International Journal of Innovation and Scientific Research
ISSN 2336-0046 Vol. 81 No. 2 Nov. 2025, pp. 113-119

http://www.ijisr.issr-journals.org/

Comparative analysis of machine learning and deep learning approaches for predicting student dropout in higher education

Achi Harrisson Thiziers and Koné Moussa

ISN Training and Research Unit, Virtual University of Côte d'Ivoire, Abidjan, Côte d'Ivoire

ABSTRACT: Student dropout rates are a major challenge for higher education. Both parents and academic institutions are seeking to reduce this phenomenon by investigating its root causes, as it has economic, social and institutional consequences. Based on an academic dataset enriched by SMOTE balancing, one-hot encoding and scaling, this article explores the application of machine learning (Random Forest, Gradient Boosting) and deep learning (MLP and CNN) techniques to predict this phenomenon and consider possible solutions. Exploring various data, such as grades, absences, failures, study time and family support, the models were compared through these metrics: Accuracy, F1-score, AUC-ROC, PR-curve. The results reveal the effectiveness of Gradient Boosting and Random Forest models (with an F1-score close to 1) over those of Multilayer Perceptron (F1-score = 0.84) and Convolutional Neural Networks (F1-score = 0,82). Analysis of the variables confirms the importance of mid-term marks (G2), absences and previous failures as key predictors. The article provides recommendations, including the parameters to be taken into account in early prediction, and opens up prospects for future work.

KEYWORDS: student dropout, machine learning, early prediction, random forest, gradient boosting.

1 INTRODUCTION AND STATE-OF-THE ART

1.1 OBJECTIVES

The phenomenon of dropping out of higher education remains a serious problem worldwide and has significant repercussions for individuals and society. It not only affects students' academic careers, but also generates economic and institutional losses for higher education institutions. In this context, the development of highly accurate and useful predictive indicators for identifying at-risk students over time is necessary in order to plan targeted actions.

In the academic environment, the application of artificial intelligence, more specifically machine learning and deep learning, has shown a certain degree of effectiveness in the science of prediction. These models are capable of tracing patterns from a wealth of varied information, including socio-demographic parameters, academic data and behaviours inherited from information systems. Although several studies have proven the effectiveness of classical supervised algorithms such as Random Forests (RF), Support Vector Machines (SVM) or ensemble methods such as XGBoost or LightGBM, more recent work has explored more sophisticated models, including Deep Neural Networks and even Graph Neural Networks (GNN) to capture latent relationships between students. However, there is a lack of studies offering a systematic and in-depth comparison of the performance of these different approaches on real-world datasets in higher education.

This research aims to fill this gap by conducting a comprehensive evaluation of the performance of various machine learning (ML) and deep learning (DL) models in predicting university dropout. We analyse their performance on the models in question, evaluating various assessment metrics and investigating the effects of class imbalance and feature selection techniques. The goal is to find the most robust and interpretable approaches to aid decision-making in student retention policy [1], [2], [3].

1.2 RELATED WORKS

Traditional approaches to machine learning Traditional machine learning (ML) algorithms remain common tools for predicting university dropout rates due to their robustness and interpretability. Methods such as decision trees (DT), random forests (RF), support

vector machines (SVM), and boosting algorithms such as XGBoost, CatBoost, and LightGBM are widely used. Villar and Andrade [4] and Demirtürk, B. [5] compared several of these algorithms on an unbalanced dataset, demonstrating the superiority of boosting models, particularly LightGBM optimised with Optuna. Other recent work has highlighted the effectiveness of stacking traditional models to improve predictive accuracy while managing noisy or unbalanced datasets [6], [7]. Deep neural networks (DNNs) are increasingly used to capture non-linear dependencies between predictive variables. However, their success depends heavily on data quality and regularisation techniques, such as dropout [8]. Some studies show that neural networks outperform traditional models in specific configurations, particularly when optimised through grid search or when they incorporate temporal or behavioural data [9], [10]. Despite their potential, their lack of interpretability is a barrier to their adoption in institutional educational contexts. Hybrid models, particularly those based on stacking, are gaining ground in approaches to predicting dropout. The study by Niyogisubizo et al. [11] proposed a two-layer model combining RF, XGBoost, Gradient Boosting and neural networks, achieving superior performance to single models. These approaches offer a compromise between accuracy and robustness by aggregating the strengths of different types of algorithms, while mitigating their individual weaknesses. With the widespread use of digital learning environments, data from LMS platforms (such as Moodle) has become a valuable source for detecting early warning signs of dropout. Vaarma and Li [12] showed that LMS data (connection frequency, activity on Moodle) has predictive power comparable to, or even superior to, demographic and performance data. Marcolino et al. [13] used CatBoost to analyse weekly Moodle logs, with notable results in terms of recall (F1 score ≈ 0.8 for the minority class). A more recent advance is the use of Graph Neural Networks (GNN) to model implicit relationships between students. Almeida et al. [11] compared GCN and GraphSAGE models to tabular models (RF, XGBoost, TabNet), showing that certain GNN configurations outperformed classical models when a relevant graph structure was derived using clustering techniques (PCA + KMeans). However, graph construction remains a sensitive and decisive step in the performance obtained. Given that dropout data is often imbalanced (many more students succeed than drop out), several techniques have been proposed to improve the representativeness of the minority class. Among the most recent are SMOTE, ADASYN and multi-objective optimization using NSGA-II. Marcolino demonstrated the value of combining ADASYN with CatBoost, while Villar and Andrade used Optuna to refine the hyperparameters. These studies confirm that pre-processing and optimization have a direct and significant effect on model performance.

2 METHODOLOGY

This study offers a comparative analysis of several machine learning (ML) and deep learning (DL) models for predicting higher education dropout rates. The methodology adopted is structured around the following key stages: data preparation, pre-processing, rebalancing, modelling, and evaluation. The complete pipeline is illustrated in Figure 1.

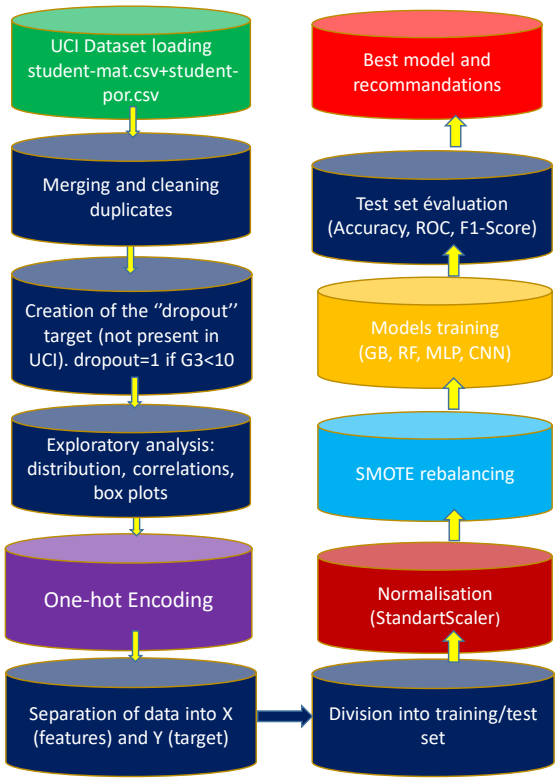

Fig. 1. Pipeline of our Methodology

2.1 DATASET

The online database used is that of the UCI repository [14]. This multivariate data comes from the public UCI Student Performance Dataset, comprising two CSV files: `student-mat.csv` and `student-por.csv`. These files describe the academic performance of Portuguese students in mathematics and Portuguese, including sociodemographic, educational, and behavioural variables. The two files are merged and duplicates are removed to form a single analysis corpus. A dropout target variable was derived: a student is considered to be a 'dropout' (dropout = 1) if their final G3 grade is strictly less than 10, and 0 otherwise. We paid special attention to the following variables: such as grades, absences, failures, study time, higher and family support.

2.1.1 PRE-PROCESSING

One-hot encoding was applied to qualitative variables, and numerical features were normalized using a StandardScaler. The dataset was divided into a training set (80%) and a test set (2%), stratifying the target variable in order to maintain class proportions.

2.1.2 CLASS REBALANCING

As the dropout target variable was unbalanced, the SMOTE (Synthetic Minority Over-sampling Technique) was used on the training set to oversample the minority class.

2.1.3 MODELING

Four models were trained on the pre-processed and re-balanced data:

- Gradient Boosting Classifier: n_estimators=100, learning_rate=0.1, max_depth=6
- Random Forest Classifier: n_estimators=100
- MLPClassifier (classical neural network): 3 hidden layers (64-32-16), ReLU activation, Adam optimizer
- Convolutional Neural Network (CNN): one Conv1D layer, followed by a GlobalMaxPooling1D layer, then two dense layers with dropout. The CNN architecture is adapted to tabular inputs thanks to data formatting via reshape (1D).

2.1.4 DATASET OVERVIEW AFTER PRE-PROCESSING

Dataset size: (1044, 34)

RangeIndex: 1044 entries, 0 to 1043

Distribution of target variable:

Dropout

2.1.5 NAME: PROPORTION, DTYPE: FLOAT64

Size after balancing: (1302, 42)

Distribution after SMOTE: [651 651]

Figure 2 summarize abandon distribution

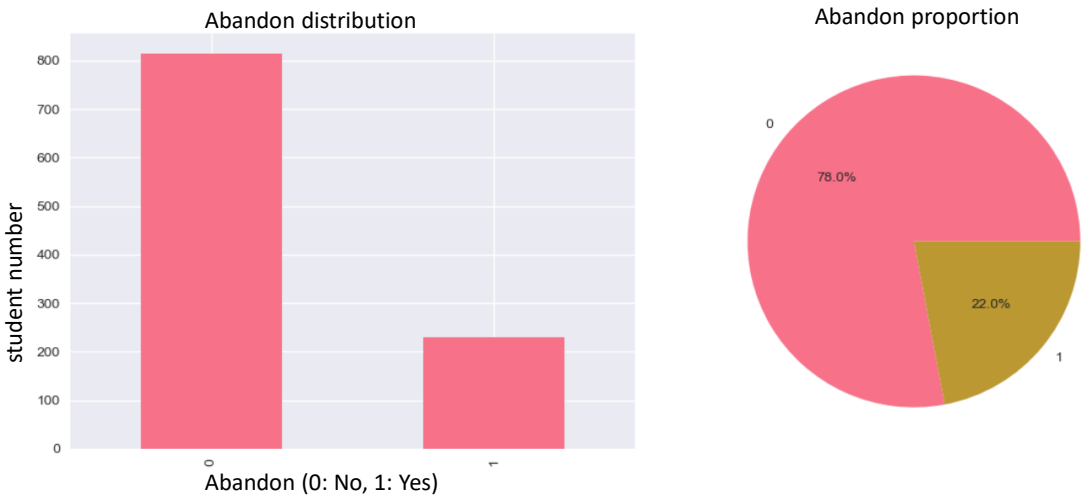

Fig. 2. Abandon distribution

2.2 MATHEMATICAL MODELING

Considering our dataset, the notation and target variable could be presented by formulas below:

$$D = \{(xi, yi)\}_{i=1}^{n}, xi \in Rp, yi \in \{0, 1\}$$

The target variable is derived from the final grade *G*3:

$$yi = 1\{G3i < 10\}$$

3 RESULTS

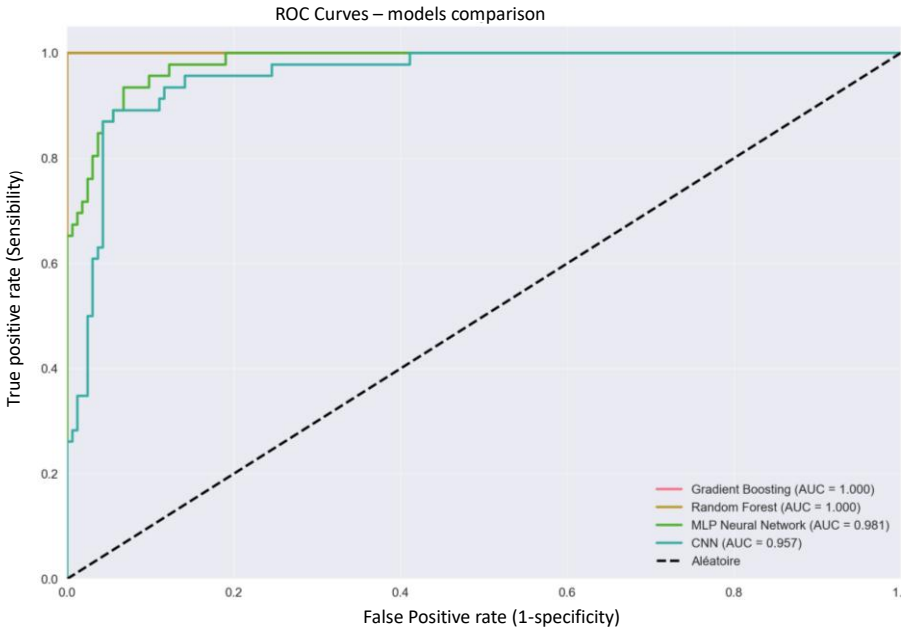

Fig. 3. ROC curves, models comparison

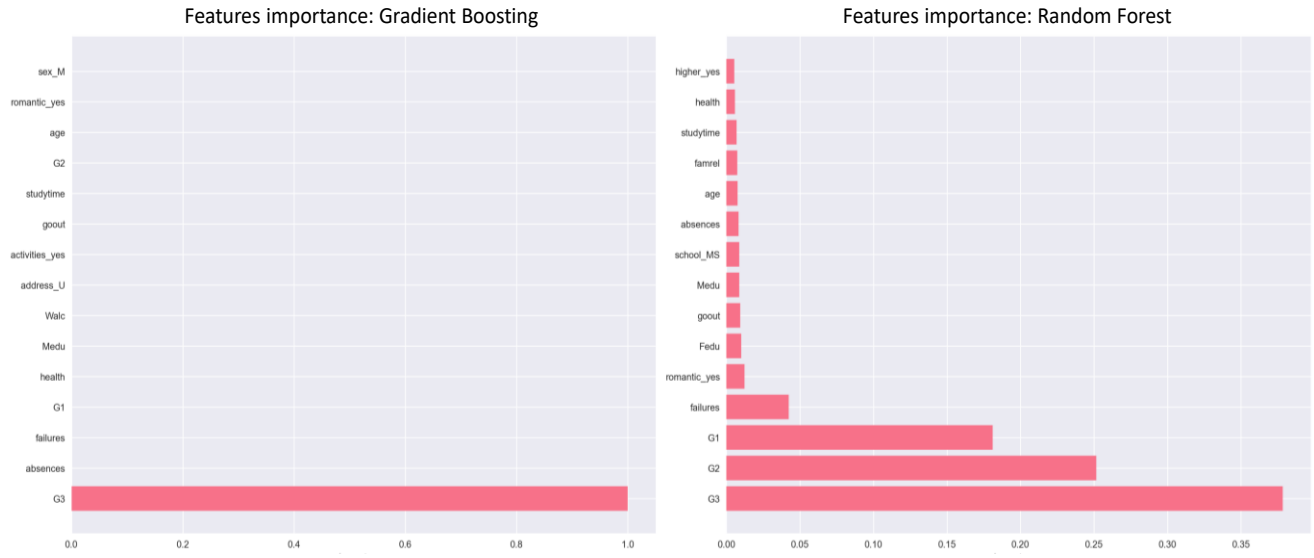

Fig. 4. Features importance For Gradient Boosting and Random Forest models

Table 1. Models performance

Models	Accuracy	F1-score	AUC-ROC
Random Forest	1.000	1.000	1.000
Gradient Boosting	1.000	1.000	1.000
MLP Classifier	0.9330	0.8444	0.9812
CNN	0.9187	0.8283	0.9575

Table 2. Classification report for CNN

Random Forest	Precision	Recall	F1-score	Support
No Abandon	0.97	0.93	0.95	163
Abandon	0.77	0.89	0.83	46
Accuracy			0.92	209
Macro avg	0.87	0.91	0.89	209
Weight avg	0.93	0.92	0.92	209

Table 3. Classification report for MLP Neural Network

Random Forest	Precision	Recall	F1-score	Support
No Abandon	0.95	0.96	0.96	163
Abandon	0.86	0.83	0.84	46
Accuracy			0.93	209
Macro avg	0.91	0.89	0.90	209
Weight avg	0.93	0.93	0.93	209

Table 4. Classification report for Random Forest

Random Forest	Precision	Recall	F1-score	Support
No Abandon	1.00	1.00	1.00	163
Abandon	1.00	1.00	1.00	46
Accuracy			1.00	209
Macro avg	1.00	1.00	1.00	209
Weight avg	1.00	1.00	1.00	209

Table 5. Classification report for Gradient Boosting

Random Forest	Precision	Recall	F1-score	Support
No Abandon	1.00	1.00	1.00	163
Abandon	1.00	1.00	1.00	46
Accuracy			1.00	209
Macro avg	1.00	1.00	1.00	209
Weight avg	1.00	1.00	1.00	209

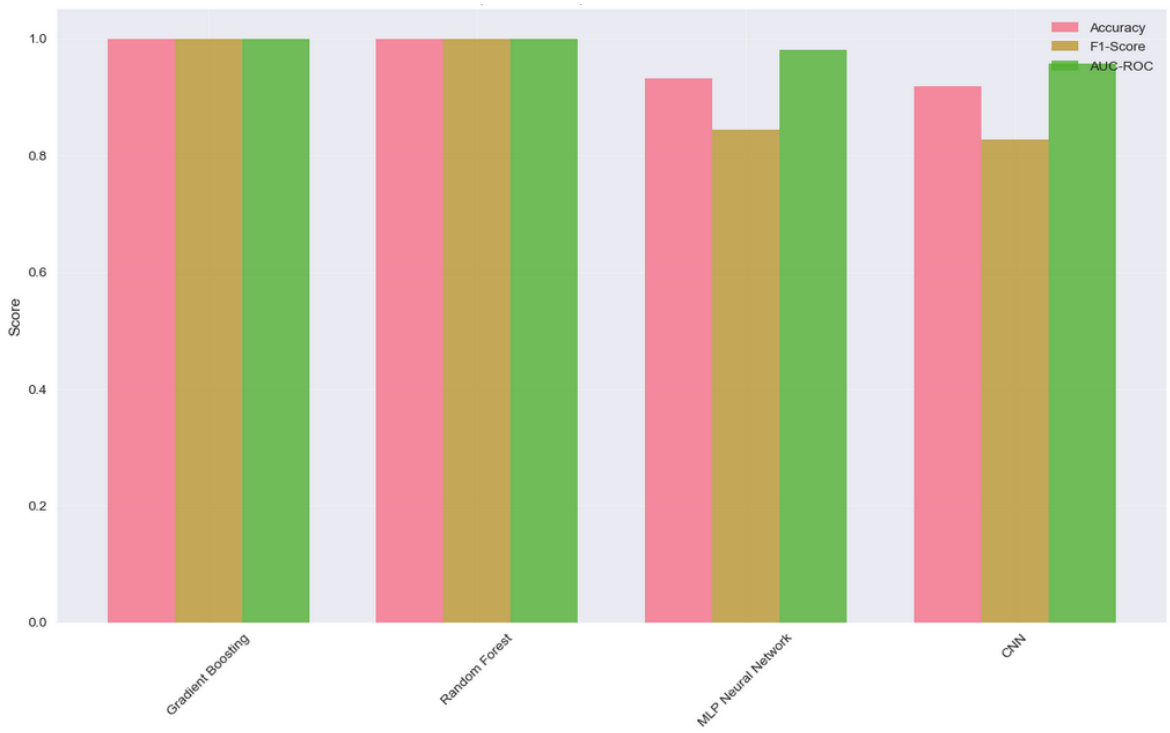

Fig. 5. Models performance comparison

4 DISCUSSION

- Considering Cross-validation (5-Fold)

  Gradient Boosting: mean F1-score: 1.0000 (+/- 0.0000)

  Random Forest: mean F1-score: 1.0000 (+/- 0.0000)

  MLP Neural Network: mean F1-score: 0.9691 (+/- 0.0069)

- Considering error analysis giving:

  Gradient Boosting:

  Number of false positives: 0

Number of false negatives: 0

The best model is Gradient Boosting, with best AUC-ROC:

- F1-Score: 1.0000
- Accuracy: 1.0000
- AUC-ROC: 1.0000

Models ranking:

1. Gradient Boosting (F1-score: 1.0000)
2. Random Forest (F1-score: 1.0000)
3. MLP Neural Network (F1-score: 0.8444)
4. CNN (F1-score: 0.8283)

5 CONCLUSION

Student dropout is a phenomenon that has many social, economic and institutional consequences. Through this study, we sought to provide a solution for the early prediction of student dropouts, using grades, class absences and failures as key variables. The results of our study showed the importance of balancing the classes in our dataset. Comparing the four prediction models—Random Forest, Gradient Boosting, MLP neural network, and Convolutional neural network—Gradient Boosting yielded the best scores, with an F1-score, ROC, and AUC equal to 1, and a much better error rate than the other models. We recommend that other variables such as family support and study time should be monitored for student retention.

ACKNOWLEDGMENT

We would like to take this opportunity to thank the management of the Virtual University of Côte d'Ivoire, which provided us with the material resources to carry out this study.

REFERENCES

[1] Agrusti, F., Bonavolontà, G., & Mezzini, M. «University Dropout Prediction through Educational Data Mining Techniques: A Systematic Review». *Journal of e-Learning and Knowledge Society,* 15 (3), pp. 161-182. 2019.

[2] Almeida, P. G., Silva, G. A. L., Santos, V., Moreira, G., Silva, P., & Luz, E. (2025). «Deep learning for school dropout detection: A comparison of tabular and graph-based models for predicting at-risk students». *arXiv preprint arXiv: 2508.14057,* pp. 1 – 15, 2025.

[3] Park, H., Yoo, S., & Gu, Y. «Deep learning-based early dropout prediction in university online learning». *International Journal on Informatics Visualization (JOIV),* 9 (3), pp. 1218–1225, 2025.

[4] Villar, A., & Robledo Velini de Andrade, C. «Supervised machine learning algorithms for predicting student dropout and academic success: a comparative study». *Discover Artificial Intelligence,* 4 (2), pp. 1-15, 2024.

[5] Demirtürk, B. «A comparative analysis of different machine learning methods for predicting student academic success using genetic algorithm-based hyperparameter optimization». *Applied Sciences, 15* (11), 5879, pp. 1–20, 2025.

[6] Andrade-Girón, D., Sandivar-Rosas, J., Marín-Rodriguez, W., Susanibar-Ramirez, E., Toro-Dextre, E., Ausejo-Sanchez, J., Villarreal-Torres, H., & Angeles-Morales, J. «Predicting student dropout based on machine learning and deep learning: A systematic review». *EAI Endorsed Transactions on Scalable Information Systems,* 10 (5), pp. 1-15, 2023.

[7] Skudai, K. T. Chong, N. I., & Huspi, S. H. «A systematic review of machine learning techniques for predicting student engagement and performance in online higher education». *Journal of Information Technology Education,* 24, pp. 109–130, 2025.

[8] Salehin, I., & Kang, D.-K. «A review on dropout regularization approaches for deep neural networks within the scholarly domain». Electronics, 12 (3106), pp. 1-23, 2023.

[9] Goren, O., Cohen, L., & Rubinstein, A. «Early prediction of student dropout in higher education using machine-learning models». *Proceedings of the 17th International Conference on Educational Data Mining,* pp. 349–359, 2024.

[10] Cheng, J., Yang, Z.-Q., Cao, J., Yang, Y., & Zheng, X. «Predicting student dropout risk with a dual-modal abrupt behavioral changes approach». *arXiv preprint arXiv: 2505.11119,* pp. 1–15, 2025.

[11] Niyogisubizo, J., Liao, L., Nziyumva, E., Murwanashyaka, E., & Nshimyumukiza, P. C. «Predicting student's dropout in university classes using two-layer ensemble machine learning approach: A novel stacked generalization». *Computers and Education: Artificial Intelligence,* 3, 100066, pp. 1-22, 2022.

[12] Vaarma, M., & Li, H. «Predicting student dropouts with machine learning: An empirical study in Finnish higher education». *Technology in Society,* 76, 102474, 1-14, 2024.

[13] Marcolino, M. R., Porto, T. R., Primo, T. T., Targino, R., Ramos, V., Queiroga, E. M., Muñoz, R., & Cechinel, C. «Student dropout prediction through machine learning optimization: Insights from Moodle log data». *Scientific Reports,* 15, 9840, pp. 1-12, 2025.

[14] UCI repository 2025. [Online] Available: https://archive.ics.uci.edu/ml/machine-learning-databases/00320/student.zip