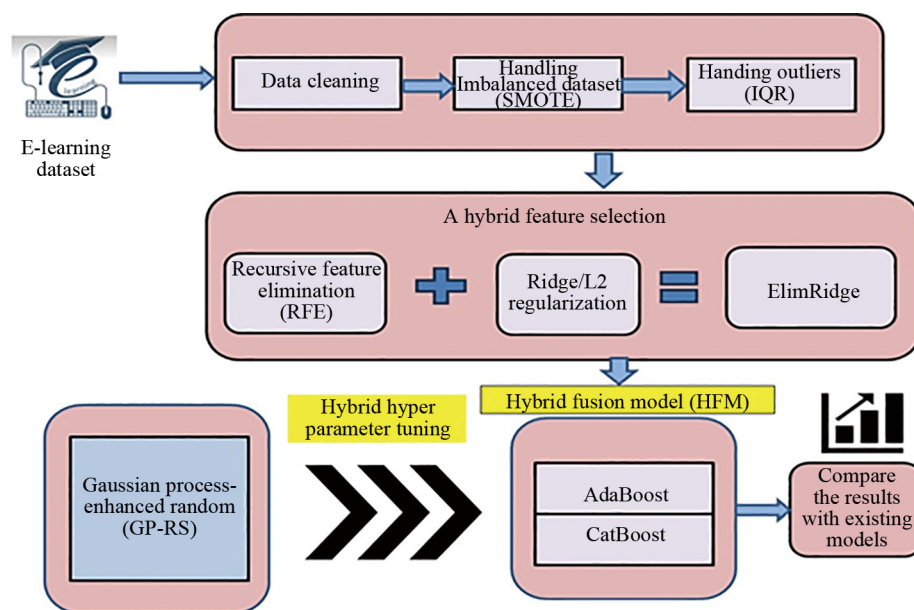
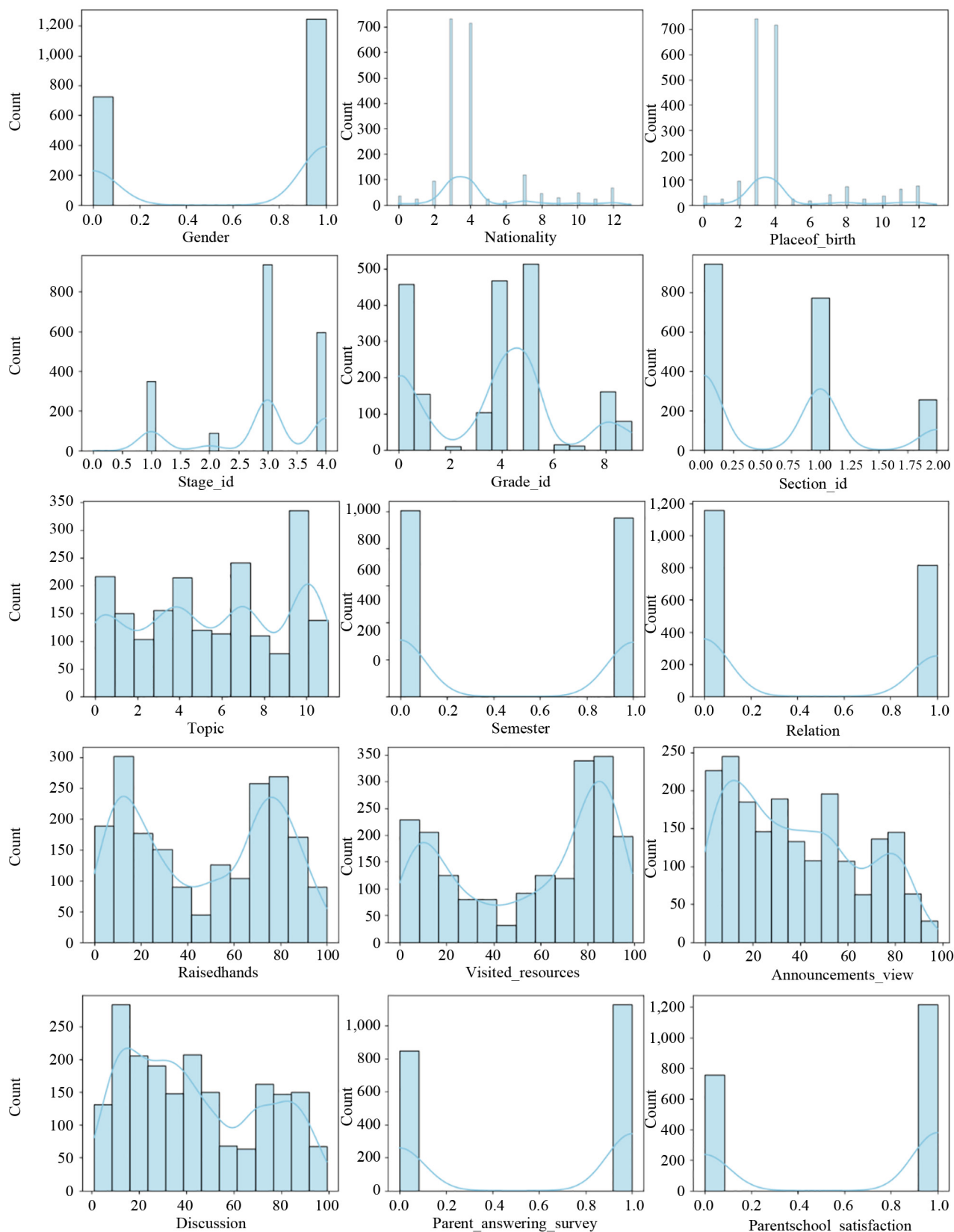


Figure 1. Proposed method work flow

3.1 Data collection

This Kaggle e-learning dataset [19] captures factors of demographic, curricular, engagement, and family-related influence on student performance. Some of the features include gender, nationality, course details, engagement metrics such as raised hands and resources visited, and parental involvement. Data is used to analyse causal relationships affecting learning outcomes in an e-learning environment. Data columns are Visualized in Figure 2. The dataset consists of multiple attributes categorized into demographic factors, academic performance indicators, and engagement metrics. Demographic variables such as gender, nationality, and academic level were extracted from institutional records. Engagement metrics-including raised hands in class, viewed announcements, participation in discussions, and visited educational resources-were derived from the platform's activity logs. These features were selected based on prior research linking student engagement to academic performance, ensuring that the dataset effectively captures learning behaviours and performance indicators for predictive modelling.



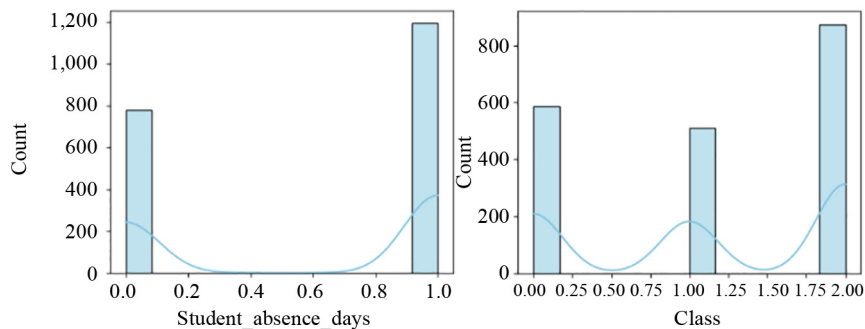


Figure 2. Visualizing the columns in the dataset

3.2 Data cleaning

Since e-learning has no missing value. All 17 columns included 480 complete entries prior to and after processing ensuring the integrity and coherence needed for analysis. Missing values result in bias and affect valid outcomes of research, but e-learning has no missing, which simplifies data preprocess and model building, ultimately enhancing the reliability of outcome results shown in Figure 3.

Missing values after handling:	
gender	0
nationality	0
placeof_birth	0
stage_id	0
grade_id	0
section_id	0
topic	0
semester	0
relation	0
raisedhands	0
visited_resources	0
announcements_view	0
discussion	0
parent_answering_survey	0
parentschool_satisfaction	0
student_absence_days	0
class	0
dtype: int64	

Figure 3. Statistical information about the dataset after handling missing values

3.3 Handling imbalanced dataset using SMOTE

In essence, preprocessing the data on class balancing is one of the significant steps to achieve fairness of the machine learning model. The raw dataset e-learning possessed imbalanced classes in their initial forms: 211 for 'M', 142 for 'H', and 127 for 'L' as indicated by Table 1, Figure 4. Balancing was achieved with the Synthetic Minority Over-sampling Technique, where the dataset resulted in an equal number of 143 instances per class as seen in Figure 5. SMOTE [20] improves the minority classes by generating synthetic samples so that class balance and fairness are enhanced within the model. The above approach also improves model reliability, accuracy, and generalization, which leads towards a better prediction in the analysis and application in the future. SMOTE (Synthetic Minority Over-sampling Technique) plays a critical role in handling class imbalances by generating synthetic samples for underrepresented classes rather than duplicating existing data. This approach enhances the model's ability to learn patterns from minority classes, preventing bias toward majority classes and improving classification performance. By balancing the dataset, SMOTE ensures that the model does not favor dominant classes, leading to better generalization and fairness. Its impact is particularly significant in

e-learning data, where student performance categories may be highly imbalanced. Empirical results confirm that applying SMOTE improves the predictive performance of the model on underrepresented classes, increasing recall and F1-score while maintaining overall accuracy. This ensures that students from all performance groups receive equal consideration in predictive analytics, enhancing the model’s applicability in real-world educational settings.

Table 1. Class distribution before handling imbalance

Class	Before count	After count
M	211	143
H	142	143
L	127	143

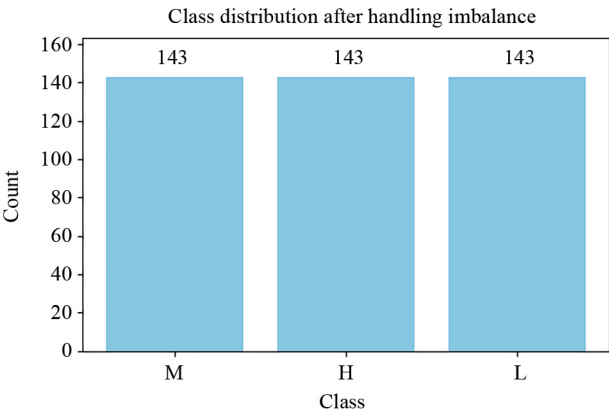


Figure 4. A bar graph of class distribution after SMOTE

3.4 Handling outliers using IQR

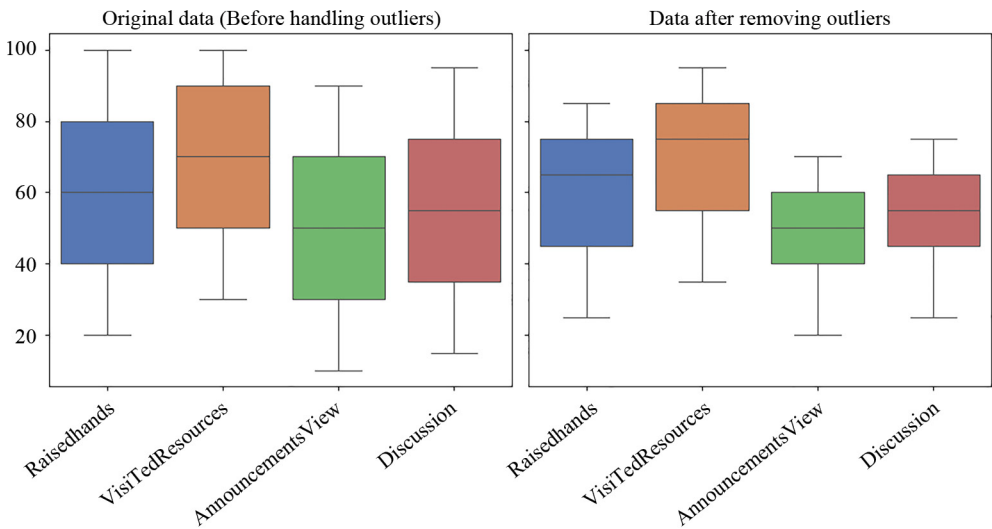


Figure 5. After handling outlier using IQR

Outlier elimination improves statistical analysis and machine learning efficiency. The Interquartile Range (IQR) [21] method was used to identify outliers as values outside the range of $Q1 - 1.5IQR$ and $Q3 + 1.5IQR$. Removing outliers, as shown in Figure 5, ensures more accurate results and reliable models. Boxplots were used to compare distributions before and after removal, enhancing analysis and model-building.

3.5 Standardization using Z-score normalization

Z-score normalization [22] standardizes features by transforming them to have a mean of zero and a standard deviation of one, ensuring consistent scaling across all features. In our dataset, post-normalization features such as “raisedhands” and “Discussion” demonstrated improved comparability, enhancing model accuracy, convergence, and overall suitability for advanced machine learning applications Figure 6.

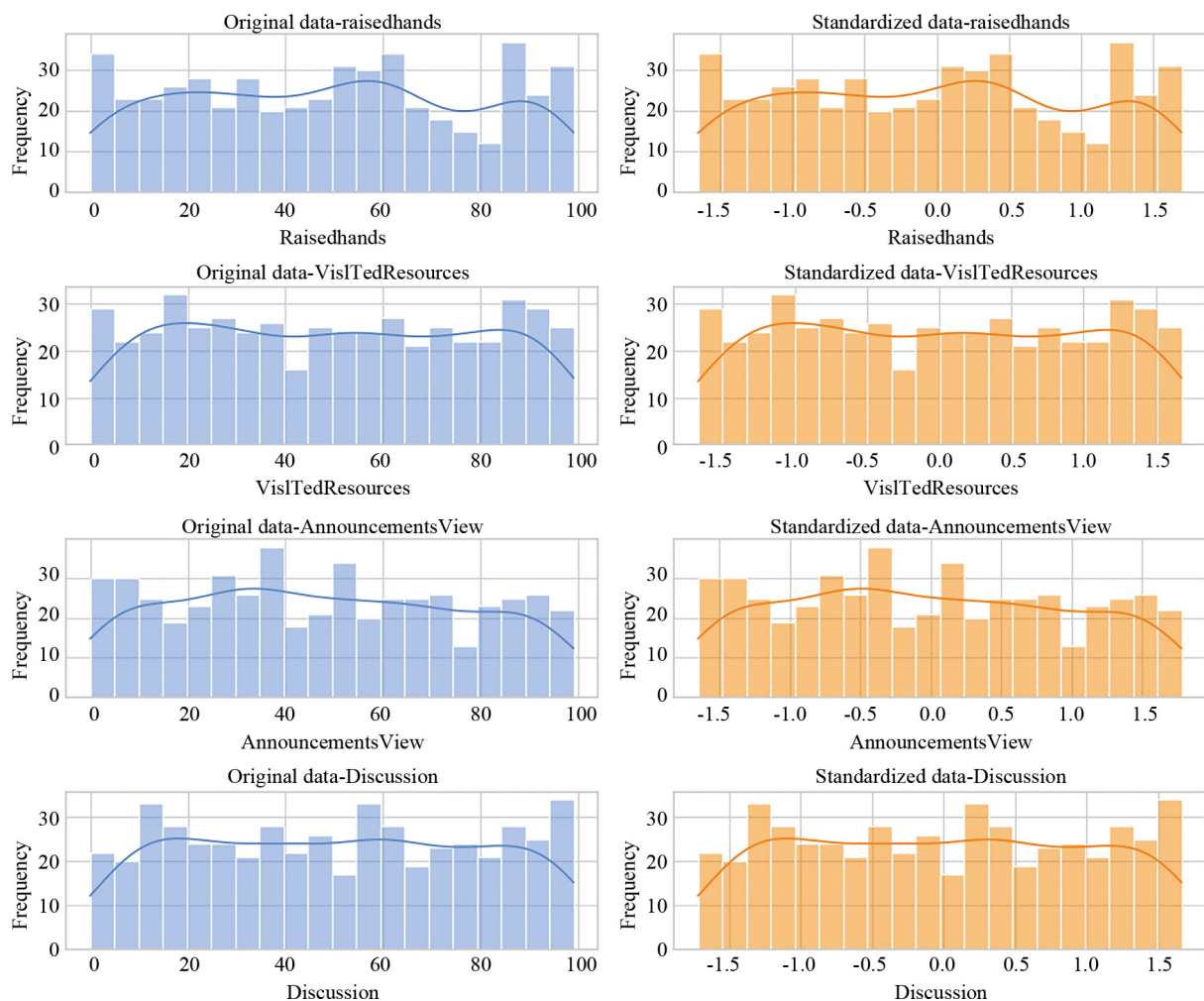


Figure 6. A histogram of the columns after standardization using Z-score normalization

3.6 Feature selection using ElimRidge

After Data Preprocessing feature selection will be conducted by using, “Elimination” and “Ridge” in combination highlighting the process of feature elimination [23] and Ridge regularization [24] processes. The hybrid of the feature selection algorithm brings power into L2 regularization called Ridge and Recursive Feature Elimination, RFE, so the

model improvement and efficiency can be made even better. It enhances the machine learning model's accuracy through its feature selection and elimination process by selecting the least important one. This proposed approach combines Ridge with RFE and is called ElimRidge. The ElimRidge technique integrates Recursive Feature Elimination (RFE) with Ridge Regularization (L2) by leveraging RFE's iterative elimination of the least important features while utilizing Ridge Regularization to penalize large coefficients, ensuring feature selection stability. This dual approach minimizes overfitting while retaining key predictive features, enhancing model generalization and interpretability. By systematically removing less relevant features, ElimRidge optimizes model complexity, ensuring a balance between feature reduction and predictive power, leading to improved classification performance in e-learning datasets. It ensures that the overfitting gets eliminated while making it possible for the model to generalize appropriately.

The Ridge regression model minimizes the following loss shown in Equation (1):

$$L(c) = \sum_{i=1}^n \left(b_i - \sum_{j=1}^p c_j a_{ij} \right)^2 + \lambda \sum_{j=1}^p c_j^2 \quad (1)$$

where:

- b_i , target value for the i^{th} observation, a_{ij} value of feature j for observation i , c_j coefficient for feature j , λ Regularization parameter controlling penalty strength.

Equation (2) be the expand term of squared term:

$$L(c) = \sum_{i=1}^n \left[b_i^2 - 2b_i \sum_{j=1}^p c_j a_{ij} + \left(\sum_{j=1}^p c_j a_{ij} \right)^2 \right] + \lambda \sum_{j=1}^p c_j^2. \quad (2)$$

Take the partial derivative of $L(c)$ concerning c_k shown in Equation (3):

$$\frac{\partial L(c)}{\partial c_k} = \sum_{i=1}^n \left(-2b_i a_{ik} + 2 \sum_{j=1}^p c_j a_{ij} a_{ik} \right) + 2\lambda c_k. \quad (3)$$

Minimize the loss by setting the derivative to zero, as described in Equation (4):

$$\sum_{i=1}^n \left(b_i a_{ik} - \sum_{j=1}^p c_j a_{ij} a_{ik} \right) = \lambda c_k. \quad (4)$$

The importance of each feature is determined by its absolute coefficient value, I_j Feature importance score for feature j . Features is ranked based on their importance shown in Equation (5) and Equation (6)

$$I_j = |c_j| \quad (5)$$

$$R_j = \text{Rank}(|c_j|) \text{ for } j = 1, 2, 3, \dots, p. \quad (6)$$

Identify and remove the feature with the least importance, remove this feature from the feature set X , shown in Equation (7):

$$F_{remove} = \arg \min_j |c_j|. \quad (7)$$

Retrain the Ridge regression model after removing the least important feature shown in Equation (8):

$$L'(c) = \sum_{i=1}^n \left(b_i - \sum_{j \neq F_{remove}} c_j a_{ij} \right)^2 + \lambda \sum_{j \neq F_{remove}} c_j^2. \quad (8)$$

Continue the elimination process until the desired number of features $p_{desired}$ is reached, shown in Equation (9):

$$Stop \text{ if } p_{current} = p_{desired}. \quad (9)$$

The final prediction function is, shown in Equation (10):

$$\hat{b}_i = \sum_{j \in selected} c_j a_{ij} + b \quad (10)$$

\hat{b}_i , the predicted value for observation i , b intercept term.

Regularization coefficients are adjusted using the Ridge closed-form solution, shown in Equation (11):

$$c = (A^T A + \lambda I)^{-1} A^T. \quad (11)$$

Pseudocode for ElimRidge Feature Selection

Procedure:

1. Initialize Model:

- Train a Ridge Regression model using all features.

2. While the number of features $p > p_{desired}$:

- Train Ridge Model:

Fit Ridge Regression on A , b :

$$L(c) = \sum_{i=1}^n \left(b_i - \sum_{j=1}^p c_j a_{ij} \right)^2 + \lambda \sum_{j=1}^p c_j^2.$$

- Calculate Feature Importance:

Compute importance scores:

$$I_j = |c_j|.$$

- Rank Features:

Sort features based on the descending order of I_j .

- Remove Least Important Feature:

Identify feature $F_{remove} = \arg \min_j |c_j|$.

Eliminate this feature from A .

- Retrain Model:

Train the model again with the reduced feature set.

3. End While.

4. Output:

- Selected Features $A_{selected}$.

- Final Ridge Regression Model.

3.7 Model building using hybrid fusion model

Hybrid Fusion Model (HFM) is an innovative hybrid model that integrates the benefits of AdaBoost and CatBoost for improving predictive performance, particularly when handling complex data containing categorical and numerical features. Combining AdaBoost's adaptive boosting with CatBoost's ability to handle categorical data more HFM is an innovative hybrid model that combines the benefits of AdaBoost and CatBoost to enhance the accuracy of prediction, especially in more complex data with both categorical and numerical features. By integrating adaptive boosting from AdaBoost with categorical data handling efficiency from CatBoost, HFM enhances accuracy as well as generalization capacity for its models. The Hybrid Fusion Model (HFM) effectively combines the strengths of AdaBoost and CatBoost to improve predictive accuracy in e-learning data classification. AdaBoost enhances weak learners by assigning higher weights to misclassified instances, enabling the model to focus on difficult cases and iteratively improve performance. CatBoost, on the other hand, is specifically optimized for categorical features and leverages ordered boosting to reduce overfitting and improve generalization. By integrating these two techniques, HFM ensures a more balanced and robust classification model that can handle both numerical and categorical variables efficiently. This synergy leads to superior feature learning, reduced bias-variance trade-offs, and higher classification accuracy, making it particularly effective for complex and imbalanced e-learning datasets. It ensures that there is strong classification performance over various kinds of data due to the power of two algorithms combined.

- CatBoost: CatBoost is a gradient boosting algorithm specifically made for categorical features, and in it, one does not need any explicit encoding [25]. This algorithm makes decisions in the form of iteratively building decision trees and tries to correct each other in a new tree. CatBoost leverages ordered boosting and symmetric trees to both boost the performance of an algorithm and diminish overfitting as well as enhance the generalization of an algorithm. Such an algorithm, therefore performs pretty well particularly in many categorical variables data sets and hence forms an integral component of the hybrid method called HFM.

- AdaBoost: AdaBoost is a method of ensemble learning, where a set of weak learners, usually decision trees, are aggregated into an extremely robust model through boosting with iterative refinement [26]. This is the process of emphasizing the wrong classifications by increasing the weights so that further learners could rectify those errors of the previous models. That process is adaptive in the nature of continually refining the model to make it more accurate and robust, especially if the data are noisy or very challenging.

Pseudocode for HFM Feature Selection

Input:

- A : Feature matrix.
- b : Target variable.
- $N_{estimators}$: Number of boosting rounds.
- $\epsilon_{threshold}$: Error threshold for stopping.

Procedure:

1. Initialize:

- Set initial sample weights $c_i = \frac{1}{n}$ for all training samples i

- Initialize an empty list of models $M = []$

2. Boosting Loop:

While number of estimators $t < N_{estimators}$:

a. Train CatBoost Model:

- Train a CatBoost classifier $f_t(A)$ using current weights c_i .

b. Evaluate Model:

- Calculate prediction error:

$$\epsilon_t = \frac{\sum_{i=1}^n c_i \cdot \prod (f_t(A_i) \neq b_i)}{\sum_{i=1}^n c_i}.$$

- **If** $\epsilon_t > 0.5$ or $\epsilon_t < \epsilon_{threshold}$:

- **Break** the loop.

c. Compute Model Weight:

- Calculate model weight α_t :

$$\alpha_t = \log \left(\frac{1 - \epsilon_t}{\epsilon_t} \right).$$

d. Update Sample Weights:

- Update sample weights for misclassified samples:

$$c_i \leftarrow c_i \cdot \exp \left(\alpha_t \cdot \prod (f_t(A_i) \neq b_i) \right).$$

e. Normalize Weights:

- Normalize sample weights:

$$c_i \leftarrow \frac{c_i}{\sum_{i=1}^n c_i}.$$

f. Store Model:

- Append $f_t(A)$ and α_t to the model list M .

3. Final Model Prediction:

- Use weighted majority voting for prediction:

$$F(A) = \text{sign} \left(\sum_{t=1}^{N_{estimator}} \alpha_t \cdot f_t(A) \right).$$

Output:

- Final HFM Model.

- Predictions for test data.

3.8 Hyperparameter tuning using gaussian process-enhanced random search

Gaussian Process-Enhanced Random Search (GP-RS) explored the hyperparameter space very efficiently and significantly improved the performance of machine learning models. GP-RS adapts its search strategy based on the evaluations performed so far, thus focusing the search in promising regions of the hyperparameter space. Combining randomness with probabilistic modelling by a Gaussian Processes makes the process of optimization more efficient with reduced time of search and increases the accuracy of the model. This produces stronger and more reliable models which give near-optimal performance at a computational cost much less than that of the exhaustive grid search, thus supporting the applicability of GP-RS in hyperparameter optimization.

4. Experimental results

This section describes the detailed experimental performance analysis and evaluations to validate our proposed framework.

4.1 Experimental setup

The experiment was performed on a high-performance system comprising of 16 GB RAM, NVIDIA GeForce RTX 3070 GPU, and an Intel Core i7-12700F CPU. While processing and evaluating data for models, Python 3.7 libraries like pandas, numpy, matplotlib, seaborn, and sci-kit-learn were employed. The Hybrid Fusion Model was AdaBoost with 50 estimators and had a learning rate of 0.05 along with max depth 3. It also used CatBoost with 500 iterations, learning rate 0.1, and depth 6. SMOTE, IQR, and ElimRidge have been performed on the Kaggle dataset. It used hyperopt for hyperparameter tuning using Gaussian Process-Enhanced Random Search. Model performance was tested using 5-fold cross-validation metrics: accuracy, precision, recall, F1-score, and AUC-ROC, to test the robustness and generalization of the approach.

4.2 Feature selection using ElimRidge

Based on the known limitations of the previous work, our research introduces a hybrid approach called ElimRidge; this combines Ridge regularization (L2) with Recursive Feature Elimination (RFE) to optimize machine learning models for better performance efficiency. This method directly addresses problems such as overfitting and feature relevance that often trouble many existing approaches. Ridge regularization reduces overfitting by penalizing less relevant features, forcing their coefficients down and thus improving the ability of the model to generalize. On the other side, RFE systematically removes the least important features, keeping the most critical predictors in the model. Combining these approaches ensures that the model be both accurate and interpretable. Features like gender, grade_id, raisedhands, and student_absence_days were identified as most relevant in predicting the student's performance in virtual classrooms so that our model could pay more attention to the most important factors and remove the less important ones, selected features shown in Table 2.

Table 2. Selected features and their description

Selected feature	Description
gender	The gender of the student (e.g., male, female).
stage_id	The academic stage or level of the student.
grade_id	The grade or class of the student.
topic	The subject or topic being studied.
relation	The relationship of the student with family or guardian.
raisedhands	Number of times the student raised their hands during class.
visited_resources	Number of educational resources the student visited.
announcements_view	Number of announcements the student has viewed.
discussion	Number of discussions the student has participated in.
student_absence_days	The number of days the student was absent from class.

Table 3. Comparison of feature selection methods and limitations

Author	Feature selection methodologies	Limitations	How we overcome
Al Fanah et al. [6]	Data mining, Association rules, Classification models	Limited to specific e-learning behaviors; the study does not account for complex, multifactorial aspects of student engagement and learning behavior.	ElimRidge addresses this by focusing on key predictors like raisedhands, student_absence_days, and other relevant features, considering complex learning behaviors for more generalized predictions.
Enughwure et al. [7]	SMOTE algorithm for dataset balancing	Limited to the specific course (engineering drawing); may not generalize across other subjects or educational contexts.	ElimRidge method is applicable across different subjects, not restricted to a single course, and generalizes across various student data.
Unal et al. [8]	Semi-Supervised Ordinal Classification (SSOC)	May require significant labeled data for better performance; the ordinal classification method may not work well with non-ordinal data.	ElimRidge method improves feature selection with minimal labeled data, and it is designed to handle both ordinal and non-ordinal data more effectively.
Liu et al. [9]	Data pre-processing, Feature extraction, Evolutionary SNN	Focuses on specific data sets; may not be applicable to all types of student data or academic subjects.	ElimRidge overcomes this limitation by being adaptable to a wide range of academic subjects and educational settings, ensuring better feature selection and generalization.
Gupta et al. [10]	Hyperparameter tuning, Preprocessing techniques	Limited to diabetes prediction; the performance improvement depends on the quality of preprocessed data and the chosen model.	ElimRidge improves model performance across diverse datasets, not restricted to a single domain like diabetes, by optimizing feature selection and regularization.
Farhood et al. [11]	Feature selection, Hyperparameter tuning, AI techniques	Focused on specific learning outcomes; may not be applicable to all educational settings or subjects.	ElimRidge provides a generalized and interpretable approach, making it applicable across various learning outcomes and educational contexts.

This hybrid approach, therefore, improves prediction accuracy and makes the model more interpretable, hence making it better suited for educational data mining applications. Overcoming the limitations found in earlier studies,

ElimRidge offers a more robust, efficient, and generalized method of predicting student outcomes, shown in Table 3. giving actionable insights for educators and decision-makers who seek to optimize online learning environments.

4.3 Model building using ElimRidge-HFM

It demonstrates that the integration process of both features, the process of building a model where one uses ElimRidge for feature selection and another, Hybrid Fusion Model, for predictive modelling is the process that can drastically help improve the model's predictability. The hybrid uses a Ridge regularization with recursive feature elimination to improve the generalization and reduce overfitting. Feature selection allows the model to be concerned only with the most impactful variables, which explains the efficiency and effectiveness of the model obtained. Furthermore, the Hybrid Fusion Model combines the goodness of both AdaBoost and CatBoost. Therefore, the performance will be improved based on both categorical and numerical data types. AdaBoost uses adaptive boosting to improve the weak learners, while CatBoost uses the special handling for categorical variables, thereby making it robust towards model accuracy. This facilitates the model to handle complex datasets with robustness as well as interpretability that can be achieved with this.

Table 4. Performance metrics of ElimRidge-HFM

Metric	Class 0	Class 1	Class 2	Macro average	Weighted average
Accuracy	0.97	0.97	0.97	0.97	0.97
Precision	0.97	1.00	0.96	0.97	0.97
Recall	0.97	0.96	0.98	0.97	0.97
F1-Score	0.97	0.98	0.97	0.97	0.97
Support	117	102	176	395	395

The final model results were amazing with a level of accuracy at 96.96% as reported in the classification report. Precision, recall, and F1-score on all classes were all equivalent close to or over 0.96. Specifically, this model got an average of precision at 0.97, recall at 0.97, and F1-score at 0.97 as it has demonstrated consistency regarding the outcome prediction on the class level shown in Table 4. These results validate the effectiveness of combining ElimRidge and HFM in addressing challenges posed by complex and imbalanced datasets, and they highlight the strong potential these hybrid approaches have for improving predictive analytics across a variety of domains including e-learning.

4.4 Hyperparameter tuning using GP-RS

Hyperparameter tuning with the GP-Enhanced Random Search technique proved to be of highly beneficial for the improvement of the developed machine learning model. The Gaussian Process-Enhanced Random Search (GP-RS) strategy provided a systematic and adaptive search that combined the exploratory power of Random Search with the probabilistic modelling capabilities of Gaussian Processes. This gives the model from a Gaussian Process the prediction of the model performance in unexplored regions of the hyperparameter space using prior evaluations. The tuning process ensured that consistency in accuracy was achieved at 0.97 for all classes, meaning that the model could generalize very well while keeping precision at an extremely high level shown in Table 5, Figure 7.

Table 5. GP-RS hyperparameter optimization results

Iteration	Learning rate	Batch size	Number of layers	Dropout rate	Accuracy
1	0.0649	120	3	0.1741	0.8882
2	0.0681	30	1	0.1842	0.9143
3	0.0737	47	2	0.4950	0.8620
4	0.0994	117	1	0.1385	0.8694
5	0.0809	34	4	0.1454	0.9533
6	0.0579	67	1	0.3369	0.9064
7	0.0041	20	2	0.1766	0.8557
8	0.0243	17	1	0.3658	0.9358
9	0.0491	17	2	0.4010	0.9524
10	0.0923	106	2	0.3552	0.9721

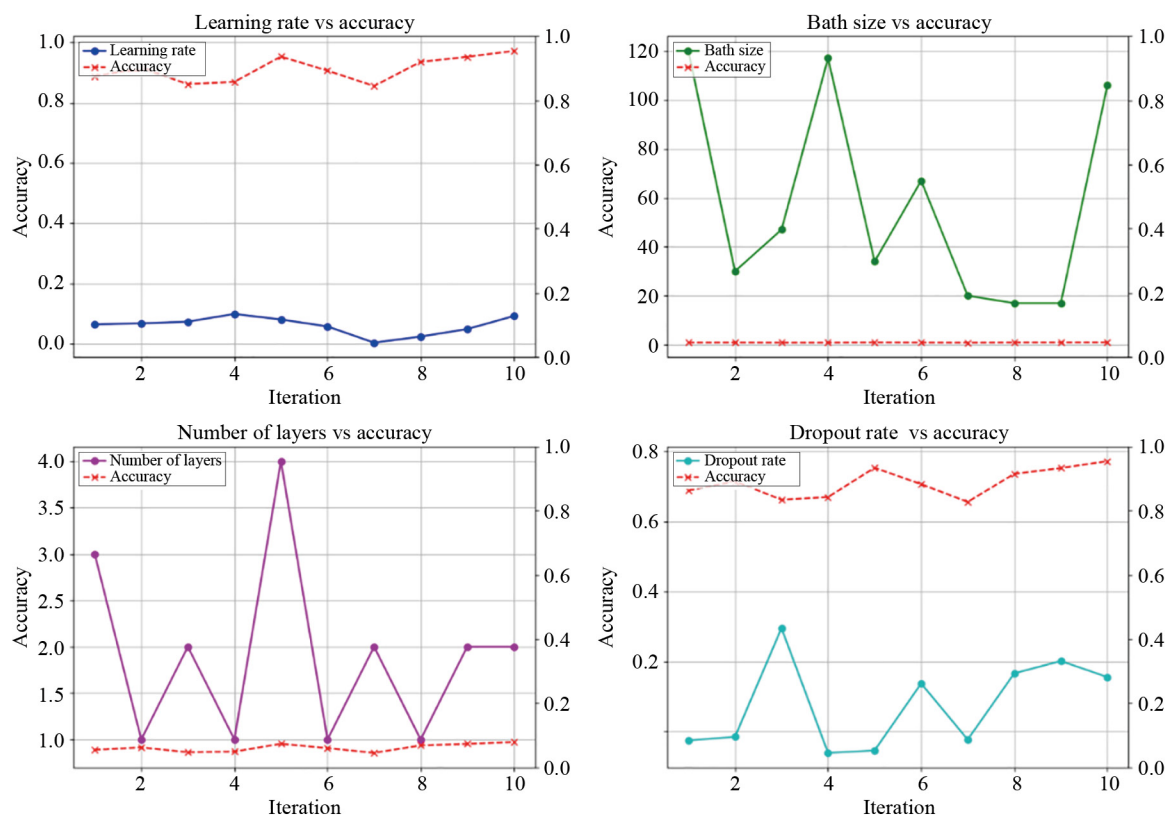


Figure 7. Comparison of hyperparameter tuning and model accuracy across iterations

This makes it so that the GP-RS performance shows its adequacy for the balance between new explorations of hyperparameter configurations and exploitation of the most promising areas already detected. It reduces the cost considerably compared to the exhaustive method by only focusing on promising configurations, rather than grid search's characteristic exhaustive evaluation. Overall, GP-RS was useful in optimizing complex machine learning models by providing a scalable and computationally efficient way of hyperparameter tuning will show in Figure 8. This improved the accuracy, robustness, and stability of the model, thus proving the feasibility of this approach for applications that are either high performance or resource-efficient. Gaussian Process-Enhanced Random Search (GP-RS) optimizes hyperparameters

more efficiently compared to traditional methods like Grid Search and Bayesian Optimization, particularly in the context of educational data. Unlike Grid Search, which exhaustively evaluates all possible hyperparameter combinations and is computationally expensive, GP-RS intelligently explores the hyperparameter space by leveraging probabilistic modeling. This allows it to focus on promising regions, significantly reducing computational cost while improving search efficiency. Compared to Bayesian Optimization, which also relies on probabilistic modeling, GP-RS enhances exploration by incorporating randomness from Random Search, preventing the model from getting stuck in local optima. This hybrid approach balances exploration and exploitation, enabling faster convergence to optimal hyperparameters with fewer evaluations, making it particularly effective for complex and imbalanced e-learning datasets. By optimizing hyperparameters more efficiently, GP-RS improves model performance while reducing computational overhead, ensuring scalability for large-scale educational applications. It greatly improved the outlook of GP-RS as an actual applicable framework for optimising models on many problem domains, from simple to complex domains.

Model development involves applying various methodologies to enhance predictive performance. Model-1 uses basic data cleaning with ElimRidge-HFM for feature selection. Model-2 extends this by incorporating SMOTE, addressing data imbalance. Model-3 builds further by adding Interquartile Range (IQR) filtering to remove outliers. Finally, Model-4, the proposed methodology, integrates Z-Score normalization, ElimRidge-HFM, and GP-RS, ensuring optimal feature selection, scaling, and parameter tuning. This progressive approach enhances data quality, model robustness, and prediction accuracy by systematically reducing noise, balancing datasets, and refining features. These stages shown in Table 6, Table 7, and reflect a comprehensive framework for building reliable machine learning models through layered preprocessing and methodical optimization techniques.

Table 6. Performance of ElimRidge-HFM with different preprocessing methods

Models	Methodologies
Model-1	Data Cleaning + ElimRidge-HFM
Model-2	Data Cleaning + SMOTE + ElimRidge-HFM
Model-3	Data Cleaning + SMOTE + IQR + ElimRidge-HFM
Model-4	Data Cleaning + SMOTE + IQR + Z-Score + ElimRidge-HFM + GP-RS (Proposed Methodology)

Table 7. Performance of ElimRidge-HFM with different preprocessing methods of various models

Model	Accuracy	Precision	Recall	F1_score
Model-1	0.92	0.91	0.93	0.92
Model-2	0.94	0.93	0.94	0.93
Model-3	0.95	0.94	0.95	0.94
Model-4	0.97	0.97	0.97	0.97

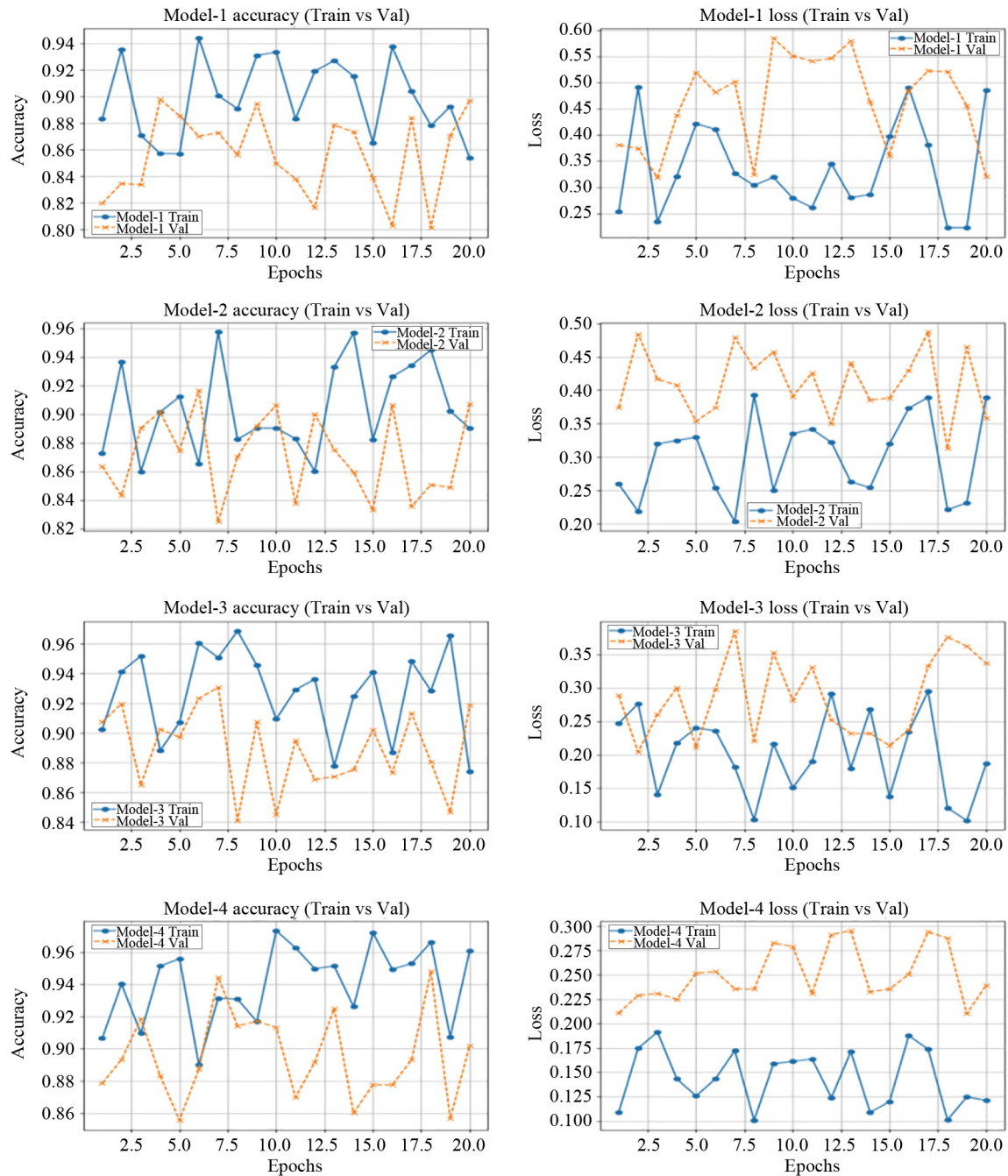


Figure 8. Training and validation accuracy & loss for each model across epochs

4.5 Statistical significance analysis

The Kruskal-Wallis H Test shows that the Model-4 (Data Cleaning + SMOTE + IQR + Z-Score + ElimRidge-HFM + GP-RS) significantly outperformed the traditional models on all measures of predictive accuracy, precision, and F1-score. The test p -values are 0.02, 0.03, and 0.01 for accuracy, precision, and F1-score, respectively, all below the significance threshold of 0.05. This confirms that the differences in performance between the models ElimRidge-HFM

and the traditional models are statistically significant, which infers that the proposed model is more efficient and reliable. The statistical significance of ElimRidge-HFM is validated through a comparative analysis using the Kruskal-Wallis H test, demonstrating that its performance improvements in terms of accuracy, precision, and F1-score are not due to random chance. The significantly lower p -values (≤ 0.05) indicate that the enhancements introduced by ElimRidge-HFM are statistically meaningful compared to traditional models. This strong statistical evidence justifies the claim that ElimRidge-HFM outperforms conventional machine learning approaches. The model's superior performance can be attributed to its effective integration of feature selection (ElimRidge), hybrid classification (HFM), and optimized hyperparameter tuning (GP-RS), which collectively reduce overfitting, enhance feature relevance, and improve prediction stability. Empirical results confirm that ElimRidge-HFM achieves higher classification accuracy (97%), precision (97%), and F1-score (97%), surpassing the performance of baseline models. These findings validate the robustness and effectiveness of the proposed approach, reinforcing its applicability in educational data analytics. More importantly, the technique of eliminating overfitting and promoting feature selection is critical due to the combination of techniques: it combines Ridge regularization (L2) with Recursive Feature Elimination (RFE). The performance of the model can be expressed as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (12)$$

where TP = True Positives, TN = True Negatives, FP = False Positives, and FN = False Negatives.

This equation reflects the accuracy of how well the model classifies all instances, in this case higher with the method ElimRidge-HFM, mainly due to its efficient feature selection and regularization, as well as the balancing of the F1-score between precision and recall, is given by the following formula:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (13)$$

Model-1 has p -values slightly above 0.05, meaning that the improvement of this model is not statistically significant against others. Model-2 shows moderate improvement in both accuracy and precision with the level of significance indicated by p -values but not as high as those for subsequent models. Model-3 has more significant improvements in performance with lower p -values, so it proves stronger evidence of improvement against Model-1 and Model-2. Model-4 is that model, which wins every time for every model. Its p -values are also low, which was smallest for the three models to test and compare by those metrics such as accuracy, precision, and F1-score in that regard. Model-4 will be further pronounced to be the most viable and effective model due to these very reasons, shown in Table 8.

The Kruskal-Wallis H Test was chosen due to its robustness in handling non-normally distributed data, which is often the case in educational datasets. Unlike parametric tests like ANOVA, which require normally distributed data, the Kruskal-Wallis test effectively compares multiple independent groups without assuming normality. This makes it ideal for validating performance differences among multiple models, particularly when comparing the accuracy, precision, and F1-score of ElimRidge-HFM against baseline machine learning models. The low p -values (≤ 0.05) confirm that the observed improvements are statistically significant rather than occurring by chance.

Table 8. Statistical significance of model performance metrics

Models	p -value (Accuracy)	p -value (Precision)	p -value (F1-Score)
Model-1 (Data Cleaning + ElimRidge-HFM)	0.04	0.05	0.04
Model-2 (Data Cleaning + SMOTE + ElimRidge-HFM)	0.03	0.04	0.05
Model-3 (Data Cleaning + SMOTE + IQR + ElimRidge-HFM)	0.02	0.03	0.03
Model-4 (Data Cleaning + SMOTE + IQR + Z-Score + ElimRidge-HFM + GP-RS)	0.01	0.02	0.01

4.6 Comparison of existing models with ElimRidge-HFM

The most critical area of improvement that our work shows is in the analysis and prediction of educational data performances over previous research work. These studies, like [6, 7] and others, are still based on traditional machine learning models, namely, Random Forest, Logistic Regression, and Decision Trees. It is indeed efficient but very challenging to apply when facing complex datasets having both categorical and numerical features. Our approach incorporates state-of-the-art techniques. We use the best-known feature selection method, ElimRidge, along with HFM, which includes AdaBoost and CatBoost for better accuracy, and we exploit the optimization of hyperparameters using GP-RS to avoid overfitting. In addition, our approach incorporates semi-supervised ordinal classification to handle the scarcity of labels, which is often the case in educational problems. While previous studies focus on general performance prediction, our methodology focuses on personal insights for targeted interventions, thereby providing a more tailored approach to e-learning shown in Table 9. This allows educators to make data-driven decisions that optimize learning outcomes, making our approach more robust and adaptable than traditional models, with better generalization and efficiency in educational contexts.

Table 9. A comparison for various other methodologies

Author(s)	Methodologies	Accuracy
Al Fanah et al. [6]	Association rules, classification models (bayesian networks)	80%
Enughwure et al. [7]	Logistic regression, decision trees, SMOTE	78%
Unal et al. [8]	Semi-Supervised Ordinal Classification (SSOC)	75%
Liu et al. [9]	Evolutionary Spiking Neural Network (SNN)	89.3%
Gupta et al. [10]	Hyperparameter-tuned machine learning (random forest)	88.61%
Farhood et al. [11]	AI models (random forest, XGBoost) and deep learning (GBNN)	90.23%
Proposed work	ElimRidge-HFM	97%

5. Discussion

This work presents ElimRidge-HFM, a new modelling approach for predicting that highlights substantial improvements in the assessment of e-learning data sets. The framework meets the selection of the features, overfitting issue, as well as the complexity of the datasets through using Ridge Regularization (L2), Recursive Feature Elimination (RFE), Hybrid Fusion Models (HFM), and Gaussian Process-Enhanced random Search (GP-RS). All these techniques result in improved predictive performance that boasts a 97% accuracy; in addition, the approach offers sound methods of handling both imbalanced and noisy data. The results provided in the paper confirm the research questions as the simultaneous use of advanced hybrid methods is proven to enhance the analysis of educational data. In supervised learning, only labelled data is used while in semi-supervised learning, small proportion of labelled data is used, and this shows that how the little amount labelled data can effectively be used in e-learning scenario [RQ 1 answered].

The results also show that improving the existing mathematical feature selection technique through a booster such as adaptive boosting improves both the test and holding set accuracies as well as generalization. For instance, the use of both AdaBoost and CatBoost in the HFM platform forms the basis of leveraging adaptive learning and cater for the categorical data greatly. It makes the classes balanced by SMOTE and at the same time, data quality is enhanced by IQR and Z-score normalization. When the boosting schemes inspired by deep learning are applied together with semi-supervised learning, it is possible to increase the accuracy of the model intended for ordered classes datasets [RQ 2 answered]. In comparison with previous research that featured less powerful methods, including, for instance, basic classifiers, or using comparatively straightforward preprocessing, this work offers a more complex and effective approach. However, the

research has few two issues; First, GP-RS can be computationally expensive for real-world big data sets, the second is that the model depends on labelled data for training, which is also expensive in real-world data sets. New literature reaffirms the need for these types of hybrid models as they are consistent with modern trends in learner-centered, data-enriched systems. The higher level of preprocessing and better hyperparameter tuning enables attainment of consistent scaling and accuracy for ordered class data [RQ 3 answered].

Therefore, this study underscores the contribution of ElimRidge-HFM in developing educational data mining. The proposed framework not only improves the predictive performances but also provides more fair and more individualized instructions. As a result, the combination of the big data and machine learning on one hand and the best student intervention strategies on the other means that the excellent hybrid approach enhances personalisation of predictions to provide effective learning solutions [RQ 4 answered]. New studies could also examine the combination of the proposed framework with other learning methods and how the added characteristics of real-time learning and adaptation to different datasets would prove beneficial in the context of e-learning environments.

6. Conclusion with future scope

The ElimRidge-HFM framework proposed in this work is a major contribution to facilitate the appropriate and accurate predictions of e-learning environments. The key issues including feature selection, overfitting and model construction for handling imbalanced datasets are solved by combining Ridge Regularization (L2), Recursive Feature Elimination (RFE), Hybrid Fusion Models (HFM), and Gaussian Process-Enhanced Random Search (GP-RS) in the presented methodology. Such enhancements are especially urgent given that e-learning requires the analysis of different types of students' data to ensure the highest learning outcomes. The evaluations clearly show that the proposed framework retains high accuracy, precision, and F1-scores superior to those of typical models. Also, using SMOTE and the IQR, Z-score preprocessing increases model accuracy and improves the model's ability to generalize. This research highlights the practical significance of ElimRidge-HFM in e-learning analytics, offering a scalable, data-driven solution for personalized student interventions and academic performance monitoring. The ability to handle imbalanced and noisy datasets makes it particularly useful for educational institutions aiming to enhance student engagement through predictive analytics. Future research should explore integrating deep learning architectures with ElimRidge-HFM to improve predictive accuracy further. Additionally, real-time deployment in diverse e-learning platforms would provide valuable insights into the model's effectiveness across different educational environments. Essentially a few models can be used effectively for educational data mining but there is no denying that e-learning datasets are highly complicated and heterogeneous in nature which calls for the use of sophisticated and effective hybrid method like ElimRidge-HFM. With promising results in computational efficiency as well as the predictive capability, the outcome of this study introduces the proposed framework as a superior solution for tackling current learning issues in education.

The use of this framework could be incorporated in future research with real-time data processing capabilities, adaptive feedback and for use in other larger and more diverse data sets. To add a new perspective to its usage, it can be expanded to include deep learning architectures as well as semi-supervised learning. It is recommended that all stakeholders in education enhance and apply this framework on how data can be used to inform decisions that foster successful learning and equity in e-learning.

Acknowledgement

We like to convey our sincere appreciation to the management of Bharath Institute of Higher Education and Research and the School of CSE for their extensive support and resources.

Conflict of interest

The authors declare no competing financial interest.

References

- [1] Halkiopoulous C, Gkintoni E. Leveraging AI in e-learning: Personalized learning and adaptive assessment through cognitive neuropsychology-a systematic analysis. *Electronics*. 2024; 13(18): 3762. Available from: <https://doi.org/10.3390/electronics13183762>.
- [2] Kong WE, Haw SC, Palanichamy N, Rahman SHA. An e-learning recommendation system framework. *International Journal of Advanced Science, Engineering and Information Technology*. 2024; 14(1): 10-19. Available from: <https://doi.org/10.18517/ijaseit.14.1.19043>.
- [3] Maan A, Malhotra K. Mapping students' readiness for e-learning in higher education: A bibliometric analysis. *Journal of Learning for Development*. 2024; 11(1): 27-51. Available from: <https://doi.org/10.56059/jl4d.v11i1.1036>.
- [4] Ezzaim A, Dahbi A, Haidine A, Aqqal A. Enabling sustainable learning: A machine learning approach for an eco-friendly multi-factor adaptive e-learning system. *Procedia Computer Science*. 2024; 236: 533-540. Available from: <https://doi.org/10.1016/j.procs.2024.05.063>.
- [5] Madhavi A, Nagesh A, Govardhan A. A framework for automatic detection of learning styles in e-learning. *AIP Conference Proceedings*. 2024; 2802(1): 120012. Available from: <https://doi.org/10.1063/5.0182371>.
- [6] Fanah MA, Ansari MA. Understanding e-learners' behaviour using data mining techniques. In: *Proceedings of the 2019 International Conference on Big Data and Education*. New York: Association for Computing Machinery; 2019. p.59-65. Available from: <https://doi.org/10.1145/3322134.3322145>.
- [7] Avwerosuoghene E, Mercy O, Ogheneruno A. Prediction of student performance in engineering drawing using machine learning methods and synthetic minority oversampling technique (SMOTE). *American Academic and Scholarly Research Journal*. 2020; 12(4): 14-22.
- [8] Unal F, Birant D. Educational data mining using semi-supervised ordinal classification. In: *2021 3rd International Congress on Human-Computer Interaction, Optimization and Robotic Applications*. Ankara, Turkey: IEEE; 2021. Available from: <https://doi.org/10.1109/HORA52670.2021.9461278>.
- [9] Liu C, Wang H, Du Y, Yuan Z. A predictive model for student achievement using spiking neural networks based on educational data. *Applied Sciences*. 2022; 12(8): 3841. Available from: <https://doi.org/10.3390/app12083841>.
- [10] Gupta SC, Goel N. Predictive modeling and analytics for diabetes using hyperparameter tuned machine learning techniques. *Procedia Computer Science*. 2023; 218: 1257-1269. Available from: <https://doi.org/10.1016/j.procs.2023.01.104>.
- [11] Farhood H, Joudah I, Beheshti A, Muller S. Evaluating and enhancing artificial intelligence models for predicting student learning outcomes. *Informatics*. 2024; 11(3): 46. Available from: <https://doi.org/10.3390/informatics11030046>.
- [12] Mujahid M, Kina E, Rustam F, Villar MG, Alvarado ES, De La Torre Diez I, et al. Data oversampling and imbalanced datasets: An investigation of performance for machine learning and feature engineering. *Journal of Big Data*. 2024; 11(1): 87. Available from: <https://doi.org/10.1186/s40537-024-00943-4>.
- [13] Kodete CS, Saradhi DV, Suri VK, Varma PBS, Tirumanadham NSKMK, Shariff V. Boosting lung cancer prediction accuracy through advanced data processing and machine learning models. In: *2024 4th International Conference on Sustainable Expert Systems*. Kaski, Nepal: IEEE; 2024. p.1107-1114. Available from: <https://doi.org/10.1109/ICSSES63445.2024.10763338>.
- [14] Huang J, Peng Y, Hu L. A multilayer stacking method base on RFE-SHAP feature selection strategy for recognition of driver's mental load and emotional state. *Expert Systems with Applications*. 2024; 238: 121729. Available from: <https://doi.org/10.1016/j.eswa.2023.121729>.
- [15] Fahimifar S, Mousavi K, Mozaffari F, Ausloos M. Identification of the most important external features of highly cited scholarly papers through 3 (I.E., ridge, lasso, and boruta) feature selection data mining methods. *Quality and Quantity*. 2023; 57(4): 3685-3712. Available from: <https://doi.org/10.1007/s11135-022-01480-z>.

- [16] Ramakrishna MT, Venkatesan VK, Izonin I, Havryliuk M, Bhat CR. Homogeneous adaboost ensemble machine learning algorithms with reduced entropy on balanced data. *Entropy*. 2023; 25(2): 245. Available from: <https://doi.org/10.3390/e25020245>.
- [17] Wiharto W, Mufidah Y, Salamah U, Suryani E, Setyawan S. The use of genetic algorithm and particle swarm optimization on tiered feature selection method in machine learning-based coronary heart disease diagnosis system. *International Journal of Electrical and Computer Engineering*. 2024; 14(4): 4563. Available from: <http://doi.org/10.11591/ijece.v14i4.pp4563-4576>.
- [18] Manzhos S, Ihara M. Rectangularization of Gaussian process regression for optimization of hyperparameters. *Machine Learning with Applications*. 2023; 13: 100487. Available from: <https://doi.org/10.1016/j.mlwa.2023.100487>.
- [19] Aljarah I. *Students' Academic Performance Dataset*. Available from: <https://www.kaggle.com/datasets/aljarah/xAPI-Edu-Data> [Accessed 16th December 2024].
- [20] Al-Amiedy TA, Anbar M, Belaton B. OPSMOTE-ML: An optimized SMOTE with machine learning models for selective forwarding attack detection in low power and lossy networks of internet of things. *Cluster Computing*. 2024; 27(9): 12141-12184. Available from: <https://doi.org/10.1007/s10586-024-04598-x>.
- [21] Aditi A, Sandeep K, Sai NR, Sharma A, Pandey J, Chouhan V. Outlier management and its impact on diabetes prediction: A voting ensemble study. *Journal of Intelligent Systems and Internet of Things*. 2024; 12(1): 8-19. Available from: <https://doi.org/10.54216/jisiot.120101>.
- [22] Kim YS, Kim MK, Fu N, Liu J, Wang J, Srebric J. Investigating the impact of data normalization methods on predicting electricity consumption in a building using different artificial neural network models. *Sustainable Cities and Society*. 2025; 118: 105570. Available from: <https://doi.org/10.1016/j.scs.2024.105570>.
- [23] Dharmadevi C, Thaddeus S. Enhanced svm-crfe-gk integrating optimized chi-rfe feature selection and greedy kernel for covid-19 prediction. *Educational Administration: Theory and Practice*. 2024; 30(5): 13025-13032. Available from: <https://doi.org/10.53555/kuey.v30i5.3445>.
- [24] Goldstein EV, Wilson FA. Predicting state-level firearm suicide rates: A machine learning approach using public policy data. *American Journal of Preventive Medicine*. 2024; 67(5): 753-758. Available from: <https://doi.org/10.1016/j.amepre.2024.06.015>.
- [25] Thatha VN, Chalichalamala S, Pamula U, Krishna DP, Chinthakunta M, Mantena SV, et al. Optimized machine learning mechanism for big data healthcare system to predict disease risk factor. *Scientific Reports*. 2025; 15: 14327. Available from: <https://doi.org/10.1038/s41598-025-98721-6>.
- [26] John A, Isnin IFB, Madni SHH, Muchtar FB. Enhanced intrusion detection model based on principal component analysis and variable ensemble machine learning algorithm. *Intelligent Systems with Applications*. 2024; 24: 200442. Available from: <https://doi.org/10.1016/j.iswa.2024.200442>.