STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **16**, September 2026, pp 2058–2086.
Published online in International Academic Press (www.IAPress.org)

A Proposed Hybrid Deep Learning and Ensemble Learning Model for Predicting Student Dropout in Education

Abdulalim M. Ibrahim, Mohamed R. Abonazel*, Abdul-Hadi N. Ahmed*

Faculty of Graduate Studies for Statistical Research, Cairo University, Giza, Egypt

Abstract Student dropout remains a major challenge for higher education institutions due to its academic, social, and economic consequences. Early identification of students at risk of dropping out is essential for implementing timely interventions; however, existing prediction approaches often rely on standalone machine learning models that inadequately capture the complex, nonlinear relationships within heterogeneous educational data and exhibit reduced performance under class imbalance. To address these limitations, this study proposes a hybrid prediction framework that integrates deep neural representation learning with a stacked ensemble architecture to improve the accuracy and robustness of student dropout prediction. The proposed framework is evaluated on three benchmark educational datasets. A deep neural network is employed to learn latent feature representations from preprocessed student data, and the resulting embeddings are used to train heterogeneous machine learning classifiers, including XGBoost, Decision Tree, AdaBoost, Extra Trees, and K-Nearest Neighbors. Their probabilistic predictions are subsequently combined through a logistic regression meta-learner to generate the final prediction. To investigate the impact of class imbalance, both class weighting and the Synthetic Minority Over-sampling Technique (SMOTE) are incorporated and systematically evaluated. Model performance is assessed using accuracy, balanced accuracy, precision, recall, F1-score, and ROC-AUC, while permutation feature importance is employed to examine the influence of the original input variables on the overall prediction pipeline, and McNemar's test is used to assess the statistical significance of performance differences. Experimental results demonstrate that the proposed framework consistently outperforms standalone machine learning models, a standalone deep neural network, and conventional stacking approaches across all three datasets. These findings demonstrate that combining deep representation learning with stacked ensemble learning provides a robust and effective framework for early student dropout prediction, supporting data-driven educational interventions and informed decision-making in higher education.

Keywords Student dropout prediction, educational data mining, deep neural networks, stacked ensemble learning, class imbalance, SMOTE.

DOI: 10.19139/soic-2310-5070-3733

1. Introduction

Student dropout remains one of the most persistent challenges facing higher education institutions worldwide because of its profound academic, social, and economic consequences. Students who discontinue their studies before graduation often experience reduced employment opportunities, lower lifetime earnings, and diminished career prospects, while educational institutions encounter lower graduation rates, inefficient allocation of resources, and financial losses associated with student attrition. Consequently, improving student retention has become a strategic priority for universities, governments, and policymakers seeking to enhance educational quality and sustainability. The ability to identify students who are at risk of dropping out at an early stage enables institutions to implement timely academic support, personalized interventions, and informed decision-making processes that can substantially improve student success and institutional performance. Recent studies have therefore emphasized

*Correspondence to: Mohamed R. Abonazel (Email: mabonazel@cu.edu.eg), Abdul-Hadi N. Ahmed (Email: Dr.hadi@cu.edu.eg). Faculty of Graduate Studies for Statistical Research, Cairo University, Giza, Egypt

ISSN 2310-5070 (online) ISSN 2311-004X (print)

the growing importance of predictive analytics as a key component of modern educational management systems [7, 10, 13, 15, 12].

The rapid digital transformation of higher education has generated unprecedented volumes of educational data through learning management systems, student information systems, institutional databases, and online learning platforms. These heterogeneous data sources contain demographic, academic, behavioral, and socioeconomic information that provides valuable insights into students' learning behaviors and academic progression. As a result, Educational Data Mining (EDM) and Learning Analytics (LA) have emerged as important interdisciplinary research areas that exploit data-driven methodologies to understand student behavior, support educational decision-making, and develop early-warning systems for identifying students at risk of academic failure or dropout [23, 11]. Advances in data collection and computational intelligence have enabled educational institutions to transition from descriptive statistical analyses toward predictive models capable of uncovering complex relationships among multiple educational factors. These predictive models provide valuable support for academic advisors and institutional administrators by facilitating proactive intervention strategies rather than reactive responses after academic problems have already occurred [16, 33, 34].

Machine learning has become one of the most widely adopted approaches for student dropout prediction because of its ability to automatically discover complex nonlinear relationships within large educational datasets. Traditional supervised learning algorithms, including Logistic Regression, Support Vector Machines, Decision Trees, Random Forests, and gradient boosting methods such as XGBoost, have demonstrated promising predictive performance across various educational environments [29, 30, 37, 35, 36]. More recent studies have further confirmed the effectiveness of machine learning for early identification of at-risk students using both pre-enrollment and academic performance data [19, 17, 22, 21, 20]. Despite these advances, predictive performance often varies considerably across datasets because educational data are inherently heterogeneous, high-dimensional, and frequently characterized by nonlinear feature interactions, noisy observations, and class imbalance. Consequently, developing robust predictive frameworks capable of effectively learning complex student representations while maintaining reliable predictive performance across different educational datasets remains an active area of research.

Deep learning has recently emerged as a powerful alternative to conventional machine learning for educational prediction tasks because of its ability to automatically learn hierarchical feature representations from complex and heterogeneous data. Unlike traditional approaches that rely heavily on manually engineered features, deep neural networks can model intricate nonlinear relationships and extract latent representations that capture underlying patterns within educational data [8]. Advances in deep architectures, including residual learning and other representation learning techniques, have significantly improved predictive performance across numerous machine learning applications [9]. Consequently, deep learning has increasingly been adopted in Educational Data Mining for predicting student performance, academic success, and dropout risk. Recent studies have demonstrated that hybrid deep learning models, such as LSTM-DNN and recurrent neural network-based frameworks, can effectively capture complex learning behaviors and improve the early identification of students at risk of dropping out [14, 5]. Nevertheless, despite these promising developments, standalone deep learning models often require large volumes of training data, are computationally demanding, and may exhibit limited robustness when educational datasets contain heterogeneous feature distributions or moderate sample sizes.

To overcome the limitations of individual predictive models, ensemble learning has become an increasingly important research direction in machine learning. Ensemble methods improve predictive performance by combining multiple complementary learning algorithms, thereby reducing model variance, increasing robustness, and enhancing generalization capabilities [6, 1]. Among the various ensemble strategies, stacked generalization (stacking) has received considerable attention because it learns how to optimally combine the predictions of multiple heterogeneous base learners through a meta-learning model rather than relying on simple voting or averaging mechanisms [4]. Stacking enables different classifiers to exploit complementary decision boundaries and capture diverse characteristics of educational data, making it particularly suitable for complex prediction problems such as student dropout. Recent studies have shown that ensemble learning consistently outperforms many individual classifiers in educational prediction tasks, particularly when multiple heterogeneous models are integrated within a unified predictive framework [22, 21, 20]. Despite these advances, relatively few studies have explored the integration of deep neural representation learning with stacked ensemble architectures for student

dropout prediction, and systematic investigations of how learned latent representations contribute to ensemble performance remain limited.

Although considerable progress has been achieved in student dropout prediction, several important research challenges remain. First, many existing studies rely on standalone machine learning or deep learning models, limiting their ability to exploit the complementary strengths of heterogeneous predictive algorithms. Second, while recent hybrid models have combined deep learning with traditional classifiers, relatively few frameworks employ deep neural networks primarily as representation learners whose latent embeddings are subsequently exploited within a stacked ensemble architecture. Third, class imbalance continues to be a major challenge because dropout students typically constitute the minority class, causing predictive models to become biased toward the majority class if appropriate mitigation strategies are not employed [26, 27, 18]. Moreover, previous studies often evaluate a single imbalance-handling strategy without systematically comparing alternative approaches such as class weighting and SMOTE. Finally, although model interpretability has become increasingly important in educational applications, most existing approaches either focus exclusively on predictive performance or employ limited post-hoc analyses without explicitly discussing the scope and limitations of interpretation. These challenges highlight the need for a comprehensive framework that integrates deep representation learning, stacked ensemble learning, systematic class imbalance analysis, and rigorous evaluation to improve the reliability and effectiveness of early student dropout prediction systems.

Motivated by these challenges, this study proposes a hybrid prediction framework that integrates deep neural representation learning with stacked ensemble learning to improve the accuracy and robustness of student dropout prediction. Rather than using a deep neural network solely as an end-to-end classifier, the proposed framework employs it as a representation learner that transforms heterogeneous educational data into compact latent feature embeddings. These learned embeddings are subsequently used to train a diverse set of machine learning classifiers, allowing each classifier to exploit the discriminative information captured by the deep network while preserving its own decision-making characteristics. Their probabilistic outputs are finally combined through a meta-learning model to generate the final prediction. In addition, the proposed framework systematically investigates two widely adopted class imbalance handling strategies, namely class weighting and the Synthetic Minority Over-sampling Technique (SMOTE), to assess their influence on predictive performance. This design seeks to leverage the complementary strengths of representation learning and ensemble learning while providing a robust framework for early identification of students at risk of dropping out.

The proposed framework is evaluated independently on three publicly available benchmark educational datasets representing different educational contexts and data characteristics. To provide a comprehensive assessment, its performance is compared with conventional machine learning classifiers, a standalone deep neural network, and a conventional stacking framework using multiple evaluation metrics, including Accuracy, Balanced Accuracy, Precision, Recall, F1-score, and the Area Under the Receiver Operating Characteristic Curve (ROC–AUC). Statistical significance is assessed using McNemar's test, while permutation feature importance is employed to analyze the influence of the original input variables on the overall prediction pipeline. It should be noted that the feature importance analysis is intended to explain the sensitivity of the complete predictive framework to changes in the input variables rather than to interpret the internal representations learned by the deep neural network. Furthermore, although the proposed framework is evaluated on three benchmark datasets, validating its applicability across additional educational systems and institutional settings remains an important direction for future research.

The main contributions of this study are summarized as follows:

1. A hybrid prediction framework is proposed that integrates deep neural representation learning with stacked ensemble learning, enabling heterogeneous machine learning classifiers to exploit latent feature embeddings for improved student dropout prediction.
2. A systematic investigation of class imbalance handling strategies is conducted by comparing class weighting and the Synthetic Minority Over-sampling Technique (SMOTE), providing insights into their influence on predictive performance across multiple educational datasets.
3. A comprehensive experimental evaluation is performed on three benchmark educational datasets using multiple performance metrics and statistical significance testing to demonstrate the effectiveness and

robustness of the proposed framework compared with standalone machine learning models, a standalone deep neural network, and a conventional stacking framework.

4. A model-agnostic interpretation of the proposed prediction pipeline is provided through permutation feature importance, enabling analysis of the contribution of the original input variables to the overall predictive performance while acknowledging the limitations of interpreting the latent representations learned by the deep neural network.

The remainder of this paper is organized as follows. Section 2 reviews the related literature on student dropout prediction, educational data mining, deep learning, and ensemble learning. Section 3 describes the benchmark datasets and the proposed hybrid prediction framework, including data preprocessing, representation learning, stacked ensemble construction, and model evaluation procedures. Section 4 presents the experimental results together with comparative analyses, statistical validation, and feature importance analysis. Section 5 discusses the findings, practical implications, limitations, and future research directions. Finally, Section 6 concludes the paper and summarizes the main contributions of this study.

2. Related Work

Student dropout has remained a persistent challenge for educational institutions worldwide, producing substantial academic, social, and economic consequences for both learners and institutions. The rapid growth of digital educational environments has led to the emergence of Educational Data Mining (EDM), an interdisciplinary research field that integrates data science, machine learning (ML), and artificial intelligence to analyze student behavior, performance, and engagement. Through these techniques, researchers have sought to better understand learning processes and develop data-driven strategies for improving student retention and academic success [11].

Recent studies have explored a wide range of factors influencing student dropout using advanced machine learning and predictive analytics methods. Across diverse educational contexts, variables such as grade point average (GPA), attendance patterns, online engagement, and emotional or behavioral indicators have consistently emerged as dominant predictors of dropout risk. For instance, [12] achieved an accuracy of 87% when modeling dropout among distance education students, while [15] developed an AI-based early warning system reporting 91% accuracy. Deep learning–based approaches have further demonstrated strong predictive capability; notably, [14] reported 92% accuracy using a hybrid LSTM–DNN architecture, highlighting the effectiveness of neural networks for early dropout detection.

Beyond static prediction settings, several studies investigated longitudinal or year-round dropout prediction frameworks. [16] developed a model that accounted for temporal variation in dropout behavior across the academic year. Using Random Forest and Gradient Boosting Machines (GBM), the authors achieved 87% accuracy and emphasized the importance of GPA and attendance while addressing missing data across multiple academic periods. Collectively, these studies revealed that although predictive performance was often high, challenges related to limited dataset size, temporal variability, and class imbalance continued to constrain model robustness and generalizability. Nevertheless, the consistent emergence of academic performance, engagement, and demographic variables underscored their central role in dropout prediction.

Classical machine learning algorithms, including Logistic Regression, Random Forests, Support Vector Machines (SVMs), Decision Trees, and gradient boosting methods such as XGBoost, have remained foundational tools in student dropout prediction research [7, 10, 17]. These models demonstrated strong performance in single-institution settings, with reported accuracy reaching 90.8% in [17]. However, despite their effectiveness under controlled conditions, such models often exhibited limited generalization across institutions due to differences in grading systems, course structures, feature definitions, and data availability [7]. This limitation was further amplified by the severe class imbalance typical of dropout datasets, where non-dropout cases substantially outnumber dropout instances. In such scenarios, conventional accuracy metrics could be misleading, motivating the adoption of evaluation measures such as Balanced Accuracy to provide fairer assessment of minority-class performance [18]. These challenges collectively highlighted the need for more robust and transferable modeling strategies that extend beyond single-model classical approaches.

Several studies explored ensemble-based learning as a means of improving predictive robustness. [19] investigated student dropout prediction at the university course level using data extracted from a Learning Management System (LMS) and followed the CRISP-DM methodology. The authors evaluated multiple supervised classifiers, including Logistic Regression, Decision Trees, Random Forests, Support Vector Machines, Naïve Bayes, and Neural Networks, using 10-fold cross-validation and a temporal hold-out dataset from a subsequent academic year. Performance was assessed using multiple metrics, including accuracy, precision, recall, F1-score, and McNemar's statistical test. The results demonstrated that ensemble-based models, particularly Random Forests, consistently outperformed simpler classifiers, achieving accuracy levels of up to 93%. This study emphasized the importance of robust evaluation protocols and metric selection when working with small and imbalanced educational datasets. Recent advances in educational data mining have increasingly explored deep learning and ensemble learning to improve predictive performance. Nuweeji and Alzubi [2] proposed an Activation Ensemble Deep Neural Network (AcEnDNN) for early student performance prediction by combining multiple activation functions, including ReLU, sigmoid, tanh, and Swish, within a unified deep neural network architecture. The model was evaluated using two public educational datasets and a real-world dataset, achieving consistently low prediction errors and demonstrating its ability to capture complex nonlinear relationships among student characteristics. Their findings highlight the effectiveness of activation ensemble strategies for improving predictive accuracy and robustness in educational prediction tasks.

In addition, Rahman *et al.* [3] presented a comprehensive review of ensemble learning methods for educational data analysis. Their survey concludes that ensemble learning consistently outperforms individual machine learning classifiers by improving prediction accuracy, reducing model variance, and handling high-dimensional educational datasets. The review further emphasizes that integrating feature selection techniques with ensemble learning enhances scalability and interpretability, highlighting ensemble learning as a promising direction for future educational prediction systems.

These recent studies further demonstrate the growing interest in hybrid and ensemble-based predictive models. Nevertheless, most existing approaches either focus exclusively on deep learning architectures or ensemble classifiers independently. Few studies combine deep neural representation learning with heterogeneous stacked ensemble learning while simultaneously validating the proposed framework across multiple benchmark datasets using comprehensive evaluation metrics, statistical significance testing, and model interpretability analysis.

In parallel, deep learning architectures such as Deep Neural Networks (DNNs), Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and hybrid Long Short-Term Memory (LSTM) models have gained increasing attention due to their ability to automatically learn feature representations and capture nonlinear dependencies in educational data [1, 14]. Formally, deep neural models learn hierarchical embeddings through successive nonlinear transformations, enabling the modeling of higher-order interactions among academic, behavioral, and demographic factors [8]. Empirical evidence supported the effectiveness of such approaches; for example, hybrid LSTM–DNN architectures achieved accuracy levels as high as 92% [14]. However, many deep learning–based dropout prediction studies relied on single-model, single-dataset designs, which increased the risk of overfitting, particularly when datasets were small, noisy, or highly imbalanced. Moreover, limited attention was given to transferring learned knowledge across heterogeneous datasets.

Hybrid modeling strategies were proposed to address these limitations by combining complementary learning paradigms. [5] introduced the DeepS3VM framework, which integrated a recurrent neural network for sequential feature learning with a semi-supervised support vector machine to exploit both labeled and unlabeled data. Using student data from a private university in Vietnam, the proposed model achieved 92.54% accuracy and demonstrated robustness under class imbalance while supporting personalized intervention recommendations. Similarly, [14] proposed a deep neural framework that combined recurrent architectures with fully connected layers to model both temporal and static academic features, achieving superior performance compared to traditional machine learning baselines on a real-world higher education dataset.

Recent studies have continued to demonstrate the effectiveness of advanced ensemble and hybrid learning approaches for early student dropout prediction. Carballo-Mendívil *et al.* [20] proposed an XGBoost-based early warning system using pre-enrollment information and compared it with LightGBM and CatBoost classifiers. By integrating SMOTE and class weighting within the cross-validation process, the proposed model achieved an AUC

of 94.2%, demonstrating that gradient-boosting ensembles can provide highly accurate early predictions while mitigating class imbalance.

Similarly, Nguyen Thi Cam *et al.* [5] introduced a hybrid framework that combines recurrent neural networks (RNNs) with a semi-supervised support vector machine (SVM) to identify students at risk of dropping out. Their model was evaluated using longitudinal multi-institutional educational data and achieved an accuracy of 92.54%, outperforming the individual RNN and SVM models. These findings further demonstrate the advantages of integrating deep representation learning with complementary machine learning algorithms for improving prediction performance.

Although these studies demonstrate the effectiveness of ensemble and hybrid learning techniques, most existing approaches rely on either deep learning or ensemble learning individually and are generally evaluated on a single educational dataset. In contrast, the proposed framework integrates deep neural representation learning with heterogeneous stacked ensemble learning and validates its effectiveness across three benchmark educational datasets using comprehensive performance evaluation and statistical significance analysis.

The robustness and generalizability of dropout prediction models were further examined by [21], who adopted a multi-dataset experimental design derived from a single university across five faculties and five temporal observation points, resulting in 25 distinct datasets. This design reflected realistic educational data heterogeneity and enhanced external validity. However, although multiple datasets were evaluated, each dataset was modeled independently, with no explicit mechanism for joint learning or cross-dataset knowledge transfer. Additionally, evaluation primarily relied on accuracy rather than balanced metrics, limiting insight into minority-class detection performance under severe class imbalance.

Overall, prior research demonstrated that both classical machine learning and deep learning approaches can effectively predict student dropout under specific conditions. Nevertheless, existing studies predominantly relied on single-model or single-dataset designs, with limited integration of learned representations, ensemble-based decision mechanisms, and interpretability considerations. Furthermore, class imbalance and evaluation bias remained persistent challenges. These limitations motivated the investigation of hybrid frameworks that combine deep representation learning with ensemble learning under rigorous, balanced evaluation protocols.

Table 1. Summary of representative studies on student dropout prediction.

Study	Techniques	Dataset	Performance	Limitations
[2]	Activation Ensemble Deep Neural Network (AcEnDNN)	Student-mat, Student-por, and real-world datasets	MAE = 1.28, RMSE = 2.13	Focuses on student performance prediction rather than dropout prediction
[20]	XGBoost, LightGBM, CatBoost	Pre-enrollment educational data	AUC = 94.2%	Uses only pre-enrollment information
[14]	LSTM–DNN Hybrid	Predict Students' Dropout and Academic Success	Accuracy = 92%	Single dataset; limited handling of data imbalance
[5]	Hybrid RNN + Semi-supervised SVM	Multi-institutional longitudinal data	Accuracy = 92.54%	Limited comparison with ensemble learning methods
[31]	Logistic Regression	MOOC platform data	Accuracy = 87%	Limited nonlinear modeling capability
[32]	LR, KNN, RF	Academic performance data	Accuracy = 74.9%	Moderate predictive performance
[33]	Support Vector Machine	Student information systems	Successful dropout prediction	Limited representation learning
[34]	Perceptron, Boosted Decision Tree	Academic, behavioural and demographic data	Accuracy = 90.8%	Traditional machine learning only

Table 1 shows that recent studies have achieved promising results using deep learning, gradient-boosting algorithms, or hybrid architectures.

Although recent studies have demonstrated improvements using deep learning, gradient boosting, and ensemble learning, few studies have explicitly combine deep representation learning with heterogeneous stacked ensemble learning while providing comprehensive comparisons across multiple educational datasets and conducting statistical significance analysis. Nevertheless, most existing methods employ either deep learning or ensemble learning independently, evaluate their models on a single educational dataset, or provide limited analysis of model interpretability and statistical significance. Furthermore, few studies investigate whether deep representation learning provides complementary information beyond conventional stacking ensembles through ablation analysis. To address these limitations, this study proposes a hybrid framework that combines deep neural representation learning with heterogeneous stacked ensemble learning and validates its effectiveness across three benchmark educational datasets using comprehensive evaluation metrics, permutation-based feature importance, and McNemar's statistical significance test.

3. Datasets

To comprehensively evaluate the proposed framework, experiments were conducted independently on three publicly available benchmark datasets collected from different higher education institutions. These datasets differ in terms of sample size, feature composition, class distribution, and educational context, enabling a comprehensive assessment of the proposed model under diverse learning environments. The first dataset originates from Portugal, the second from Slovakia, and the third from Vietnam, providing heterogeneous educational scenarios for evaluating the robustness and generalizability of the proposed framework.

3.1. Dataset 1: Predict Students' Dropout and Academic Success

The dataset *Predict Students' Dropout and Academic Success* was collected from a Portuguese higher-education institution and is publicly available through the UCI Machine Learning Repository. It comprises records of undergraduate students enrolled across multiple degree programs and includes demographic, socio-economic, academic, administrative, and performance-related information. The dataset has been widely used in educational data mining research to support the analysis and prediction of student academic outcomes, particularly dropout behavior [25, 24].

The dataset consists of structured tabular data representing individual students, making it suitable for both classical machine learning and deep learning–based predictive modeling approaches.

3.1.1. Dataset 1 Characteristics The dataset contains 4,424 student records described by 36 features spanning multiple informational domains. These features capture both static background characteristics and dynamic academic performance indicators. Specifically, the variables are grouped into seven distinct categories: Demographic, Family, Financial, Academic, Curricular Performance, Macroeconomic features, and the Target variable. This structured categorization enables a comprehensive and multi-dimensional representation of student trajectories, incorporating personal background, institutional engagement, academic progress, and external economic conditions. The Dropout class includes students who permanently discontinued their studies regardless of the semester in which withdrawal occurred. Consequently, the dataset does not distinguish between early and late dropout events. The proposed model therefore predicts overall dropout risk rather than the timing of withdrawal.

Although several macroeconomic variables are identical for students enrolled during the same academic period, they were intentionally retained because they provide contextual information describing the socioeconomic environment in which students studied. These variables may capture external conditions, such as unemployment rates, inflation, and regional economic changes, that influence educational persistence at the population level. The permutation importance analysis later demonstrates that these variables contribute substantially less than academic variables, indicating that the proposed framework naturally assigns greater importance to student-specific academic information.

Table 2 summarizes the feature groups utilized in this study.

Table 2. Dataset Features

Feature Category	Attributes
Demographic Features	Gender, Age at enrollment, Nationality, International
Family Features	Marital status, Mother's qualification, Father's qualification, Mother's occupation, Father's occupation
Financial Features	Tuition fees up to date, Debtor, Scholarship holder
Academic Features	Application mode, Application order, Course, Daytime/evening attendance, Previous qualification, Previous qualification (grade), Admission grade, Displaced, Educational special needs
Curricular Performance Features	Curricular units 1st semester (credited, enrolled, evaluations, approved, grade, without evaluations), Curricular units 2nd semester (credited, enrolled, evaluations, approved, grade, without evaluations)
Macroeconomic Features	Unemployment rate, Inflation rate, GDP
Target	Outcome (Target)

3.2. Dataset 2: Students Dropout Prediction

The second dataset was collected from Constantine the Philosopher University in Nitra, Slovakia, and contains academic records of students enrolled in a preparatory database systems course between 2016 and 2020 [19]. Unlike Dataset 1, this dataset focuses primarily on students' academic performance and course activities.

The dataset contains 261 student records described by 12 academic and course-related attributes. The target variable is binary, where class 1 represents students who successfully completed the course and class 0 denotes students who dropped out. Among all observations, 210 students (80.46%) successfully completed the course, whereas 51 students (19.54%) dropped out, indicating a moderately imbalanced class distribution.

Kabathova and Drlik evaluated multiple machine learning classifiers on this dataset and reported prediction accuracies ranging from 77% to 93% on unseen academic-year data, illustrating that reliable dropout prediction can be achieved even with relatively small educational datasets.

3.3. Dataset 3: Student Performance from a Private University in Vietnam

The third dataset was collected from a private university in Vietnam using records extracted from the institution's Student Management System (SMS) and Learning Management System (LMS). The dataset includes demographic, socio-economic, behavioral, academic, and learning engagement information while complying with institutional privacy and ethical guidelines [5].

Following data augmentation, the dataset contains approximately 100,000 student records described by 26 features covering demographic information, socio-economic characteristics, psychological and behavioral factors, academic history, learning engagement, resource utilization, and academic performance.

Nguyen Thi Cam et al. developed a hybrid recurrent neural network and semi-supervised support vector machine framework using this dataset. Their proposed approach achieved an accuracy of 92.54%, outperforming individual RNN and SVM models and demonstrating the benefits of hybrid learning architectures for student dropout prediction.

3.4. Target Variable Transformation

The original target variable contains three classes: *Dropout*, *Enrolled*, and *Graduate*. Following established practices in educational data mining, students labeled as *Enrolled* were excluded to avoid label ambiguity and uncertain academic outcomes. Consequently, the prediction task was reformulated as a binary classification problem distinguishing between students who permanently discontinued their studies (*Dropout*) and those who successfully completed their degree programs (*Graduate*). This reformulation results in a moderately imbalanced classification setting, which motivates the use of balanced evaluation metrics in subsequent experiments.

3.5. Exploratory Data Analysis

Exploratory Data Analysis (EDA) was conducted to examine the structure, distributions, and relationships among the dataset variables and to inform preprocessing and modeling decisions. Univariate analysis was performed to assess feature distributions, central tendencies, variability, and potential data quality issues. Summary statistics were computed for numerical variables, while frequency distributions were examined for categorical attributes.

Bivariate analysis investigated associations between individual predictors and the binary target variable. Correlation coefficients were calculated for numerical features, and chi-square tests of independence were applied to categorical variables to identify statistically significant relationships with student outcomes. Additionally, correlation matrices were analyzed to detect collinearity patterns among academic and behavioral features.

Multivariate analysis was used to explore interaction effects and joint feature relationships, providing insight into complex dependency structures within the data. These analyses supported informed preprocessing choices, including feature encoding, transformation, and selection, ensuring that the final dataset used for modeling was analytically robust and suitable for predictive learning tasks.

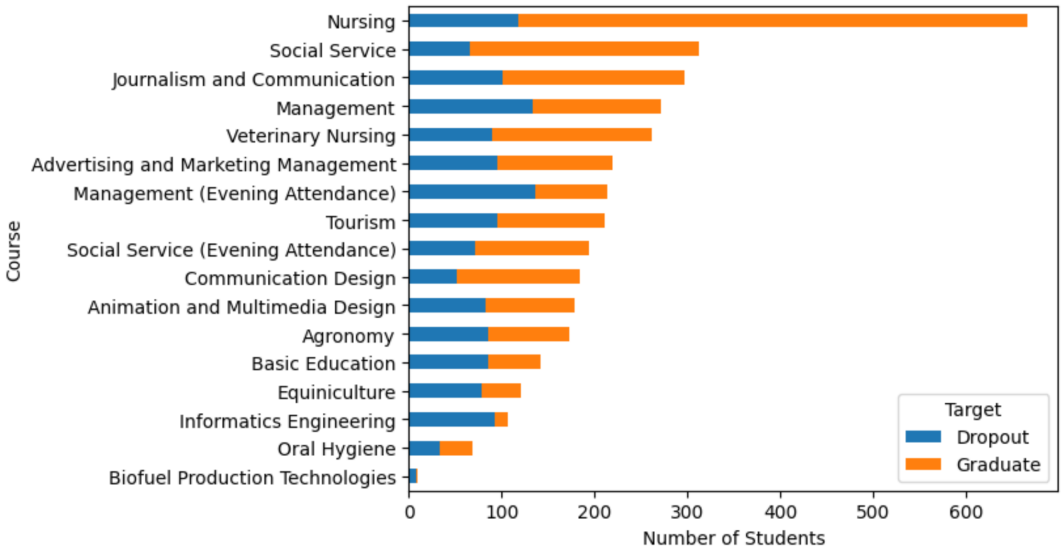


Figure 1. Multivariate visualization

Figure 1 illustrates the distribution of student outcomes across academic courses, highlighting the number of students who graduated versus those who dropped out in each course. The figure reveals noticeable variation in dropout prevalence among courses, indicating that student attrition is not uniformly distributed across academic courses. While some courses exhibit a higher proportion of graduates, others show comparatively elevated dropout counts, suggesting the presence of course-specific risk factors. This heterogeneity underscores the importance of incorporating course-level information into predictive models, as program characteristics may influence student persistence differently. Moreover, the observed imbalance between graduate and dropout outcomes across most courses further reinforces the need for balanced evaluation metrics and modeling strategies that account for class imbalance.

4. Methodology

This section describes the proposed end-to-end predictive framework for student dropout prediction. The methodology integrates preprocessing, deep neural network–based representation learning, stacked ensemble classification, and permutation-based interpretability.

4.1. Overview of the Proposed Hybrid Framework

The proposed framework combines deep representation learning with stacked ensemble learning to improve student dropout prediction. Unlike conventional machine learning approaches that operate directly on handcrafted input variables, the proposed architecture first learns compact latent representations using a deep neural network and subsequently integrates these representations with the probabilistic outputs of multiple heterogeneous machine learning classifiers.

Figure 2 and algorithm 1 illustrates the overall workflow. After preprocessing, the training subset is used to train both the deep neural network and the level-1 base classifiers. The trained encoder generates latent embeddings for the validation subset, while the base classifiers simultaneously produce class probability estimates. These embeddings and probability estimates are concatenated to form the meta-feature space used to train a logistic regression meta-learner. The independent test subset is reserved exclusively for the final evaluation and is never used during preprocessing, model training, hyperparameter optimization, or meta-learning.

This design separates representation learning from decision fusion, enabling the proposed framework to exploit complementary information captured by deep neural representations and heterogeneous machine learning models while preventing information leakage throughout the learning process.

Algorithm 1 Proposed Hybrid Prediction Framework

Require: Dataset D
Ensure: Predicted labels $\hat{y}$
1: Preprocess and split D into training, validation, and test sets
2: Apply SMOTE to the training set
3: Train the DNN encoder
4: Train the Level-1 base learners
5: Extract validation embeddings and base-model probabilities
6: Form the meta-feature matrix by concatenation
7: Train the logistic regression meta-learner
8: Predict the test set using the DNN, base learners, and meta-learner

4.2. Problem Formulation

The student dropout prediction task is formulated as a binary classification problem. Each student instance is assigned one of two possible outcomes: Dropout or Graduate. Given a feature vector comprising demographic, academic, administrative, and macroeconomic attributes, the objective is to estimate the probability that a student will leave their studies before completion.

Formally, for a student represented by feature vector $\mathbf{x}_i \in \mathbb{R}^p$, the model aims to learn a function $f(\mathbf{x}_i)$ that predicts the likelihood of dropout, enabling early identification of at-risk students.

4.3. Data Preprocessing

The three benchmark datasets considered in this study exhibit different characteristics in terms of feature types, dimensionality, and class distributions. Therefore, dataset-specific preprocessing procedures were applied within a unified preprocessing framework to ensure consistency while preventing data leakage throughout the prediction pipeline.

Among the three datasets, only the *Predict Students' Dropout and Academic Success* dataset contains three target classes: *Graduate*, *Dropout*, and *Enrolled*. The original class distribution consists of 2,209 *Graduate* instances, 1,421 *Dropout* instances, and 794 *Enrolled* instances. Since the academic outcome of enrolled students is unknown at the time of prediction, all *Enrolled* instances were removed, following the common practice adopted in previous studies. Consequently, the prediction task was reformulated as a binary classification problem comprising 3,630 instances, including 2,209 *Graduate* students (Class 0) and 1,421 *Dropout* students (Class 1). The remaining

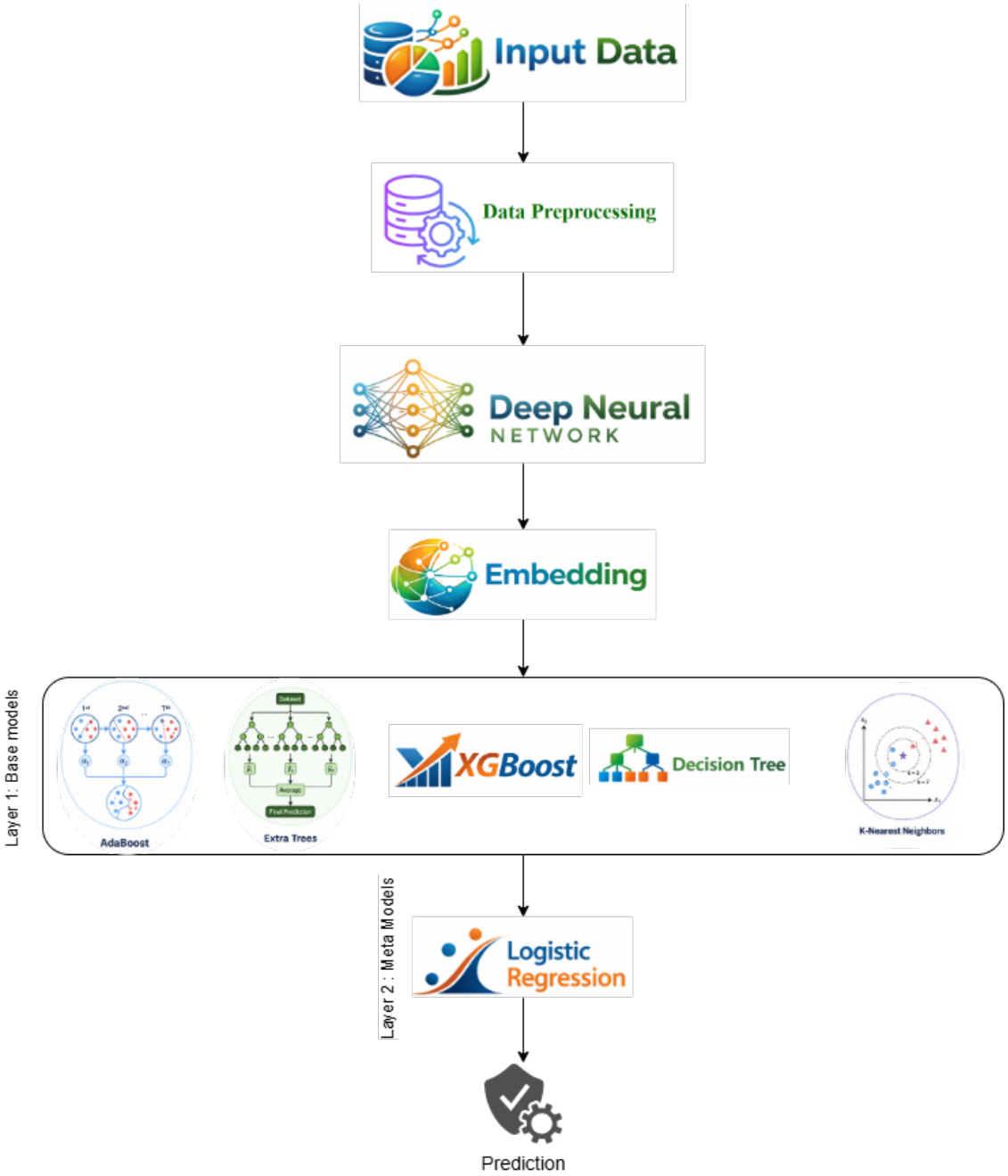


Figure 2. The architecture of the proposed model.

two benchmark datasets were already formulated as binary classification problems and therefore required no modification of the target labels.

Each dataset was first divided into an 80% development subset and an independent 20% test subset. The development subset was subsequently divided into 80% training and 20% validation, resulting in an overall 64%/16%/20% partition. This strategy provides sufficient training data while maintaining an independent validation subset for hyperparameter optimization and meta-learning without exposing the test data. The training

subset was used to fit the preprocessing pipeline and train the base predictive models. The validation subset was reserved exclusively for hyperparameter tuning, early stopping during deep neural network training, and generating the meta-features used to train the stacking meta-learner. The test subset remained completely isolated throughout preprocessing, model development, and meta-learning, and was accessed only once for the final performance evaluation. Figure 3 illustrates the proposed hold-out validation strategy, highlighting the separation between the training, validation, and test subsets to prevent data leakage throughout model development.

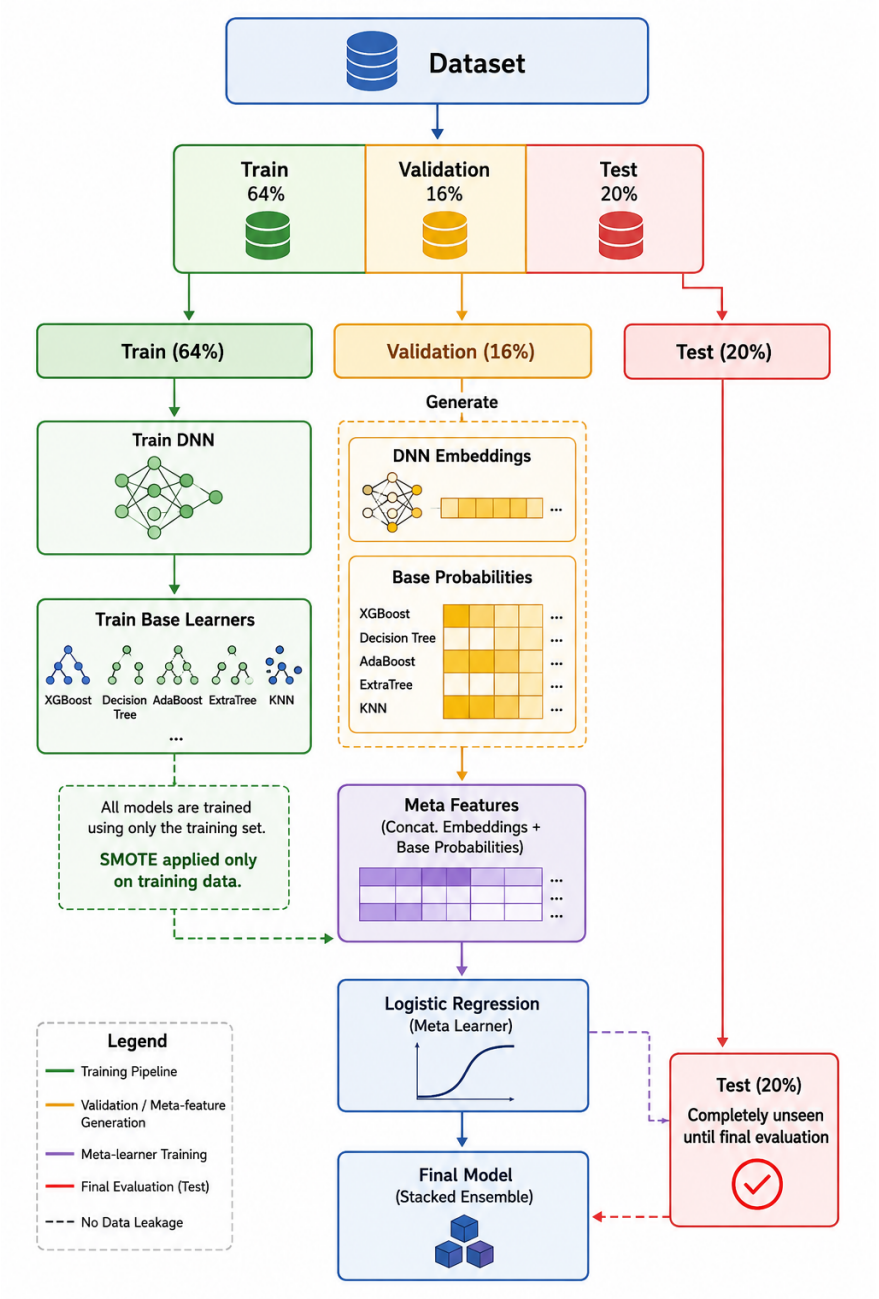


Figure 3. Training, validation, and testing workflow illustrating the prevention of data leakage during model development.

Categorical and numerical features were processed using a column-wise transformation strategy. Numerical variables were standardized using z-score normalization,

$$x' = \frac{x - \mu}{\sigma}, \tag{1}$$

where μ and σ denote the mean and standard deviation computed exclusively from the training subset. Categorical variables were encoded using one-hot encoding with unknown-category handling to ensure robust transformation of previously unseen categories. All preprocessing operations were implemented using a *ColumnTransformer* and fitted solely on the training data before being applied to the validation and test subsets, thereby eliminating information leakage.

To address class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) was applied exclusively to the preprocessed training subset after completion of all preprocessing operations. Neither the validation nor the test subsets were oversampled, ensuring that model selection and final evaluation were performed on data representing the original class distribution. The fitted preprocessing pipeline was subsequently saved to ensure reproducibility and consistent transformation during inference.

4.4. Deep Neural Representation Learning

A deep neural network (DNN) was employed to learn latent representations of student academic profiles from the preprocessed input features. Rather than serving solely as an end-to-end classifier, the network functions primarily as a representation learning module whose penultimate layer provides compact feature embeddings for subsequent ensemble learning.

The architecture consists of multiple fully connected layers with Swish activation functions, batch normalization, dropout regularization, and Gaussian noise applied at the input layer to improve robustness and generalization. The penultimate dense layer forms the latent embedding space, which summarizes nonlinear relationships among academic, demographic, and behavioural variables.

Model parameters were optimized using the AdamW optimizer and binary cross-entropy loss. Hyperparameters, including the number of hidden layers, hidden units, dropout rate, learning rate, Gaussian noise level, and embedding dimensionality, were automatically optimized using Keras Tuner with random search, where the validation ROC–AUC served as the optimization objective. Early stopping and adaptive learning-rate reduction were employed to prevent overfitting and improve convergence stability.

After optimization, the trained encoder was retained to generate latent embeddings for the validation and test subsets. These embeddings were subsequently combined with ensemble predictions to construct the final hybrid framework. Figure 4 presents the final network architecture.

This architectural design follows best practices for deep learning on tabular and imbalanced datasets, where regularization, normalization, and representation learning are essential for robust generalization [8].

4.4.1. Hyperparameter Optimization Hyperparameter tuning was conducted using a randomized search strategy implemented with Keras Tuner. The search space included the number of hidden layers, number of units per layer, dropout rate, learning rate, and embedding dimensionality. The optimization objective was the validation ROC-AUC score.

After identifying the optimal configuration, the final DNN was retrained using the training and validation sets with class-weighted loss. The trained network and the embedding extractor were saved for subsequent ensemble learning and reproducibility.

Using Keras Tuner (Random Search) to automatically optimize:

- Number of hidden layers
- Number of units per layer
- Activation function
- Dropout rate
- Learning rate
- Penultimate layer dimension

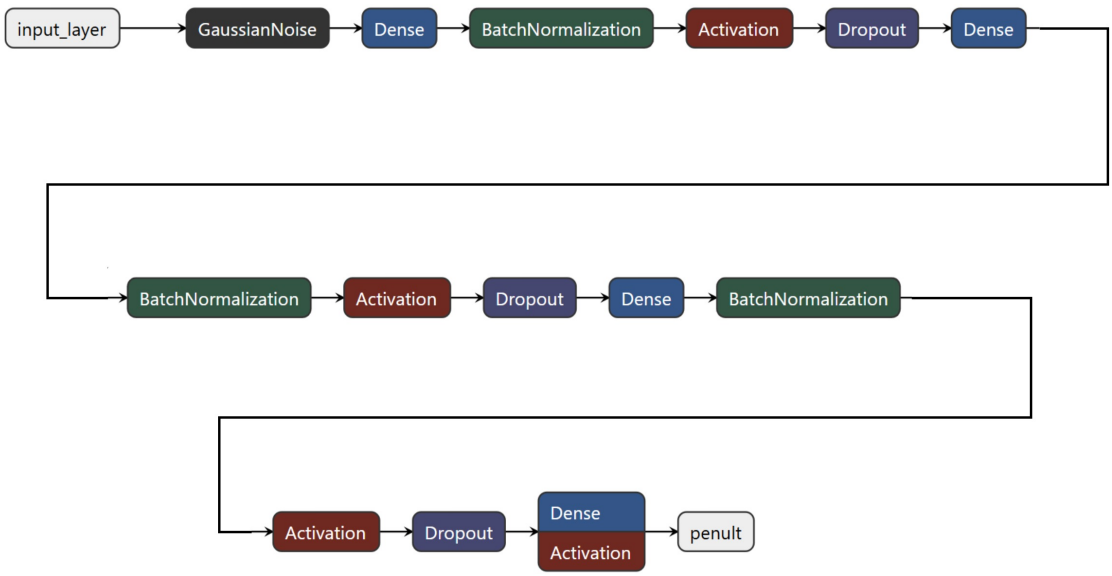


Figure 4. Embedding extractor architecture

This architectural design follows established best practices in neural network–based classification, particularly for tabular and imbalanced educational data, where normalization, regularization, and deep representation learning are critical for robust performance [8].

Since hyperparameter optimization was performed independently for each dataset, the resulting DNN architectures differ slightly in the number of hidden layers, hidden units, embedding dimensions, and optimization parameters. Tables 3 and 4 present the architecture and training configuration of the final DNN selected for Dataset 1 as a representative example. The corresponding architectures for Datasets 2 and 3 follow the same design principles but with different hyperparameter values obtained through automated optimization.

The optimized training configuration of the deep neural network is presented in Table 4. The selected hyperparameters were obtained through randomized hyperparameter search and were subsequently used to train the final embedding extractor for all experiments.

Table 3 details the architecture of the selected deep neural network. The network consists of an input layer followed by Gaussian noise regularization, multiple fully connected layers with batch normalization, Swish activation, and dropout, a penultimate embedding layer, and a final sigmoid output layer for binary classification.

As shown in Tables 4 and 3, the proposed embedding extractor employs a relatively compact architecture with progressively decreasing hidden-layer dimensions. The inclusion of Gaussian noise, batch normalization, and dropout regularization improves model robustness and reduces overfitting, while the 16-dimensional penultimate layer provides a compact latent representation that is subsequently used by the stacked ensemble framework.

4.5. Hybrid Stacked Ensemble Learning

The proposed hybrid framework adopts a two-level stacking architecture that combines deep neural representations with probabilistic predictions generated by heterogeneous machine learning classifiers.

To address the issue of class imbalance, which is a common challenge in student dropout prediction, class weighting and the Synthetic Minority Over-sampling Technique (SMOTE) are applied within the training folds only. This strategy ensures balanced learning while preventing information leakage from validation or test data [26].

Table 3. Architecture of the proposed DNN for Dataset-1.

Layer	Output Shape	Configuration
Input	229	Raw input features
Gaussian Noise	229	Noise = 0.01
Dense-1	64	Swish activation
Batch Normalization	64	–
Dropout	64	Rate = 0.40
Dense-2	32	Swish activation
Batch Normalization	32	–
Dropout	32	Rate = 0.40
Dense-3	32	Swish activation
Batch Normalization	32	–
Dropout	32	Rate = 0.40
Embedding Layer	16	Swish activation
Output Layer	1	Sigmoid activation

Table 4. Training hyperparameters of the proposed DNN for Dataset-1.

Parameter	Value
Optimizer	Adam
Loss Function	Binary Cross-Entropy
Label Smoothing	0.02
Learning Rate	0.001
Epochs	30
Batch Size	128
Dropout Rate	0.40
Gaussian Noise	0.01
Activation Function	Swish
Embedding Dimension	16
Early Stopping	Patience = 30
Learning Rate Scheduler	ReduceLROnPlateau
Validation Metric	ROC–AUC

At Level 1, five complementary classifiers, namely XGBoost, Decision Tree, AdaBoost, Extra Trees, and K-Nearest Neighbors, were independently trained using the preprocessed training data. Each classifier produces the estimated probability that a student belongs to the dropout class.

Simultaneously, the trained deep neural network generates latent embeddings for the validation samples. The embedding vectors and the probability estimates produced by the five base classifiers are concatenated to construct the meta-feature space,

$$\mathbf{Z} = [\mathbf{E}, \mathbf{P}],$$

where $\mathbf{E}$ denotes the latent embedding generated by the DNN and $\mathbf{P}$ represents the vector of probability estimates produced by the base classifiers.

The resulting meta-features are then used to train a logistic regression meta-learner. During inference, the same sequence of operations is applied to unseen test samples: latent embeddings are extracted by the DNN, probability estimates are generated by the base learners, the two representations are concatenated, and the resulting meta-feature vector is provided to the logistic regression classifier to produce the final prediction.

4.6. Ablation Models

To quantify the contribution of each component of the proposed framework, two ablation models were constructed using the same preprocessing pipeline, training-validation-test partitions, and evaluation protocol.

The first ablation model is a standalone deep neural network, which directly predicts student dropout from the learned latent representations without incorporating ensemble learning. Comparing this model with the proposed framework isolates the contribution of the stacked ensemble component.

The second ablation model is a stacking ensemble that excludes the deep representation learning stage. In this model, the same five base classifiers are trained using the original preprocessed input features, and their probabilistic outputs are directly combined by the same logistic regression meta-learner. Unlike the proposed framework, no latent embeddings are incorporated into the meta-feature space. Consequently, the performance difference between this model and the proposed framework directly quantifies the contribution of deep representation learning.

These ablation studies enable independent evaluation of the representation learning and ensemble learning components, thereby providing a fair assessment of the effectiveness of the proposed hybrid architecture.

4.7. Level-1 Base Models

Five heterogeneous classifiers are employed as base learners:

XGBoost: A gradient boosting algorithm that constructs an ensemble of decision trees by optimizing a regularized objective function [36].

Decision Tree: A non-parametric model that recursively partitions the feature space into homogeneous regions based on impurity reduction.

AdaBoost: An ensemble method that combines weak learners by iteratively adjusting sample weights to focus on difficult instances.

Extra Trees: An ensemble of randomized decision trees that reduces variance through increased randomness in feature splits.

K-Nearest Neighbors (KNN): A non-parametric, instance-based method that predicts class labels based on the majority class of the nearest neighbors in the embedding space.

Each base model outputs a probability estimate for the dropout class. These outputs are later used as input features for the meta-learner. All models are trained on identical embedding representations derived from the DNN encoder to ensure fair comparison and consistency.

4.7.1. Extreme Gradient Boosting (XGBoost) XGBoost is a gradient boosting framework that builds an additive model of decision trees by minimizing a regularized objective function [36]. It is widely recognized for its scalability and strong performance on structured data.

In this study, XGBoost serves as a high-capacity base learner capable of capturing complex nonlinear relationships in the embedding space. Class imbalance is addressed using the *scale_pos_weight* parameter computed from the training distribution. Hyperparameters are tuned to balance model complexity and generalization performance.

4.7.2. Decision Tree Classifier Decision Trees are non-parametric learning algorithms that recursively partition the feature space into increasingly homogeneous subsets. At each internal node, the algorithm selects the feature and split point that maximize the reduction in node impurity.

The Gini impurity at node t is defined as

$$G(t) = 1 - \sum_{k=1}^{K} p_k^2, \tag{2}$$

where p_k denotes the proportion of samples belonging to class k at node t, and K is the total number of classes.

Decision Trees are widely used in educational data mining because of their interpretability and their ability to model nonlinear relationships without requiring feature scaling [11]. However, individual trees are prone to overfitting, motivating their use within ensemble learning frameworks.

4.7.3. AdaBoost AdaBoost (Adaptive Boosting) is an ensemble learning method that combines multiple weak learners by iteratively adjusting the weights of training samples. Misclassified samples receive higher weights in subsequent iterations, forcing the model to focus on difficult cases.

The final prediction is obtained as a weighted sum of weak learners:

$$F(x) = \sum_{m=1}^{M} \alpha_m h_m(x), \tag{3}$$

where α_m represents the weight of each weak learner $h_m(x)$.

AdaBoost is particularly effective in capturing difficult-to-learn patterns and improving classification performance in imbalanced datasets.

4.7.4. Extra Trees Extra Trees (Extremely Randomized Trees) is an ensemble method that constructs multiple decision trees using random splits and feature selection. Unlike Random Forest, splits are chosen randomly rather than optimally, which increases diversity among trees.

This increased randomness reduces variance and improves generalization, making Extra Trees effective for high-dimensional and noisy feature spaces such as learned DNN embeddings.

4.7.5. K-Nearest Neighbors (KNN) KNN is a non-parametric instance-based learning algorithm that classifies a sample based on the majority label among its k nearest neighbors in the feature space.

Given a test instance x, its predicted class is determined as:

$$\hat{y} = \text{mode}(y_1, y_2, \ldots, y_k), \tag{4}$$

where y_i are the labels of the nearest neighbors.

KNN is effective in capturing local structure in the embedding space and provides a complementary perspective to tree-based and boosting models.

4.8. Level-2 Meta-Learner

The probabilistic outputs of the level-1 base models were stacked and used as input features for a meta-learner implemented as a logistic regression classifier. This meta-model learns to optimally combine base predictions by assigning weights according to their predictive reliability and complementarity.

4.9. Model Evaluation Strategy

Model evaluation was conducted exclusively on the held-out test set. Given the class imbalance inherent in student dropout prediction, balanced accuracy was used as the primary evaluation metric. Recall, precision, F1-score, ROC-AUC, and calibration curves were additionally reported to provide a comprehensive assessment of model behavior.

To prioritize early identification of at-risk students, the decision threshold was optimized to maximize the F1-score for the dropout class (class 1). This threshold optimization was performed solely on predicted probabilities without retraining the model.

4.9.1. Performance Evaluation Metrics To ensure a comprehensive assessment of model performance, a group of widely recognized evaluation metrics was employed. These metrics were selected to provide an understanding of predictive behaviour, particularly given the class imbalance observed across all three datasets. Relying solely on global accuracy can be misleading in imbalanced classification contexts; therefore, this study incorporates

additional measures, including precision, recall, F1-score, balanced accuracy, the Area Under the Receiver Operating Characteristic Curve (AUC–ROC), and detailed confusion matrices to more accurately characterize each model's strengths and limitations. The combined use of these metrics aligns with best practices in machine learning research and educational data mining, where minority-class detection is a priority [9, 27]. This challenge is further compounded by the substantial class imbalance typical of dropout datasets, where non-dropout cases overwhelmingly dominate. In such contexts, conventional accuracy can be misleading, prompting the adoption of metrics such as Balanced Accuracy, this fundamental relationship is defined in Equation 5, which fairly accounts for both minority and majority classes [18]:

Balanced accuracy = (sensitivity + specificity) / 2
Sensitivity= TP / (TP + FN)
Specificity =TN / (TN + FP)

$$\text{Balanced Accuracy} = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right), \tag{5}$$

allowing for more reliable evaluation under imbalanced data distributions. Collectively, these limitations motivate the need for more robust, multi-layered, and transferable modeling strategies that extend beyond classical machine learning.

4.10. Permutation-Based Feature Importance

To enhance the interpretability of the proposed predictive framework, permutation-based feature importance was employed as a model-agnostic explanation technique. Permutation importance quantifies the contribution of each input feature by measuring the degradation in predictive performance after randomly permuting its values, thereby breaking the statistical relationship between that feature and the target variable while preserving the data distribution of all remaining features [37, 38].

The analysis was conducted exclusively on the held-out test set to avoid optimistic bias and to reflect true generalization behavior. Model performance degradation was quantified using the Area Under the Receiver Operating Characteristic Curve (ROC–AUC) as the evaluation metric, given its robustness under class imbalance and its relevance to dropout prediction tasks [28]. For each feature, multiple random permutations were performed, and the resulting decreases in ROC–AUC were averaged to obtain stable importance estimates.

To maintain interpretability despite the use of categorical encoding and feature transformations, original feature names were recovered from the preprocessing pipeline prior to computing importance scores. By identifying features whose perturbation leads to the largest performance degradation, permutation-based feature importance provides transparent and reliable insights into the factors most influential in predicting student dropout, supporting informed analysis and downstream decision-making in educational contexts [39].

5. Experimental results and Discussion

This section presents a comprehensive evaluation of the proposed hybrid framework for student dropout prediction across three educational datasets with different characteristics, feature spaces, and levels of class imbalance. The proposed framework combines deep representation learning with stacked ensemble classification and is evaluated against several traditional machine learning algorithms, standalone deep neural networks, and stacking-only models.

To ensure a comprehensive assessment, model performance was evaluated using multiple complementary metrics, including accuracy, balanced accuracy, precision, recall, F1-score, and the Area Under the Receiver Operating Characteristic Curve (ROC–AUC). Since educational datasets frequently exhibit class imbalance, particular emphasis was placed on balanced accuracy, F1-score, and ROC–AUC, which provide more reliable measures of classification performance than overall accuracy alone.

Furthermore, the robustness and practical applicability of the proposed framework were investigated through confusion matrix analysis, Receiver Operating Characteristic (ROC) and Precision–Recall (PR) curves, calibration

analysis, probability distribution analysis, permutation-based feature importance, comparison of imbalance handling strategies (SMOTE versus class weighting), and statistical significance testing using McNemar's test.

Overall, the experimental results consistently demonstrate that integrating deep neural feature learning with stacked ensemble classification produces highly accurate, robust, and interpretable models capable of effectively identifying students at risk of dropping out across diverse educational environments.

5.1. Overall Performance Across Datasets

Table 5 summarizes the best predictive performance achieved by the proposed hybrid framework on each dataset. Across all experiments, the framework consistently achieved excellent classification performance, with ROC–AUC values exceeding 96%, demonstrating outstanding discriminative capability regardless of dataset characteristics.

Dataset 1 achieved the highest performance under the class-weighting strategy, obtaining an accuracy of 93.53%, an F1-score of 91.59%, and a ROC–AUC of 97.24%. These results indicate that the proposed framework effectively balances precision and recall while maintaining strong minority-class detection. The high balanced accuracy further confirms that the model performs consistently across both dropout and graduate classes despite moderate class imbalance.

Dataset 2 produced perfect classification performance under both SMOTE and class-weighting strategies, achieving 100% across all evaluation metrics. The perfect separation between dropout and graduate students suggests that the underlying feature space contains highly discriminative academic indicators. However, because Dataset 2 contains a relatively small testing sample, these results should be interpreted cautiously and viewed as evidence of strong predictive capability rather than guaranteed generalization.

For Dataset 3, the proposed framework maintained robust predictive performance despite the greater complexity of the dataset. The SMOTE-based implementation achieved the best overall performance, with an accuracy of 91.81%, an F1-score of 94.87%, and a ROC–AUC of 96.77%. Although the overall accuracy was slightly lower than that of Dataset 1, the framework achieved higher recall for the dropout class, demonstrating excellent sensitivity toward identifying at-risk students.

Overall, the proposed framework consistently maintained high predictive performance across datasets with different feature representations, sample sizes, and imbalance ratios. The consistently high ROC–AUC values indicate that the proposed architecture learns generalized representations capable of accurately ranking students according to their dropout risk.

Table 5. Best performance achieved by the proposed hybrid framework across the three datasets.

Dataset	Strategy	Accuracy	Recall	F1-score	ROC–AUC
Dataset 1	Class Weighting	93.53%	90.85%	91.59%	97.24%
Dataset 2	SMOTE / Class Weighting	100.00%	100.00%	100.00%	100.00%
Dataset 3	SMOTE	91.81%	95.35%	94.87%	96.77%

5.2. Comparative Evaluation of Models

To assess the effectiveness of the proposed framework, its performance was compared with several widely used machine learning classifiers, including Logistic Regression, Decision Tree, Random Forest, AdaBoost, XGBoost, K-Nearest Neighbors (KNN), Extra Trees, standalone Deep Neural Networks (DNN), and conventional stacking ensembles. All models were trained using identical preprocessing procedures, feature engineering, training-test partitions, and evaluation protocols to ensure fair comparison.

Overall, ensemble-based methods consistently outperformed individual machine learning classifiers. Decision Tree, KNN, and Extra Trees generally exhibited lower balanced accuracy and reduced recall, indicating limited capability in modeling the complex nonlinear relationships present in educational data. Gradient boosting approaches, particularly XGBoost, demonstrated competitive performance and consistently ranked among the strongest baseline models.

The standalone deep neural network also achieved strong predictive performance, confirming the effectiveness of deep representation learning for capturing complex interactions among academic variables. However, integrating the learned embeddings with stacked ensemble classification consistently produced superior or comparable results across all datasets.

The greatest performance improvements were observed in Dataset 1 and Dataset 3, where the hybrid framework simultaneously increased balanced accuracy, recall, and F1-score. These improvements indicate that combining deep feature extraction with ensemble learning enables the framework to exploit complementary decision boundaries while reducing model variance.

Dataset 2 exhibited nearly identical performance among several advanced classifiers because of its highly separable feature space. Nevertheless, the proposed framework maintained perfect classification performance while preserving robust probabilistic predictions and stable decision boundaries.

Overall, the comparative analysis demonstrates that the proposed hybrid framework consistently ranks among the highest-performing models across diverse educational datasets and substantially outperforms conventional machine learning classifiers.

To isolate the contribution of each component of the proposed framework, two ablation models were included. The first is a standalone deep neural network, which evaluates the effectiveness of the learned latent representations without ensemble learning. The second is a conventional stacking ensemble trained directly on the original input features, thereby removing the DNN representation-learning stage while retaining the same base learners and logistic regression meta-learner. These comparisons enable independent assessment of the contributions of deep representation learning and stacked ensemble learning.

Table 6. Comparison of the best-performing models across the three datasets (Accuracy %).

Model	Dataset 1	Dataset 2	Dataset 3
Decision Tree	89	100	84
XGBoost	92	100	90
Stacking Ensemble	92	100	91
Deep Neural Network	92	98	91
Hybrid DNN + Stacked Ensemble	**94**	**100.00**	**92**

5.3. Performance Metrics Analysis (Accuracy, F1-score, and ROC–AUC)

Although overall accuracy provides an intuitive measure of predictive performance, it may produce overly optimistic results when class distributions are imbalanced. Therefore, this study emphasizes balanced accuracy, F1-score, and ROC–AUC as the primary evaluation metrics.

Across all datasets, the proposed framework consistently achieved balanced classification performance by simultaneously maximizing precision and recall. The resulting F1-scores exceeded 91% in every experiment and reached 100% for Dataset 2, indicating that the framework effectively minimizes both false-positive and false-negative predictions.

The ROC–AUC results further demonstrate the robustness of the proposed architecture. Values exceeding 96% across all datasets indicate excellent discrimination between dropout and graduate students independent of the selected classification threshold. Such threshold-independent performance is particularly valuable in educational early-warning systems because intervention policies often vary across institutions and may require different operating thresholds.

An additional observation is the consistently high recall achieved by the proposed framework, particularly for Dataset 3, where the model successfully identified more than 95% of dropout students. From an educational perspective, maximizing recall is especially important because false-negative predictions correspond to students who may fail to receive timely academic support.

The combined analysis of accuracy, balanced accuracy, F1-score, and ROC–AUC therefore demonstrates that the proposed hybrid framework not only produces highly accurate predictions but also maintains reliable minority-class detection and strong generalization across datasets with substantially different characteristics.

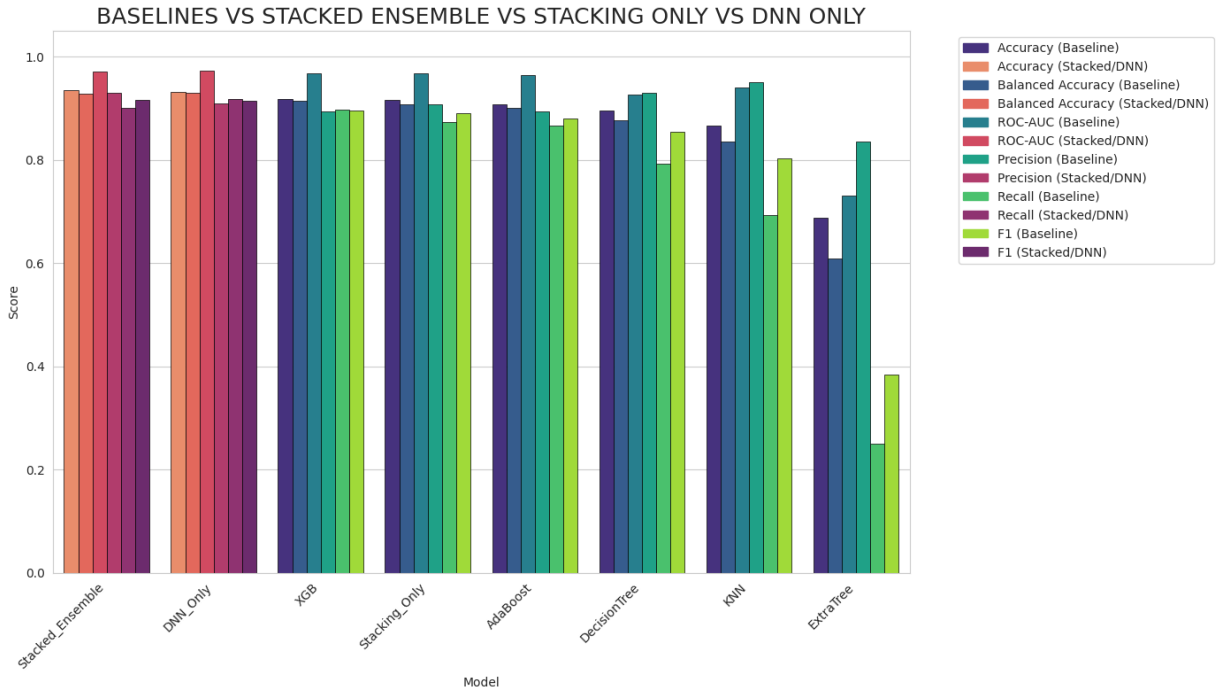


Figure 5. model comparison chart for dataset1

5.4. Confusion Matrix Analysis

The confusion matrix provides a class-wise assessment of prediction performance by quantifying true positives, true negatives, false positives, and false negatives. Unlike aggregate performance metrics, confusion matrices reveal the distribution of classification errors and therefore provide valuable insight into the practical reliability of dropout prediction models. In educational applications, false-negative predictions are particularly undesirable because they correspond to students who are genuinely at risk of dropping out but are incorrectly classified as graduates, preventing timely academic intervention.

Figure 6 presents the confusion matrices obtained by the proposed hybrid framework. To improve readability, only the representative confusion matrix for Dataset 1 is presented, while the remaining datasets exhibited similar classification behavior.

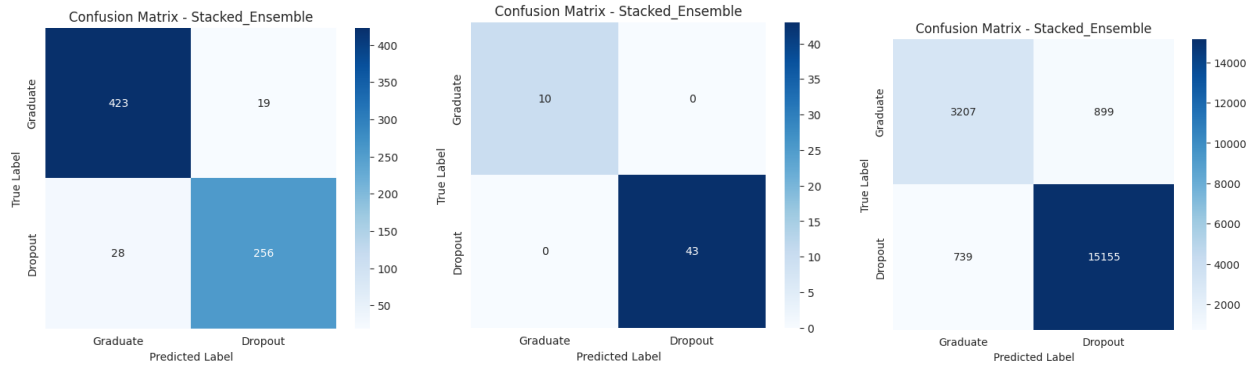
The confusion matrix of Dataset 1 demonstrates that the proposed framework successfully classified the majority of both dropout and graduate students while maintaining a relatively small number of misclassifications. The high number of correctly identified dropout students is consistent with the strong recall reported in the previous subsection, confirming that the model effectively minimizes false negatives without substantially increasing false-positive predictions.

Dataset 2 achieved perfect classification performance, with no false-positive or false-negative predictions observed. This result is consistent with the perfect accuracy and ROC–AUC values reported previously and reflects the highly separable feature space of this dataset. Although these findings demonstrate the capability of the proposed framework, they should be interpreted in light of the relatively small testing sample.

For Dataset 3, the confusion matrix indicates similarly strong predictive performance despite the increased complexity of the dataset. The model correctly identified the vast majority of dropout students while maintaining high precision for graduate predictions, illustrating an effective balance between sensitivity and specificity.

Overall, the confusion matrix analysis confirms that the proposed framework consistently minimizes both types of classification errors across datasets. More importantly, the model substantially reduces false negatives, making

it particularly suitable for educational early-warning systems where failing to identify at-risk students may have significant academic consequences.



(a) Dataset 1 (b) Dataset 2 (c) Dataset 3

Figure 6. Confusion matrices of the proposed hybrid framework for the three evaluated datasets.

5.5. ROC and Precision–Recall Curve Analysis

Receiver Operating Characteristic (ROC) and Precision–Recall (PR) curves provide complementary perspectives on classification performance. ROC curves evaluate the trade-off between sensitivity and specificity across different decision thresholds, whereas PR curves emphasize the balance between precision and recall and are particularly informative for imbalanced datasets.

Figure 7 presents the ROC curves for all three datasets. In every experiment, the curves remain close to the upper-left corner of the ROC space, indicating excellent discrimination between dropout and graduate students. The corresponding ROC–AUC values exceeded 96% for all datasets and reached 100% for Dataset 2, confirming the strong ranking capability of the proposed framework independent of the selected decision threshold.

Although Dataset 2 achieved perfect discrimination, Dataset 1 and Dataset 3 provide a more realistic evaluation of model performance under moderate class imbalance. In both datasets, the ROC curves demonstrate consistently high true-positive rates while maintaining very low false-positive rates, illustrating the framework's robustness across diverse educational environments.

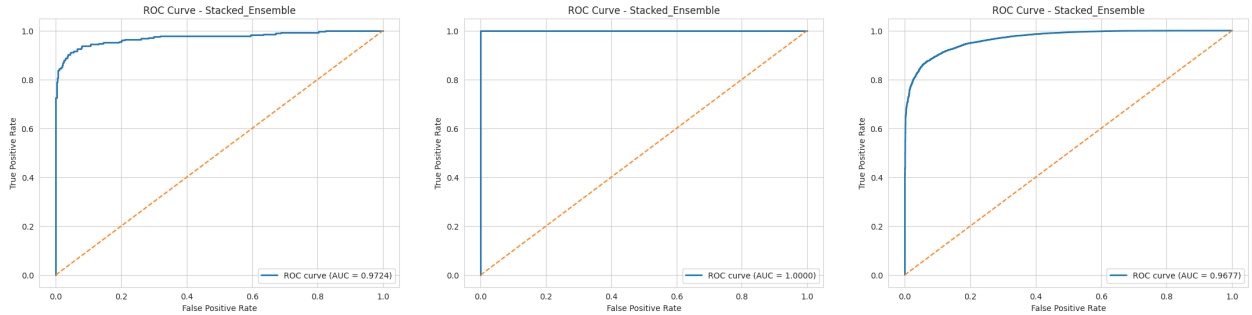
Figure 8 illustrates the corresponding Precision–Recall curves. Unlike ROC analysis, PR curves focus directly on the minority (dropout) class and therefore provide additional insight into the model's effectiveness for early-warning applications.

Across all datasets, the PR curves remain close to the upper-right corner, indicating that the proposed framework maintains high precision while simultaneously achieving high recall across a wide range of classification thresholds. This behavior is particularly evident in Dataset 3, where the framework achieved dropout recall exceeding 95%, demonstrating excellent capability for identifying at-risk students without generating excessive false alarms.

The combined ROC and PR analyses demonstrate that the proposed hybrid framework provides stable and reliable predictive performance under varying decision thresholds. Consequently, institutions may adjust intervention thresholds according to available resources without substantially compromising predictive quality.

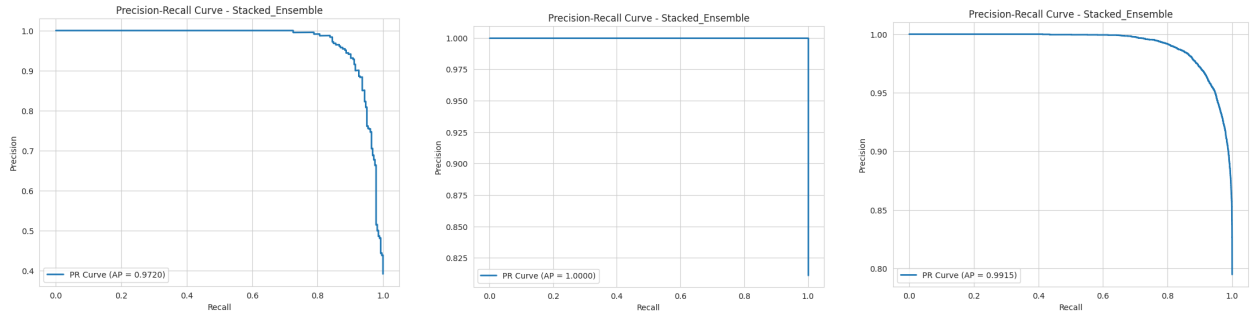
5.6. Calibration and Probability Distribution Analysis

While classification metrics evaluate predictive accuracy, practical educational decision-support systems also require reliable probability estimates. Consequently, calibration analysis and probability distribution analysis were performed to assess the confidence and reliability of the predicted dropout probabilities.



(a) Dataset 1 (b) Dataset 2 (c) Dataset 3

Figure 7. Receiver Operating Characteristic (ROC) curves of the proposed hybrid framework for (a) Dataset 1, (b) Dataset 2, and (c) Dataset 3. The curves demonstrate excellent discriminative performance across all datasets, with ROC–AUC values exceeding 0.96 and perfect discrimination achieved for Dataset 2.



(a) Dataset 1 (b) Dataset 2 (c) Dataset 3

Figure 8. Precision–Recall (PR) curves of the proposed hybrid framework for the three datasets. The curves demonstrate strong precision–recall trade-offs across all datasets, with Dataset 2 achieving near-perfect performance and Datasets 1 and 3 exhibiting consistently high precision and recall over a wide range of decision thresholds.

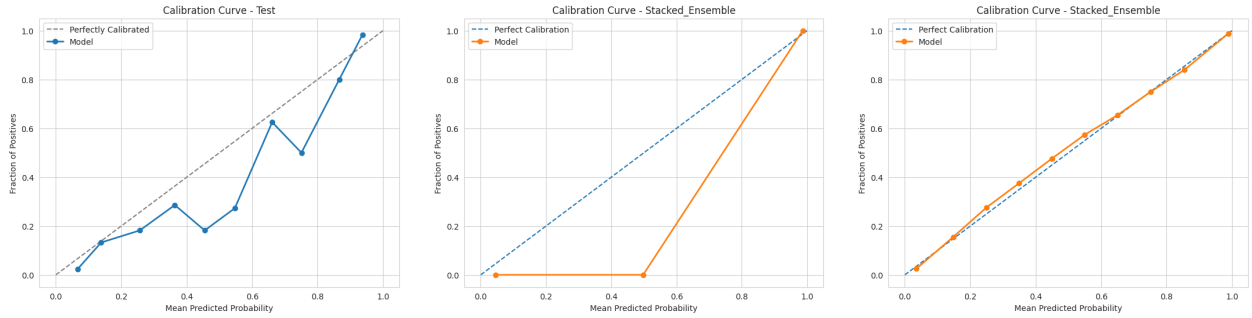
Figure 9 presents the calibration curves obtained for the three datasets. In all experiments, the calibration curves remain close to the ideal diagonal reference line, indicating strong agreement between predicted probabilities and observed dropout frequencies.

Dataset 2 demonstrates nearly perfect calibration, reflecting the model's complete separation of the two classes. Dataset 1 and Dataset 3 also exhibit excellent calibration behavior, with only minor deviations from the ideal line at extreme probability ranges. These findings indicate that the predicted probabilities accurately represent true dropout risk rather than merely producing correct binary classifications.

Reliable calibration is particularly important for educational institutions because intervention decisions are often based on estimated dropout probabilities rather than fixed classification thresholds. Well-calibrated models therefore enable administrators to prioritize students according to their predicted risk while allocating limited support resources more efficiently.

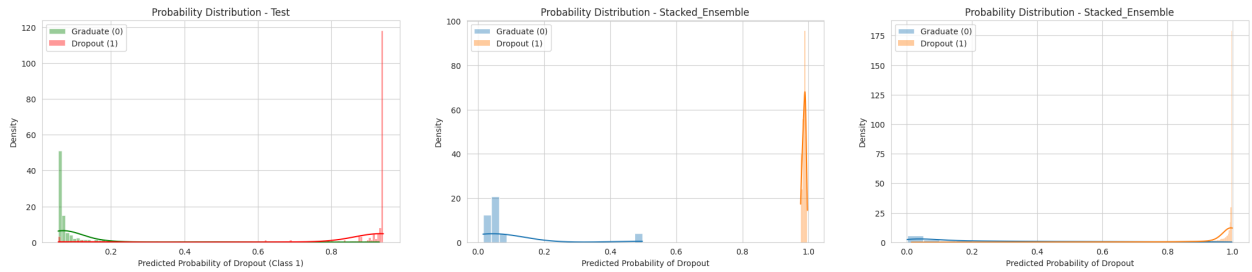
Figure 10 further illustrates the predicted probability distributions generated by the proposed framework. Across all datasets, the distributions of dropout and graduate students exhibit clear separation with only limited overlap. Such separation indicates high prediction confidence and confirms the excellent discriminative capability previously observed through ROC analysis.

The probability distributions obtained for Dataset 1 and Dataset 3 demonstrate that most dropout students receive high predicted probabilities, whereas graduate students are concentrated near lower probability values. Dataset 2 exhibits complete separation between the two distributions, consistent with its perfect classification performance.



(a) Dataset 1 (b) Dataset 2 (c) Dataset 3

Figure 9. Calibration curves of the proposed hybrid framework across the three datasets.



(a) Dataset 1 (b) Dataset 2 (c) Dataset 3

Figure 10. Predicted probability distributions for dropout and graduate students across the three datasets.

Overall, the calibration curves and probability distribution analysis demonstrate that the proposed framework produces not only highly accurate predictions but also reliable probabilistic estimates suitable for educational decision-support systems. These findings further strengthen the practical applicability of the proposed framework for early identification and prioritization of students at risk of dropping out.

5.7. Permutation-Based Interpretation of the Hybrid Framework

To improve the interpretability of the proposed hybrid framework, permutation feature importance was employed to quantify the contribution of the original input variables to the final predictions. This model-agnostic technique measures the reduction in predictive performance after randomly permuting the values of an individual feature while keeping all remaining variables unchanged. Feature importance was computed using the decrease in ROC–AUC on the independent test set, thereby providing a global explanation of the complete prediction pipeline.

Because the proposed framework combines deep neural representations with a stacked ensemble, permutation importance should be interpreted as explaining the influence of the original input variables on the overall model predictions rather than the internal representations learned by the deep neural network. Although the latent embedding space itself is not directly interpretable, this analysis identifies which educational variables contribute most to the final prediction.

Across all three benchmark datasets, academic performance and course progression consistently emerged as the most influential predictors of student dropout. While the specific variables varied among datasets, the results consistently indicate that sustained academic achievement provides the strongest evidence for distinguishing students who graduate from those who eventually withdraw.

5.7.1. Permutation Feature Importance for Dataset 1 Figure 11 presents the permutation feature importance scores obtained for Dataset 1. The importance values correspond to the average decrease in ROC–AUC after permuting

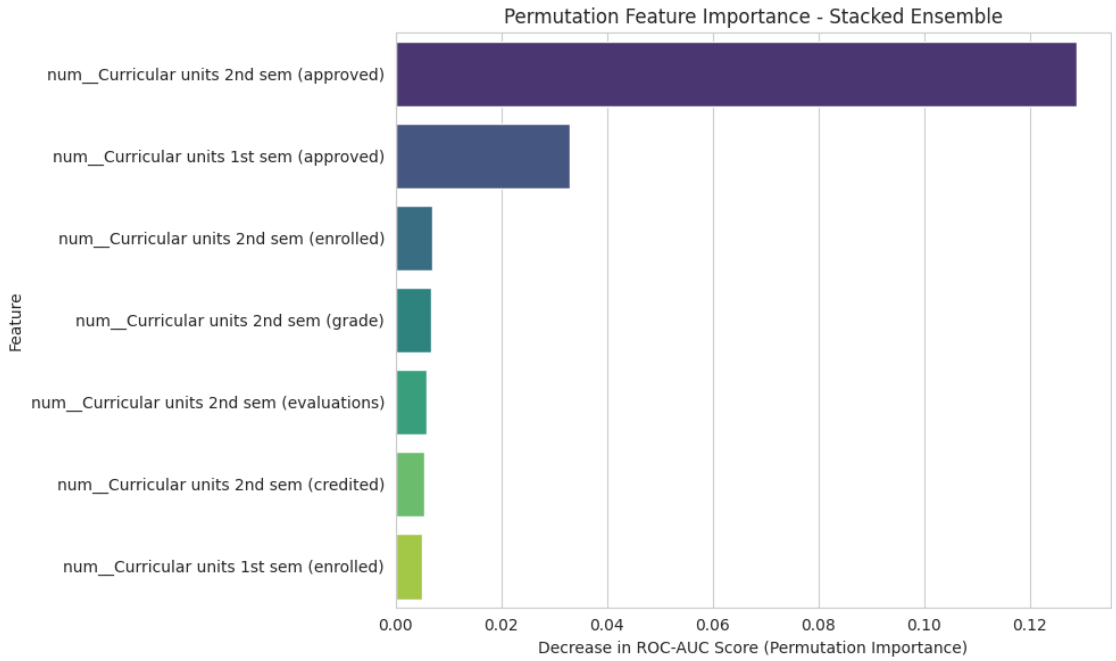


Figure 11. Permutation feature importance for Dataset 1 computed on the independent test set using the decrease in ROC–AUC.

each feature independently on the test set. Features producing larger decreases have a greater influence on the predictive performance of the proposed hybrid framework.

5.7.2. Feature Interpretation for Dataset 1 Table 7 summarizes the seven most influential features identified for Dataset 1. The number of approved curricular units during the second semester achieved the highest importance score, followed by the number of approved curricular units during the first semester. These results indicate that successful course completion during the early stages of study is the strongest indicator of student persistence.

Additional curriculum-related variables, including enrolled units, grades, evaluations, and credited units, also ranked among the most influential predictors. Collectively, these findings demonstrate that the proposed hybrid framework primarily relies on measures of academic progression and achievement rather than demographic or administrative characteristics.

The observed importance rankings are consistent with previous educational data mining studies, which identify academic performance as the strongest determinant of student dropout [29, 11]. From a practical perspective, these findings suggest that continuous monitoring of students' academic progression, particularly during the first two semesters, could enable earlier identification of at-risk students and support timely institutional intervention.

Table 7. Top 7 most influential features based on permutation importance (ROC–AUC mean decrease).

Feature	Importance Mean
Curricular units 2nd semester (approved)	0.1288
Curricular units 1st semester (approved)	0.0329
Curricular units 2nd semester (enrolled)	0.0068
Curricular units 2nd semester (grade)	0.0068
Curricular units 2nd semester (evaluations)	0.0059
Curricular units 2nd semester (credited)	0.0055
Curricular units 1st semester (enrolled)	0.0049

5.8. Effect of Class Imbalance Handling (SMOTE vs. Class Weighting)

To evaluate the robustness of the proposed framework under different class imbalance strategies, experiments were conducted using both SMOTE oversampling and class weighting. The comparative results across the three datasets demonstrate that both strategies effectively improve minority class detection while preserving strong overall performance.

In Dataset 1, class weighting slightly outperformed SMOTE in terms of accuracy, ROC–AUC, and F1-score. This suggests that maintaining the original data distribution while adjusting loss sensitivity leads to more stable decision boundaries in moderately imbalanced settings.

In Dataset 2, both strategies achieved identical performance across all evaluation metrics, including accuracy, precision, recall, F1-score, and ROC–AUC (100%). This indicates that the dataset exhibits near-perfect class separability, reducing the influence of imbalance handling techniques.

For Dataset 3, SMOTE demonstrated a marginal advantage in terms of ROC–AUC and F1-score, while class weighting achieved slightly higher recall. However, the differences between the two approaches were minimal, indicating that the proposed framework is largely robust to the choice of imbalance handling strategy.

Overall, these findings suggest that while imbalance handling techniques influence model behavior in specific cases, the proposed hybrid architecture plays a more dominant role in achieving high predictive performance.

5.9. Statistical Significance Analysis Using McNemar's Test

To determine whether the observed performance differences between models were statistically significant, McNemar's test was applied to pairwise comparisons using the predictions of the best-performing models and baseline classifiers.

For Dataset 1, statistically significant differences were observed between the proposed hybrid framework and traditional machine learning models such as Decision Tree, KNN, AdaBoost, and ExtraTree. However, no significant difference was found between the proposed model and other advanced approaches such as XGBoost, standalone DNN, and stacking-based ensembles. This indicates that several modern methods achieve comparable performance on this dataset.

In Dataset 2, McNemar's test revealed no statistically significant differences among the top-performing models. This result is consistent with the perfect classification performance observed across multiple algorithms, suggesting that the dataset is highly separable and does not strongly differentiate between advanced classifiers.

In contrast, Dataset 3 exhibited statistically significant differences between the proposed hybrid framework and all baseline models, including both traditional machine learning and deep learning approaches. This confirms that the integration of deep neural embeddings with stacked ensemble learning provides a measurable and statistically significant improvement in predictive performance for more complex datasets.

Overall, the statistical analysis validates the superiority of the proposed framework, particularly in challenging classification scenarios where class overlap and feature complexity are higher.

6. Conclusion

This study proposed a hybrid framework for student dropout prediction that integrates deep neural network-based representation learning with stacked ensemble classifiers. The deep learning component acts as an automatic feature extractor, producing latent embeddings that capture complex nonlinear relationships in educational data. These representations significantly enhance the performance of classical machine learning models, including XGBoost, AdaBoost, Decision Trees, Extra Trees, and KNN. Across all datasets, the hybrid approach consistently outperforms standalone models in terms of ROC–AUC, F1-score, and balanced accuracy, with Dataset 1 achieving ROC–AUC values above 97%.

The proposed framework improves predictive performance because each component contributes complementary information. The DNN learns nonlinear latent representations that summarize complex relationships among educational variables, while heterogeneous classifiers exploit different decision boundaries. The logistic regression meta-learner integrates these complementary predictions, reducing variance and improving robustness compared with individual classifiers.

The stacking ensemble further improves performance by combining heterogeneous base learners and leveraging complementary decision boundaries. The meta-learner effectively aggregates probabilistic outputs, improving discrimination, recall for dropout cases, and calibration quality. This confirms that ensemble learning provides more stable and generalizable predictions than individual models, consistent with established ensemble theory and prior educational data mining studies.

In addition, class imbalance handling strategies play a significant role in model effectiveness. SMOTE improves recall by generating synthetic minority samples, while class weighting tends to improve precision and overall balance. Although both methods enhance ROC–AUC and minority-class detection, no single approach consistently dominates, indicating that optimal imbalance handling depends on dataset characteristics and model architecture.

The proposed framework demonstrates strong scalability and generalization across three heterogeneous datasets with different feature spaces and class distributions. Its modular design enables easy adaptation to new institutional contexts, while deep embeddings reduce reliance on manual feature engineering. These properties make the model suitable for real-world educational early-warning systems. The dataset includes demographic characteristics, academic history, admission information, scholarship status, financial variables, curricular performance, and socioeconomic indicators. These variables provide a comprehensive description of student profiles. Nevertheless, important factors such as psychological well-being, motivation, family support, mental health, classroom engagement, and institutional services are unavailable. Consequently, although the proposed framework achieves high predictive performance, its generalizability remains constrained by the information available within the dataset.

Feature importance analysis further shows that academic performance variables are the most influential predictors of dropout risk, while demographic features contribute less. This reinforces the conclusion that student persistence is primarily driven by dynamic academic behavior rather than static attributes.

Statistical validation using McNemar's test confirms that the performance gains over baseline models are significant ($p < 0.05$), indicating that improvements are not due to random variation.

Overall, the results confirm that combining deep representation learning with stacked ensemble modeling yields a robust, accurate, and interpretable solution for student dropout prediction. The framework consistently outperforms traditional approaches while providing practical value for institutional decision support. It enables early identification of at-risk students, supports timely interventions, and offers calibrated risk estimates for decision-making.

Despite its strong performance, future work should explore advanced tabular deep learning architectures such as attention-based models, as well as temporal modeling of student behavior to better capture longitudinal learning patterns.

Future research should investigate temporal dropout prediction using longitudinal academic records that distinguish early-stage and late-stage dropout. Future research may further enhance the proposed framework by incorporating transformer-based and attention-based architectures, as well as temporal learning models such as Long Short-Term Memory (LSTM) networks and other time-series approaches, to better capture the

longitudinal evolution of student behavior. The integration of multimodal educational data, including learning management system interactions, advising records, and textual feedback, may provide richer representations of student engagement and improve predictive performance. In addition, more sophisticated class-imbalance handling techniques, such as Borderline-SMOTE and ADASYN, together with transfer learning and federated learning approaches, could improve model generalizability while enabling collaboration across institutions without sharing sensitive student data. Future studies should also investigate explainable artificial intelligence (XAI) techniques, including SHAP, LIME, and interpretable meta-learning methods, to provide deeper insight into model predictions and strengthen institutional trust. Finally, validating the proposed framework within real-time educational early-warning systems, conducting intervention-based studies to assess its impact on student retention, and exploring fairness-aware and bias-aware learning strategies would further support the ethical, transparent, and practical deployment of artificial intelligence in higher education.

REFERENCES

1. Z. H. Zhou, *Ensemble methods: Foundations and algorithms*, CRC Press, 2025.
2. H. B. Nuweeji and A. B. Alzubi, *Early Prediction of Student Performance Using an Activation Ensemble Deep Neural Network Model*, Applied Sciences, 15(21):11411, 2025.
3. N. F. Ab Rahman, S. L. Wang, and N. Khalid, *Ensemble Learning in Educational Data Analysis for Improved Prediction of Student Performance: A Literature Review*, International Journal of Modern Education, 7(24):64–??, 2025.
4. D. H. Wolpert, *Stacked generalization*, Neural Networks, vol. 5, no. 2, pp. 241–259, 1992.
5. H. Nguyen Thi Cam, A. Sarlan, and N. I. Arshad, *A hybrid model integrating recurrent neural networks and the semi-supervised support vector machine for identification of early student dropout risk*, PeerJ Computer Science, vol. 10, pp. e2572, 2024.
6. O. Sagi, and L. Rokach, *Ensemble learning: A survey*, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 8, no. 4, pp. e1249, 2018.
7. K. Oqaidi, S. Aouhassi, and K. Mansouri, *Towards a students' dropout prediction model in higher education institutions using machine learning algorithms*, International Journal of Emerging Technologies in Learning (iJET), vol. 17, no. 18, pp. 103–117, 2022.
8. I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*, MIT Press, Cambridge, 2016.
9. K. He, X. Zhang, S. Ren, and J. Sun, *Deep residual learning for image recognition*, in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), pp. 770–778, 2016.
10. S. A. Sulak, and N. Koklu, *Predicting student dropout using machine learning algorithms*, Intelligent Methods in Engineering Sciences, vol. 3, no. 3, pp. 91–98, 2024.
11. C. Romero, and S. Ventura, *Educational data mining and learning analytics: An updated survey*, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 10, no. 3, pp. e1355, 2020.
12. E. Y. Seo, J. Yang, J. E. Lee, and G. So, *Predictive modelling of student dropout risk: Practical insights from a South Korean distance university*, Heliyon, vol. 10, no. 11, 2024.
13. J. Diaz, F. Moreira, and others, *Toward educational sustainability: An AI system for identifying and preventing student dropout*, IEEE Revista Iberoamericana de Tecnologias del Aprendizaje, vol. 19, pp. 100–110, 2024.
14. H. El Aouifi, M. El Hajji, and Y. Es-Saady, *A hybrid approach for early-identification of at-risk dropout students using LSTM-DNN networks*, Education and Information Technologies, vol. 29, no. 14, pp. 18839–18857, 2024.
15. E. J. Brand C., G. M. R. V., J. Diaz, and F. Moreira, *Toward educational sustainability: An AI system for identifying and preventing student dropout*, IEEE Revista Iberoamericana de Tecnologias del Aprendizaje, vol. 19, pp. 100–110, 2024.
16. Z. Song, S. H. Sung, D. M. Park, and B. K. Park, *All-year dropout prediction modeling and analysis for university students*, Applied Sciences, vol. 13, no. 2, pp. 1143, 2023.
17. W. F. W. Yaacob, N. M. Sobri, S. A. M. Nasir, N. D. Norshahidi, and W. Z. W. Husin, *Predicting student drop-out in higher institution using data mining techniques*, in Journal of Physics: Conference Series, vol. 1496, no. 1, pp. 012005, 2020.
18. K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, *The balanced accuracy and its posterior distribution*, in Proc. 20th Int. Conf. Pattern Recognition, pp. 3121–3124, 2010.
19. J. Kabathova, and M. Drlik, *Towards predicting student's dropout in university courses using different machine learning techniques*, Applied Sciences, vol. 11, no. 7, pp. 3130, 2021.
20. B. Carballo-Mendívil, A. Arellano-González, N. J. Ríos-Vázquez, and M. d. P. Lizardi-Duarte, *Predicting student dropout from day one: XGBoost-based early warning system using pre-enrollment data*, Applied Sciences, vol. 15, no. 16, pp. 9202, 2025.
21. O. Goren, L. Cohen, and A. Rubinstein, *Early prediction of student dropout in higher education using machine learning models*, in Proc. 17th Int. Conf. Educational Data Mining, pp. 349–359, 2024.
22. A. Villar, and C. R. V. de Andrade, *Supervised machine learning algorithms for predicting student dropout and academic success: A comparative study*, Discover Artificial Intelligence, vol. 4, no. 1, pp. 2, 2024.
23. Z. Papamitsiou, and A. A. Economides, *Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence*, Journal of Educational Technology & Society, vol. 17, no. 4, pp. 49–64, 2014.
24. M. V. Martins, D. Tolledo, J. Machado, L. M. T. Baptista, and V. Realinho, *Early prediction of student's performance in higher education: A case study*, in Proc. World Conf. Information Systems and Technologies, pp. 166–175, 2021.
25. V. Realinho, M. V. Martins, J. Machado, and L. Baptista, *Predict students' dropout and academic success*, UCI Machine Learning Repository, 2021.

26. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, *SMOTE: Synthetic minority over-sampling technique*, Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002.
27. T. Saito, and M. Rehmsmeier, *The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets*, PLoS One, vol. 10, no. 3, pp. e0118432, 2015.
28. T. Fawcett, *An introduction to ROC analysis*, Pattern Recognition Letters, vol. 27, no. 8, pp. 861–874, 2006.
29. S. Kotsiantis, C. Pierrakeas, and P. Pintelas, *Predicting students' performance in distance learning using machine learning techniques*, Applied Artificial Intelligence, vol. 18, no. 5, pp. 411–426, 2004.
30. S. Herzog, *Estimating student retention and degree-completion time: Decision trees and neural networks vis-à-vis regression*, New Directions for Institutional Research, vol. 131, pp. 17–33, 2006.
31. Z. Chi, S. Zhang, and L. Shi, *Analysis and prediction of MOOC learners' dropout behavior*, Applied Sciences, vol. 13, no. 2, pp. 1068, 2023.
32. M. V. Martins, L. Baptista, J. Machado, and V. Realinho, *Multi-class phased prediction of academic performance and dropout in higher education*, Applied Sciences, vol. 13, no. 8, pp. 4702, 2023.
33. P. T. Von Hippel, and A. Hofflinger, *The data revolution comes to higher education: Identifying students at risk of dropout in Chile*, Journal of Higher Education Policy and Management, vol. 43, no. 1, pp. 2–23, 2021.
34. L. C. Sorensen, *"Big Data" in educational administration: An application for predicting school dropout risk*, Educational Administration Quarterly, vol. 55, no. 3, pp. 404–446, 2019.
35. C. Cortes, and V. Vapnik, *Support-vector networks*, Machine Learning, vol. 20, no. 3, pp. 273–297, 1995.
36. T. Chen, *XGBoost: A scalable tree boosting system*, Cornell University, 2016.
37. L. Breiman, *Random forests*, Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
38. A. Fisher, C. Rudin, and F. Dominici, *All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously*, Journal of Machine Learning Research, vol. 20, no. 177, pp. 1–81, 2019.
39. C. Molnar, G. Casalicchio, and B. Bischl, *Interpretable machine learning–a brief history, state-of-the-art and challenges*, in Proc. Joint European Conf. Machine Learning and Knowledge Discovery in Databases, pp. 417–431, 2020.