INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

ORIGINAL ARTICLE

OPEN ACCESS

Received: 17/12/2025

Accepted: 03/01/2026

Published: 22/01/2026

Citation: Arulvani SV, Dr. C Jayanthi (2026) A Robust Statistical Inference-Driven Subspace Ensemble Classifier for Student Dropout detection in Higher Education. Indian Journal of Science and Technology 19(2): 71-87. https://doi.org/10.17485/IJST/v19i2.1870

* **Corresponding authors**.

arulphdcs@gmail.com

jayanthibabu117@gmail.com

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

A Robust Statistical Inference-Driven Subspace Ensemble Classifier for Student Dropout detection in Higher Education

S V Arulvani[1]*, Dr. C Jayanthi[2]*

1 Research Scholar, PG & Research Department of Computer Science, Government Arts College (Autonomous), Karur-5, Affiliated to Bharathidasan University,Tiruchirappalli, Tamil Nadu, India

2 Associate Professor, PG & Research Department of Computer Science, Government Arts College (Autonomous), Karur-5, Affiliated to Bharathidasan University, Tiruchirappalli , Tamil Nadu, India

Abstract

Background/Objectives: Student drop-out is the most burning issues in higher education that induce considerable social and economic costs. The main objective of this study is to predict accurately the dropout risks due to the complexity and diversity of student behaviors. **Method:** This study has proposed a novel Robust Statistical Inference Fuzzified Random Subspace Ensemble Classifier (ROSIF-RASE). The proposed ROSIF-RASE model focusses on improving prediction accuracy while minimizing time consumption. This model includes several key processes, namely data acquisition, preprocessing, feature selection, and classification. In the data acquisition stage, raw student data samples are collected from the dataset with help of IoT devices. In the second stage, a preprocessing step is performed to convert the input data into a usable format. This step involves two main processes namely imputing missing values and eliminating outliers. To address missing data, the model utilizes robust nonparametric estimator, which estimates the missing values based on known data points. For identifying and managing outliers, statistical residual test is applied for effectively removed from the dataset. After preprocessing, Statistical Inference feature re-weighting algorithm is employed for significant feature selection to reduce the student dropout prediction time. Finally, the classification process is performed using the Fuzzified preferential Random subspace ensemble Classifier to improve the stability and accuracy of student dropout prediction. This process helps to enhance the accuracy of dropout prediction. Experimental assessments are conducted using various evaluation metrics, student dropout prediction accuracy, and precision, sensitivity, and student dropout prediction time. **Findings:** The proposed ROSIF-RASE model shows measurable performance gains, with increases of 6% in accuracy, 4% in precision, 2% in sensitivity, and 3% in F1-score, along with a minimum 10% decrease in student dropout prediction time. **Novelty:** The proposed model improves the student dropout prediction accuracy with lesser time compared to different existing methods.

Keywords: Student dropout prediction; Robust nonparametric estimator; Statistical residual test; IoT devices; Statistical Inference feature re-weighting algorithm; Fuzzified preferential Random subspace ensemble Classifier

1 Introduction

Education Data Mining (EDM) refers to the application of data mining techniques to educational data for discovering patterns, relationships, and imminent that improves educational practices, outcomes, and decision-making. The main aim is to analyze student data to enhance teaching and learning processes, predict student performance, and optimize educational resources.

An ensemble model was proposed in [1] with the aim of predicting student dropout based on characterization and knowledge discovery process. The designed model minimized the financial losses of educational institutions for higher education student dropout analysis. However, it was not possible to consider features related to market demands, economic activity in the country, and evolution of the social profile, among others to improve model performance, which aimed at achieving higher accuracy while minimizing time. To overcome above issue, an eXtreme Gradient Boosting with Adaptive Guided best SinhCosh Optimizer(XG-AGbSCHO) algorithm was designed in [2] for multi-class classification of student academic performance prediction. However, it did not focus on AI-driven solutions for addressing student dropout leading to increasing the time complexity.

To overcome above drawback, a novel XGBoost algorithm that reduced the time complexity and performed in [3] to evaluate students' academic progress and dropout risk. However, advanced algorithms have not been applied to enhance the prediction precision. The Various Artificial Intelligence algorithms overcome the above issue, the precision was improved and introduced in [4] for student dropout prediction. Although better accuracy was achieved, it was considered that including socioeconomic characteristics of the student. A clustering-based approach was designed in [5] for categorizing the students based on their academic performance and focusing on their communications. However, the performance was significantly reduced when dealing with extremely large dataset.

A CatBoost algorithm was designed in [6]with the aim of Student dropout prediction. However, the algorithm was not scalable or effective to enhance student dropout prediction. To overcome this issue, a Machine learning models were developed in [7]to predict higher secondary education dropout. However, the time consumption of the dropout prediction was high. A multinomial logistic regression analysis introduced in [8] integrated with finite mixture models, for predicting the student dropout by reducing the time consumption. However, the logistic regression analysis was inefficient for large scale student data samples. In order to overcome this issue, a comprehensive machine learning (ML) and explainable AI (XAI) model were introduced in [9] for dropout prediction by solving the imbalanced dataset problem and selecting the significant features. Neural Network based model was developed in [10] for solving the issues of the student dropout over time. However, it did not consider the different factors contributing to predict student dropout and enhance the accuracy of the model's decisions.

Particle Swarm Optimization (PSO) integrated with Weighted Ensemble Framework to overcome above drawback and designed in [11] for student dropout prediction with high accuracy. However, it did not incorporate deep learning architectures and exploring explainable AI techniques to enhance scalability, and robustness of the framework in different educational settings. In [12], uncertainty related to the decision-making process was considered to detect dropout students among first and second-year students. However, it was not possible to guarantee higher accuracy. To overcome this issue, relationship between student engagement and dropout risk assessment was performed in [13] to ensure the better accuracy. However, complexity of dropout risk assessment was not reduced.

Machine learning (ML) techniques were developed in [14]to predict early student dropout using real-world data by considering the factors, including demographic, academic, and behavioral information via minimizing the dropout risk assessment of complexity level. However, the dimensionality reduction remains challenging issue. Multilayer Perceptron classifier was developed in [15] to predict student suspension and dropout with the aim of improving the precision. However, feature augmentation and cost-sensitive learning issues remained unresolved, limiting further improvement in class separability and prediction precision.

1.1 Major Contributions

To overcome the issues from the existing methods, a ROSIF-RASE model is developed with the new contributions are summarized as follows,

- To enhance the accuracy of student dropout prediction accuracy, novel ROSIF-RASE model is introduced, based on processes namely pre-processing, feature selection, and classification.
- To minimize the student dropout prediction, the ROSIF-RASE model integrates data pre-processing to attain a structured dataset. It also utilizes the Statistical Inference feature re-weighting algorithm to reduce the dimensionality of the dataset by choosing the more relevant features based on a similarity measure and removing the other redundant features for dropout prediction.

- A Fuzzified preferential Random subspace ensemble Classifier model is developed in ROSIF-RASE model for student dropout prediction using selected more relevant features. This approach increases the accuracy, precision, recall and minimizes incorrect classifications.
- Extensive analysis is conducted to measure the performance of the ROSIF-RASE model and other related works with various evaluation metrics.

Binary Classification Models were developed in [16] for student dropout prediction with high accuracy and F1 score. However, the models did not achieve higher positive class ratios. A hybrid neural network model (CNN-LSTMAE) model was developed in [17] for accurate student dropout prediction by extracting local features of students' behaviors. However, multimodal data was not explored to enhance the practical applicability of the model. An In-Session Stacked Ensemble Learning model was introduced in [18] for Dropout Prediction during course sessions by integrating the multiple base learners. The model improves accuracy, but computational efficiency was not enhanced. A Stacked Classifier Approach was developed in [19] for early detection of dropout risks. However, it did not explore methods to simplify the stacked classifier while maintaining predictive accuracy. A random forest (RF) and feature tokenizer transformer model were developed in [20] to predict academic attrition. However, the issue of time complexity in prediction performance remained unaddressed.

Regularized Regression Models were developed in [21] for predicting the higher education student dropout. However, it did not explore advanced techniques to significantly improve predictive performance in educational data mining. Support Vector Machine (SVM) machine learning method were developed in [22] to enhance the Dropout Prediction Performance through the oversampling. However, issues related to missing data handling and outlier removal in the student records were not addressed. A predictive model using ensemble machine learning method designed in [23] to improve the accuracy using student dropout in higher education. A method of predicting dropout introduced in [24] to enhance the dropout prediction and reduce the incidence of dropouts resulting from loss of scholarships. A novel approach that uses clustering as preprocessing to overcome data ambiguity and inconsistencies designed in [25].

The research often focuses on predictive accuracy but less on translating these findings into concrete, actionable intervention strategies that can be integrated into existing educational management systems. The limitation of Student dropout detection method incomplete/low-quality data, missing key factors (socio-economic, health), class imbalance in datasets, model complexity (feature engineering, interpretability), data privacy/ethics, focus on binary outcomes (drop/stay) instead of nuanced statuses, and challenges in integrating diverse data sources like LMS activity. These restrict generalizability, real-time application, and understanding complex, real-world influences beyond academic metrics, hindering effective, timely interventions. The Student drop-out is the most burning issues in higher education that induce considerable social and economic costs. This method faces challenges in accurately predicting dropout risks due to the complexity and diversity of student behaviors. To overcome this limitation, a novel RObust Statistical Inference Fuzzified RAndom Subspace Ensemble Classifier (ROSIF-RASE) model is developed to enhance the prediction accuracy with minimum time consumption.

2 Methodology

The proposed ROSIF-RASE model, illustrated in Figure 1, consists of four primary stages namely data acquisition, data preprocessing, feature selection, and classification. These interconnected phases work together to effectively differentiate the students into dropout and graduates. This section provides an in-depth explanation of the proposed model, including its ability to analyze student data, select meaningful features, and accurately classify student outcomes as dropout and graduate.

Figure 1 illustrates the structure of the proposed ROSIF-RASE model developed to deliver precise predictions of student dropout from their education. The model is composed of four key components such as data acquisition, preprocessing, feature selection, and classification. Each stage contributes significantly to refining the input data and improving the model overall predictive accuracy. The following subsections offer a process of these components to highlight their individual roles and implication within the overall framework.

2.1 Data Acquisition

The data acquisition phase serves as the base of the proposed ROSIF-RASE model. In this stage, relevant student-related data is collected from Predict Students' Dropout and Academic Success dataset from the kaggle repository https://www.kaggle.com/datasets/mattop/predict-students-dropout-and-academic-success.The dataset utilized in this paper is gathered from a higher education institution and integrates data sourced from multiple independent databases. It comprises the information about students enrolled in various undergraduate programs, including fields such as agronomy, design, education, nursing, journalism, management, social services, and technology. The dataset collects details available at the time of enrollment,

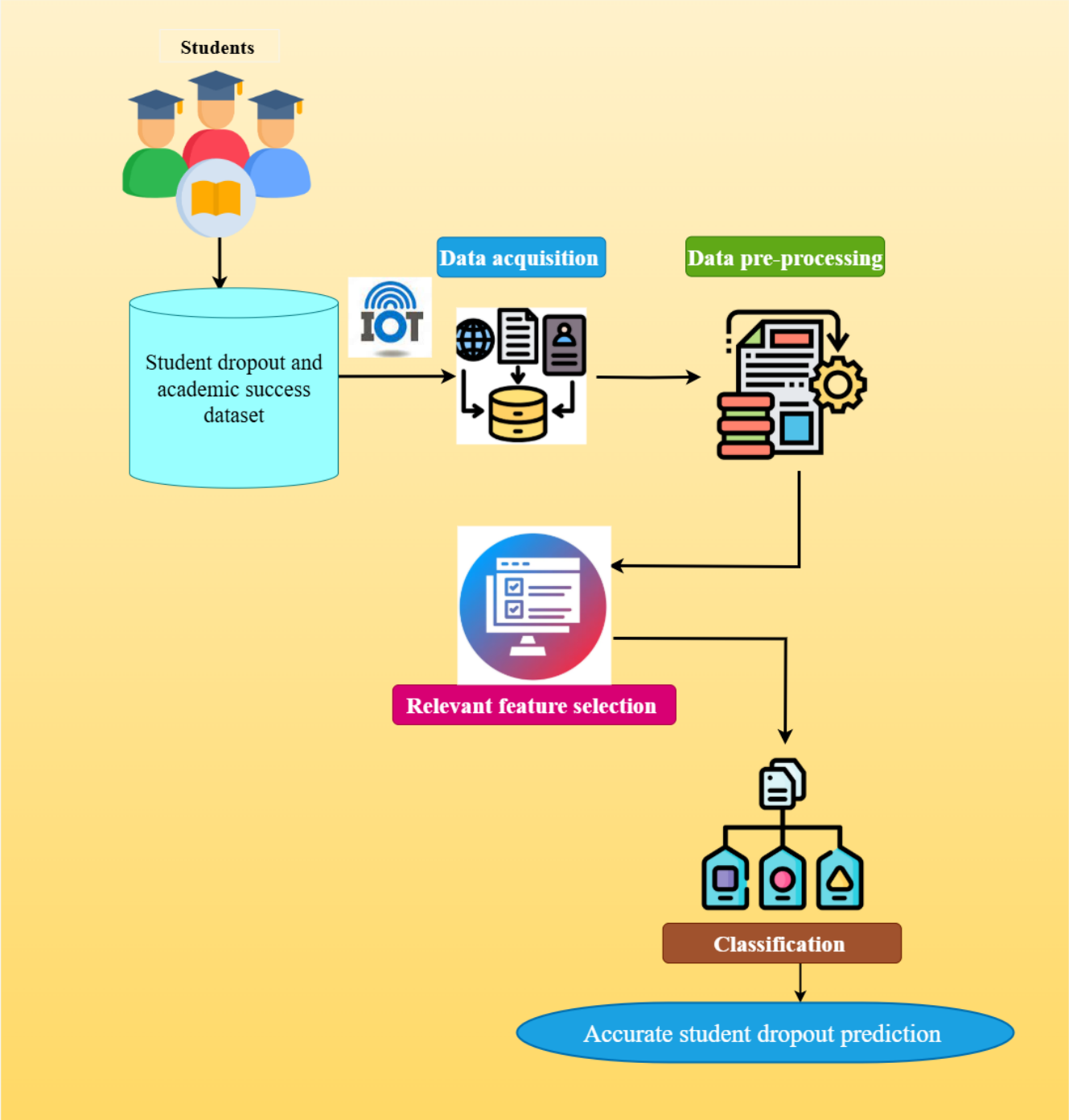

Fig 1. Architecture Diagram of Proposed ROSIF-RASE Model

such as academic background, demographic profiles, and socio-economic status. In addition, it includes academic performance outcomes from the first and second semesters. This data serves as the basis for developing classification models aimed at predicting student dropout risk and academic achievement. The dataset consists of 4424 records of students, with 37 attributes.

2.2 Data Preprocessing

The second process of proposed ROSIF-RASE model is a preprocessing. It is a fundamental step in preparing raw data for effective analysis and model training. In the context of the proposed ROSIF-RASE model, this phase involves cleaning, transforming, and organizing the collected student data to ensure, accuracy, and usability. The input student dataset 'DS' is organized in the form of matrix as follows,

$$D = \begin{bmatrix} A_1 & A_2 & \dots & A_m \\ Ds_{11} & Ds_{12} & \dots & Ds_{1n} \\ Ds_{21} & Ds_{22} & \dots & Ds_{2n} \\ \vdots & \vdots & \dots & \vdots \\ Ds_{m1} & Ds_{m2} & \dots & Ds_{mn} \end{bmatrix} \tag{1}$$

Where, input matrix 'S' is organized based on 'm' number of features or attributes $A_1, A_2, \dots A_m$ and it represented in the column and the overall student data samples 'Ds' presented in the 'n' row respectively. The pre-processing step includes two major processes namely missing data imputation and outlier data removal.

Robust nonparametric estimator is employed to address missing values within the input dataset. This estimator is a robust statistical technique that calculates a pseudo median based on the available data points. The process begins by detecting any missing value within a specific cell of the dataset. Once identified, the remaining values in the corresponding column are sorted in ascending order. Following this, the pair wise average of observations is calculated, which is defined as follows:

$$Q = \frac{Ds_i + Ds_j}{2} \quad \text{Where,} \quad i \neq j \tag{2}$$

Where, Q denotes pair wise average, Ds_i and Ds_j denotes a two non missing data points. Once the pairwise average from the available (non-missing) values is computed, the robust nonparametric estimator is applied to find the median of the pair wise average.

$$Ds_{miss} = med\ \{Q\} \tag{3}$$

Where, Ds_{miss} denotes a missing data, $med\ \{Q\}$ denotes a median of the pairwise average values. The identified missing value is estimated and filled using the existing data within the dataset. This approach ensures that all missing entries are addressed by relevant data points.

After handing in the missing data, statistical residual test is carried out in for outlier data removal from the input dataset. The statistical residual test is an applied statistical method used to identify noise or outliers within a dataset. It works by analyzing individual data points in a dataset and determining whether any values deviate significantly from the overall distribution. In this context, a data point is determined as an outlier if its normalized residual value considerably deviates from the majority of other values. The statistical approach for conducting the statistical residual test is formulated as follows:

$$\varphi_{Mean} = \frac{1}{n}\sum_{i=1}^{n} Ds_i \tag{4}$$

$$SRT = \left[\frac{Ds_i - \varphi_{Mean}}{\sigma}\right]; Z = \arg\max\ SRT \tag{5}$$

Where, 'SRT' indicates a statistical residual test, Ds_i indicates a data samples, φ_{Mean} refers to a mean value of samples, σ specifies a deviation between the data samples and mean. From the analysis, the maximum statistical residual test output '$\arg\max\ SRT$' of the particular data samples are called as anoutlier. The algorithmic step of data preprocessing is described as given below,

Algorithm 1: Data pre-processing
Input: Dataset 'D', features $A_1, A_2, \ldots A_m$, data samples or records $Ds_1, Ds_2, \ldots Ds_n$,
Output: Pre-processed dataset
Begin
Step 1: Number of features 'A_m'and data samples ' Ds_j' collected from the dataset
Step 2:**For** each data samples 'Ds' with features 'A'
Step 3: Formulate input matrix 'D' using (1)
Step 4: **If** missing data **then**
Step 5: Fill missing value in respective row using (3)
Step 6:End if
Step 7: Compute statistical residual test 'SRT" using (5)
Step 8: **if** $(\arg\max SRT')$ **then**
Step 9: Detect outliers data
Step 10: **else**
Step 11: Detect normal data
Step 12: **End if**
Step 13: Return (preprocessed dataset)
End

Algorithm 1 outlines the preprocessing steps used to improve the accuracy of student dropout prediction. Initially, the dataset is and it contains multiple samples and associated features. The first step involves detecting any missing values, which are then imputed by replacing them with the median of the corresponding feature values. Next, outliers are identified using the statistical residual test, which helps eliminate outlier data that negatively impacts model performance. These preprocessing steps collectively contribute to improving the precision and reliability of student dropout prediction.

2.3 Feature Selection

The third process of the proposed ROSIF-RASE model focuses on feature selection, which aims to enhance the model's accuracy by reducing the dimensionality of the input dataset. This step is essential for improving both the accuracy and speed of student dropout prediction. To achieve this, the ROSIF-RASE model uses a novel algorithm known as the Statistical Inference feature re-weighting algorithm, a similarity learning algorithm identifies and selects the most relevant features from the dataset, thereby minimizing computational complexity and ensuring faster prediction.

Initialize a feature vector in the distributed dataset.

$$A = \{ A_1, A_2, A_3, \ldots A_m \} \tag{7}$$

After that, weight vector 'W' is assigned for each features A_j with zeros.

$$W\ (A_j) = 0 \tag{8}$$

Where, $W\ (A_j)$ denotes a weight vector assigned to each features A_j with zero. In each iteration, the algorithm randomly selects feature vectors from the dataset and computes similarity using Laplace kernel functions.

$$K = \exp\left[-\sum_{j=1}^{m} \left(\frac{|A_j - T_c|}{D} \right) \right] \tag{9}$$

Where, K denotes a Laplace kernel, A_j indicates a features in the dataset, T_c indicates a target output, D symbolizes deviation between the attributes and its target. The Laplace kernel function returns the output value from 0 to 1. Based on kernel output, the nearest hit and miss value is computed as

$$K = \begin{cases} 1, & hit \\ 0, & miss \end{cases} \tag{10}$$

The kernel 'K' provides the output as 1 represents the nearest hit, kernel provides the output as 0 represents the nearest miss. Next, the weight of each feature is adjusted based on its relationship with the nearest hit and nearest miss samples. This update

reflects the feature differentiates between the same class and those of different classes. The weight vector for each feature is recalculated by considering the squared differences between the feature value of the current sample and those of its nearest hit and nearest miss as follows,

$$W_{t+1} = W_t - \left(A_j - hit_j\right)^2 + \left(A_j - miss_j\right)^2 \tag{11}$$

Where, W_{t+1} designates an updated weight vector of feature 'A_j', W_t represents the current weight vector, A_j denotes a j^{th} feature vector, hit_j represents a nearest hit value of feature vector, $miss_j$ designates a nearest missing value of feature vector. This process is repeated for a set number of iterations. After the final iteration, features are sorted based on their updated weights. The feature with the highest weight is retained, while lower weights are discarded. Removing these less significant features helps to minimize the overall time complexity of the student dropout prediction.

Algorithm 2: Statistical Inference feature re-weighting algorithm
Input: Preprocessed dataset, features $A_1, A_2, \ldots A_m$, data samples or records $Ds_1, Ds_2, \ldots Ds_n$,
Output: Select significant features
Begin
Step 1: Initialize maximum iteration '$t_{\max}$**'**,weight vector '$W\ (A_j)= 0'$
Step 2: Collect the features $A_1, A_2, \ldots A_m$ and data samples or records $Ds_1, Ds_2, \ldots Ds_n$
Step 3: While (t<$t_{\max}$**)do**
Step 4: For each features A_j
Step 5: Initialize the weight vector '$W\ (A_j)= 0'$
Step 6: End for
Step 7: For eachA_j
Step 8: Compute the similarity 'K' using (9)
Step 9: End for
Step 10: if$(K = 1)$**then**
Step 11: A feature is considered a nearest hit
Step 12: else
Step 13: A feature is considered a nearest miss
Step 14: End if
Step 15: Update weight (W_{t+1}) using (11)
Step 16: Increment t=t+1
Step 17: Go to step 3
Step 18: End while
Step 19: Sort the features according to weights
Step 20: Retain features with the most significant weight scores
Step 21: Remove Features with the less weight scores
End

Algorithm 2 outlines the procedure for selecting the more relevant features from the dataset to enhance the accuracy of student dropout prediction and reduce computational time. The process begins by utilizing the pre-processed dataset. Each feature is initially with the weight of zero. The algorithm then evaluates the similarity between feature vectors and target to identify the nearest hit (same class) and nearest miss (different class) for each instance. Using this information, feature weights are updated iteratively over a predefined number of iterations. Once the iterations are completed, all features are sorted in descending order based on their final weight scores. The features with higher weights are retained, while those with lower weights are removed, thereby enhancing prediction performance.

2.4 Fuzzified Preferential Random Subspace Ensemble Classification

The final step of proposed ROSIF-RASE model is to perform the student dropout prediction through the classification by using Fuzzified preferential Random subspace ensemble method. Random subspace ensemble is an ensemble machine learning technique that integrates numerous weak classifiers to achieve strong predictive performance. It operates by sequentially training each weak classifier, where each new model focuses more on the instances that were misclassified by previous ones. A weak classifier is only slightly better classification results, while a strong classifier accurately distinguishes between students as graduate or dropout.

The proposed model adopts this ensemble strategy to effectively identify the graduate or dropout students, thereby enhancing prediction accuracy. Contrary to other ensemble methods, the advantage of Random subspace ensemble includes high accuracy and robustness by training on smaller feature sets, and better performance on high-dimensional data.

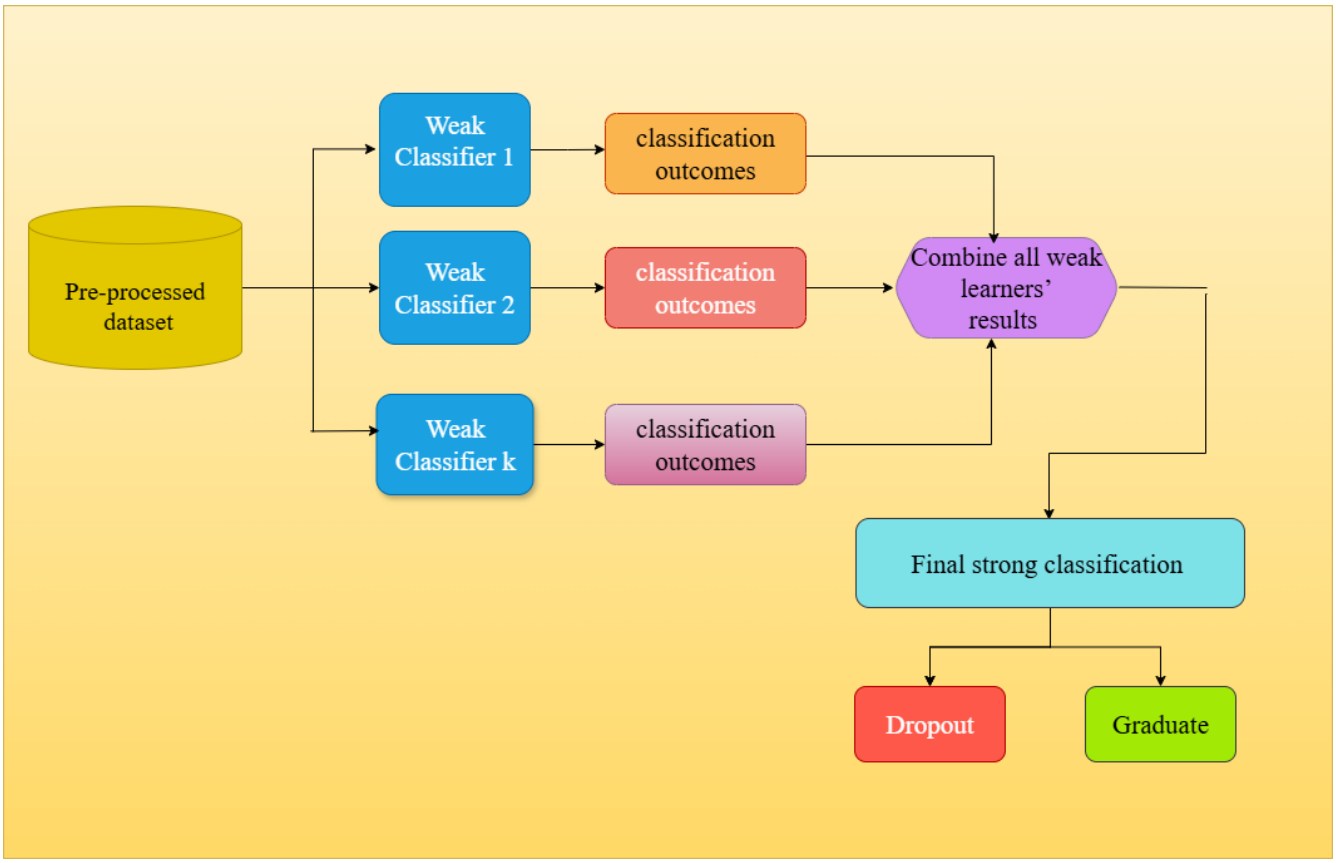

Fig 2. Schematic Construction of Fuzzified Preferential Random Subspace Ensemble

Figure 2 illustrates the schematic illustration of Fuzzified preferential Random subspace ensemble for prediction the student dropout or graduate with higher accuracy and minimum time consumption. The ensemble technique considers the training set $\{Ds_i, Y_i\}$ where Ds_i indicate the selected features with training student samples and Y_i indicate the ensemble classification output. First, the ensemble technique constructs 'k' number of weak learners $L_1, L_2, L_3, \ldots . L_k$ as decision stump classifier. A decision stump is a machine learning model, often described as a one-level decision tree. It makes a classification decision based on a single threshold or rule. The model selects the student data samples and best separates the data into different classes such as graduate, dropout and enrolled.

Let us consider the number of selected features $A = \{A_1, A_2, \ldots, A_b\}$ with student data samples $Ds_i = \{Ds_1, Ds_2, \ldots, Ds\}$. A Tamas indexive Decision Stump is a one-level decision tree consists of root node that directly connected to the leaf node.

Figure 3 shows the structure of the Tamas indexive Decision Stump classifier for accurate student dropout prediction. A decision stump is a simple form of a decision tree model that consists of a single decision node (the root) and leaf nodes. It is primarily used for classification tasks and represents a one-level decision tree, commonly referred to as a 'stump'. Each decision stump uses student data samples to split the data into different classes based on whether they satisfy a certain condition. In this context, the decision stump classifier employs Tamas index matching coefficient is a statistical technique used to evaluate the relationship between students' training and testing data samples. By applying matching coefficient, the root node classifies student data samples into three classes namely dropout, graduate and enrolled.

$$MC = 1 - \left(\frac{Ds_i \, \Delta \, Ds_t}{n} \right) \tag{12}$$

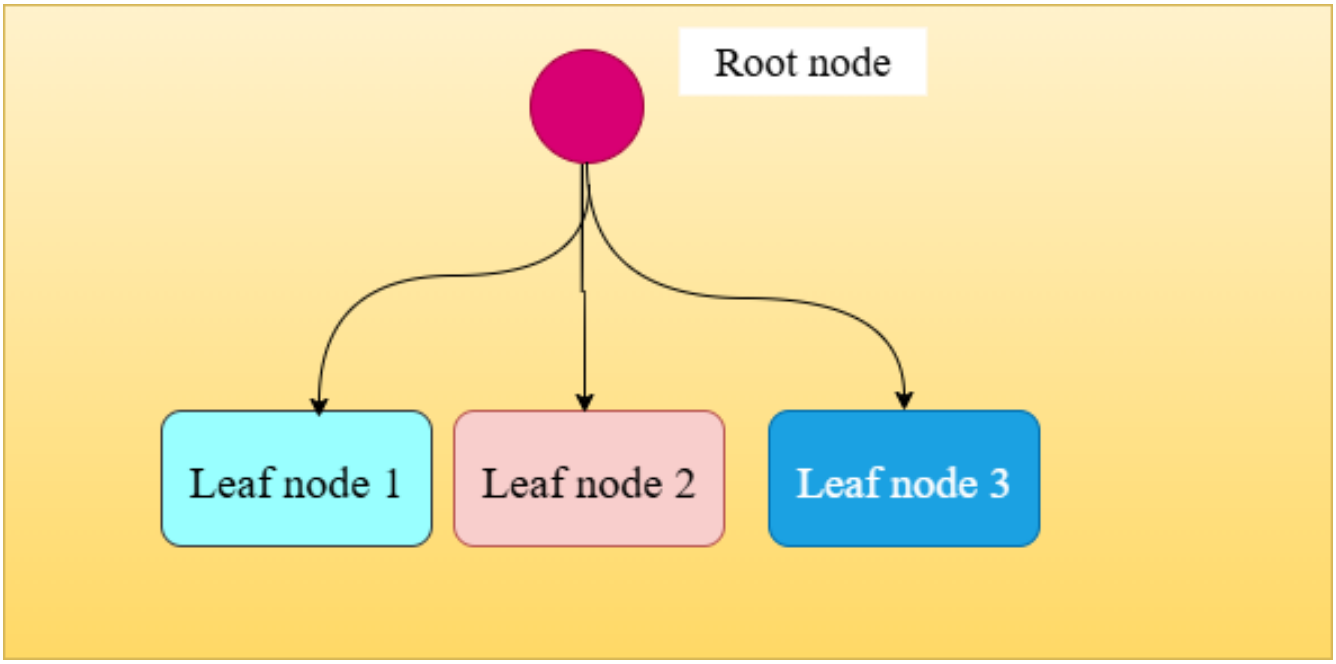

Fig 3. Tamas Indexive Decision Stump Based Classification

Where, MC indicates a matching coefficient,Ds_i Indicates the training data samples, Ds_trepresents a testing student data, $Ds_i \, \Delta \, Ds_t$ refers to a variance between the two data samples, 'n' indicates a total number of data samples. The coefficient 'MC' provides the output values remains the ranges from 0 to 1. Based on the outcomes, student dropout is predicted. However, the weak learner has some training error in the student dropout prediction. From the observed results, the classification error rate 'CE' is measured as follows

$$CE = \frac{number\ of\ misclassified\ samples}{Total\ number\ of\ samples} \tag{13}$$

In order to minimize the error, the weak learner results are combined and to create the strong output.

$$Y = \sum_{k=1}^{b} Z_k(L_k) \tag{14}$$

Where 'Y' indicates the output of a strong learner, $Z_k\,(L_k)$ indicates the output of the weak learners. Then the Fuzzy-weighted preferential voting method is used in ensemble learning to integrate the predictions from multiple weak classifiers. Preferential voting is a voting method that allows expressing multiple preferences rather than selecting just a single option. In this system, each preference is typically assigned a weight, indicating the level of support for each class. By using weighted preferences, the system gives a more accurate result, where the weights act as preference scores.

Initially, each classifier in the system assigns a vote to one or more possible class outcomes based on its analysis. By aggregating votes from multiple classifiers, the algorithm improves the accuracy and robustness of the final classification decision.

$$\vartheta_k \rightarrow Z_k(L_k) \tag{15}$$

Where, ϑ_k represents votes assigned to the classification results 'Z_k' of the base classifier 'L_k'. After that, assign the weights 'τ' based on voting count.

$$\tau = \left(\frac{\vartheta_{k_{class}}}{L_k} \right) \tag{16}$$

Where,L_k denotes a total number of classifiers, $\vartheta_{k_{class}}$ denotes a number of votes received by classifier. After computing the normalized weights for each class by means of the voting results from multiple classifiers, the next stage integrates fuzzy logic

to enhance decision-making in random subspace ensemble classifier. Triangular membership functions for each input value (weights) are computed as follows,

$$\mu_s(\tau) = \begin{cases} 0 & \tau < a \; or \; \tau \geq c \\ \dfrac{\tau - a}{b - a} & a < \tau \leq b \\ \dfrac{\tau - a}{b - a} & b < \tau < c \end{cases} \tag{17}$$

Where τ denotes a weight for each class, a, b, c denotes a parameter of the triangular base and peak. The fuzzy membership functions provide the output range [0, 1].

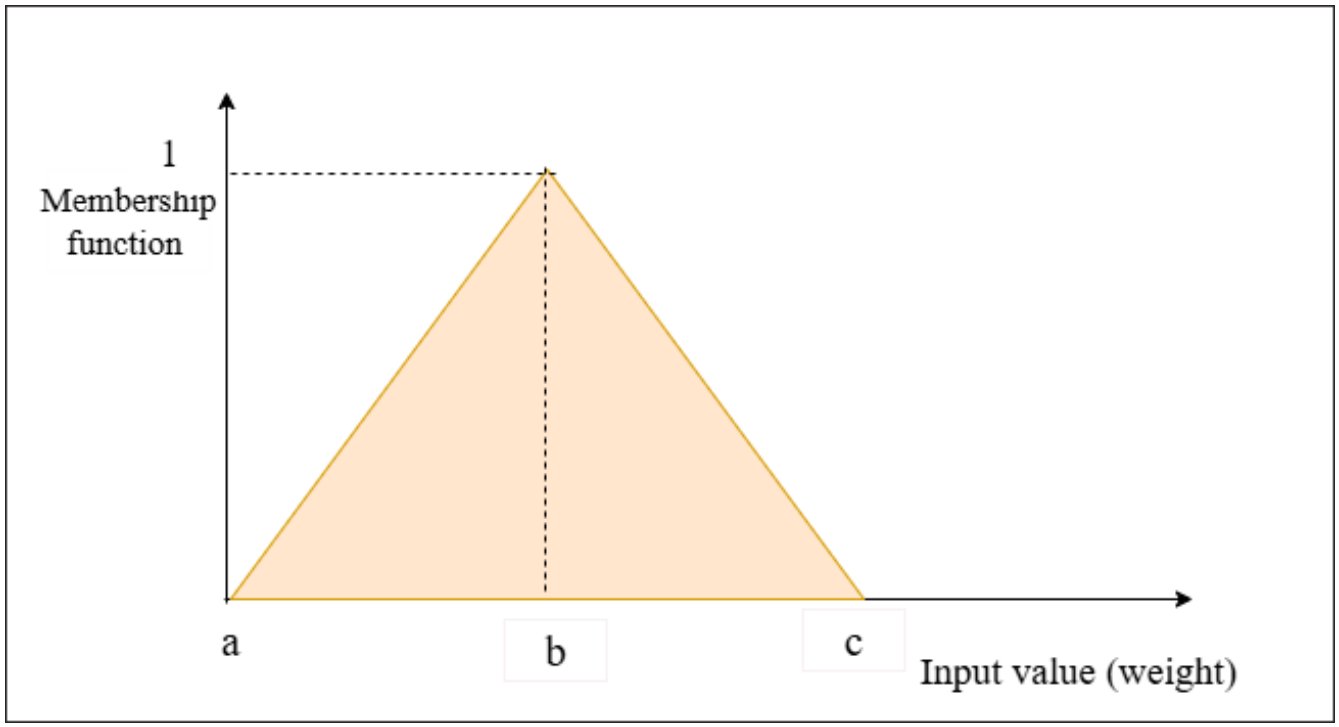

Fig 4. Fuzzy Triangular Membership Function

Figure 4 illustrates the triangular fuzzy membership function used in fuzzy logic systems. It is defined by three key parameters such as a, b, and c, which determine the shape of the triangle. The horizontal axis represents the input values or weights, while the vertical axis represents the membership degree, ranging from 0 to 1. The membership value starts at 0 when the input is equal to a, increases linearly to reach the maximum value of 1 at b, and then decreases linearly back to 0 at c. This simple and computationally efficient shape makes the triangular function ideal for representing fuzzy concepts. The function provides a smooth and interpretable way to map crisp input values into degrees of belonging, which is essential for fuzzy inference and decision-making processes. At this stage, the final classification decision is made by evaluating the fuzzy membership scores assigned to each class. The system compares these scores, and the class with the highest membership value is selected as the final output. This involves using a fuzzy membership function to assess the weight of each possible class. The class whose membership value increases a certain predefined threshold is finally chosen as the classification result.

$$Y = \arg\max \; \mu_s(\tau) \;\; i.e \;\; \mu_s(W) > T \tag{18}$$

Where, $\arg\max$ denotes an argument of maximum, $\mu_s(W)$ denotes fuzzy membership function of weight, T denotes a threshold. Finally, the classification result with the highest weight and maximum fuzzy membership value is accurately observed as output.

Algorithm 3: Fuzzified preferential Random subspace ensemble classifier
Input: selected features $A_1, A_2, \ldots A_r$, data samples or records $Ds_1, Ds_2, \ldots Ds_n$
output: Increase student dropout prediction accuracy
Begin
Step 1:foreach training data samples'Ds' with selected features
Step 2: Construct 'k' number of weak learners as decision tree
Step 3: Measure similarity between testing and training data samples using (12) in root node
Step 4: **If**$(MC = 1)$**then**
Step 5: The data samples are categorized into a particular class
Step 6: **else**
Step 7: The data samples are categorized into a another class
Step 8: **End if**
Step 9: **End for**
Step 10: Obtain weak learner results 'Z'
Step 11: For each L_k
Step 12: Calculate error 'CE using (13)
Step 13: Combine all weak learners $Y = \sum_{k=1}^{b} Z_k(L_k)$
Step 14: Apply the votes to weak learners using (15)
Step 15: Assign the weights based on vote counts using (16)
Step 16: End For
Step 17:For each weighted class
Step 18: Compute membership grade using (17)
Step 19: End for
Step 20: For each membership value
Step 21: **if** $(\mu_s(W) > T)$ **then**
Step 22: **Obtain the final** accurate classification results
Step 23: End if
Step 24: End for
Step 25: **Return** (student dropout, graduate, enrolled classification)
End

Algorithm 3 outlines the step-by-step procedure for predicting student dropout in higher education. Initially, the input training student data samples are fed into a subspace ensemble classification model. Within this ensemble, a weak learner known as the Tamas indexive Decision Stump classifier is employed to process both the training and testing data. The classification begins at the root node of the tree, where decisions are made based on similarity scores between data samples. If two samples exhibit high similarity, the model assigns a coefficient of '1', otherwise the coefficient returns '0'. This mechanism allows for accurate classification of the student samples into particular class either dropout class or graduate class. Following this, the outputs from all weak learners are aggregated. Each learner's training error is then calculated to impact on the overall model. To further improve the classification outcome, a preferential voting mechanism is applied, followed by the weight is computed based on the vote's counts for each class output. The triangular fuzzy membership function is utilized to determine which value exceeds a specified threshold. Among the many results, the class associated with the highest weight and the greatest membership degree is selected as the final classification output. The ensemble model then combines these outcomes to select the most reliable result with the minimal classification error. Based on this classification process, the algorithm successfully predicts student dropout with improved accuracy.

Experimental Setup: The ROSIF-RASE model is evaluated through experiments to assess its performance in predicting student dropout. The evaluation process includes several critical stages, including data collection, data preprocessing, feature selection, and classification. This analysis is performed using a Predict Students' Dropout and Academic Success dataset from the kaggle repository. The initial step involves collecting patient data samples, as shown in Figure 5.

Figure 6 illustrates the distribution of student academic success and dropout rates across three categories namely Graduate, Dropout, and Enrolled. The Graduate group, represented by the blue bar, has the highest number of students, with over 2,000 individuals successfully completing their studies. The Dropout category, shown in red, represents a moderate number of students who left their studies before completion. Finally, the Enrolled group, depicted in green, is the smallest, indicating

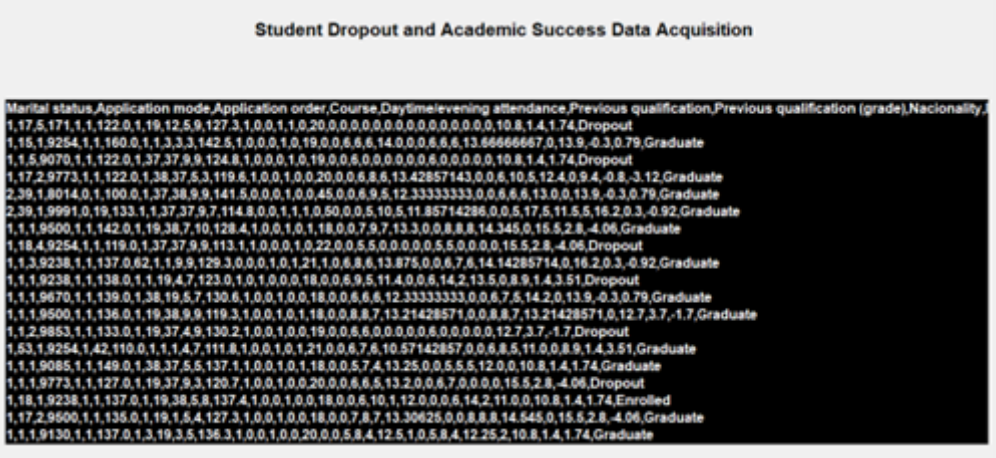

Fig 5. Sample Data Collection

the number of students still actively pursuing their education. This distribution highlights the relative success and challenges in student retention.

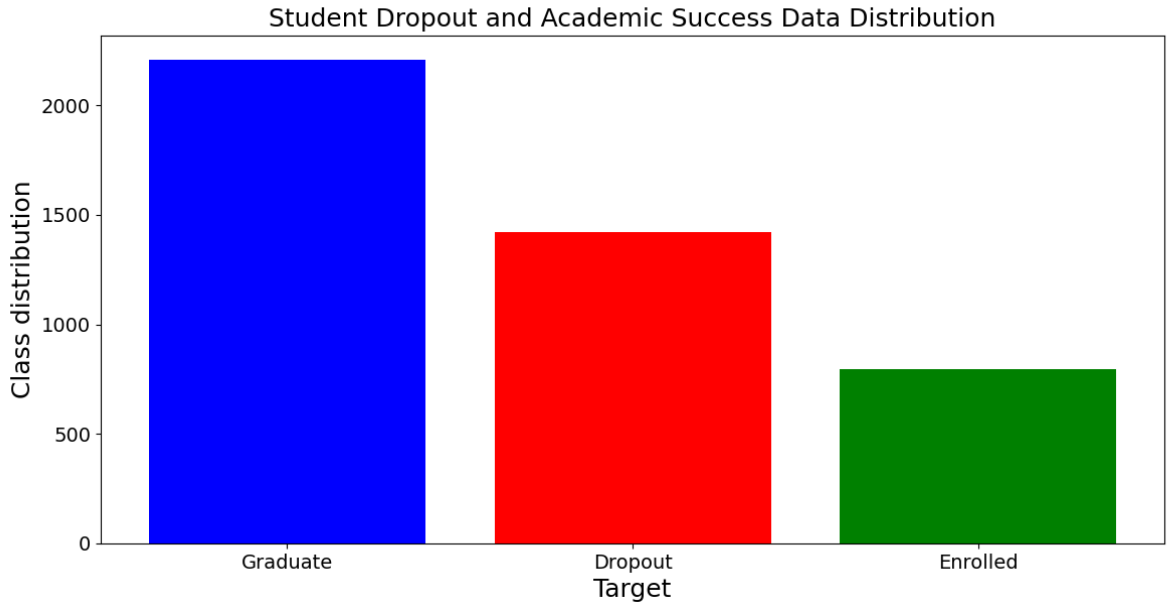

Fig 6. Class Distribution in Dataset

After data collection, data processing is carried out to handle missing data and outlier data removal for analysis. The original dataset had a size of 501 KB with number of instances is 4424. After completing the preprocessing steps, the dataset size was reduced to 482 KB with number of instances is 4259 due to the outlier data removal.

After preprocessing, identifying the most relevant and informative features from a dataset while eliminating redundant ones. The goal is to enhance model accuracy and lower computational time by minimizing the feature space. In this process, ROSIF-RASE model selects twenty one (21) features among 37 features in the given dataset.

Finally, the Classification task involves classifying the category or class label such as graduate, dropout and enrolled with selected features.

3 Results and Discussion

The evaluation of proposed ROSIF-RASE model is compared with different existing methods namely Ensemble **model [1], XG-AGbSCHO [2] and predictive** model [23] conducted on different parameters namely, student dropout prediction accuracy, precision, sensitivity, and student dropout prediction time.

Student dropout prediction accuracy: It is referred as the proportion of correctly predicted dropout cases by the algorithm relative to the total number of input student data samples. The accuracy for predicting dropout is mathematically expressed as follows:

$$SDPA = \frac{TP+TN}{TP+TN+FP+FN} * 100 \tag{19}$$

Where, '$SDPA$denotes student dropout prediction accuracy, 'TP' denotes atrue positive, TN denotes a true negative, 'FP' denotes a false positive, FN denotes a false negative. It is measured in terms of percentage (%).

Precision: It is determined by dividing the count of correctly identified positive cases (true positives) by the total number of cases predicted as positive, which includes both true positives and false positives. The precision metric is expressed as follows

$$PS = \left(\frac{TP}{TP+FP}\right) \tag{20}$$

Where, PS represents a precision, TP indicates the true positive, FP represents the false positive.

Sensitivity: It is also referred to as Recall, measures the ability of a model to correctly identify all actual positive cases. It is calculated as the ratio of true positive predictions to the total number of actual positive instances, which includes both true positives and false negatives.

$$ST = \left(\frac{TP}{TP+FN}\right) \tag{21}$$

Where, ST indicates a sensitivity, TPrepresents the true positive, FNdenotes the false negative.

F1-score: This metric combines both precision and recall into one value, offering a balanced assessment of model effectiveness.

$$F1 = 2 * \left(\frac{PS * ST}{PS+ST}\right) \tag{22}$$

Where,$F1$ denotes an F1score, PS indicates a precision, ST denotes a sensitivity.

Student dropout prediction time: It is measured as the amount of time consumed by the algorithm to predict the dropout students in the given input data samples. The prediction time is formulated as follows,

$$SDPT = \sum_{i=1}^{n} Ds_i * TME(SDP)] \tag{23}$$

Where '$SDPT$' a student dropout prediction time, n indicates the number of student data samples 'Ds', $TME(SDP)$ indicates a time for predicting the dropout with one data sample. It is measured in terms of milliseconds (ms). Table 1 compares student dropout prediction accuracy using three different methods namely the proposed ROSIF-RASE model and two existing approaches, Ensemble model [1], XG-AGbSCHO [2] and predictive model [23].

In the Table 1, the horizontal axis represents the number of student data samples, ranging from 440 to 4400, while the vertical axis indicates the student dropout prediction accuracyas a percentage. Among the three methods, the ROSIF-RASE modeldemonstrates a significant improvement in dropout prediction accuracy. For instance, with 440 data samples, the ROSIF-RASE modelachieved an accuracy of 94.31%, whiles the existing methods [1] and [2], [23] showed accuracies of 90.90% and 88.63%, 93.45% respectively. The overall comparison of results reveals that the ROSIF-RASE model outperforms the other models, with an accuracy improvement of approximately 4% compared to [1] and 7% compared to [2] and 1% compared to [23]. This enhancement is achieved due to the use of a Fuzzified preferential Random subspace ensemble Classifier with selected relevant features. These features with training samples are analyzed by constructing the weak classifiers' and determine the

Table 1. Comparison of student dropout prediction accuracy

Number of student data	Student dropout prediction accuracy (%)			
	Proposed ROSIF-RASE	Ensemble model [1]	XG-AGbSCHO [2]	Predictive model [23]
440	94.31	90.90	88.63	93.45
880	94.63	90.85	88.56	93.87
1320	94.45	90.75	88.46	93.66
1760	94.05	90.63	88.23	93.12
2200	93.89	89.89	88.74	92.90
2640	94.06	90.78	88.66	93.48
3080	94.78	90.12	88.44	93.75
3520	94.63	90.46	88.42	93.59
3960	94.87	90.96	88.56	93.91
4400	94.06	90.36	88.47	93.08

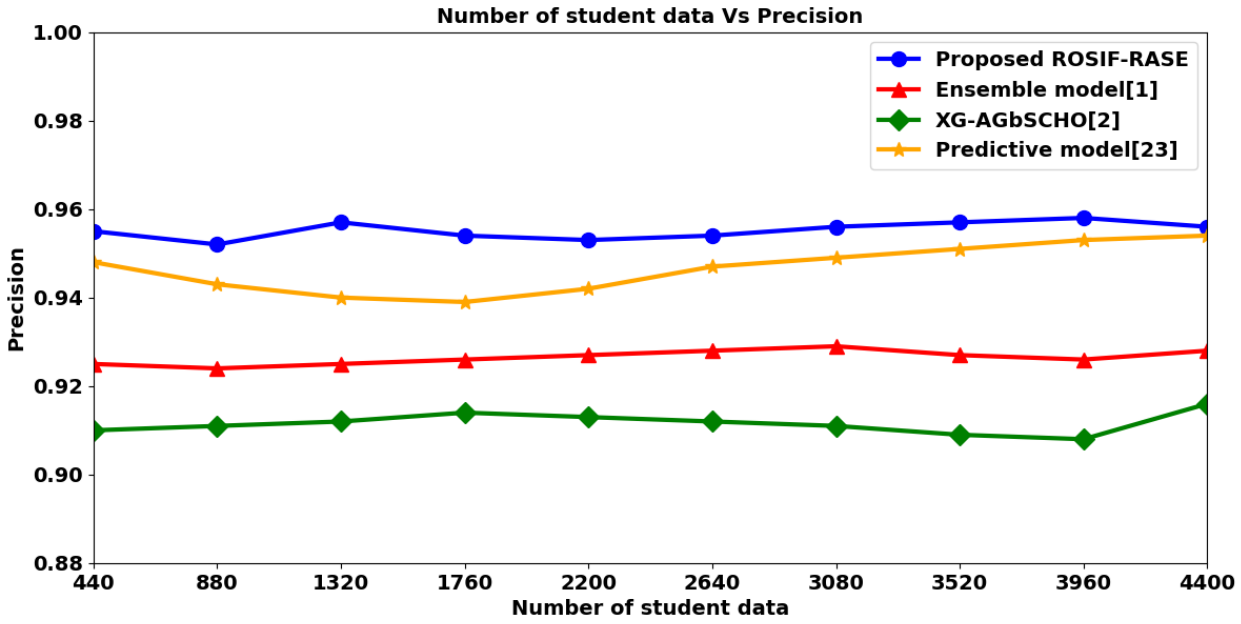

Fig 7. Comparison of Precision among different models

student dropout. The ensemble classifier apply preferential voting scheme and fuzzy logic for accurately classifying the student data samples which enhances the true positive and true negative rate, thereby reducing both false positives and false negatives.

Figure 7 presents comparison of precision against the number of student data samples, ranging from 440 to 4440. The performance of three methods namely ROSIF-RASE model and three existing approaches such as, Ensemble model [1] and XG-AGbSCHO [2] and predictive model [23] is evaluated for accurate student dropout predictions. The x-axis represents the number of student data samples collected, while the y-axis shows the precision performance. The results demonstrate that the ROSIF-RASE model outperforms the other two models in terms of precision. For example, when using 440 data samples, the precision of the ROSIF-RASE model was 0.955, 0.925 for [1] and 0.910 for [2] and 0.948 for [23]. The performance varied with different sample sizes, but the ROSIF-RASE model consistently showed superior results. Overall, the precision performance of ROSIF-RASE modelis increased by 3% higher than that of [1] and 5% higher than [2] and 1% higher than [23]. This improvement is achieved due to the utilization of a random subspace ensemble model, which integrates decision tree classifier analysis across the data samples. This approach enhances student dropout prediction by increasing the true positive rate and minimizing false positives.

Table 2. Comparison of Sensitivity

Number of student data	Sensitivity			
	Proposed ROSIF-RASE	Ensemble model [1]	XG-AGbSCHO [2]	Predictive model [23]
440	0.969	0.953	0.938	0.960
880	0.965	0.952	0.937	0.959
1320	0.964	0.951	0.936	0.957
1760	0.963	0.948	0.933	0.955
2200	0.966	0.95	0.937	0.958
2640	0.967	0.954	0.938	0.963
3080	0.965	0.953	0.937	0.961
3520	0.964	0.957	0.936	0.960
3960	0.967	0.958	0.935	0.962
4400	0.966	0.955	0.937	0.964

Table 2 illustrates the experimental analysis of sensitivity against the number of student data samples. From the graphical data, it is evident that the ROSIF-RASE model demonstrates a significant improvement in recall performance compared to the other three methods, Ensemble model [1] and XG-AGbSCHO [2] and predictive model [23]. For instance, when using 440 student data samples in the initial run, the sensitivity for the ROSIF-RASE model was 0.969 while for [1], [2] and [23], the sensitivity values were 0.953 and 0.938, 0.956 respectively. As the number of data samples increases, each method exhibits varying recall outcomes. A comparison of the results demonstrates that the ROSIF-RASE model outperforms [1] by 1% and [2] by 3% and [23] by 1% in terms of sensitivity. This improvement is accomplished due to ROSIF-RASE model ability to increase the true positive rate while reducing false negatives in the student dropout prediction process. This is achieved through the use of an ensemble classifier.

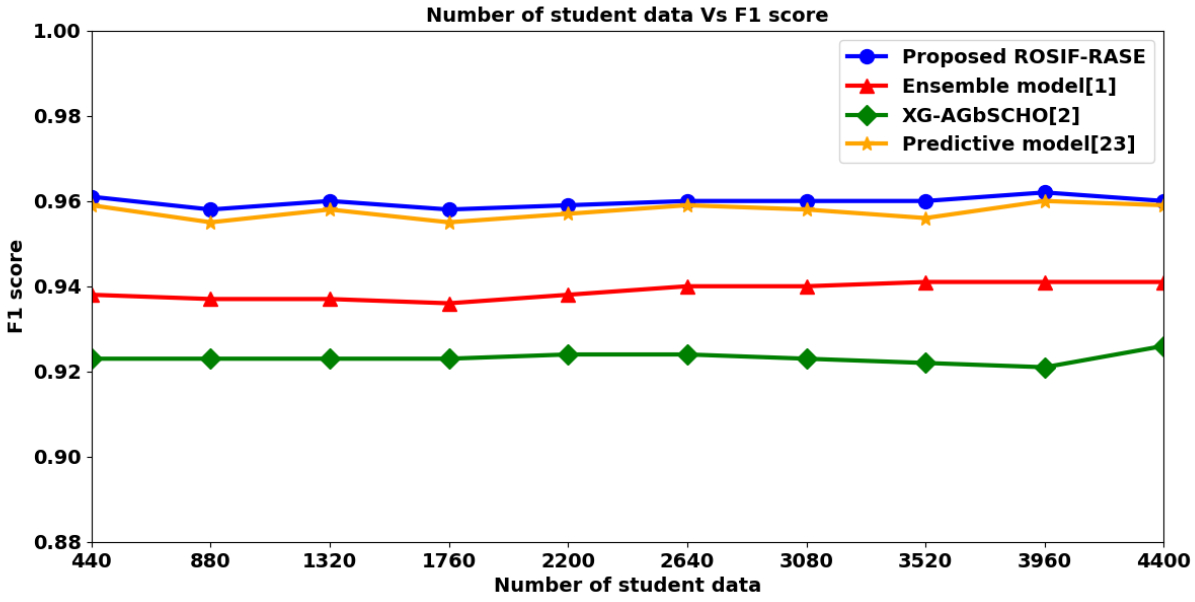

Fig 8. Comparison of F1 score among different models

Figure 8 illustrates the F1 score performance in predicting student dropout across varying data sample sizes, ranging from 440 to 4440. The x-axis represents the number of student data samples, while the y-axis shows the corresponding F1 score achieved by the three evaluated methods. The results indicate that the ROSIF-RASE model outperforms the existing methods Ensemble model [1] and XG-AGbSCHO [2] and predictive model [23] in terms of F1-score. This performance improvement of the ROSIF-RASE model is achieved due to its enhanced ability to balance both precision as well as recall (sensitivity), resulting in more

reliable and accurate predictions of student dropout. Finally, after ten rounds of comparison, the average results indicates that the ROSIF-RASE model demonstrated an increase in the F1 score by 2%, 4%, 1% compared to [1] and [2] and [23], respectively.

Table 3. Comparison of student dropout prediction time

Number of student data	Student dropout prediction time (ms)			
	Proposed ROSIF-RASE	Ensemble model [1]	XG-AGbSCHO [2]	Predictive model [23]
440	26.4	30.8	33	29.5
880	28.8	32.7	35.8	30.8
1320	30.2	33.8	37.8	33.7
1760	32.7	35.7	38.9	35.6
2200	34.5	36.9	40.5	37.4
2640	36.8	38.4	42.6	39.2
3080	38.9	40.2	43.8	39.5
3520	40.2	42.8	46.7	41.3
3960	42.8	45.9	49.8	44.6
4400	44.7	48.5	52.6	48.9

Table 3 presents the overall performance of student dropout prediction time versus the number of student data samples collected from the dataset. The prediction time for student dropout increases with the number of data samples across all three methods namely ROSIF-RASE model and two existing approaches, Ensemble model [1] and XG-AGbSCHO [2] and predictive model [23]. However, the ROSIF-RASE model shows a significant reduction in student dropout prediction time compared to the other three techniques. For example, when 440 data samples were used, the student dropout prediction time for the ROSIF-RASE model was 26.4ms, while 30.8ms consumed for [1] and 33ms consumed for [2] and 29.5ms consumed for [23]. This trend continued as the number of data samples increased, with ROSIF-RASE model consistently providing faster predictions. The results show that the ROSIF-RASE model method reduces prediction time by 8% compared to [1] and by 15% compared to [2] and 7% compared to [23]. This efficiency is achieved due to the ROSIF-RASE model integrated approach, which combines data preprocessing through the robust nonparametric estimator and statistical residual test. Additionally, the use of Fuzzified preferential Random subspace ensemble Classifier utilizes the selected features and removes others to further increases the prediction process while reducing the time complexity.

4 Conclusion

This study has proposed a novel ROSIF-RASE model for student dropout prediction in education data mining, incorporating a series of distinct processes. A novel process of data acquisition is utilized to collect raw student data samples taken from the dataset. Then, a novel preprocessing method to change the input data into a usable format and it comprises the two significant processes such as, imputing missing values and eliminating outliers. In missing data, the preprocessing utilizes robust nonparametric estimator to measure the missing values based on identified data points. For identifying and managing outliers, statistical residual test is determined for effectively removing from the dataset. The novelty of Statistical Inference feature re-weighting algorithm using feature selection process achieved to minimum student dropout prediction time. At last, the Fuzzified preferential Random subspace ensemble Classifier to enhance the stability and accuracy of student dropout prediction. A comprehensive set of experiments was conducted using various performance metrics, including accuracy, precision, sensitivity, F1 score and prediction time. The results show that the ROSIF-RASE model significantly enhances the 6%, 4%, 2% and 3% of accuracy, precision, sensitivity and F1 score, while also reducing 10% student dropout prediction time when compared to traditional approaches. The improving student dropout detection comprises the integrating richer data using advanced AI/ML for earlier & more accurate predictions. We focus on the future enhancement of student dropout detection in higher education is carried on integrating diverse data sources, developing interpretability method and equality as well evaluating the real-world impact of interventions rather than just predictive accuracy and throughput .

References

1) Rabelo AM, Zarate LE. A model for predicting dropout of higher education students. *Data Science and Management*. 2025;8(1):72–85. 10.1016/j.dsm.2024.07.001.
2) Radic G, Jovanovic L, Bacanin N, Stankovic MS, Simic V, Antonijevic M. Identifying and Understanding Student Dropouts Using Metaheuristic Optimized Classifiers and Explainable Artificial Intelligence Techniques. *IEEE Access*. 2024;12:122377–122400. 10.1109/ACCESS.2024.3446653.

3) Ridwana A, Priyatnob AM, Ningsih L. Predict Students' Dropout and Academic Success with XGBoost. *Journal of Education and Computer Applications*. 2024;1(2):1–8. 10.69693/jeca.v1i2.13.
4) Martínez J, Castillo D. Prediction of student dropout using Artificial Intelligence algorithms. *Procedia Computer Science*. 2024;251:764–770. 10.1016/j.procs.2024.11.182.
5) Pecuchova J, Drlik M. Enhancing the Early Student Dropout Prediction Model through Clustering Analysis of Students' Digital Traces. *IEEE Access*. 2024;12:159336–159367. 10.1109/ACCESS.2024.3486762.
6) Marcolino MR, Porto TR, Primo TT, Targino R, Ramos V, Queiroga EM, et al. Student dropout Prediction through machine learning optimization: insights from moodle log data. *Scientific Reports*. 2025;15:1–16. 10.1038/s41598-025-93918-1.
7) Psyridou M, Prezja F, Torppa M, Lerkkanen MK, Poikkeus AM, Vasalampi K. Machine learning predicts upper secondary education dropout as early as the end of primary school. *Scientific Reports*. 2024;14:1–14. 10.1038/s41598-024-63629-0.
8) Forti M, Bilancia M, Cafarelli B, del Gobbo E, Nigri A. Predicting dropout from higher education with multinomial finite mixture models: empirical evidence from Italy. *Quality & Quantity*. 2025. 10.1007/s11135-025-02135-5.
9) Mustofa S, Emon YR, Mamun SB, Akhy SA, Ahad MT. A novel AI-driven model for student dropout risk analysis with explainable AI insights. *Computers and Education: Artificial Intelligence*. 2025;8:1–15. 10.1016/j.caeai.2024.100352.
10) Ghosh P, Charit A, Banerjee H, Bandhu D, Ghosh A, Pal A, et al. DropWrap: A Neural Network Based Automated Model for Managing Student Dropout. *International Journal of Networked and Distributed Computing*. 2025;13:1–14. 10.1007/s44227-025-00058-z.
11) Jain A, Dubey AK, Khan S, Panwar A, Alkhatib M, Alshahrani AM. A PSO weighted ensemble framework with SMOTE balancing for student dropout prediction in smart education systems. *Scientific Reports*. 2025;15:1–28. 10.1038/s41598-025-97506-1.
12) Galve-Gonzalez C, Bernardo AB, Castro-Lopez A. Understanding the dynamics of college transitions between courses: Uncertainty associated with the decision to drop out studies among first and second year students. *European Journal of Psychology of Education*. 2024;39:959–978. 10.1007/s10212-023-00732-2.
13) Szabo L, Zsolnai A, Fehervari A. The relationship between student engagement and dropout risk in early adolescence. *International Journal of Educational Research Open*. 2024 Jun;6:1–11. 10.1016/j.ijedro.2024.100328.
14) Borges GA, Pedro CFDS, Anjos JCSD, Rodrigues A, Boavida F, Silva JS. A Platform for Early Class Dropout Prediction of University Students. *IEEE Access*. 2025;13:109116–109133. 10.1109/ACCESS.2025.3581751.
15) Cheng YH, Shih MH, Kuo CN. Early Prediction of Student Suspension and Dropout for Counseling Resource Optimization Using Multilayer Perceptron. *IEEE Access*. 2025;13:119810–119819. 10.1109/ACCESS.2025.3587484.
16) Kim H, Park Y. Calibrating F1 Scores for Fair Performance Comparison of Binary Classification Models With Application to Student Dropout Prediction. *IEEE Access*. 2025;13:136554–136567. 10.1109/ACCESS.2025.3594735.
17) Niu K, Cai J, Zhou Y, Tai W, Feng X. Hybrid neural network model for MOOC dropout prediction. *Complex System Modeling and Simulation*. 2025;p. 1–10. 10.23919/CSMS.2025.0003.
18) Alghamdi S, Soh B, Li A. ISELDP: An Enhanced Dropout Prediction Model Using a Stacked Ensemble Approach for In-Session Learning Platforms. *Electronics*. 2025;14(13):1–24. 10.3390/electronics14132568.
19) Noviandy TR, Zahriah Z, Yandri E, Jalil Z, Yusuf M, Yusof NISM, et al. Machine Learning for Early Detection of Dropout Risks and Academic Excellence: A Stacked Classifier Approach. *Journal of Educational Management and Learning*. 2024;2(1):28–34. 10.60084/jeml.v2i1.191.
20) Zanellati A, Zingaro SP, Gabbrielli M. Balancing Performance and Explainability in Academic Dropout Prediction. *IEEE Transactions on Learning Technologies*. 2024;17:2086–2099. 10.1109/TLT.2024.3425959.
21) Bouihi B, Bousselham A, Aoula E, Ennibras F, Deraoui A. Prediction of Higher Education Student Dropout based on Regularized Regression Models. *Engineering, Technology & Applied Science Research*. 2024;14(6):17811–17815. 10.48084/etasr.8644.
22) Liliana, Santosa PI, Hartanto R, Kusumawardani SS. Chi-Square Oversampling to Improve Dropout Prediction Performance in Massive Open Online Courses. *International Journal of Technology*. 2025;16(4):1220–1231. 10.14716/ijtech.v16i4.7047.
23) Olive U, Bosco MJ, Enan NM. Predicting Student Dropout in Higher Education: An Ensemble Learning Approach with Feature Importance Analysis. *Journal of Information and Technology*. 2025;5(4):31–40. 10.70619/vol5iss4pp31-40.
24) Rabelo AM, Zárate LE. A model for predicting dropout of higher education students. *Data Science and Management*. 2025;8(1):72–85. 10.1016/j.dsm.2024.07.001.
25) Rani N, Pachigolla V, Kumar A. Clustering based pre-processing for feature reduction and robust student dropout classification. *International Journal of Information Technology*. 2025;17:2001–2013. 10.1007/s41870-025-02450-y.