iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** IV **Month of publication:** April 2026

DOI: https://doi.org/10.22214/ijraset.2026.80129

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IV Apr 2026- Available at www.ijraset.com

Students Dropout Prediction Based on their Previous Score Using Various Machine Learning Algorithms

K. Sai Shri[1], D. Anil Kumar[2]
Department of Computer Science and Engineering, Gandhi Institute of Engineering and Technology University, Gunupur, Odisha, India

Abstract: ***Student dropout prediction is an important task in educational data analytics that can be used to intervene early and achieve improved academic outcomes. This paper introduces an effective prediction model using the past performance of the students in seven subjects, which include Mathematics, Physics, Chemistry, Biology, English, Social Science and Computer Science. To ensure uniformity and improve the model performance, preprocessing of the data is done through data cleaning, normalization and feature scaling. A number of baseline ML models are implemented such as LR, DT, SVM to be compared. A multi-layer neural network model is suggested to improve predictive power, which is expected to reflect intricate nonlinear relationships among scores of the subjects. Standard measures of performance of all models include accuracy, precision, recall and F1-score. Through experimentation, it is observed that the proposed neural network model will attain a peak accuracy of 89.60% in comparison to the baseline models. The results demonstrate the efficiency of deep learning methods to properly recognize at-risk students and provide educational intervention in a timely manner.***

"Keywords: Student Dropout Prediction, Neural Networks, Educational Data Mining, Academic Performance Analysis, Classification Models, Machine Learning."

I. INTRODUCTION

Since it allows schools to identify at-risk students and carry out prompt interventions to improve academic achievements, student dropout prediction has become a crucial field of research in educational data mining. The increasing accessibility of information about students particularly the records of academic achievement has simplified predictive models that may look at trends associated with student achievement or failure. The initial research provided the basis of future research in this field, including the work of Dekker et al. [1], that demonstrated that data mining could be applied to predict student dropout. Similarly, Márquez-Vera et al. [2] highlighted the importance of academic indicators in prediction activities and demonstrated the effectiveness of classification algorithms in the early detection of dropouts based on high school data.

Academic achievement has been proven severally to be a good predictor of student retention. Although newer studies by Realinho et al. [4] have validated the significance of integrating academic and behavioral data to enhance the prediction effects, Nagy and Molontay [3] showed that performance in the secondary school has a very strong effect on predicting school drop-out in higher education. Combined, these studies emphasize the importance of examining previous educational data to help predict student performances more accurately. Moreover, the increasing use of data-driven decision-making in learning institutions has only intensified the role of predictive analytics in increasing student success rates and streamlining the performance of institutions.

Although this has occurred, common machine learning algorithms often cannot learn complex and nonlinear relationships among different academic factors. In the high-dimensional data, tree-based algorithms are prone to overfitting, and linear models, such as logistic regression, might fail to extract complex associations among scores in the subjects. Additionally, these models often heavily depend on manual feature engineering, potentially not exploiting the inherent patterns in the data. The increased complexity of educational data as input features demands more and more powerful methods capable of learning deep representations.

The neural networks provide a potential solution to these issues due to the ability to automatically learn feature hierarchies and capture nonlinear relationships. By applying multiple layers of neural networks and nonlinear activation functions, hidden patterns and interactions between subject-wise features can be efficiently captured by neural networks. They are particularly appropriate in addressing complex academic data due to this flexibility, in which various interconnected variables influence student achievement.

©IJRASET: All Rights are Reserved

This study presents a neural network-based predictive model to predict the dropout of students because of their previous performance in different subjects. The methodology will be used to process data on performance by subject, and classify students more accurately into pass or fail groups. The proposed approach tries to enhance prediction performance by effectively extracting interdependencies among subject-wise features. To demonstrate the effectiveness and robustness of the model, it is also compared to conventional machine learning techniques. This study found that the neural network models have the potential to improve the accuracy and reliability of the student dropout prediction system, which is useful in proactive academic interventions and efficient decision-making in the academic setting.

II. RELATED WORK

The idea of predicting student dropout has become popular in the last several years, and multiple studies have been conducted to determine how machine learning and artificial intelligence techniques can be used to predict student dropout more accurately. Lee and Chung suggested a machine learning-based early warning system that identifies children who are at risk using a range of categorization techniques [5]. Their approach demonstrated that supervised learning models can significantly enhance prediction effectiveness when multiple features are used to predict academic performance, and timely academic interventions can be implemented. Another importance of automated technologies in the study was the ability of the teachers to monitor the progress of the students. Applying such ML algorithms in the context of educational data mining, Del Bonifro et al. [6] investigated the possibility of predicting student dropout. Their study has emphasized the importance of data preparation and feature selection to increase the efficacy of the model. Comparing several classification models, they showed that no one of the models is efficient in all cases and the choice of an algorithm depends on the characteristics and features of the dataset. Their results highlight the need to customize prediction models for certain educational settings. Similarly, Tenpipat and Akkarajitsakul [7] presented a case study, which employed institutional data to predict the rates of student dropout. To classify students based on their likelihood of dropping out, their work has been based on traditional ML algorithms such as Decision Trees and Support Vector Machines. The results indicated that although these models can achieve an acceptable level of accuracy, they often fail when dealing with complex and nonlinear interactions among input features. This limitation underscores the necessity of more advanced models that are able to process multidimensional academic data. Recent developments have been on the use of more complex models to improve the performance of prediction. Christou et al. In a study on applying the machine learning procedures to estimate early dropout rates in higher education, [8] found that a few features using advanced classification procedures gave greater accuracy. Their results indicate that the combination of various data sources and the tuning of the parameters of the models can greatly contribute to the accuracy of forecasts. This points to the significance of utilizing data in depth to come up with reliable prediction systems.

Song et al. introduced another comprehensive dropout prediction method, which analyzes performance of students over an academic year [9]. Their model involves sophisticated learning plans and time based information to constantly monitor the progress of the students. The paper highlights the importance of the dynamic model to capture the dynamic trends in academics and improve the accuracy of the forecasts as time goes by. These strategies are quite useful in real life scenarios whereby there is change in performance by students in different semesters.

Moreover, Kim et al. [10] introduced a high-performance dropout prediction model based on the latest machine learning methods to enhance accuracy and recall. Their approach demonstrates that the size of the advancements in the accuracy of predictions that can be obtained through the refinement of training processes and shaping the model design can be enormous. It is also revealed in the study that to achieve high-performance results, it is necessary to perform hyperparameter optimization and model refinement.

Besides these strategies, the combination of hybrid and ensemble learning methods in enhancing resilience in prediction has emerged as a key area of emerging research endeavors. It has been established that integration of various models enhances better generalization and reduces variance. Moreover, since deep learning methods have the capability of automatically learning feature representations without much manual feature engineering, they are gaining popularity. These gains are indicative of a shift in traditional models and towards more complex systems capable of handling large-scale, high-dimensional educational data.

Although these studies have been based on major contributions to the field, most of the existing techniques are based on traditional machine learning algorithms that may not be sufficient in reflecting the complexity of the relationships among numerous scholarly factors. Also, not all models can be flexible and scaled to work with diverse educational datasets. These limitations influence the increasing demand of more advanced models such as neural networks that are able to learn nonlinear relationships and improve the overall dropout prediction system of students. By utilizing a neural network-based strategy to increase prediction accuracy and dependability, the suggested work fills up these gaps.

III. METHODOLOGY

A. Proposed System

The proposed model is a neural network model that predicts the student dropout (pass/fail outcome) based on previous academic performance. The system receives the scores of students in the seven subjects, mathematics, physics, chemistry, biology, English, social science and computer science. These inputs are processed with the help of a multi-layer neural network because it is necessary to reflect complex relationships between academic attributes. The model is split into two groups: Pass and Fail. The proposed approach will maximize the prediction accuracy using nonlinear interactions among features and model parameters optimized through gradient-based learning and backpropagation.

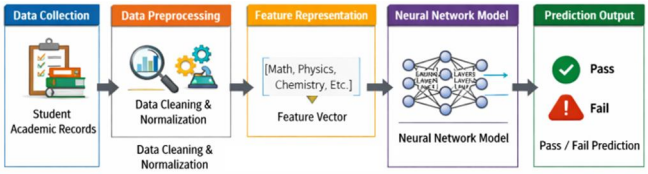


Fig.1 Proposed Architecture

Fig. 1 shows the overall architecture of the system. Some of its stages include the data input, preprocessing, feature representation, training the model and prediction. Raw student data is preprocessed in order to eliminate any inconsistencies. After processing, the input is turned into feature vectors and processed through neural network and machine learning models. Besides classification, the proposed neural network produces the ultimate prediction result. The architecture ensures that there is effective model training and a structured data flow.

B. Dataset Description

One thousand student records gathered from academic performance data make up the dataset used in this study. Scores from seven subjects—mathematics, physics, chemistry, biology, English, social science, and computer science—are included in each record. The information will show how well pupils did on internal and yearly exams. The target variable is used in presenting the final score of each student which is either Pass or Fail. The dataset is balanced between the two classes to allow the unbiased training and testing of the model. It is this structured information that enables the analysis of academic trends and prediction of student outcomes.

C. Data Pre-processing

Data preprocessing is necessary to process raw academic data to enable model training and evaluation. It ensures that the data is of quality, consistent and can be used in the machine learning procedures by controlling the missing values, outliers, scaling as well as encoding.

1) *Missing Value Handling:* Missing values are a common occurrence in real-life educational datasets due to missing records or issues with data collection. This is dealt with by using mean imputation of the numerical attributes which replaces the mean of the subject scores of the missing scores. The process ensures that there is no loss of important records since the general distribution of data is preserved. The data is made complete by taking care to fill the gaps and that will be processed further. This measure will ensure that null or undefined values will not lead to a malfunction of machine learning models.
2) *Outlier Detection and Treatment:* Outliers can significantly affect the performance of a model since they add bias and noise to the data. Based on the distribution of subject evaluations, statistical techniques are used in this study to identify the outliers. In order to maintain the integrity of data, extreme values which have a huge deviation with the normal range are monitored and dealt with closely. This may involve the regularization of the abnormal changes or clipping of values within a reasonable range. The fewer the outliers the more stable and reliable the dataset is, and the model will be less affected by the irregular data values, and more likely to find meaningful patterns.
3) *Feature Scaling:* In order to ensure that each input variable makes an equal contribution to the model, feature scaling is essential. Min-Max normalization is applied to convert the subject scores in this work to a standardized range between 0 and 1. This technique guarantees uniformity of all the inputs by rescaling every feature based on its minimum and maximum values. Normalization also accelerates the convergence of optimization algorithms in addition to avoiding the characteristics with broader number ranges to dominate the learning process. This means that the neural network model is able to learn more effectively and is more effective with regard to prediction.

4) *Label Encoding:* The aim variable is categorical (Pass/Fail) that illustrates the student outcome. These categories are then coded to a number (label encoding) to enable them to be subjected to classification techniques. In this the code of pass and fail is 1 and 0 respectively. This binary shape assists the model train, and satisfies the needs of the binary classification problem. Using an appropriately encoded output variable, the output will be understandable and measurable in terms of accuracy, precision, recall and F1-score.

D. Feature Representation

A feature vector that consists of subject-scores characterizes each student record. The input feature vector will be as follows:

$$X = [x_1, x_2, x_3, x_4, x_5, x_6, x_7]$$

where x1 is used as a representation of the scores of the seven subjects. Academic data can be effectively processed by this numerical form through which the model works. The characteristics are continuous and standardized to ensure that the input is consistent throughout the training. The binary output variable is shown in the following way:

$$Y \in \{0,1\}$$

In which 1 represents prosperity and 0 represents failure. This structured representation enables the effective classification based on the machine learning and neural network models.

E. Algorithms

This study employs a number of machine learning methods to identify the performance and compare the results with the proposed neural network structure. Due to their effectiveness, computational efficiency, and usage, the models are selected to be utilized in performing binary classification activities in educational data research. The different insights that each algorithm brings about the trends in academic data learning makes it possible to have a comprehensive comparison.

1) *Logistic Regression:* A statistical classification model called logistic regression uses a sigmoid function to forecast the likelihood of a binary result. It estimates coefficients to maximize probability and therefore it is a model of the relationship between input features and output. It is easy, ordinary and effective in comparing baseline in the classification tasks.

$$P(y = 1 \mid X) = \frac{1}{1 + e^{-(W^T x + b)}} \quad (1)$$

Logistic regression is a simple baseline model due to its simplicity and interpretability. It is assumed that the relationship between the dependent variable and independent variables are linear. Although it cannot address nonlinear interactions, it provides a convenient point of reference to evaluate more complex models. Its low computational cost and easy implementation make it very suitable in classification tasks in the early analysis and performance evaluation.

2) *Decision Tree:* Supervised learning technique called Decision Tree is a method which divides the dataset into subsets in accordance to feature values. It forms a tree-like structure with the decisions at the central nodes and the outcomes at the peripheral nodes. During the process of dealing with complex data, overfitting may take place even though it is easy to interpret.

$$I(i) = 1 - \sum_{i=1}^{k} p_i^2 \quad (2)$$

Decision tree is highly interpretable and can represent nonlinear relationships in the feature space by dividing it recursively. They are capable of working with both numeric and non-numeric data and provide some criteria to be applied to decisions. Moreover, they are susceptible to overfitting, especially as the tree becomes deep and picks noise in the data. Pruning and parameter readjustment is often required to maintain the stability of the model and improve its generalization.

3) *Support Vector Machine (SVM):* The high-dimensional space powerful classification method SVM identifies the best hyperplane to divide classes. It is most effective when the maximum distance between classes of data points is maximized and in cases where a set of separation limits are available.

$$minimize \ \frac{1}{2} ||W||^2 + C \sum_{i=1}^{n} \xi_i \quad (3)$$

SVM can be effectively applied to data of high dimensions and has the advantage of being able to utilize kernel functions to replicate the appearance of a complicated decision boundary. It aims at maximizing the distance between classes in order to improve on generalization. But in big data, SVM may be computationally expensive and require a tuned choice of kernel parameters. Even

with these challenges, it remains a strong categorizer of structured data in which the separation of classes is evident despite these challenges.

4) *Proposed Neural Network:* The proposed model is a MLP consisting of an output layer that has a sigmoid activation function, nonlinear activation functions (ReLU) in the hidden layers, and seven neurons in the input layer. Backpropagation with binary cross-entropy loss and the Adam optimizer is used to train the model. This is due to the fact that compared to the input features, it is more effective in predicting performance because it can measure the non linear relationships between the two.

The proposed neural network model will help to overcome the disadvantages of the traditional approach by automatically learning the complex associations between features. The model is able to use multiple hidden layers to offer hierarchical features representations based on subject-wise scores. The sigmoid function makes the binary classification to be properly outputted and the ReLU activation function improves the training performance. Adam optimizer also promotes stability and speed of convergence of the training. The model has the capability to identify the intricate patterns within the academic data and this is the reason why it is able to operate at a higher level.

F. Evaluation Metrics

In order to ensure that the models are fully explored, the performance of the models is evaluated based on the conventional categorization measures. These measures gauge various elements of prediction accuracy and reliability.

1) *Accuracy:* Accuracy by dividing the number of correctly predicted cases by the total number of cases, determines the overall accuracy of the model. It does generalize the signal of the model performance in comparison between 2 classes.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (4)$$

It is among the most frequently used evaluation metrics since accuracy is used to determine the overall performance in classification tasks of a model. It serves well with balanced datasets as it considers both the well-anticipated positive and negative instances. Accuracy, however, does not distinguish among the various types of classification errors, and thus may not provide a complete picture when dealing with unbalanced data. Consequently, it is often used together with other measures to yield a more reliable evaluation of the performance of models.

2) *Precision:* Precision is the percentage of correctly predicted positive cases of all the positive cases predicted. It depicts the percentage of the correctly predicted positive cases.

$$Precision = \frac{TP}{TP + FP} \quad (5)$$

Precision is extremely important when it is necessary to minimize false positives. When it comes to predicting student dropout, a high precision means that students who are anticipated to be at-risk are indeed likely to fail, which minimizes needless interventions. It demonstrates the extent to which the model can be used to make favorable predictions. The higher the value of precision, the more the prediction system can be sure to detect true positive situations and few false positive classifications.

3) *Recall:* Recall is the percentage of correctly predicted positive cases in the actual positive cases. It assesses how well the model can find all pertinent positive cases.

$$Recall = \frac{TP}{TP + FN} \quad (6)$$

In applications where missing affirmative cases could have dire repercussions, recall is essential. Recalling in this study indicates the ability of the model to identify students who are indeed at risk of failure. An increase in the recall value implies that the model reduces false negatives by picking most of the true positive cases. This is particularly important in educational systems as a failure to identify the at-risk pupils may lead to the loss of the opportunity to receive timely assistance and help.

4) *F1-Score:* The harmonic mean of the precision and recall is the F1- Score which provides a fair evaluation of the model.

$$F1\ Score = 2 * \frac{Precision \times Recall}{Precision + Recall} \quad (7)$$

The F1-score is particularly helpful when it is necessary to take into account both false positives and false negatives at the same time since it offers a single statistic that strikes a compromise between precision and recall. It is particularly effective with datasets that are not precisely balanced. With a greater F1-score, the model can be assured to maintain a stable and reliable performance because it indicates that the model has a good balance between recall and precision. Due to this reason, it is an important pointer to gauge classification models in real-world applications such as student dropout prediction.

IV. RESULTS AND DISCUSSION

The results of the proposed neural network model are compared and assessed with the traditional machine learning algorithms, such as LR, DT, and SVM. The assessment is conducted based on typical measures of classification like Accuracy, Precision, Recall, and F1-Score. Table 1 shows the comparative results.

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	81.20	79.85	80.40	80.12
Decision Tree	83.75	82.10	83.30	82.69
Support Vector Machine (SVM)	85.10	84.25	85.00	84.62
Proposed Neural Network	89.60	88.90	89.20	89.05

Table.1 Performance Comparison of Different Models

The proposed neural network model has accuracy (89.60), precision (88.90), recall (89.20) and F1-score (89.05) which is better than any other model that was considered as shown in Table 1. In comparison, Logistic Regression has an accuracy of 81.20, Decision Tree, 83.75, and SVM, 85.10. Although SVM still performs better than Decision Trees and Logistic Regression, it is not as good as the proposed model in all aspects of evaluation. The performance improvement can be attributed to the ability of the neural network to determine complex and nonlinear relationships between the scores of a number of subjects. Although DT can be overfitted and unstable, the older models, such as Logistic Regression, are limited to linear decision boundaries. Despite SVM having better generalization, it may fail to model complex interaction of features in the case of educational data.

Also, the proposed model has balanced accuracy and recall rates, which contribute to the increased F1-score. This indicates that the model is always able to point out the instances of pass and fail as well as predicting the results of the students accurately. This is necessary in educational uses, where false positives and false negatives can be of critical importance.

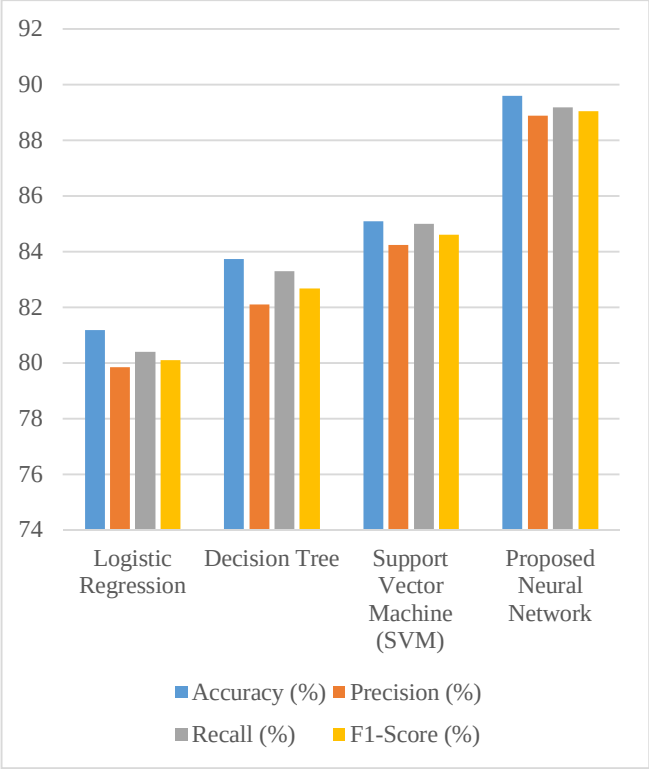


Fig. 2 is a graphical comparison of all the models using a bar chart approach of all the evaluation measures. The proposed neural network is more effective compared to the baseline approaches, which is apparent in the visualization. The effectiveness and strength of the proposed solution are validated by the gradual increase in all parameters.

In general, the results of the experiments prove that the proposed neural network model is a reliable and efficient means of predicting student dropout. It can better handle complex feature correlations than traditional machine learning algorithms, making it more suitable to academic performance analysis.

A. Discussion

According to the results, since the proposed neural network model can model the complex nonlinear interactions among subject-wise information, it is always effective in comparison with the traditional methods of machine learning. The difference of the performance of the proposed model and the baseline methods highlights the inefficiency of conventional algorithms in dealing with multi-dimensional academic content. In particular, the neural network is effective to capture topic interdependencies, facilitating a balanced metric performance and classification accuracy. Decision trees and logistic regression are rather weak in their performance, which means that they are sensitive to data distribution and feature interactions. Moreover, the stability of the precision and recall values testifies to the possibility of the proposed model to make correct predictions in both classes, which can justify its application in the real educational decision-making systems.

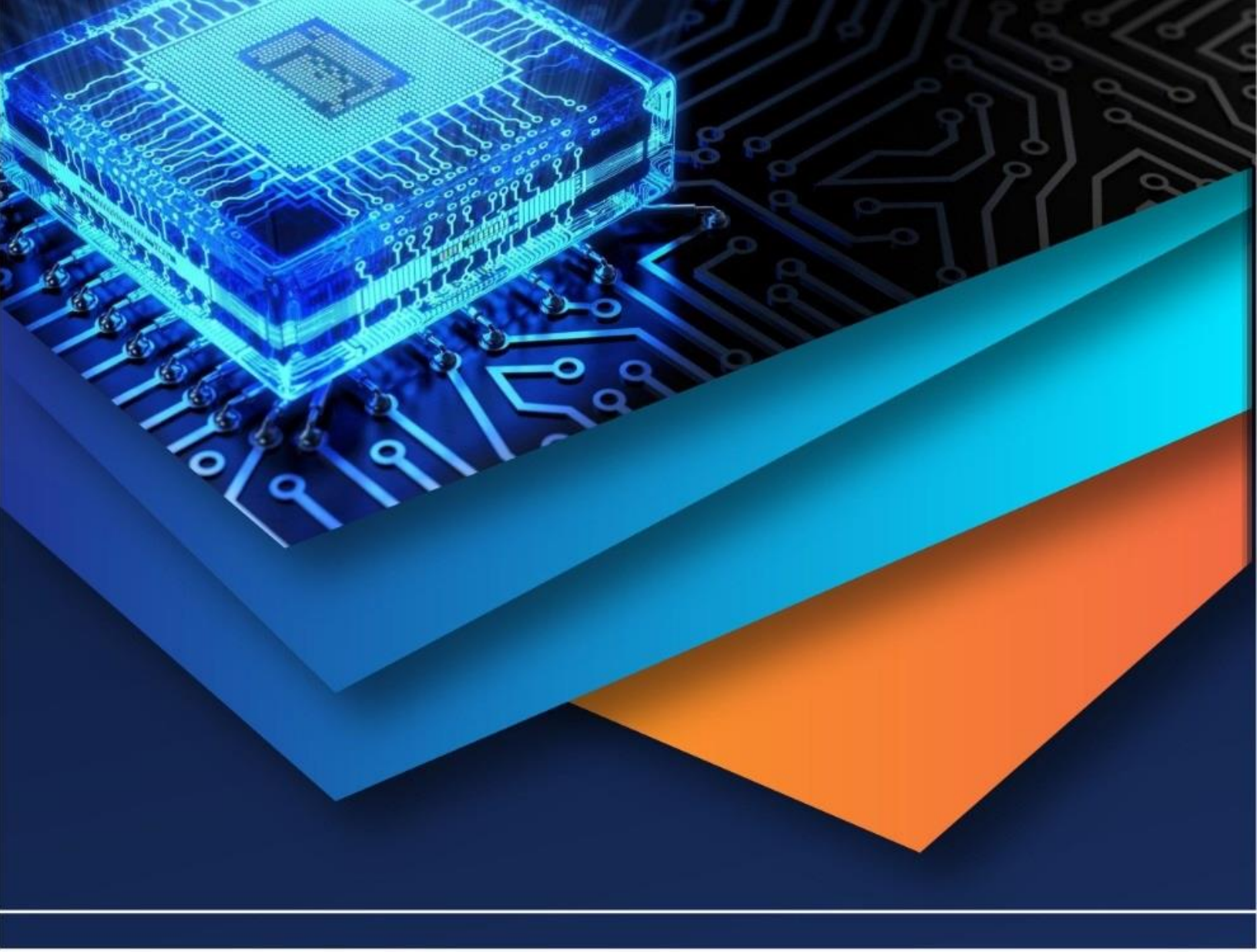
V. CONCLUSION

In this research, a set of data regarding academic performance of various participants was used to create a neural network-based predictor of student dropout. The authors compared the effectiveness of the proposed model to traditional ML methods, which are LR, DT, and SVM, where subject-wise scores are used as input data. The results of the experiment proved the ability of the proposed neural network model to classify students as pass or fail correctly by demonstrating that it outperforms in all the assessment parameters. The model is able to capture the complex and nonlinear interactions between academic variables, which traditional models often fail to capture, as reflected in the higher accuracy, precision, recall, and F1-score. Moreover, the equal level of performance according to a wide range of measures underlines the reliability and stability of the proposed approach. The results highlight how crucial it is to use advanced learning strategies in educational data analysis in order to facilitate the early detection of pupils who are at danger. These prediction algorithms would assist educators and institutions to make informed decisions and implement timely interventions in order to enhance the student success rates and reduce the rate of dropouts.

To further improve the performance of prediction, future studies can focus on the further development of the proposed model by incorporating other features such as attendance history, behavioral profiles, and socioeconomic factors. More complex deep learning structures like recurrent neural networks and hybrid systems can be added to establish temporal relationships in student performance data. Moreover, the implementation of the model in the real-time educational systems can help to support the proactive approach to intervention, as it will help to monitor and predict dynamically. Future research could also explore model interpretability measures in order to provide instructors with useful information.

REFERENCES

[1] G. W. Dekker, M. Pechenizkiy, and J. M. Vleeshouwers, “Predicting students drop out: A case study,” in Proc. 2nd Int. Conf. Educational Data Mining (EDM), Cordoba, Spain, Jul. 2009, pp. 41–50.
[2] C. Márquez-Vera, A. Cano, C. Romero, A. Y. M. Noaman, H. Mousa Fardoun, and S. Ventura, “Early dropout prediction using data mining: A case study with high school students,” Expert Systems, vol. 33, no. 1, pp. 107–124, 2016.
[3] M. Nagy and R. Molontay, “Predicting dropout in higher education based on secondary school performance,” in 2018 IEEE 22nd Int. Conf. Intelligent Engineering Systems (INES), 2018, pp. 389–394.
[4] V. Realinho, J. Machado, L. Baptista, and M. V. Martins, “Predicting student dropout and academic success,” Data, vol. 7, no. 11, p. 146, 2022.
[5] S. Lee and J. Y. Chung, “The machine learning-based dropout early warning system for improving the performance of dropout prediction,” Applied Sciences, vol. 9, no. 15, p. 3093, 2019.
[6] F. Del Bonifro, M. Gabbrielli, G. Lisanti, and S. P. Zingaro, “Student dropout prediction,” in Proc. Int. Conf. Artificial Intelligence in Education (AIED), Cham, Switzerland: Springer, Jun. 2020, pp. 129–140.
[7] W. Tenpipat and K. Akkarajitsakul, “Student dropout prediction: A KMUTT case study,” in 2020 1st Int. Conf. Big Data Analytics and Practices (IBDAP), 2020, pp. 1–5.
[8] V. Christou et al., “Performance and early drop prediction for higher education students using machine learning,” Expert Systems with Applications, vol. 225, p. 120079, 2023.
[9] 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, p. 1143, 2023.
[10] S. Kim, E. Choi, Y. K. Jun, and S. Lee, “Student dropout prediction for university with high precision and recall,” Applied Sciences, vol. 13, no. 10, p. 6275, 2023.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)