International Journal of Basic and Applied Sciences, 14 (SI-1) (2025) 258-263

International Journal of Basic and Applied Sciences

Website: www.sciencepubco.com/index.php/IJBAS
https://doi.org/10.14419/z3vc6t60
Research paper

An Improved Ensemble Based Technique for Handling Noisy Class Imbalnced Education Data for Prediction of Students Dropout in Hei

Ms. K. Sangeetha [1] *, Dr. N. Shanmugapriya[2]

[1] *Research Scholar, Department of Computer Science, Dr. SNS Rajalakshmi College of Arts & Science, Coimbatore*
[2] *Associate Professor & Head, Department of Computer Applications (PG), Dr. SNS Rajalakshmi College of Arts & Science, Coimbatore*
**Corresponding author E-mail: advsciresearch2025@gmail.com*

Received: May 28, 2025, Accepted: June 27, 2025, Published: June 7 2025

Abstract

Increasingly, the sector of education is becoming more interested in the creation of intelligent technology. The fast rise of educational data suggests that standard processing methods may have limitations and may even result in distortion. It is for this reason that the process of reconstructing the research technique of data mining in the field of education has become increasingly important. The amount of information about students that is stored in educational databases is growing on a daily basis; thus, the information that is extracted from these databases needs to be updated on a consistent basis. In a scenario in which there is a requirement to manage a constant flow of student data, there is a challenge of figuring out how to manage this enormous volume of data into the information and how to adapt new knowledge that is introduced with the new data. When working with classes that have few instances, a class imbalance issue is crucial. The machine learning classification of classes is significantly impacted by noisy, class-unbalanced datasets. This research proposes an enhanced hybrid bag-boost model using a suggested resampling technique. A suggested resampling method for addressing noisy, unbalanced datasets is included in this model. The suggested resampling method includes Edited Nearest Neighbor (ENN) and K-Means SMOTE (Synthetic Minority Oversampling Technique) as an oversampling method. The technique of Undersampling is employed to eliminate noise. Three levels of noise reduction are achieved with this resampling technique: first, datasets are clustered using the K-Means clustering technique; second, imbalance is handled by SMOTE inside clusters, which introduces synthetic instances of the class in the minority; and third, instances that generate noise are removed using the ENN technique. The suggested model outperforms the others, according to experimental data. Furthermore, it has been verified that the suggested method works better in binary unbalanced datasets when the noise proportion is raised.

Keywords: *Educational Data Mining (EDM); Student Performance; Imbalanced Data; Class Imbalance; Oversampling; Prediction, and Ensemble Model.*

1. Introduction

Educational Data Mining (EDM) is a new field. In today's culture, predicting student performance is a major cause of anxiety. Rapid technical breakthroughs and the use of diverse machine learning methodologies have resulted in the creation of good models that have improved the accuracy of student performance progress. As a result, it's vital to create machine learning algorithms that can reliably anticipate student success. [1]. As a result of educational systems, institutions and universities acquired more skills. Over the past ten years, numerous studies on educational systems have been carried out. However, educational systems continue to face issues with student performance, dropout rates, learning outcomes, the impact on society, etc.

Higher education administrators are concerned with predicting a student's performance. Predicting students' performance in the definite world is a complex Endeavour. Any higher educational institution's most importanttarget is to develop the excellence of managerial choices and to presentexpert education. One policy to attain the best extent of eminence in the higher education system is to have aexcellent prediction of a student's achievement in a higher learning institution. There are a variety of prediction models available, each with a unique methodology [2]. The researcher reports on student performance, but it's unclear whether there are any variables that can reliably forecast whether a kid will be an academic genius, a dropout, or an ordinary performer.

Different ways to predicting student academic success have been used. Introduce a new machine learning algorithm and ensemble technique to forecast the student's success in the proposed work. The suggested system's goal is to detect students who may be at risk. The suggested work's key contribution is:

1) Compare the accuracy of predictive models to choose the best model which gives best accuracy.
2) Developed a monitoring and prediction system for accurate results.

The below is a summary of the paper's formation: The part two covers a number of related studies and demonstrates why a new technique is required. Part three describes our methodology's processes and what they will contribute in the creation of the novel prediction model. Part four focuses into the experimental studies, outcomes, and perspectives, followed by Part five conclusions and future research recommendations.

2. Related work

Predicting student academic success has become a much more important part of enhancing academic education, assisting students in their studies, and providing tutors with more flexibility when it comes to teaching their students. [3]. In recent years, a large number of works on this subject have been released. Several literature reviews were found here that looked at student academic performance modelling from various angles.

T. Devasia, T. Vinushree, T. P. Vinushree, and V. Hegde (2016) discuss how, as higher education has progressed, the number of students dropping out has increased, harming not just the students' prospects as well as the institution's credibility [4]. The proposed system is an automated application that utilizes the Naive Bayesian mining method to extract significant data. The present system merely saves and fetches the data it comprises and preserves academic information in the form of quantitative numbers. The proposed work is to maximize the outcome probability of students by using Naive Bayesian and a platform that records all student admission records, curriculum design, curriculum relevant information, student scores relevant information, attendance relevant information, and so on..

Adekitan, A. I., and Salau, O. (2020) stated that the ability to predict failure is a useful learning resource that may be used to efficiently educate students, as well as to develop and design adequate academic initiatives aimed at minimising failure and lower grades [5]. Students were enrolled in practice terms of academic capability, as assessed by their entrance criteria scores. The purpose of this study is to see if there is a link among admission requirements scores and graduation grades, as well as to see how ethnicity affects the anticipated accuracy of models built using a Nigerian university as a case study.The greatest classification performance found was 53.2%, indicating that pre-admission scores alone would be unsatisfactory for forecasting students' graduation results, although they can be used as a guide. The efficiency was enhanced to 79.8% by using the over-sampling methodology.

Shingari, I., Kumar, D., and Khetan, M. (2017) discuss how a student's performance is important in any institution [6]. It is the standard by which any prestigious educational institution's academic excellence is measured. The student is concerned not just with the institute, but also with his or her academic success. There are different approaches for projecting a student's performance at this time, but data mining is the most scientific and reliable method for doing so. Data mining techniques can be effective for sifting through data that is tailored to a certain requirement. The proposed research focuses on identifying the gaps in existing performance prediction approaches. These patterns assist educators in forecasting a student's achievement.

The challenge of quickly gathering data in education was addressed by Han, M., et al., (2017), who discovered that researchers can anticipate students' academic achievement [7]. The machine learning model stands out from the rest of the prediction models. The model AdaBoost was proposed in the proposed study to forecast student courses using an ensemble approach. Experiments were conducted to compare the model to others such as the decision tree, neural network, random forests, and SVM. The model had the best prediction performance, according to the findings.

MLAs are generally used in these references to categorise student data. However, because the student data set comprises a significant number of features and data records, applying these methods to it is inefficient in most cases [10].

3. Proposed methodology

The coreprinciple of this research effort is the identification and exploration ofsupervised learning algorithms coupled with standard computing techniques that canlead to the development of robust approach for students' academicperformance prediction. In order to achieve this, a number of machine learningalgorithms were experimented and some investigations were explored.

The students' performance analysis is one of the significant activities of anyeducational organization, particularly in the domain of higher education. In traditional educationalsystem the prediction of students' performance is not popular (formal), but in moderneducation era, the prediction of the success of students in educational institutions hasbecome most crucial issues in academic organizations across the globe [11]. It isnecessary to build up an effective model to forecast the performance of the students'in all the courses so that instructors can identify weak learners, advance learners andaverage learners in the class. Another difficulty occurs when educator with limitedunderstanding about data analysis techniques tries to forecast the students'performance.

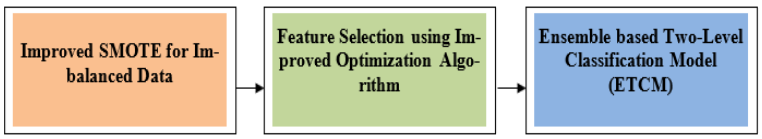

Fig. 1.1: Workflow of Proposed Approach.

With these requirements, it is necessary to develop a solid approach that can be used by teachers as well as students to anticipate academic performance.

3.1. Improved SMOTE (I-SMOTE) for preprocessing imbalanced data

If the classes that make up the dataset are not roughly equally represented, the dataset is said to be imbalanced. Resampling is a common method for dealing with very unbalanced datasets. Over-sampling and under-sampling are two types of sampling [12]. Over-sampling involves adding additional samples to the minority class whereas under-sampling involves removing samples from the majority class. SMOTE is an oversampling method in which the minority class is over-sampled using synthetic examples rather than replacement over-sampling.

3.1.1. I-SMOTE algorithm

Depending on the abovementioned, we propose an enhanced preprocessing technique (I-SMOTE) for imbalanced training dataset that attempts to properly quantify the borderline and develop various synthetic samples using SMOTE generalization. The following are the two elements of our proposed scheme:
First, use the SMOTE technique to create a synthetic instance based on the equations below:

$$N = 2 * (r - z) + z \ldots \text{Equ.} \quad (1)$$

There are r samples from the mainstream class and z samples from the fringe class. The number of majority class samples is r, the number of minority class samples is z, and the initial synthetic instance number (newly generated) is N.
Synthetic cases that are closer to the SMOTE boundary and synthetic samples that are more similar to the majority class than the minority should be rejected in the second stage. A thorough description of the I-SMOTE technique is provided below:

- Step 1: SMOTE's synthetic instances may be approved or refused based on two criteria, which correspond to the first level:
- Step 2: With both the acceptable synthetic instances in hand, the following is done to remove the noise.
- Step 3: Now compute the difference between I and each initial majority S_a in the same way,

The I-SMOTE methodology is used to build a solid preprocessing method for imbalance learning.

3.2. Feature selection using modified optimization algorithm

The selection of features is a crucial step in improving prediction accuracy. Feature selection eliminates attributes from data sets that aren't important. Since many attributes in the student's performance prediction dataset do not play a significant role in performance prediction, feature selection techniques must be used to remove them [13]. Feature extraction is a technique for identifying only the attributes in a dataset that have the greatest influence on the prediction variable or outcome. The accuracy of many models can be compromised by irrelevant data gathering attributes.
The HAFSCSO technique is used to choose the qualities in the proposed work. The proposed strategy, which contained the two connected components of redundant and irrelevant attribute reduction, improved attribute selection efficiency. For getting the relevant properties, AFSCSO is offered. This research work discovered the redundant attributes that are shown in the relevant attributes after choosing the relevant features.
Once their search region is narrowed through levy flight optimization, the suggested HAFSCSO is used to attain a high convergence rate. The number of fishes, features, iterations, and other parameters in this method are all important. In the first iteration, the entire fish will choose an attribute subset of m attributes at random. To bring the visual position up to date and control the attribute subsets of the following iteration, the best subsets (k <nf) are used. The entire fish begins with m - p attributes that are chosen at random from the previously selected k best subsets, where p might be an integer ranging from 1 to m - 1. This sequence of attribute selections may cause all systems to fail to select the appropriate attributes. As a result, the CSO method is used to choose the optimal attribute. By based on this, the attributes arecharacterized the optimal k subsets have more probability to present within the attributesof the next iteration. But, it will obtain the entire fish for considering other features. For aknown fish j, those attributes are the ones that finish the optimal cooperation among theprevious understanding consists of visual position and the current best of cuckoo search.
The redundant attributes are displayed within the relevant attributes after the relevant attributes have been achieved. As a result, the properties are clustered using NMF techniques. By selecting a representative attribute within the clusters, the redundant attributes are removed from the relevant attributes, and the final attribute subset is generated.

3.3. Ensemble two-level classification model (ETLCM)

Provide a student's conceptual framework in the specified task, using ensemble methodologies. Ensemble learning is a sort of learning during which an issue has been resolved using many models [14].
In contrast to conventional instructional methods, typically train data using an unique training model, ensemble methods aim to train data using a range of models and then integrate them to vote on their outcomes. In most cases, ensemble estimates are more effective than single-model projections. This technique's purpose is to give a reliable estimation of the factors that may determine a student's academic advancement.
Dependent and independent procedures are the two categories into which ensemble approaches fall. The next learner is developed using the work of the previous one in a dependent approach. One instance of a dependent technique is boosting. Every student works autonomously and independently, and the results are combined using a vote process. Bagging and random forest are two different processes. By resampling the original data, these methods provide samples that are then trained independently using a different classifier. Support Vector Machines (SVM), Decision Trees (DT), and Nave Bayesian (NB). Each classifier's output is aggregated after a voting process; the class chosen by the most classifiers is referred to as the ensemble decision.

4. Performance evaluation

4.1. Dataset details

The EDM dataset comes from the machine learning repository at UC Irvine (UCI). It is a time-series dataset and a log data of students' learning activities and type of Complex from the Systems Laboratory in Italy. It is part of the Eindhoven University of Technology's Department of Industrial Design. This dataset was collected for the purpose of this study [15]. The attributes in this dataset have been reorganised into a more user-friendly manner. Applying the dataset directly to the classification and clustering method is difficult. A second operation is required to turn the dataset into a multidimensional numeric dataset.
In the suggested work, four Intel CPUs running at 2.67GHz apiece and a PC with 6GB of RAM were used for the studies. In our tests, WEKA was used to assess the suggested classification models and comparisons. Additionally, we used 10-fold cross validation to separate the dataset into training and testing sections.

4.2. Evaluation metrics

There are five different methods for evaluating classification quality in our trials [16]. Measures derived from Equations 2, 3, 4, and 5, respectively, using a classification confusion matrix.

4.2.1. Correlation coefficient

The correlation coefficient is a statistical performance criterion for determining whether actual and expected values are correlated. The correlation between the actual and expected values is measured using this criterion. The value of the correlation coefficient for a perfect statistical correlation is 1, while the value for no correlation is 0.

Table 4.1: Evaluation of Correlation Efficient

S. No.	Class Index	Existing – 1 (SMOTE) (Mean Value)	Existing – 2(ADASYN) (Mean Value)	Proposed(I-SMOTE) (Mean Value)
1.	0	0.838	0.844	0.995
2.	1	0.821	0.835	0.989
3.	2	0.817	0.835	0.980
4.	3	0.817	0.842	0.980
5.	4	0.803	0.842	0.961
6.	5	0.803	0.840	0.949
7.	6	0.792	0.832	0.923
8.	7	0.784	0.824	0.913.
9.	8	0.776	0.813	0.907
10	9	0.763	0.807	0.894

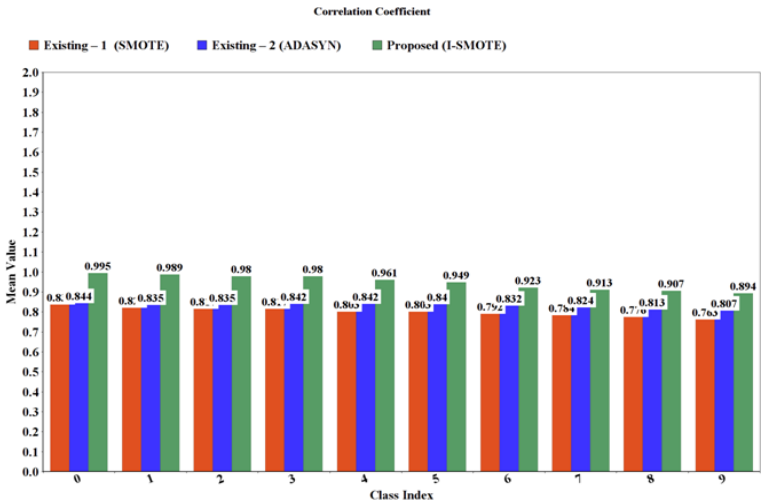

Fig. 4.1: Comparison of Correlation Coefficient.

4.3.2. Mean absolute error (MAE)

The MAE is a measurement that measures the average magnitude of errors in a series of estimates without considering their distribution into account.

Table 4.2: Evaluation of Error Rate

S. No.	Class Index	Existing – 1 (SMOTE) (Error Rate)	Existing – 2 (ADASYN) (Error Rate)	Proposed(I-SMOTE) (Error Rate)
1.	0	0.149	0.13	0.087
2.	1	0.149	0.125	0.088
3.	2	0.153	0.154	0.087
4.	3	0.157	0.124	0.093
5.	4	0.159	0.126	0.095
6.	5	0.179	0.156	0.103
7.	6	0.189	0.189	0.104
8.	7	0.215	0.171	0.125
9.	8	0.237	0.203	0.149
10	9	0.269	0.222	0.171

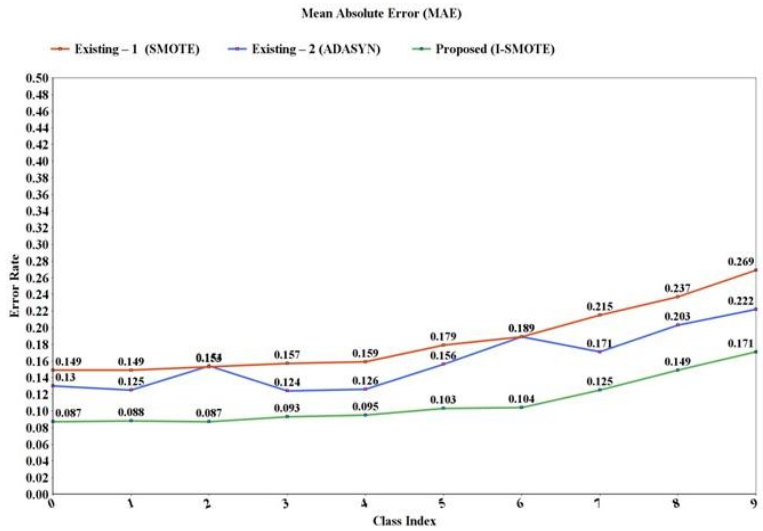

Fig. 4.2: Comparison of Error Rate.

4.3.3. Accuracy

Considering that n_{ij} is the number of models in section i that were separated into section j, the overall classification accuracy is as follows:

$$accu_{all} = \sum_i n_{ii} \; / \; \sum_{ij} n_{ij} \; \ldots equ \tag{4.1}$$

The following formula shows the accuracy of each classification:

$$accu_i = n_{ii} \; / \; \sum_j n_{ij} \; \ldots equ \tag{4.2}$$

The average time it takes to interpret a picture and generate a classification result is known as run time.

Table 4.3: Evaluation of Accuracy

S. No.	Class Index	Existing – 1 (SMOTE) (%)	Existing – 2 (ADASYN) (%)	Proposed (I-SMOTE) (%)
1.	0	88	91	98
2.	1	86	89	96
3.	2	84	87	94
4.	3	82	85	92
5.	4	80	83	90
6.	5	78	81	88
7.	6	76	79	86
8.	7	74	77	84
9.	8	72	75	82
10	9	70	73	80

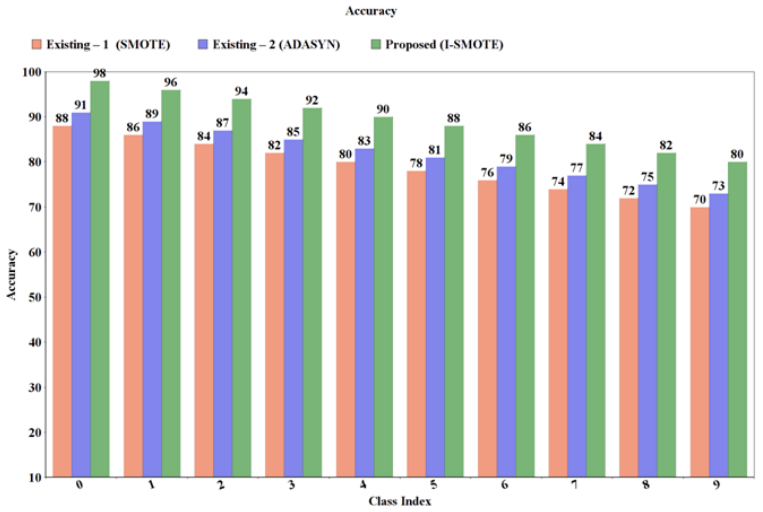

Fig. 4.3: Comparison of Accuracy.

5. Conclusion

Every educational institution presently requires an effective method for predicting student academic achievement. Furthermore, in a student's performance prediction model, overcoming data quality problems is typically the much more significant task. This research developed a two-level classifier model for predicting student performance using the supervised learning technique. This research presents a

better learning approach for forecasting student performance that entails learning a collection of two-level classifiers, each of which trains a subset of the entire set of labels. The proposed ensemble model, which includes the Random Forest (RF) classifier, was 98% accurate.. Researchers will test the suggested new ensemble model on a student dataset with much more training instances and larger label spaces in the future. The model can generate significant outcomes if the dataset and label space are sufficiently large.

References

[1] Amrieh, E. A., Hamtini, T., & Aljarah, I. (2016). Mining educational data to predict student's academic performance using ensemble methods. *International Journal of Database Theory and Application*, *9*(8), 119–136. https://doi.org/10.14257/ijdta.2016.9.8.13.

[2] Aggarwal, D., Mittal, S., & Bali, V. (2021). Significance of non-academic parameters for predicting student performance using ensemble learning techniques. *International Journal of System Dynamics Applications (IJSDA)*, *10*(3), 38–49. https://doi.org/10.4018/IJSDA.2021070103.

[3] Pandey, M., &Taruna, S. (2018). An ensemble-based decision support system for the students' academic performance prediction. In ICT Based Innovations (pp. 163-169). Springer, Singapore. https://doi.org/10.1007/978-981-10-6602-3_16.

[4] Devasia, T., Vinushree, T. P., &Hegde, V. (2016, March). Prediction of students performance using Educational Data Mining. In 2016 International Conference on Data Mining and Advanced Computing (SAPIENCE) (pp. 91-95). IEEE. https://doi.org/10.1109/SAPIENCE.2016.7684167.

[5] Adekitan, A. I., &Salau, O. (2020). Toward an improved learning process: the relevance of ethnicity to data mining prediction of students' performance. SN Applied Sciences, 2(1), 1-15. https://doi.org/10.1007/s42452-019-1752-1.

[6] Shingari, I., Kumar, D., & Khetan, M. (2017). A review of applications of data mining techniques for prediction of students' performance in higher education. Journal of Statistics and Management Systems, 20(4), 713-722. https://doi.org/10.1080/09720510.2017.1395191.

[7] Han, M., Tong, M., Chen, M., Liu, J., & Liu, C. (2017, July). Application of Ensemble Algorithm in Students' Performance Prediction. In 2017 6th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI) (pp. 735-740). IEEE. https://doi.org/10.1109/IIAI-AAI.2017.73.

[8] Livieris, I. E., Drakopoulou, K., Mikropoulos, T. A., Tampakas, V., &Pintelas, P. (2018). An ensemble-based semi-supervised approach for predicting students' performance. In Research on e-Learning and ICT in Education (pp. 25-42). Springer, Cham. https://doi.org/10.1007/978-3-319-95059-4_2.

[9] Rao, B. M., & Murthy, B. R. (2020). Prediction of student's educational performance using machine learning techniques. In Data Engineering and Communication Technology (pp. 429-440). Springer, Singapore. https://doi.org/10.1007/978-981-15-1097-7_36.

[10] Ade, R. (2019). Students performance prediction using hybrid classifier technique in incremental learning. International Journal of Business Intelligence and Data Mining, 15(2), 173-189. https://doi.org/10.1504/IJBIDM.2019.101255.

[11] Kumari, P., Jain, P. K., &Pamula, R. (2018, March). An efficient use of ensemble methods to predict students academic performance. In 2018 4th International Conference on Recent Advances in Information Technology (RAIT) (pp. 1-6). IEEE. https://doi.org/10.1109/RAIT.2018.8389056.

[12] Pandey, M., &Taruna, S. (2014). A comparative study of ensemble methods for students' performance modeling. International Journal of Computer Applications, 103(8). https://doi.org/10.5120/18095-9151.

[13] Hassan, H., Anuar, S., & Ahmad, N. B. (2019, May). Students' performance prediction model using meta-classifier approach.In International Conference on Engineering Applications of Neural Networks (pp. 221-231). Springer, Cham. https://doi.org/10.1007/978-3-030-20257-6_19.

[14] Ajibade, S. S. M., Ahmad, N. B. B., &Shamsuddin, S. M. (2019, August). Educational data mining: enhancement of student performance model using ensemble methods. In IOP Conference Series: Materials Science and Engineering (Vol. 551, No. 1, p. 012061). IOP Publishing. https://doi.org/10.1088/1757-899X/551/1/012061.

[15] Nespereira, C. G., Elhariri, E., El-Bendary, N., Vilas, A. F., & Redondo, R. P. D. (2016). Machine learning based classification approach for predicting students performance in blended learning. In The 1st International Conference on Advanced Intelligent System and Informatics (AISI2015), November 28-30, 2015, BeniSuef, Egypt (pp. 47-56). Springer, Cham. https://doi.org/10.1007/978-3-319-26690-9_5.

[16] Adejo, O. W., & Connolly, T. (2018). Predicting student academic performance using multi-model heterogeneous ensemble approach. Journal of Applied Research in Higher Education. https://doi.org/10.1108/JARHE-09-2017-0113.

[17] Abdullah, D. (2020). A linear antenna array for wireless communications. National Journal of Antennas and Propagation, 2(1), 19–24. https://doi.org/10.31838/NJAP/02.01.04.

[18] Barhoumi, E. M., Charabi, Y., & Farhani, S. (2024). Detailed guide to machine learning techniques in signal processing. Progress in Electronics and Communication Engineering, 2(1), 39–47.

[19] Parizi, L., Dobrigkeit, J., & Wirth, K. (2025). Trends in software development for embedded systems in cyber-physical systems. SCCTS Journal of Embedded Systems Design and Applications, 2(1), 57–66.