Al-Zaytoonah University Journal of Business, Vol. 02 No. 01, March 2026
ISSN Number: Print: 3107- 2178 / Online: 3107- 2186

A Comparative Study of Machine Learning and Deep Learning Algorithms for Student Dropout Prediction in Higher Education

Huthaifa Aljawazneh[1], Raed Alqirem[2] and Nour Al-Adwan[3]

[1-3]Department of Business Intelligence
Al-Zaytoonah University of Jordan, Amman, Jordan
huthaifa.rj@zuj.edu.jo, drraed@zuj.edu.jo, Nooradwan70@gmail.com

Abstract Student dropout in higher education represents a significant academic and economic challenge. This study investigates the effectiveness of machine learning and deep learning techniques for early identification of at-risk students. A two-stage experimental framework is proposed. In the first stage, three machine learning algorithms (Random Forest, Support Vector Machine, and XGBoost) are compared with two deep learning models (Deep Neural Networks and TabNet) using the original dataset. In the second stage, the impact of data balancing techniques, namely SMOTE and Borderline-SMOTE, is evaluated. Model performance is assessed using accuracy, precision, recall, and specificity. The results demonstrate that XGBoost consistently achieves superior performance across both imbalanced and balanced datasets, while data balancing techniques significantly improve recall, enhancing the detection of at-risk students. These findings provide valuable insights into the role of data balancing in improving predictive performance in student dropout prediction.

Keywords: Student dropout, classification, machine learning, XGBoost, TabNet.

1 Introduction

In recent years, student dropout in higher education has become a significant research issue in Educational Data Mining (EDM), as researchers aim to analyze its causes and better understand the factors that influence it (Goren et al, 2024). Student dropout occurs when a student leaves from an educational program before completion for any reason (Hegde and Prageeth, 2018). In higher education, especially in traditional classroom settings, dropout can occur in several forms, including higher education dropout, where the student decides to stop pursuing their degree before its completion (Rabelo and Zárate, 2024).

Despite the growing research interest in this issue, student dropout remains a major concern for educational institutions and policy makers (Aulck et al., 2017). In fact, more than one million students drop out of schools or educational institutions each year, averaging nearly 8,000 students per day during the academic year (Hegde and Prageeth, 2018). Moreover, according to some research, approximately 40% of students pursuing bachelor's degrees fail to complete their study within six years (National Center for Education Statistics, 2015), resulting in universities losing tens of billions of dollars in revenue each year (Aulck et al., 2017).

However, student dropout is influenced by various internal or external factors including personal, academic, and socio-economic factors (Rabelo and Zárate, 2024). This extends beyond the student and the educational institutions to the general economy, where a financial cost to the taxpayer results for those students who do not complete their educational programs. As a result, many studies have been conducted to identify students who are at risk of dropping out (Goren et al, 2024).

Therefore, the aim of this study is to determine the most effective approach for identifying students at risk of dropping out and enabling early intervention using two methodological approaches. The first approach compares the performance of three machine learning algorithms, namely, Random Forest (RF) (Sun et al., 2015), Support Vector Machine (SVM) (Cortes and Vapnik, 1995), and XGBoost (Islam, Sholahuddin and Abdullah, 2021) with two powerful deep learning models, i.e., Deep Neural Network (DNN) (Miikkulainen et al., 2019) and TabNet (Gorishniy et al., 2021), using the original dataset. The second approach examines the effect of two data balancing techniques: SMOTE (Mansourifar and Shi, 2020) and Borderline-SMOTE (Han, Wang and Mao, 2005), on the performance of these models. Model effectiveness was evaluated using four performance metrics: accuracy, precision, recall, and specificity.

ISSN Number: Print: 3107- 2178 / Online: 3107- 2186

The remainder of this paper is organized as follows. Section 2 reviews the related works. Section 3 describes the dataset considered. Section 4 outlines the proposed methodology. Section 5 presents and discusses the results. Finally, Section 6 concludes the paper and suggests directions for future work.

2 Related works

Several studies have explored the application of machine learning techniques to address the problem of student dropouts and identify at-risk students in higher education, specifically when dealing with tabular data (Goren et al, 2024). Accordingly, Kim et al. (2023) used RF, Logistic Regression, SVM and XGBoost to predict student dropout based on academic, demographic, and socioeconomic data. They found that academic data is the most influential factor, and RF performing best and enabling early identification of at-risk students. Moreover, Kemper et al. (2020) developed dropout prediction models based on academic transcript data using logistic regression and decision trees. They found that DT was more interpretable and practical than LR. Furthermore, Sülak and Köklü (2024) compared the performance of Artificial Neural Network, Decision Tree and RF to predict student dropout using data from 4,424 students. Based on their results, ANN achieved the best performance in terms of precision, recall, and F-score.

On the other hand, with the increasing interest in the use of predictive analytics in the educational domain, deep learning has gained attention in recent years as an effective technique in modeling complex patterns in large-scale datasets (Basnet et al, 2022). For example, Agrusti et al. (2020) have utilized Convolutional Neural Networks (CNN) in predicting dropouts. Their experiment has demonstrated that CNN performs better in accuracy than Bayesian Networks, especially in academic data. In the same year, Baranyi et al. (2020) have utilized DNN and gradient boosted trees and have demonstrated that deep learning performs slightly better than XGBoost with 72.4% accuracy and 77.1% AUC. Moreover, Alruwais (2023) has proposed an explainable DeepFM model that uses factorization machines and deep neural networks in predicting student dropouts with 99% accuracy and performs better than Random Forest.

However, despite the widespread application of machine learning and deep learning methods for student dropout prediction, it has been observed in the literature that there exists a significant gap in the field, especially in terms of addressing the problem of class imbalance, which is commonly encountered in such problems. Another gap in the literature is related to the comparative analysis of traditional machine learning methods and deep neural network methods in terms of imbalanced as well as balanced data. Therefore, this study aims to address these limitations by conducting a systematic comparison between machine learning and deep neural network models using the original dataset as well as balanced datasets generated through SMOTE and Borderline-SMOTE techniques, in order to provide deeper insights into the impact of data balancing on predictive performance

3 Dataset description

The dataset used in this study originates from a Portuguese higher education institution and downloaded from Kaggle[1]. It contains information about 4,424 students with 37 features in eight-degree programs. The original class feature has three classes: enrolled, graduate, and dropout. For the purposes of this study, the enrolled class was excluded from the analysis, as required for the binary classification algorithm. This was achieved by excluding all the records associated with the enrolled class from the dataset. The target variable was therefore redefined with two classes: graduate, which was assigned the class label (0), and dropout, which was assigned the class label (1).

4 Methodology

This section presents the methodology used to predict student dropouts. In the first stage, three machine learning algorithms including RF, SVM, and XGBoost are compared to two powerful deep learning models, i.e., DNN and TabNet, using a higher-education dataset. Although the dataset is relatively balanced, the class distribution is not identical. Therefore, the second stage analyzes the impact of applying balancing techniques on the models' performance using SMOTE and Borderline-SMOTE techniques.

ISSN Number: Print: 3107- 2178 / Online: 3107- 2186

4.1 Random Forest (RF):

This classifier combines several decision trees into a single model. The output is determined by averaging the predictions made by each decision tree on a particular data point (Sun et al., 2015).

4.2 Support Vector Machine (SVM):

It is a supervised learning algorithm used for classification. SVM maps input data into a higher-dimensional space through a non-linear transformation. In this new space, it develops an optimum boundary to distinguish classes while maximizing generality (Cortes and Vapnik, 1995).

4.3 Extreme Gradient Boosting (XGBoost):

XGBoost is a supervised learning method applicable to both classification and regression tasks. It constructs a robust prediction system using weak learning models with the help of ensemble learning techniques, where each weak learner is trained to correct the mistakes made by the previous ones (Islam, Sholahuddin and Abdullah, 2021).

4.4 Deep Neural Networks (DNN):

DNNs are composed of multiple hidden layers of interconnected neurons that are able to learn complex nonlinear relationships between data points using weighted connections between the neurons. The depth of the DNN enables them to learn complex patterns in the data (Miikkulainen et al., 2019).

4.5 TabNet:

TabNet is a deep learning model developed for tabular data that uses a step-wise attention mechanism to identify the most relevant features during the learning process. This approach enhances prediction performance while preserving model interpretability (Gorishniy et al., 2021).

4.6 Advanced balancing techniques:

Balanced and precise data are important for reliable machine learning predictions. Moreover, imbalanced classes can result in bias toward the majority class (Jain et al., 2025), thus reducing the precision of dropout detection. In this study, two resampling techniques have been used to assess their influence on the classifiers' performance while avoiding data leakage.

4.6.1 Synthetic Minority Over-sampling Technique (SMOTE):

SMOTE generates new minority class instances by synthesizing data points between existing instances in the feature space. This improves the balance between the classes and the model's ability to identify minority class instances correctly (Mansourifar and Shi, 2020).

4.6.2 Borderline-SMOTE (BL-SMOTE):

Borderline-SMOTE addresses class imbalance by selectively generating synthetic minority-class samples around boundary regions in the feature space, where instances are more prone to misclassification (Han, Wang and Mao, 2005).

5 Results:

In this section, a comparative evaluation of three models of machine learning and two models of deep learning will be presented. The classification models were trained and tested using three datasets. The datasets used were the original imbalanced dataset and two balanced datasets generated using SMOTE and Borderline-SMOTE. The performance of the models was also evaluated using four performance metrics.

ISSN Number: Print: 3107- 2178 / Online: 3107- 2186

5.1 Evaluation metrics:

In this study, dropout prediction is treated as a binary classification problem, where dropout students represent the positive class and graduates represent the negative class. Model performance is assessed using the confusion matrix, which includes true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN). Based on these values, several evaluation metrics are calculated to measure predictive performance (Cho et al, 2023).

- Accuracy: represents the proportion of correctly classified instances, including both dropout and graduate students.

$$Accuracy = (TP + TN) / (TP + TN + FP + FN)$$

- Precision: represents the proportion of predicted dropout cases that are correctly identified.

$$Precision = TP / (TP + FP)$$

- Recall: measures the model's ability to correctly identify all actual dropout cases.

$$Recall = TP / (TP + FN)$$

- Specificity: measures the proportion of actual graduates classified correctly as non-dropouts (Yager, 1982).

$$Specificity = TN / (TN + FP)$$

5.2 Results obtained from the original dataset:

Model	Accuracy	Precision	Recall	Specificity
Machine learning algorithms				
RF	0.9219	0.9295	0.8662	0.9578
SVM	0.9265	**0.9529**	0.8545	**0.9729**
XGBoost	**0.9284**	0.9328	**0.8803**	0.9593
Deep learning algorithms				
DNN	0.9027	0.8883	**0.8592**	0.9306
TabNet	**0.9082**	**0.9035**	0.8568	**0.9412**

Table 1. Evaluation metrics of the classifiers evaluated on the original dataset.

As shown in Table 1 machine learning algorithms outperform deep learning algorithms. XGBoost achieved the highest accuracy and recall, indicating the best performance in predicting dropout cases. SVM obtained the best precision and specificity, making it more effective in predicting graduate cases. In deep learning models, TabNet achieved the highest values in several metrics. DNN achieved higher recall compared to TabNet, indicating better capability in identifying dropout cases in deep learning models.

5.3 Results obtained from the balanced dataset using SMOTE:

Model	Accuracy	Precision	Recall	Specificity
Machine learning algorithms				
RF	**0.9210**	0.9067	**0.8897**	0.9412
SVM	0.9146	**0.9152**	0.8615	**0.9487**
XGBoost	0.9164	0.9075	0.8756	0.9427
Deep learning algorithms				
DNN	0.9017	0.8735	**0.8756**	0.9186
TabNet	**0.9082**	**0.9095**	0.8498	**0.9457**

Table 2. Evaluation metrics of the classifiers evaluated on the dataset balanced using SMOTE.

Table 2. shows that after applying SMOTE, recall improved across some classifiers, enhancing their ability to detect dropout cases, while accuracy, precision, and specificity showed a slight decrease. However, machine learning models performed better than deep learning models. RF had the highest accuracy and recall values. SVM achieved the best precision and specificity. Among deep learning models, TabNet achieved the best accuracy, precision and specificity values. DNN obtained the highest recall.

5.4 Results obtained from the balanced dataset using BL-SMOTE

Model	Accuracy	Precision	Recall	Specificity
Machine learning algorithms				
RF	0.9155	0.8902	**0.8944**	0.9291
SVM	0.9091	0.8921	0.8732	0.9321
XGBoost	**0.9192**	**0.9005**	0.8920	**0.9367**
Deep learning algorithms				
DNN	**0.8861**	**0.8528**	0.8568	**0.9050**
TabNet	0.8669	0.7970	**0.8850**	0.8552

Table 3. Evaluation metrics of the classifiers evaluated on the dataset balanced using BL-SMOTE.

According to Table 3 machine learning techniques consistently outperformed deep learning techniques. BL-SMOTE improves recall compared to both the original dataset and SMOTE. XGBoost achieved the highest accuracy, precision and recall. RF had the best recall demonstrating the best ability to identify dropout students. On the other hand, SVM showed balanced performance in all evaluation metrics. In deep learning models, DNN achieved the best values in several metrics while TabNet obtained the best recall.

5.5 Comparative Analysis of Results

Overall, machine learning techniques outperformed deep learning techniques both approaches. In the original dataset, XGBoost had the highest accuracy and recall values, showing strong performance in predicting dropout students. After applying SMOTE, recall improved and the other three metrics slightly decreased. With BL-SMOTE, recall reached its highest value, especially from RF, which showed better performance in identifying dropout cases but with a slight decrease in other metrics. The findings indicate that XGBoost combined with BL-SMOTE is the best approach to predict student dropouts.

6 Conclusion and future work

To address the problem of student dropout, this study followed two experimental stages using three versions of dataset: the original dataset, dataset balanced using SMOTE, and dataset balanced using BL-SMOTE. The first stage compares three machine learning algorithms namely RF, SVM and XGBoost with two advanced deep learning algorithms (i.e., DNN and TabNet,). In the second one, balancing techniques such as SMOTE, and Borderline-SMOTE were applied to assess their impact on model performance. Moreover, four evaluation metrics, including accuracy, precision, recall and specificity were used for a comprehensive evaluation. The results indicate that XGBoost achieved the best overall results in the original dataset, while its combination with BL-SMOTE yielded the highest performance after applying data balancing techniques. Future work will explore advanced feature engineering techniques, hyperparameter optimization, and the use of larger and more diverse datasets to further enhance predictive performance.

7 References

1. Agrusti, F., Bonavolontà, G., & Falcione, A. (2020). Deep learning approach for predicting university dropout: A case study at Roma Tre University. Journal of e-Learning and Knowledge Society, 16(3), 13–21.
2. Alruwais, N. M. (2023). Deep FM-based predictive model for student dropout in online classes. IEEE Access, 11, 96954–96970.
3. Aulck, L., Velagapudi, N., Blumenstock, J., & West, J. (2017). Predicting student dropout in higher education. arXiv. https://arxiv.org/abs/1606.06364
4. Baranyi, M., Nagy, M., & Molontay, R. (2020). Interpretable deep learning for university dropout prediction. In Proceedings of the 21st Annual Conference on Information Technology Education (pp. 1–6). ACM.
5. Basnet, R.B., Johnson, C. & Doleck, T. Dropout prediction in Moocs using deep learning and machine learning. Educ Inf Technol 27, 11499–11513 (2022). https://doi.org/10.1007/s10639-022-11068-7.

6. Cho, C. H., Yu, Y. W., & Kim, H. G. (2023). A Study on Dropout Prediction for University Students Using Machine Learning. Applied Sciences, 13(21), 12004.
7. Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273–297.
8. Goren, O., Cohen, L., & Rubinstein, A. (2024). Early prediction of student dropout in higher education using machine learning models. In B. Paaßen & C. D. Epp (Eds.), Proceedings of the 17th International Conference on Educational Data Mining (pp. 349–359). International Educational Data Mining Society.
9. Gorishniy, Y., Rubachev, I., Khrulkov, V., & Babenko, A. (2021). Revisiting deep learning models for tabular data. Advances in Neural Information Processing Systems, 34, 18932–18943.
10. Han, H., Wang, W. Y., & Mao, B. H. (2005). Borderline-SMOTE: A new over sampling method in imbalanced data sets learning. Advances in intelligent computing (Vol. 3644, pp. 878–887). Springer, Berlin, Heidelberg.
11. Hegde, V., & Prageeth, P. P. (2018). Higher education student dropout prediction and analysis through educational data mining. In *Proceedings of the 2nd International Conference on Inventive Systems and Control (ICISC)* (pp. 694–699). IEEE. https://doi.org/10.1109/ICISC.2018.8398887.
12. Islam, S. F. N., Sholahuddin, A., & Abdullah, A. S. (2021). Extreme gradient boosting (XGBoost) method in making forecasting application and analysis of USD exchange rates against rupiah. Journal of Physics: Conference Series, 1899(1), 012164.
13. Jain, A., Dubey, A. K., Khan, S., Panwar, A., Alkhatib, M., & Alshahrani, A. M. (2025). A PSO weighted ensemble framework with SMOTE balancing for student dropout prediction in smart education systems. Scientific Reports, 15(1), Article 17463.
14. Kemper, L., Vorhoff, G., & Wigger, B. U. (2020). Predicting student dropout: A machine learning approach. European Journal of Higher Education, 10(1), 1–20.
15. Kim, S., Yoo, E., & Kim, S. (2023). Why do students drop out? University dropout prediction and associated factor analysis using machine learning techniques. arXiv.
16. Mansourifar, H., & Shi, W. (2020). Deep Synthetic Minority Over-Sampling Technique. arXiv.
17. Miikkulainen, R., Liang, J., Meyerson, E., Rawal, A., Fink, D., Francon, O., Raju, B., Shahrzad, H., Navruzyan, A., Duffy, N., & Hodjat, B. (2019). Evolving deep neural networks. Artificial Intelligence, 278, 103195.
18. NCES. Fast Facts (Retrieved April. 2016). Technical report, National Center for Education Statistics, 2015.
19. Rabelo, A. M., & Zárate, L. E. (2024). A model for predicting dropout of higher education students. Data Science and Management. https://doi.org/10.1016/j.dsm.2024.07.001.
20. Sülak, S. A., & Köklü, N. (2024). Predicting student dropout using machine learning algorithms. Intelligent Methods in Engineering Sciences, 3(3), 91–98.
21. Sun, J., Zhong, G., Dong, J., & Cai, Y. (2015). Banzhaf random forests. arXiv preprint arXiv:1507.06105.
22. URL https://nces.ed.gov/fastfacts/ display.asp?id=40.
23. *Yager, R. R. (1982).* On the specificity of a possibility distribution. Fuzzy Sets and Systems, 13*(2), 107–123.*