SvedbergOpen
DISSEMINATION OF KNOWLEDGE
International Journal of Artificial Intelligence and Machine Learning
Publisher's Home Page: https://www.svedbergopen.com/
Research Paper Open Access

Using Integrated Machine Learning Models On Educational Big Data To Predict Student Dropout Risk

Sahar Alamri[1], Faisal Alamri[2]

[1]Department of English Language at Prince Sattam bin Abdulaziz University, Alkharj, Saudi Arabia, sahar.al.amri@hotmail.com
[2]Department of Computer Science & Information Technology at Jubail Industrial College (JIC), Jubail, Saudi Arabia, amrifa@rcjy.edu.sa

Abstract
Student dropout remains a significant issue in education, impacting career opportunities and economic burdens. Despite predictive analytics' growing interest, current methods struggle to handle complex datasets effectively. This research addresses that gap by proposing an integrated Machine Learning (ML) framework for predicting student dropout risk using educational big data. The model utilizes a comprehensive dataset collected from institutional academic records, attendance logs, demographic profiles, and behavioral metrics curated to ensure representation of various influencing factors. Data preprocessing involves missing values (MVs) handling, normalization, and outlier removal using robust statistical techniques. For dimensionality reduction and effective feature extraction, Kernel Principal Component Analysis (Kernel PCA) is employed, capturing non-linear relationships within the data to enhance model interpretability and performance. The core novelty of this framework is the proposed Efficient Locust Swarm-tuned Dynamic Extreme Gradient Boosting Classifier (ELS-DXGBoostC**)**, which dynamically adjusts learning rates and tree parameters, optimized through nature-inspired Locust Swarm optimization to enhance predictive accuracy and generalization. Experimental results show superior performance in terms of accuracy (98%), recall (98.4%), precision (97.5%), and F1-score (97%) compared to traditional classifiers. This work integrates swarm intelligence with gradient boosting, enhancing the academic field and supporting institutions in deploying scalable, data-driven solutions to reduce student attrition.

Keywords: Student Dropout Prediction, Educational Big Data, Career Guidance, Efficient Locust Swarm-tuned Dynamic Extreme Gradient Boosting Classifier (ELS- DXGBoostC), Predictive Analytics.

Introduction

Dropouts are children who leave the educational system before completing their academic year. A dropout leaving college doesn't mean that the student failed to receive a final mark for that year or a record stating that the student had completed that year of high school or elementary school. Dropout is when students exit school without completing their education. Dropout Rate (DR) is a necessary statistic that represents the number of students who separated early from a school system [1]. It occurs at all educational levels and is a multifaceted issue due to numerous endogenous and external factors. Endogenous variables are students' biological characteristics, whereas exogenous variables are external influences such as natural disasters, family dynamics, and economic situations [2]. High DRs and delayed completion of higher education incur significant personal and social costs. Dropping out of higher education implies a cost to the government and society, a needless burden for families, and an experience of failure for the university student [3]. Dropout is the most serious problem that higher education institutions should address to attain success. It not only reflects students' struggles but also reveals institutional failures to provide necessary academic, social, and financial support. Addressing dropout necessitates comprehensive techniques that include early diagnosis, targeted intervention, and effective data-driven technology [4].

Student dropout (SD) is known as the greatest challenge and critical issue facing educational institutions. It has a negative influence on both the individual society and students. Individual students suffer as a result

of not receiving a degree, which limits their access to excellent services and lowers their income [5]. In the education industry, the Student Dropout Prediction (SDP) problem is an important ML application in which SD is predicted using predictive models. Effective SDP systems not only identify at-risk students early on but also assist institutions in implementing timely interventions to increase retention and academic performance [6]. Predicting the DR and students' academic success is critical in higher education, especially at universities. Dropout Prediction (DP) is a prediction problem whose performance is determined by the DR of the source data used for education [7].

SD and academic failure were predicted using AI-driven ML [8]. Class imbalance was handled utilizing Moodle activity data, the CatBoost algorithm, adaptive synthetic sampling, and hyperparameter optimization with NSGA-II. The model's average F1-score was about 0.8, indicating at-risk students. Limitations included a reliance on a single platform and an inadequate, imbalanced sample. To improve SD prediction accuracy, various data balancing approaches were utilized, such as Random Forest (RF), Multi-Layer Perceptron (MLP), and Logistic Regression (LR) [9]. It had the best classification result, whereas LR accurately identified the greatest number of dropout students. The generalizability to other circumstances remains restricted.

The high secondary school dropout (SD) rate was tackled by identifying the major factors that contribute to the problem by using predictive modeling techniques [10]. To enhance the accuracy of the predictions, optimal features, algorithms and hyperparameters were identified using an Automated Machine Learning (AutoML) approach. The results showed that Decision Tree (DT), K-Nearest Neighbors (KNN), Multilayer Perceptron (MLP) and Naïve Bayes (NB) models performed well. Yet, the method proved to be quite accurate, but lacked generalizability and relied on the quality of the data. In addition to the traditional methods such as DT, GB, RF, XGBoost and SVM, boosting algorithms like Light Gradient Boosting Machine (LightGBM) and Categorical Boosting (CatBoost) were used to predict the student dropout (SD) and academic achievement [11]. To enhance the performance of the model, Synthetic Minority Over-sampling Technique (SMOTE) resampling and Optuna based hyperparameter optimization were also added. CatBoost and LightGBM outperformed the other algorithms tested. However, because of the use of one dataset the generalizability of the results was limited, but the study offered a valuable guide for selecting algorithms for higher education dropout prediction.
To predict university SD for each academic year, universal feature tables were developed based on the historical data of university from a Korean university [12]. Several machine learning models were tested and the best one was Gradient Boosting for using mean value based feature tables. The findings highlighted the significance of students' historical academic data, but the data obtained from a single university did not have the ability to be applied to other educational institutions. Likewise, SVM, DT, MLP, and RF methods were used to build classifiers to predict tertiary SD with the use of chi-squared feature selection to boost the predictive accuracy [13]. RF was the most accurate and it was able to correctly classify students at risk of dropping out. The predictive capability was still limited, however, by the size and quality of the data sets, limiting the applicability of this approach at a broader scale.

Data from a Portuguese higher-education institution [14] were analyzed regarding the major socioeconomic and academic factors that impact SD rates. Least Absolute Shrinkage and Selection Operator (LASSO), Generalized Additive Model (GAM), Random Forest (RF), XGBoost and Neural Network (NN) models were investigated and best performance was found for the XGBoost model, which has the highest F1-score. The following variables were found to be important predictors: second semester grades, scholarship status, tuition status, number of units earned, loans, and age. The GAM analysis also suggested that the dropout rates could be lowered by about 22% if scholarships were given. These findings, however, had limited cross-institutional generalizability due to the institutional specificity of the dataset used.

Higher education students were studied using machine learning methods and the predictive models' recommendations were evaluated to be more generalizable for the SD [15]. The investigated strategies showed moderate prediction accuracy, but mainly focused on the students' past academic performance and failed to include significant factors of behavior and social factors that could influence student attrition. Likewise, a study for prediction of the dropouts in an online institution located in Seoul was conducted by combining demographic, academic, and Learning Management System (LMS) data [16]. Age, Grade Point Average (GPA), residential location, occupation, and LMS-related measures were significant predictors in the longitudinal, semester-by-semester and gender analyses. Logistic Regression (LR) showed higher interpretability by identifying specific gender-based characteristics that might guide targeted interventions and support for students, while LightGBM provided a higher predictive accuracy.

In 19 Brazilian schools, prediction models were developed to gain insight into and prevent SD [17]. The use of rich datasets and model-explanation methods yielded good predictive outcomes. However, interpretability and its focus on simple datasets limited the predictive accuracy of the study. To enhance the automatic SD detection in impoverished and developing countries, five feature-selection strategies were tested [18]. Following data preparation and feature engineering, multiple ML models were assessed. Top features selected improved the F-score by about 5%, while metaheuristic feature-selection methods improved precision, but in some cases resulted in a decrease in recall, and this could not be avoided due to sensitivity and noise reduction during dropout prediction (DP).

Student engagement and academic achievement data was used to create a binary classification framework for SD in online education [19]. The Canadian database, comprising 49 sociodemographic and behavioural features, was analyzed by Python, using XGBoost and LR models. The models were able to accurately predict the results of previously unseen data, but their ability to apply the same model to other educational settings and students was constrained.
A composite model, which is a combination of RF, XGBoost, GB and Feed-Forward Neural Network (FNN), was then predicted to further predict university SD [20]. The accuracy and the Area Under the Curve (AUC) achieved by the resulting ensemble approach were better than those of the individual base learners. Its added computational complexity, however, may preclude its use in real-time educational settings. A combination of demographic and academic data with pretrained language models (PTLM) has been used to enhance the performance of the problem of freshman dropout prediction (DP) [21]. Student data was transformed into natural language representations, and presented as a Natural Language Inference (NLI) task. Fine-tuned models using continuous hypotheses achieved improved results of up to 9% F1-score when compared to the existing methods, which proves their ability to capture relationships between structured and textual student information.

Lastly, university SD was predicted while at the same time identifying potential causes of student attrition [22]. A hybridized Student Dropout Prediction (SDP) system based on Principal Component Analysis (PCA), ML and K-means clustering was developed. The proposed system outperformed the previously developed systems in terms of precision, recall and F1-score. But its performance on imbalanced datasets and in various institutions is not yet known, suggesting the importance of more powerful and widely applicable dropout prediction systems.
SD remains a persistent issue in educational systems, with long-term personal and societal effects such as fewer career prospects and a greater financial burden. Despite increased interest in predictive analytics, many current systems fall short of managing large educational datasets and providing useful insights. These restrictions prevent early identification of at-risk students, preventing timely interventions that could improve retention and academic achievement. The research addresses this gap by offering an integrated ML strategy to predict Student Dropout Risk (SDR) utilizing educational big data.

Contribution of the research

- The research aims to predict SDR using educational big data, enabling early interventions and enhanced student retention.
- The method included extensive preprocessing (MVs handling, normalization, and outlier reduction) and kernel PCA for feature extraction.
- The hybrid ELS-DXGBoostC model combines Efficient Locust Swarm with Dynamic XGBoost, allowing for adaptive learning, hyperparameter change, and robust handling of high-dimensional and imbalanced data.
- The result achieved an improved accuracy, precision, and F1-score, demonstrating its effectiveness and generalizability.

The research provides valuable information for educational institutions, enabling administrators and policymakers to identify at-risk students early on and implement targeted interventions. It is also useful for those working on predictive analytics models in the educational sector.

Research Gap

Despite advances in ML for forecasting SD, numerous gaps exist. Various studies use single-institution datasets, which limit generalizability [8]. Regardless of resampling strategies, class imbalance continues to impact minority-class prediction. Few approaches combine demographic, socioeconomic, and behavioral data extensively [16], and the interpretability and causal analysis of dropout factors are limited [14]. The article

[15] is restricted by its emphasis on previous academic records, which ignores behavioral, social, and engagement-related characteristics that have a substantial impact on SDR. The work [12] focused on single datasets, which are frequently collected from a single institution or program, limiting the generalizability of the findings to larger student groups and various educational environments. Research [17] emphasized explainability above predictive accuracy, which means that while dropout variables were explained, the models failed to consistently identify all at-risk students. Furthermore, research [22] exhibited great performance only on specific datasets, raising questions about their resilience when applied to various educational situations. Hybrid ensembles and pretrained language models are emerging solutions for structured and textual data that have received no attention [21].

The conventional ML algorithms have proven effectiveness in the prediction of students' dropout problems in previous student dropout prediction studies, such as RF, SVM, GB, XGBoost, LightGBM and CatBoost algorithms [10-22]. But there are some fundamental methodological shortcomings. First, some existing studies rely on using traditional classifiers, automatic model selection or feature selection methods specific to the datasets, and may not adequately represent the complex nonlinear relations among a broad range of student characteristics [10–13]. Boosting based methods have been shown to have very high predictive performance, but are often evaluated on institution-specific or single source data sources and so only cover a limited range of student populations and learning settings [11,12,14,16]. Second, previous research has not incorporated consistently and cohesively all of the demographic, academic, attendance, engagement, behavioral, and psychological attributes within a single predictive framework. There are a number of approaches that focus mainly on academic records and history, and less on behavioral and social factors that could be important indicators of dropping out early [15,16]. Third, feature-selection and dimensionality-reduction methods have been explored, but traditional techniques might not be adequate for capturing nonlinear relationships between the diverse student attributes [18,22] . Moreover, ensemble models can enhance predictive accuracy, and may also add some computational complexity [20]. Optimizing the classifiers is another important area of deficiency. Although some approaches like Optuna and metaheuristic feature selection have been used in existing researches [11,18], very little research has attempted to combine an improved gradient-boosting classifier with powerful metaheuristic optimization of the features. However, a single framework that is able to learn compact nonlinear representations while simultaneously optimising the parameters of classifiers is still poorly studied. To overcome these drawbacks, the proposed Efficient Locust Swarm-tuned Dynamic Extreme Gradient Boosting Classifier (ELS-DXGBoostC) uses Kernel Principal Component Analysis (Kernel PCA) to represent the features in a nonlinear manner, Dynamic Extreme Gradient Boosting (Dynamic XGBoost) to perform robust classification, and Efficient Locust Swarm (ELS) optimization to select the parameters of the model effectively. The general purpose of the framework is to enhance predictability, generalization, and binary classification of the students at risk for dropout with reliability.

Methodology

The proposed ELS-DXGBoostC framework is a systematic framework for predicting student drop-out risk by incorporating various data processing stages: data preprocessing, nonlinear feature extraction, optimized classification, and binary risk prediction. The Kaggle student dataset is initially prepared for enhancing the quality and consistency of the data. KNN imputation uses similar student records to fill in missing attribute values and z-score normalization is used to normalize a heterogeneous set of numerical attributes so that they are on a comparable scale. Additionally, IQR-based analysis can identify outliers that can distort the learning of the model, and inappropriate outliers can be removed. The attributes of the demographic, academic, attendance, engagement, behavior, and psychological attributes are then subjected to Kernel Principal Component Analysis (Kernel PCA) after preprocessing. Ker PCA preserves the non-linear relationships between the student attributes and transforms the original features to a non-linear space to create compact and informative representations of the features. The representations that are formed are passed on to the Dynamic Extreme Gradient Boosting Classifier (DXGBoostC). The enhanced boosting model is a sequential combination of decision trees that learn from the errors produced by the previous tree. It dynamically adjusts its learning rate, feature interactions and regularization which leads to higher adaptability, predictive stability and overfitting resistance. The Efficient Locust Swarm (ELS) optimizer then explores the following hyperparameter space of DXGBoostC by selective solitary and social search behaviors. The optimum configuration is chosen based on the validation performance. The optimized model then learns from the training data, is fine-tuned with the validation data, and evaluated by testing data it has never seen. Finally, the framework yields two risk groups: a 0 indicates non-dropout or continued studies, and a 1 indicates dropout. The framework therefore facilitates early identification of

students that may need prompt academic or institutional action, the entire workflow of the ELS-DXGBoostC is shown in Fig. 1.

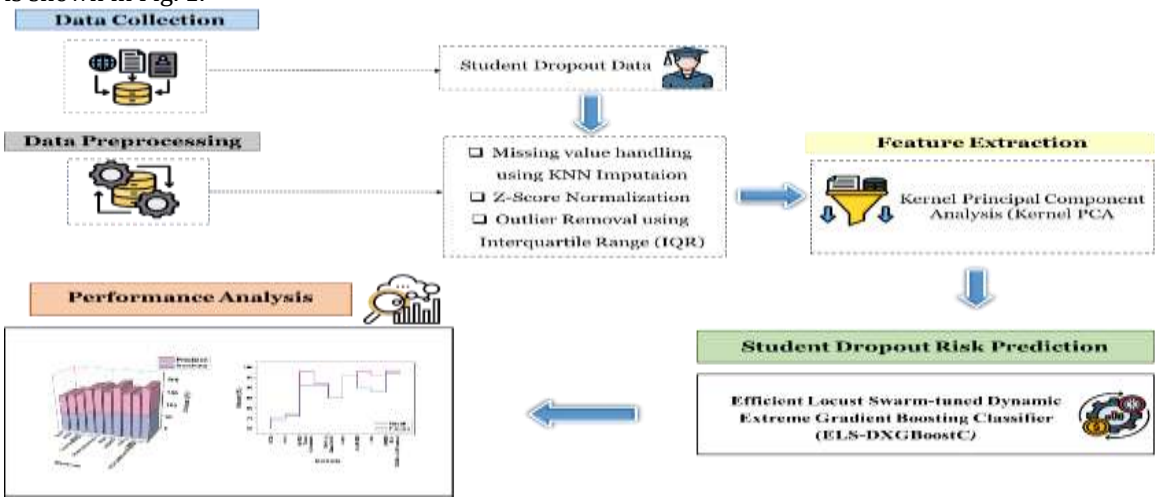


Fig.1. Overview of the ELS-DXGBoostC framework for predicting SDR

Data Collection

The Student Dropout Prediction Dataset [23] is a Kaggle-sourced dataset developed for machine-learning research on identifying students at risk of discontinuing their studies. It contains 5,000 unique student records and more than 25 attributes, integrating demographic, academic, attendance, engagement, behavioral, and psychological characteristics. Demographic variables include age, gender, socioeconomic status, parental education, family background, and first-generation status, while academic attributes comprise GPA, previous academic performance, subject-failure count, credits completed, course load, and scholarship status. Attendance and engagement are represented through attendance rate, class participation, assignment and quiz completion, learning-platform usage, login frequency, and engagement score. Behavioral factors include learning style, motivation, peer interaction, counseling sessions, and stress level. The target variable, dropout_status, is binary, where 0 indicates continued study and 1 indicates dropout. Those features are described in Table 1. The dataset is divided into 80% training (4,000 records), 10% validation (500 records), and 10% testing (500 records). The training subset is used to learn predictive patterns, the validation subset supports model tuning and hyperparameter selection, and the independent testing subset evaluates generalization performance. Thus, the dataset provides a multidimensional foundation for developing and evaluating robust student dropout prediction models.

Table 2. Description and Classification of Input and Target Features in the Student Dropout Prediction Dataset

No.	Feature	Description
1	student id	Unique identification code assigned to each student record; used for identification rather than predictive learning.
2	age	Student's age, representing demographic characteristics that may influence academic persistence.
3	gender	Student's gender category, representing demographic differences potentially associated with educational outcomes.
4	socioeconomic status	Indicates the student's socioeconomic condition, reflecting financial and social circumstances affecting continuation of studies.
5	Parental education level	Represents parents' highest educational attainment and provides information about family educational background.
6	Family background	Describes the student's family structure, such as nuclear or joint family, which may influence educational support.
7	First generation student	Indicates whether the student is the first member of their family to attend higher education.
8	Grade Point Average (GPA)	Grade Point Average representing the student's current academic achievement and performance level.

9	Previous academic performance	Represents the student's previous academic performance category before the current academic period.
10	Subject failure count	Number of subjects previously failed by the student, indicating academic difficulty.
11	Credits completed	Number of academic credits successfully completed by the student.
12	Course load	Represents the number or intensity of courses undertaken during the academic period.
13	Scholarship status	Indicates whether the student receives a scholarship, reflecting financial and academic support.
14	Attendance rate	Percentage of classes attended by the student, indicating academic participation and consistency.
15	Class participation score	Measures the student's active participation and involvement during classroom activities.
16	Assignment submission rate	Percentage of assignments submitted by the student, indicating academic responsibility and engagement.
17	Quiz completion rate	Percentage of assigned quizzes completed, representing continuous academic participation.
18	Time spent on learning platform	Measures the amount of time spent using online learning platforms for academic activities.
19	Login frequency	Indicates how frequently the student accesses the online learning platform.
20	Engagement score	Overall measure of the student's academic and digital engagement with learning activities.
21	Learning style	Represents the student's predominant learning style, such as visual, auditory, or mixed.
22	Motivation level	Measures the student's motivation toward learning and continuation of academic activities.
23	Peer interaction score	Quantifies interaction with classmates and peers, representing social and collaborative engagement.
24	Counselling sessions attended	Number of counseling sessions attended, indicating utilization of institutional support services.
25	Stress level	Measures the student's perceived stress level, which may contribute to academic disengagement and dropout risk.
26	Dropout status	Final prediction variable: 0 = continued studies (non-dropout) 1 = dropped out.

Data Pre-processing

Data pre-processing is a vital stage in handling raw educational big data, as it transforms unstructured and inconsistent information into a clean and usable format for significant outcomes and decision-making. To improve SDRP's dependability and accuracy, it is necessary to handle MVs, normalize data, and reduce noise.

- **Missing Values Handling using KNN imputation**

Missing Values (MVs) handling fills the incomplete data points to ensure a complete and reliable dataset, improving the accuracy and robustness of SDR prediction. The research uses KNN Imputation [24], which is a technique that estimates MVs in datasets using the values of the k most comparable (nearest) data points. It is used to fill in MVs in educational big data (such as attendance, grades, and demographic information) to ensure a comprehensive dataset for predicting SDR. The imputed value for a missing feature (y_{ji}) in instance Y is calculated as in Eq. (1).

$$y_j^{(imputed)} = \frac{1}{k}\sum_{i\epsilon KNN(Y)} y_{ji} \quad (1)$$

Where $y_j^{(imputed)}$ is the imputed value of the missing feature j, k is the number of nearest neighbors examined, $KNN(Y)$ is the set of k nearest neighbors to the instance Y (based on a distance metric, such as Euclidean distance), and y_{ji} is the value of feature j in neighbor i.

- **Normalization using Z-score normalization**

Normalization is a data processing technique for rescaling features to a common scale while preserving value range variance. The research employs Z-score normalization [25] to normalize data by scaling it to have a mean of 0 and a standard deviation of 1, ensuring fair comparison across features. Eq. (2) provides the formula for normalizing the Z-score.

$$z(jk) = \frac{b(jk)-\mu}{\alpha} \quad (2)$$

The input value's column mean is denoted as z, the new value as is $z(jk)$, the old value as is $b(jk)$, the mean as is μ, and the column standard deviation is given by α, $j = 1-n$ (where n indicates the number of rows or inputs), and $k = 1-l$ (where l indicates the number of bits of each input, which is 32 bits).

- **Outlier Removal using Interquartile Range (IQR)**

Outlier elimination is the process of detecting and eliminating abnormal or extreme values that impair data analysis and model quality. The research employs IQR [26] to detect and remove extreme outliers from large educational data, resulting in cleaner input for accurate SDRP. It is a statistical method for identifying outliers that computes the difference between a dataset's 25th (Q_1) and 75th (Q_3) percentiles. Eq. (3) provides the value for IQR.

$$IQR = Q_3 - Q_1 \quad (3)$$

Feature Extraction

Feature extraction is the method of transforming raw data into a set of meaningful variables that represent the most essential patterns for analysis. It aids in the reduction of high-dimensional educational big data into relevant features, which improves both efficiency and prediction accuracy in SDR. The research proposed Kernel PCA [27] for feature extraction.

PCA is a linear dimensionality reduction technique that turns correlated data into uncorrelated principal components to capture the most variance. Its application is to simplify high-dimensional educational data for models such as ELS-DXGBoostC, but its disadvantage is that it does not capture non-linear interactions between attributes. To solve these issues, the research used Kernel PCA. Kernel PCA is a form of PCA that uses kernel functions to detect nonlinear relationships between features. It reduces the dimensionality of vast volumes of educational data while preserving complex patterns in student performance, attendance, and behavioral characteristics. It achieves this by transferring the original data $y_j \in R^d$ into a higher dimensional feature space F through a non-linear transformation as described in Eq. (4).

$$z_j = \phi(y_j) \quad (4)$$

Where z_j is the converted representation of student j in the new feature space F, and $\phi(\cdot)$ is the non-linear transformation function. Eq. (5) defines a kernel function that is utilized to efficiently compute the similarity between the data points in F.

$$k(y_j, y_k) = \langle \phi(y_j), \phi(y_k) \rangle \quad (5)$$

Where $k(y_j, y_k)$ is the kernel function used here to measure the similarity between students j and k. The inner product in feature space is given by the notation $\langle \phi(y_j), \phi(y_k) \rangle$. Polynomial and radial basis function (RBF) kernels are common kernels. Using Eq. (6), it projects a new data point y onto the k^{th} principal component.

$$PC_k(y) = \sum_{j=1}^{n} \alpha_j^{(k)} k(y_j, y) \quad (6)$$

For the new data point y, $PC_k(y)$ is the value of the k^{th} principal component, $\alpha_j^{(k)}$ is the j^{th} coefficient of the eigenvector $\alpha^{(k)}$, and $k(y_j, y)$ is the kernel similarity between the new sample y and the training sample y_j.

Student Dropout Risk Prediction using Efficient Locust Swarm-tuned Dynamic Extreme Gradient Boosting Classifier (ELS-DXGBoostC)

ELS-DXGBoostC is a hybrid ML model that combines DXGBoostC, a dynamic variation of XGBoost designed for high-dimensional and imbalanced data, with ELS optimization to fine-tune hyperparameters. It improves the accuracy and robustness of SDR prediction by dynamically changing learning parameters and generalizing across varied educational datasets.

Dynamic Extreme Gradient Boosting Classifier (DXGBoostC)

XGBoost [28] is a method of learning a powerful algorithm that combines numerous weak learners to generate a robust future model. However, its shortcomings include sensitivity to noisy data, overfitting in complicated datasets, and reliance on stable hyperparameters, which restrict its usefulness in educational data analytics. To address these issues, the researchers propose DXGBoostC, which dynamically modifies learning rates and regularization parameters. This model improves robustness and adaptability, making it suitable for managing high-dimensional and unbalanced educational data.

DXGBoostC significantly improves on the original boosting architecture to better match the forecast objectives of SDR prediction. These advancements include adaptive learning rate, enhanced feature interaction modeling, and structural regularization, as explained below. DXGBoostC's target function aims to reduce prediction error for predicting tasks, as indicated in Eq. (7).

$$M(\theta) = \sum_{j=1}^{N} m\left(z_j, \hat{z}_j(\theta)\right) + \Omega(\theta) \quad (7)$$

Where $M(\theta)$ represents the total loss, $m\left(z_j, \hat{z}_j(\theta)\right)$ is the cross-entropy loss between the actual behavior (z_j) and the expected result $\hat{z}_j(\theta)$, θ is the model's parameters, and the regularization term to avoid overfitting is given by $\Omega(\theta)$. Each unique data sample in the dataset is denoted by the index j, where N signifies the total number of samples. As shown in Eq. (8), the model structure fits the residuals of earlier models using an ensemble of DTs.

$$T(y, \beta) = \sum_{l=1}^{L} f_l\,(y, \beta_l) \quad (8)$$

Where f_l is the l^{th} regression tree, L denotes the total number of the regression trees, β_l represents the structure and parameters of each tree, and $T(y, \beta)$ is the ensemble prediction. DXGBoostC determines feature relevance using a gain-based metric, as shown in Eq. (9) to assess each feature's contribution to SDR prediction.

$$Gain(S, f) = \frac{1}{|S|}\sum_{(y_j, z_j)\in S}\left[z_j - T\left(y_j, \beta\right)\right]^2 - \left(\frac{|S_M|}{|S|}H(S_M, f) + \frac{|S_Q|}{|S|}H\left(S_Q, f\right)\right) \quad (9)$$

Where S is the training sample set, S_M and S_Q are data divided by feature f, H is the sum of squared residuals, and $\left(y_j, z_j\right)$ denotes an input-output pair in S, where y_j is the input and z_j is the corresponding goal value. DXGBoostC employs an adaptive learning rate described by Eq. (10).

$$\eta_t = \eta \cdot \frac{1}{t+1} \quad (10)$$

Where the iteration count is t and the initial learning rate is η. Eq. (11) defines regularization, which is added to penalize overly complex models.

$$\Omega(\theta) = \lambda_1 \cdot \|\theta\|_1 + \lambda_2 \cdot \|\theta\|_2^2 \quad (11)$$

Where the regularization parameters are denoted by λ_1 and λ_2. The tree structure itself is optimized as follows to further regularize the model in Eq. (12).

$$F(y, \Theta) = \sum_{k=1}^{K} \gamma_k \cdot J\left(d_k(y)\right) \quad (12)$$

Where $J\left(d_k(y)\right)$ is an indicator function that yields 1, and J denotes the number of terminal nodes. $F(y, \Theta)$ is the tree ensemble's output on input y, and γ_k are the coefficients corresponding to each terminal node k if y falls into terminal node j. The regularization component includes penalties for the weights and the tree structure, as indicated in Eq. (13).

$$\Omega(\Theta) = \sum_{k=1}^{K}(\alpha|\gamma_k| \cdot \lambda \cdot \gamma_k^2) + \mu \cdot K \quad (13)$$

The regularization parameters for the $L2$ norm, $L1$ norm, and tree complexity are denoted by $\lambda, \alpha, and\ \mu$, respectively. K is the total number of leaf nodes in the DT, and $\Omega(\Theta)$ is the overall regularization term applied to the model parameters Θ to avoid overfitting. For parameter updates, the DXGBoostC model computes gradients and Hessians, as shown in Eqs. (14) and (15).

$$H_j = \partial \hat{z}_j l\left(z_j, \hat{z}_j\right) \quad (14)$$

$$G_j = \partial^2 \hat{z}_j l\left(z_j, \hat{z}_j\right) \quad (15)$$

Where $l\left(z_j, \hat{z}_j\right)$ is the loss function for the actual value z_j and the predicted value $\hat{z}_j$, and H_j and G_j stand for the gradient and Hessian, respectively. Additionally, it incorporates feature interaction modeling to capture relationships between different dropout factors, as shown by Eq. (16).

$$Interaction_{f.h(y)} = \Theta_{f.h} \cdot y_f \cdot y_h \quad (16)$$

Where y_f, y_h are the feature values, $\Theta_{f.h}$ is a parameter matrix specifying the strength and direction of the interaction, and $Interaction_{f.h}$ is the interaction effect between features f and h. Finally, Eq. (17)

illustrates how the model allows for multi-class classification to differentiate between various types of SDRs such as academic, financial, or personal factors.

$$P(z = l|y) = \frac{exp(\eta_l(y))}{\sum_{l'} exp(\eta_{l'}(y))} \quad (17)$$

Where $\eta_l(y)$ are class-specific decision functions modeled by the DXGBoostC method, and $P(z = l|y)$ reflects the likelihood of an SDR category l (such as academic failure, financial stress, or behavioral disengagement) given the input features y, the total of the exponentials of the decision functions across all potential classes l' is $\sum_{l'} exp(\eta_{l'}(y))$.

Efficient Locust Swarm (ELS)

Locust Swarm Optimization (LSO) [29] is a bio-inspired algorithm that uses the swarming and cooperative movement of locusts to tackle complicated optimization problems. Its drawbacks include the possibility of premature convergence, sensitivity to parameter changes, and the high computational cost of large-scale datasets. To overcome these issues, the research proposed ELS, which is a nature-inspired optimization technique that uses locust swarms' collective foraging and movement characteristics to efficiently explore and exploit search spaces. It is used to fine-tune the DXGBoostC model's hyperparameters, resulting in faster convergence, reduced overfitting, and higher accuracy in SDR prediction.

ELS introduces a probabilistic strategy that selects between solitary and gregarious behavior at each iteration. Unlike the original LS, it modifies the social operator by substituting many local evaluations with a likelihood based selection procedure. This permits locusts to be directed toward more potential people, which improves convergence efficiency.

- **Selecting between the Stages of Solitary and Social**

Individuals in ELS move between solitary and social stages according to the stage of the search procedure. Early iterations prioritize investigation through solitary activity, but later stages concentrate on exploitation through social connections. As the algorithm improves, social behavior becomes more prominent, reinforcing the balance between exploration and exploitation. In such cases, for each iteration k, the behavioral stage (L^k) is utilized to the population Lk is chosen as follows in Eq. (18).

$$P(L^k) = \begin{cases} Solitary & if\ rand \leq p^k \\ Social & if\ rand > p^k \end{cases} \quad (18)$$

The $rand$ signifies a random number taken from the consistently distributed interval [0,1], where value p^k, known as the probability of behaviour as determined by Eq. (19).

$$p^k = 1 - \frac{k}{itern} \quad (19)$$

where $itern$ represents the total number of iterations that are considered in the search process. It is understandable that the behavior probability p^k decreases linearly as the iterative procedure lengthens.

- **Modified Operator for the Social Stage**

During this stage, each individual l_j^k moves towards a potentially viable solution instead of doing several local evaluations. A subset of the q best solutions, $C^k = \{c_1^k, \ldots, c_q^k\} \subseteq L^k = \{l_1^k, l_2^k, \ldots, l_N^k\}$, is chosen from the population. Individuals l_j^k are probabilistically attracted to a randomly chosen solution $c_j^k \in C^k$, lowering computing effort and leveraging community knowledge. The possibility that any given solution c_j^k is selected by a particular individual l_j^k is determined by Eq. (20) and is dependent on the quality of c_j^k as well as the distance between the two individuals.

$$P_{l_j,c_j}^k = \frac{B(c_j^k)e^{-\|l_j^k - c_j^k\|}}{\sum_{n=1}^{q} B(c_n^k)e^{-\|l_j^k - c_j^k\|}} \quad (20)$$

Where a member f from these best solutions is given by $B^k(c_i^k)$, and $\|l_j^k - c_j^k\|$ denotes the Euclidian distance between the individual "j" (l_j^k), while $B(c_i^k)$ indicates the attractiveness of solution c_j^k as determined by Eq. (21).

$$B(c_j^k) = \frac{f(c_i^k) - f_{woest}(C^k)}{f_{best}(C^k) - f_{woest}(C^k) + \varepsilon} \quad (21)$$

Where the fitness value (quality) associated with c_i^k is indicated by $f(c_i^k)$, and the worst and best fitness value surrounded by the members of the C^k is represented by $f_{woest}(C^k)$ and $f_{best}(C^k)$. To avoid a division by zero, ε is a small value that usually falls between $1.0 \times 10 - 4$ and $1.0 \times 10 - 5$. Eq. (22) illustrates how each individual l_j^k within the swarm population L^k changes their location in ELS's social phase behavior.

$$l_j^{k+1} = l_j^k + 2(c_r^k - l_j^k) \cdot rand \quad (22)$$

Where l_j^k is the current position (solution vector) of the j^{th} individual (locust) in the swarm at iteration k, l_j^{k+1} *is* the updated position of the j^{th} individual for the next iteration, $rand$ is a uniformly distributed random number in [0,1], c_r^k is a randomly selected solution $c_j^k \in C^k$. Table 1 depicts the hyperparametr tuning of ELS-DXGBoostC. The pseudocode for the ELS-DXGBoostC is represented in Algorithm 2.

Table 2. Hyperparameter Tuning of DXGBoostC

Hyperparameter	Typical Range
Learning Rate (η)	0.01 – 0.3
Subsample	0.5 – 1
Max Depth	3 – 10
Number of Trees (L)	50 – 500
Colsample_bytree	0.5 – 1
$\lambda 1$	0 – 1
$\lambda 2$	0 – 1
γ_k	0 – 5
Interaction Constraints	Custom subsets
p^k	0 – 1
Population Size (N)	10 – 50
$itern$	50 – 200

Algorithm 1: Efficient Locust Swarm-tuned Dynamic Extreme Gradient Boosting Classifier (ELS-DXGBoostC)

Input:
$L\hat{}k = \{l_1^k, l_2^k, \ldots, l_N^k\}$ // *Population of candidate solutions*
itern // *Total number of iterations*
DXGBoostC model // *Base classifier*
Output: *Optimized DXGBoostC parameters for student dropout risk prediction*
Step 1: *Initialize population* L^k *randomly*
Step 2: *For iteration* $k = 1$ *to itern do*:
Step 3: *For each individual* l_j^k *in* L^k *do*:
Step 4: ***DXGBoostC***
Compute total loss $M(\theta)$ (*Equation* 7)
Fit residuals using ensemble of decision trees $T(y, \beta)$ (*Equation* 8)
Determine feature relevance $Gain(S, f)$ (*Equation* 9)
Update adaptive learning rate η_t (*Equation* 10)
Apply regularization $\Omega(\theta)$ (*Equation* 11)
Optimize tree structure $F(y, \Theta)$ (*Equation* 12)
Update overall regularization $\Omega(\Theta)$ (*Equation* 13)
Compute gradients H_j (*Equation* 14) *and Hessians* G_j (*Equation* 15)
Model feature interactions $Interaction\_(f, h(y))$ (*Equation* 16)
Compute multi $-$ *class probabilities* $P(z = l|y)$ (*Equation* 17)
Evaluate fitness of l_j^k *for student dropout risk prediction*
Step 5: *ELS Optimization*
Select behavioral phase $P(L^k)$ (*Equation* 18)
Compute behavior probability p^k (*Equation* 19)

$Generate\ rand\ \in\ [0,1]$
$If\ Social\ Phase\ is\ selected\ then$:
$Select\ subset\ of\ q\ best\ solutions\ C^k\ \subseteq\ L^k$
$Compute\ selection\ probability\ P^k_{l_j,c_j}\ (Equation\ 20)$
$Compute\ attractiveness\ B\left(c_j^k\right)\ (Equation\ 21)$
$Randomly\ select\ c_r^k\ \in C^k\ based\ on\ probability$
$Update\ position\ l_j^{k+1}\ =\ l_j^k\ +\ 2\left(c_r^k\ -\ l_j^k\right)*\ rand\ (Equation\ 22)$
$Else$
$Perform\ solitary\ movement\ as\ in\ original\ LS$
Step 6: $Update\ best\ solutions\ C^k$
$End\ For\ (iteration\ k)$
Step 7: $Return\ optimized\ DXGBoostC\ parameters$

The advantage of ELS-DXGBoostC is its capacity to handle high-dimensional and imbalanced educational data with greater accuracy and robustness. It reduces overfitting, improves feature interaction modeling, and responds dynamically to complicated patterns. It provides a dependable methodology for accurately predicting SDR.

Result

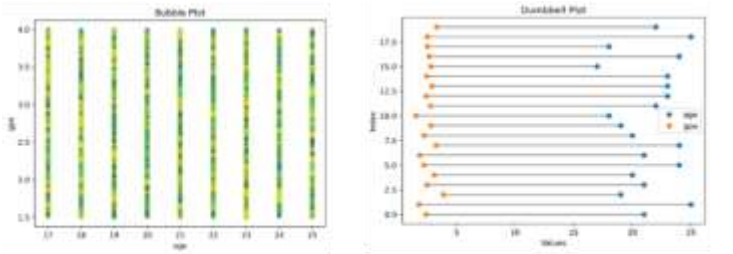
The research aims to predict SDR using educational big data. This section outlines the experimental setup, performance evaluation and comparison phase that are used for predicting SDR. The experiments used Python 3.10, 16GB RAM, and an Intel i7 processor. Educational datasets were stored and processed in CSV format, enabling fast processing of high-dimensional data for model training and evaluation.

Performance Evaluations

The research uses performance metrics such as analysis of Feature Relationships and Concentrations, Student Attributes, Data Structure and Trends by using the ELS-DXGBoostC method and dataset.

- **Analysis of Student Attributes**

Visual representations were created using age and GPA as key attributes to better understand the characteristics of students in the dropout dataset. These characteristics are among the most influential demographic and academic factors associated with dropout risk. Visualizations aid in the discovery of patterns, distributions, and correlations in data. These findings provide context for the subsequent predictive modeling. Fig. 2. (a) depicts the distribution of students' GPA across various age groups in the dataset. Fig. 2. (b) compares the age and GPA values of sampled students using paired points connected by lines.



(a) Bubble Plot of Age vs. GPA (b) Dumbbell Plot of Age and GPA

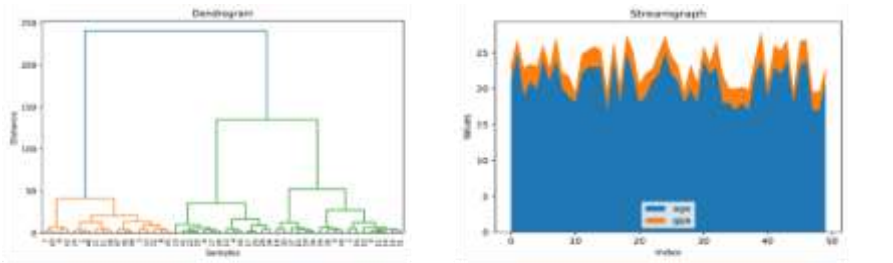
Fig.2. Visualization of Student Attributes (a) Bubble Plot and (b) Dumbbell Plot

Fig. 2 depicts each point as a student, with bubble sizes and colors representing changes in data density. It shows the concentration of students with GPAs ranging from 2.0 to 4.0 between the ages of 17 and 25, providing insight into academic performance trends important to DP. It showed that no single age group leads in terms of performance, implying that dropout risk could not be explained entirely by demographics. It also shows the range of GPA values within each age group, which is beneficial for detecting outliers or at-risk students during early intervention. Orange points represent GPA and blue points represent age. The distance between the two paired points represents the difference in demographic (age) and academic (GPA) characteristics. This type of visualization helps illustrate the balance and distribution of variables

affecting students, which are used in the associated SDR predictive modeling. Also, the dumbbell shape illustrates the multi-dimensional nature of the dataset since both the academic and demographic variables are considered at the same time. Additionally, it nicely summarizes the paired features, making it easier to notice trends or differences that indicate the likelihood of dropout.

➢ **Data Structure and Trends Analysis**

The assessment of data structure and trends provides more information about the links between student factors such as age and GPA. Clustering and stream-based representations reveal patterns of similarity and variance among students. These visualizations not only reveal hidden groupings but also track changes in academic achievement over time. Such insights are critical for understanding the factors influencing SDR and improving forecast accuracy in the proposed approach. Fig. 3. (a) illustrates the Dendrogram of Student Samples. Fig. 3. (b) depicts the Streamgraph of Age and GPA Trends.



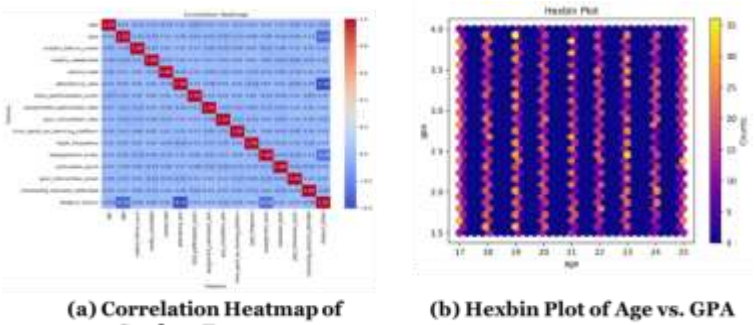
(a) Dendrogram of Student Samples (b) Streamgraph of Age and GPA Trends

Fig.3. Visualization of (a) Data Structure and (b) Trends

In Fig.3, the vertical axis reflects distance (or dissimilarity), while the horizontal axis depicts individual student IDs. The branching structure shows how students are grouped together based on age and GPA. Furthermore, the clear distinction between clusters suggests underlying patterns that correspond to dropout risk groups. The image also shows how clustering reveals hidden links between students that are not immediately apparent in raw data. The blue stream indicates age distribution, and the orange stream depicts GPA variation. The layered design allows for simultaneous adjustments in academic and demographic parameters. Furthermore, the dynamic shape of the GPA stream reveals irregular performance fluctuations across students, indicating areas of instability that contribute to dropout risk. It also shows the connection of demographic and academic developments, which reinforces the need for integrated predictive modeling.

➢ **Analysis of Feature Relationships and Concentrations**

The visualization of feature correlations and concentrations reveals dependencies between student qualities that determine dropout risk. The correlation heatmap reveals how academic, behavioral, and demographic characteristics interact, highlighting significant predictors like attendance rate and GPA. In contrast, the hexbin plot emphasizes the density of student distributions, making it simpler to identify concentration trends across age and GPA. Together, these visualizations limit the risk of misinterpretation caused by overplotting while also highlighting significant areas of academic diversity. Such investigations provide the foundation for developing effective and dependable DP models. Fig. 4. (a) depicts the Correlation Heatmap of Student Features. Fig. 4. (b) represents the Hexbin Plot of Age vs. GPA.



(a) Correlation Heatmap of Student Features (b) Hexbin Plot of Age vs. GPA

Fig. 4. Visualization of (a) Feature Relationships and (b) Concentrations

GPA and dropout status show a modestly negative relationship in Fig.4, but attendance rate and credits completed have stronger negative correlations, indicating their importance as predictors. Weak or almost nonexistent correlations between a number of variables are emphasized in the graph, demonstrating that each characteristic contributes to the DP model independently. It also stresses the importance of dimensionality reduction approaches in capturing complicated, nonlinear connections between features. The majority of students are grouped between the ages of 18 and 23 with GPAs between 2.0 and 3.5, as shown by Fig.4, where brighter hexagons denote areas with larger concentrations. Age groups with greater academic diversity are also revealed by the pattern, which is important for determining possible dropout risks. Compared to conventional scatter plots, this density-based approach also minimizes overplotting, which facilitates the detection of concentration patterns.

Comparison Phase

The research uses performance metrics such as accuracy, f1-score, recall and precision and existing methods such as DT [29], RF [29], ANN [29], Tuned XGBoost [30], Stacking Classifier [30], Naïve Bayes (NB) [31], Hybrid LR and NN (HLRNN) [31], LR [32] and SVM [32] for predicting the SDR. Table 2 depicts the performance comparison of different ML methods for SDR prediction.

Table 2. Comparative Performance Evaluation of the Proposed ELS-DXGBoostC Model against Existing Classification Methods

Methods	Accuracy (%)	Precision (%)	Recall (%)	F-score (%)
DT[32]	70.1	70	70.1	70
RF[32]	75.5	73.9	75.5	74.2
ANN[32]	77.3	76	77.3	76.2
Tuned XGBoost[33]	89	86	98	91
Stacking Classifier[33]	89	90	92	91
NB[34]	86	85	85	85
HLRNN[34]	96	96	96	96
LR[35]	93	79	98	90
SVM[35]	92	79	96	88
ELS- DXGBoostC[Proposed]	**98**	**97.5**	**98.4**	**97**

❖ **Precision**

Precision measures the proportion of accurately anticipated dropouts among all instances defined as dropouts. It is critical to reduce false alarms and verify that identified unprotected students actually need care. High precision helps educational institutions allocate resources efficiently to the students who genuinely require support.

❖ **Accuracy**

Accuracy refers to the percentage of accurately predicted cases (including dropouts and non-dropouts) out of total predictions. It indicates the overall performance of the ELS-DXGBoostC model in predicting SDR. It provides a quick overview of model reliability across the entire dataset.

❖ **F1-Score**

The F1-score is the harmonic mean of *recall* and *precision*, allowing for False Positives (FP) and False Negatives (FN). It provides a complete measure of model performance, particularly on imbalanced datasets with fewer dropouts. It is particularly valuable in educational datasets where the cost of misclassifying at-risk students has significant consequences.

❖ **Recall**

Recall calculates the fraction of actual dropouts properly detected by the model. High recall is required to ensure that unprotected students are identified and supported before dropping out. It also ensures that few at-risk students are overlooked, which is critical for proactive retention strategies. Fig.5 depicts the comparison of performance metrics for various ML methods used in SDR prediction.

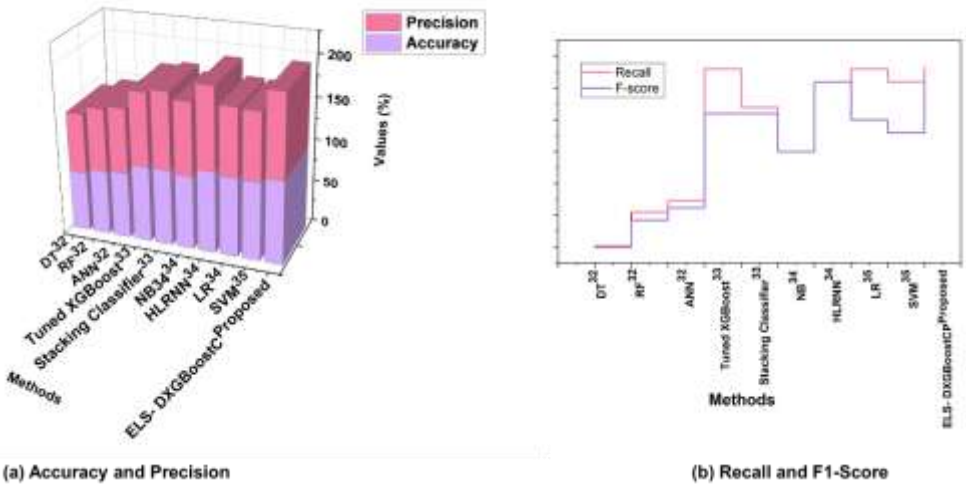


Fig.5. Comparison of performance metrics for various ML methods used in SDR prediction

The DT [32] achieved a precision of 70%, RF [32] obtained 73.9%, and ANN [32] achieved 76%, indicating moderate reliability in avoiding FP. Tuned XGBoost [33] achieved 86%, while the Stacking Classifier [33] achieved 90%, indicating better prediction of actual at-risk students. The accuracy of NB [34], HLRNN [34], LR [35], and SVM [35] was 85%, 96%, 79% and 79%, respectively. The ELS-DXGBoostC achieved a precision of 97.5%, indicating that almost all students identified as dropouts were genuinely at risk, which is critical for efficiently targeting interventions and eliminating superfluous activities. DT [32] attained an accuracy of 70.1%, suggesting moderate prediction ability, whereas RF [32] improved to 75.5% and ANN [32] to 77.3%, demonstrating incremental gains in total correctness. Tuned XGBoost [33] and Stacking Classifier [33] both scored 89%, indicating high accuracy. NB [34] scored 86%, HLRNN [34] 96%, LR [35] 93%, and SVM [35] 92%. The ELS-DXGBoostC model obtained an accuracy of 98%, outperforming all baseline models and properly categorizing nearly all students' dropout statuses, demonstrating outstanding stability across varied educational datasets. DT [32] had an F1-score of 70%, RF [32] had 74.2%, and ANN [32] had 76.2%, indicating moderate overall performance. Tuned XGBoost [33] and Stacking Classifier [33] both achieved 91%, indicating an excellent balance between FP and missed dropouts. The scores for NB [34], HLRNN [34], LR [35], and SVM [35] were 85%, 96%, 90%, and 88%, respectively. The ELS-DXGBoostC model received an F1-score of 97%, demonstrating higher overall reliability by keeping high sensitivity while decreasing false alarms, making it the most effective model for predicting SDR using educational big data. DT [32] obtained a recall of 70.1%, RF [32] obtained 75.5%, and ANN [32] obtained 77.3%, implying that many at-risk students were ignored. Tuned XGBoost [33] reached 98%, while Stacking Classifier [33] achieved 92%, demonstrating strong detection capabilities. NB [34] achieved 85% accuracy, whereas HLRNN [34] reached 96%, LR [35] reached 98%, and SVM [35] obtained 96%. The ELS-DXGBoostC has a recall of 98.4%, capturing nearly all actual dropouts. High recall guarantees that students at risk receive timely intervention, lowering DRs and enhancing educational retention across a variety of learning situations.

Discussion

The existing methods for SDR prediction exhibit several limitations when applied to large-scale and complex educational datasets. DT [32] is simple and easy to understand, but it was unstable, as slight changes in data could result in significantly different trees. RF [32] improved generalization through ensembling, but it is computationally expensive for large datasets and less effective when features are strongly correlated. Nonlinear dependencies were captured using ANN [32], but it requires a lot of training data, is difficult to understand, and is capable of overfitting. While advanced approaches such as tuned XGBoost [33] improve prediction performance, students are still sensitive to hyperparameter choices and risk overfitting if not properly calibrated. Similarly, stacking classifiers [33] incorporated many models but increased computing complexity and decreased transparency, making real-world application for education challenging. NB [34] assumes feature independence, which typically does not represent real student data, resulting in less reliable predictions. HLRNN [34] accurately captures temporal dynamics, but it is resource-intensive and challenging to scale across institutions with diverse data. LR [35] is extensively used due to its simplicity, but it struggles with nonlinear interactions and necessitates thorough preprocessing of input features. SVMs [35] are powerful for classification but more sensitive to kernel selection, parameter tuning, and

computationally expensive with large datasets. To address these issues, the ELS-DXGBoostC combines dynamic learning, adaptive regularization, and efficient locust swarm-based hyperparameter tuning, resulting in increased robustness against noisy data, scalability to high-dimensional features, and improved generalization across imbalanced educational datasets. This makes the technique extremely useful for educators, politicians, and institutions, as it provides practical insights for early interventions, reduction of DR, and student retention initiatives.

The implementation of the proposed ELS-DXGBoostC-based student dropout prediction framework involves several practical challenges. Educational datasets may contain missing, inconsistent, noisy, and heterogeneous information across academic, attendance, demographic, engagement, and behavioral attributes, making reliable preprocessing essential. KNN imputation and IQR-based outlier handling must be carefully implemented to avoid introducing bias or removing legitimate student variations, while Kernel PCA requires appropriate kernel and component selection to preserve meaningful nonlinear relationships. Another challenge is the computational cost of ELS hyperparameter optimization, particularly when repeatedly training DXGBoostC models during the search process. In addition, class imbalance, data quality differences between institutions, privacy of student information, and the generalization of a model trained on one educational environment to another can affect real-world deployment. Despite these challenges, the framework has several practical applications. It can support early identification of dropout-prone students, enabling institutions to provide targeted academic support, attendance monitoring, counseling, mentoring, financial assistance, and engagement interventions. Administrators can also use predicted dropout risk to allocate support resources more efficiently, monitor student retention patterns, and develop evidence-based retention strategies. Thus, ELS-DXGBoostC can function as a decision-support mechanism for proactive student-retention management rather than merely predicting whether a student will drop out.

Future Research Directions

The proposed ELS-DXGBoostC framework is useful for predicting student dropouts in the binary classification of students, but there are several lines of further research that can be taken to make the framework more applicable and robust. Future studies should test the framework with several sets of real-world data from various educational systems, regions and universities. The SOD-FE study showed good performance in both the Portuguese and Slovakian datasets, highlighting the need for cross-dataset performance assessments to evaluate the robustness and generalizability of a model [29]. Thus, future researches can examine if ELS-DXGBoostC is robust to changes in institutional policy, student population, academic organization, and socioeconomic factors.

Second, future work can utilize student information longitudinally and temporally, not just static student attributes. Previous studies based on administrative data have found that demographic factors like citizenship, age or residence can account for dropout prediction [30]. Further development of the model, for semester-wise or year-wise student records, might allow the model to detect evolving dropout risk and offer earlier intervention opportunities.Third, the framework can be expanded to include dynamic data from the behavioural and learning-platform, such as sequential attendance, log-in frequency, assignment activity, assessment attempts, engagement trajectories etc. In the MST-GCN study, they showed that it is possible to increase the accuracy of dropout prediction in Massive Open Online Courses (MOOCs) by fusing the short-term behavioral changes with the accumulated long-term patterns [31]. The incorporation of these temporal features with the proposed nonlinear feature extraction and optimized boosting architecture may help to enhance sensitive detection of students' quick changes in behavior.Finally, future research can focus on explainable and intervention-oriented prediction by incorporating Explainable Artificial Intelligence (XAI) techniques, notably SHapley Additive exPlanations (SHAP), which, successfully applied to improve the transparency of machine learning models in SOD-FE [29], could be incorporated into the system. Future iterations of the system might then be able to not only predict which students are at risk of dropping out but also give explanations for the risk and recommend interventions for each student, making the system ethically, transparently and institutionally helpful in early-warning systems.

Conclusion

The research proposed ELS-DXGBoostC framework for predicting SDR using educational big data. SDR prediction data was gathered from academic records, attendance logs, demographic profiles, and behavioral measures to ensure that all impacting elements were included. The data was preprocessed using advanced statistical approaches to handle MVs, normalize features, and remove outliers. Kernel PCA was used to reduce dimensionality and extract features, capturing non-linear correlations and improving model interpretability. ELS-DXGBoostC was used, with learning rates and tree characteristics being continuously modified and optimized. The experimental results outperformed standard classifiers in terms of accuracy (98%), recall (98.4%), precision (97.5%), and F1-score (97%), showing robustness and flexibility across a

wide range of educational situations. The research was limited by its dependence on institutional data, which have overlooked external social and behavioral aspects that influence dropout. Future research could combine many datasets and test real-time intervention tactics using improved AI models.

References

1. Hassan, M. A., Muse, A. H., & Nadarajah, S. Predicting student dropout rates using supervised machine learning: Insights from the 2022 National Education Accessibility Survey in Somaliland. Appl. Sci. 14, 7593 (2024). https://doi.org/10.3390/app14177593
2. Gonzalez-Nucamendi, A., Noguez, J., Neri, L., Robledo-Rella, V., & García-Castelán, R. M. G. Predictive analytics study to determine undergraduate students at risk of dropout. Front. Educ. 8, 1244686 (2023). https://doi.org/10.3389/feduc.2023.1244686
3. Alvarado-Uribe, J., Mejía-Almada, P., Masetto Herrera, A. L., Molontay, R., Hilliger, I., Hegde, V., ... & Ceballos, H. G. Student dataset from Tecnologico de Monterrey in Mexico to predict dropout in higher education. Data 7, 119 (2022). https://doi.org/10.57687/FK2/PWJRSJ
4. Realinho, V., Machado, J., Baptista, L., & Martins, M. V. Predicting student dropout and academic success. Data 7, 146 (2022). https://doi.org/10.5281/zenodo.5777339
5. Cho, C. H., Yu, Y. W., & Kim, H. G. A study on dropout prediction for university students using machine learning. Appl. Sci. 13, 12004 (2023). https://doi.org/10.3390/app132112004
6. Masood, S. W., Gogoi, M., & Begum, S. A. Optimised SMOTE-based imbalanced learning for student dropout prediction. Arab. J. Sci. Eng. 50, 7165–7179 (2025). https://doi.org/10.1007/s13369-024-09287-w
7. Pecuchova, J., & Drlik, M. Enhancing the early student dropout prediction model through clustering analysis of students' digital traces. IEEE Access (2024). https://doi.org/10.1109/ACCESS.2024.3486762
8. Rebelo Marcolino, M., et al. Student dropout prediction through machine learning optimization: Insights from Moodle log data. Sci. Rep. 15, 1–16 (2025). https://doi.org/10.1038/s41598-025-93918-1
9. Mduma, N. Data balancing techniques for predicting student dropout using machine learning. Data 8, 49 (2023). https://doi.org/10.3390/data8030049
10. Mnyawami, Y. N., Maziku, H. H., & Mushi, J. C. Enhanced model for predicting student dropouts in developing countries using automated machine learning approach: A case of Tanzanian's Secondary Schools. Appl. Artif. Intell. 36, 2071406 (2022). https://doi.org/10.1080/08839514.2022.2071406
11. Villar, A., & de Andrade, C. R. V. Supervised machine learning algorithms for predicting student dropout and academic success: A comparative study. Discover Artif. Intell. 4, 2 (2024). https://doi.org/10.1007/s44163-023-00079-z
12. Song, Z., Sung, S. H., Park, D. M., & Park, B. K. All-year dropout prediction modeling and analysis for university students. Appl. Sci. 13, 1143 (2023). https://doi.org/10.3390/app13021143
13. Dake, D. K., & Buabeng-Andoh, C. Using machine learning techniques to predict learner drop-out rate in higher educational institutions. Mob. Inf. Syst. 2022, 2670562 (2022). https://doi.org/10.1155/2022/2670562
14. Romero, S., & Liao, X. Statistical and machine learning models for predicting university dropout and scholarship impact. PLoS ONE 20, e0325047 (2025). https://doi.org/10.1371/journal.pone.0325047
15. Oqaidi, K., Aouhassi, S., & Mansouri, K. Towards a students' dropout prediction model in higher education institutions using machine learning algorithms. Int. J. Emerg. Technol. Learn. 17, 103–117 (2022). https://doi.org/10.3991/ijet.v17i18.25567
16. Seo, E. Y., Yang, J., Lee, J. E., & So, G. Predictive modelling of student dropout risk: Practical insights from a South Korean distance university. Heliyon 10, e30960 (2024). https://doi.org/10.1016/j.heliyon.2024.e30960
17. Krüger, J. G. C., de Souza Britto Jr, A., & Barddal, J. P. An explainable machine learning approach for student dropout prediction. Expert Syst. Appl. 233, 120933 (2023). https://doi.org/10.1016/j.eswa.2023.120933
18. Zapata-Medina, D., Espinosa-Bedoya, A., & Jiménez-Builes, J. A. Improving the automatic detection of dropout risk in middle and high school students: A comparative study of feature selection techniques. Math. 12, 1776 (2024). https://doi.org/10.3390/math12121776
19. Zerkouk, M., Mihoubi, M., Chikhaoui, B., & Wang, S. A machine learning based model for student's dropout prediction in online training. Educ. Inf. Technol. 29, 15793–15812 (2024). https://doi.org/10.1007/s10639-024-12500-w
20. Niyogisubizo, J., Liao, L., Nziyumva, E., Murwanashyaka, E., & Nshimyumukiza, P. C. Predicting student's dropout in university classes using two-layer ensemble machine learning approach: A novel

stacked generalization. Comput. Educ. Artif. Intell. 3, 100066 (2022). https://doi.org/10.1016/j.caeai.2022.100066
21. Won, H. S., Kim, M. J., Kim, D., Kim, H. S., & Kim, K. M. University student dropout prediction using pretrained language models. Appl. Sci. 13, 7073 (2023). https://doi.org/10.3390/app13127073
22. Kim, S., Choi, E., Jun, Y. K., & Lee, S. Student dropout prediction for university with high precision and recall. Appl. Sci. 13, 6275 (2023). https://doi.org/10.3390/app13106275
23. https://www.kaggle.com/datasets/colabsss/student-dropout-prediction-dataset
24. Singh, J., Singh, J., & Gosain, A. Enhancing medical data completeness using an iterative KNN based-Kernelized fuzzy c-means imputation method. Discover Comput. 29, 9 (2026). https://doi.org/10.1007/s10791-025-09889-4
25. Simanjuntak, C. C. R., Dakhi, S. R., & Sarumaha, Y. N. A. Hierarchical clustering sensitivity to distance metric selection based on Z-score normalization. JADEN: J. Algorithmic Digit. Eng. Netw. 1, 64–75 (2026). https://doi.org/10.65853/jaden.v1i2.122v
26. Mazarei, A., Sousa, R., Mendes-Moreira, J., Molchanov, S., & Ferreira, H. M. Online boxplot derived outlier detection. Int. J. Data Sci. Anal. 19, 83–97 (2025). https://doi.org/10.1007/s41060-024-00559-0
27. Aldrees, A., Jibrin, A. M., Dan'azumi, S., Al-Suwaiyan, M., Abba, S. I., & Yaseen, Z. M. Kernel principal component analysis-based water quality index modelling for coastal aquifers in Saudi Arabia. Sci. Rep. 15, 35097 (2025). https://doi.org/10.1038/s41598-025-18980-1
28. Ghoneim, S. S., Baz, M., Alzaed, A., & Zewdie, Y. T. Predicting the insulating paper state of the power transformer based on XGBoost/LightGBM models. Sci. Rep. 15, 17836 (2025). https://doi.org/10.1038/s41598-025-03033-4
29. Alzabidi, E., Yakut, Ö., Saad, S., & Fındık, O. SOD-FE: a supervised outlier detection and feature engineering approach for student dropout prediction. Sci. Rep. 16 (2026). https://doi.org/10.1038/s41598-026-56591-6
30. Dermol, V., & Skrbinjek, V. Predicting student dropout in higher education using Random Forest and survival analysis. Discov. Educ. 5, 655 (2026). https://doi.org/10.1007/s44217-026-01710-8
31. Duan, Y., & Chen, X. An adaptive multi-scale spatio-temporal graph network for robust MOOC dropout prediction. Sci. Rep. 16, 10966 (2026). https://doi.org/10.1038/s41598-026-40502-w
32. Sulak, S. A., & Koklu, N. Predicting student dropout using machine learning algorithms. Intell. Methods Eng. Sci. 3, 91–98 (2024). http://dx.doi.org/10.58190/imiens.2024.103
33. Arora, D. Predicting students' academic success and dropout using supervised machine learning. Int. J. Sci. Study 11, 72–78 (2023).
34. Mustofa, S., Emon, Y. R., Mamun, S. B., Akhy, S. A., & Ahad, M. T. A novel AI-driven model for student dropout risk analysis with explainable AI insights. Comput. Educ. Artif. Intell. 8, 100352 (2025). https://doi.org/10.1016/j.caeai.2024.100352
35. Kabathova, J., & Drlik, M. Towards predicting student's dropout in university courses using different machine learning techniques. Appl. Sci. 11, 3130 (2021). https://doi.org/10.3390/app11073130