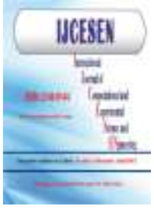
International Journal of Computational and Experimental Science and ENgineering
(IJCESEN)

Vol. 11-No.2 (2025) pp. 2135-2141
http://www.ijcesen.com



ISSN: 2149-9144

Research Article

Students Performance Prediction by EDA Analysis and Hybrid Deep Learning Algorithms

M. Kannan[1]*, K. R. Ananthapadmanaban[2]

[1]Department of Computer Science, College of Science and Humanities, SRM Institute of Science and Technology, Vadapalani Campus, Chennai, Tamilnadu
* **Corresponding Author Email:** km1200@srmist.edu.in - **ORCID:** 0000-0002-6907-3086

[2]Department of Computer Science, College of Science and Humanities, SRM Institute of Science and Technology, Vadapalani Campus, Chennai, Tamilnadu
Email: toananths@gmail.com - **ORCID:** 0000-001-5430-3355

Article Info:

DOI: 10.22399/ijcesen.1524
Received : 20 January 2025
Accepted : 23 March 2025

Keywords :

Deep Learning,
Students Performance Prediction,
Hybrid Model,
DNN-RF,
DNN-Light GBM.

Abstract:

Education is a pillar of any individual to attain success in their life. Knowledge evaluate students' performance which resulted with low accuracy and many algorithms not able to manage imbalanced dataset. This research utilized the ML algorithms, EDA development and learning makes everyone become educated person. Many universities and colleges lend graduate course of study for various disciplines, and students choose courses based on interest. At the same time many researches consider normal factors like, personal and academic features, experimented with many machine learning models and analysis and Hybrid algorithms for students' performance prediction. Exploratory data analysis performed to identify the correlation between features and features which support the evaluation of student's performance prediction. Based on the evidence from the EDA analysis this paper aims to provide a deep learning-based hybrid approach that consists of Deep Neural Network -Random Forest (DNN-RF), Deep Neural Network -Light GBM (DNN-Light GBM) algorithms to evaluate the students' performance prediction that capable of handling a wide range of datasets from small to enormous and improve the prediction accuracy. The results shows that the Deep Neural Network -Random Forest achieved an accuracy of 99.56%, precision of 97.82%, recall of 98.13%, f1 score of 98.95% and DNN-Light GBM attained an accuracy of 90.76%, 85.13%, 84.94%, 87.93%. while comparing to ML algorithms RF, Light GBM and DNN-Light GBM, DNN-RF is utmost effective algorithm for forecasting student performance.

1. Introduction

Academic performance can be evaluated from several perspectives. The extent to which a student has learned and understands the course material. Students' effort and dedication to their studies include attendance, organization, and time management. A student's behavior in the classroom includes participation, cooperation, and attention to the teacher and peers. The results of exams, quizzes, and other assessments reflect a student's knowledge and understanding of the course material. A summary of a student's academic performance is often based on a combination of the above factors. Students' ability to interact with others, manage emotions, and develop positive relationships can impact their academic success.

Content mastery is achieved through active learning, practice, and consistent engagement with the material. Here are a few strategies that can help you achieve mastery of a subject. Engage with the material through reading, writing, discussing, and solving problems. Regular practice is key to developing mastery. It could include taking practice tests, working on assignments, and solving problems. Consistent engagement with the material over an extended period is critical for mastery. Set aside dedicated time each day or each week to review and practice the material. Seek out feedback on your performance and reflect on your strengths and weaknesses. By combining these strategies, this study builds a comprehensive approach to mastering content and achieving academic success. Figure 1 shows elements affecting students

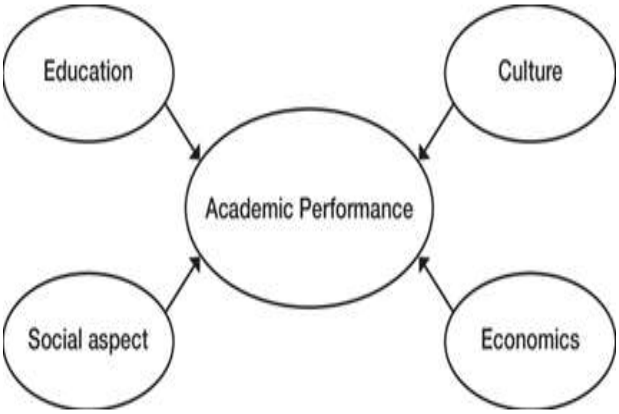


Figure 1. *Elements Affecting Students' Performance*

performance. There are several problems that educational institutions face when evaluating academic performance. BIAS in the evaluation process can lead to inaccurate assessments of student performance and result in unequal opportunities. Reliance on traditional assessment forms, such as tests and exams, may not accurately reflect a student's overall performance, leading to under or over- assessment. Different educational institutions may use different methods for evaluating academic performance, making it difficult to compare students' performance from different institutions. Many institutions may not have access to comprehensive data on student performance, making it difficult to evaluate performance accurately. Evaluating academic performance is often subjective and may be influenced by personal biases, cultural and socioeconomic background, and more. Educational institutions may not have the resources to invest in modern and sophisticated evaluation methods, leading to inaccurate evaluations of student performance. Overall, accurately evaluating academic performance is a complex and challenging task that requires addressing these issues and implementing effective methods and systems to ensure fairness, objectivity, and accuracy.

Online courses and studying through browsing increased the interest of students learning and which create high impact on their knowledge level. After COVID-19 every institution strictly follows the online teaching and learning strategy to face the global challenge in students job security. This study considers various features for students' performance prediction which includes Student, Age, Gender, GPA_Last_Semester, Attendance, Study_Hours_Per_Week, MOOC_Participation, Online_Courses_Hours, Spoken_Tutorial_Participation, Extracurricular_Activity_Score, Test Scores, GPA_Current for the experiment.

1.1 Problem Statement

The education system produces a large amount of data related to student performance, such as test scores, attendance, and other metrics. However, manually analyzing massive amounts of students' data by manual is time-consuming and tedious. Therefore, there is a need for an automated system that can analyze the data to identify patterns and predict future outcomes. This research aims to use deep learning algorithms to analyze student performance data and predict their future academic performance. The system is trained using a dataset of student performance data, and the final model is used to predict student performance, identify at-risk students, and provide insights into the factors that affect academic performance. The goal is to develop a system that can provide actionable insights to educators and administrators to help them make informed decisions to improve student outcomes.

2. Literature Review

Sankar et al., (2021), summarized deep learning techniques such as discriminative learning, unsupervised learning and hybrid learning for research which guide academic and industry development [1]. Alshamaila et al., (2024) investigated the data from Jordan UG students and analysed the features demographic information, student's majors, faculty, high school average and four semester marks. Author suggest gmean for students' performance analysis [2]. Alnasyan et al., (2024) reviewed and analysed various public and private datasets and identify the major utilizing deep learning techniques such as DNN, CNN, LSTM, CNN-LSTM for student's performance prediction. Author observed that above mentioned techniques achieved 90% accuracy and LMSs dataset is used at 44% while QULAD is mostly used dataset [3]. Kannan et al., (2023) analysed graph Neural Network for students' performance prediction and accomplished an accuracy of 91.95%. Liu et al., (2021), Proposed LSTM variants and soft attention mechanism to capture student profile-aware representation. This model tested with three different datasets and real-world dataset [4].

Khan et al., (2021) experimented the attention based BiLSTM for students' performance prediction and attained an accuracy of 90.16% [5].

Ravi et al., (2024) performed EDA analysis to visualize students' performance in different features of the LMS dataset with machine learning techniques KNN , Multiple Regression for same analysis and attained 99% accuracy for Multiple

Regression technique [6]. Srivastava et al., (2023), analyzed students demographic features for students' performance prediction [7].

2.1 Research Gab

- Many states of art method failed to consider online learning.
- Most of the experiments utilized only machine learning algorithms.

3. Methodology

This research involves the development of a predictive model that can analyze large datasets and provide insights into the factors that affect students' academic performance. The model uses hybrid deep learning algorithms to analyze data from various sources, including students past academic records, demographic information, and external factors online access resources. This study also includes identifying patterns and trends that can be used to predict students' future academic performance, as well as developing actionable insights to support teaching and learning strategies. Hence EDA analysis supports to deliver the prominent feature sof the dataset. The model can help educators identify students at risk of poor academic performance and intervene early to provide targeted support. It can also help identify individual students' strengths and weaknesses, which can inform personalized learning plans to enhance student learning outcomes.

3.1 Dataset

The dataset consists of information about the SRM Arts and Science College students and which includes Student ID, Age, Gender, GPA_Last_Semester,Attendance,Study_Hours_Per_Week,MOOC_Participation,Online_Courses_Hours,Spoken_Tutorial_Participation,Extracurricular_Activity_Score, Test_Scores, Percurrent for the experiment.

3.2 Data preprocessing

Data preprocessing is done to remove the null values, unwanted and repeated data. The data structures are converted to a single dimension to suit the spatial dimension of the data. During this process, all the features in the dataset are normalized. Data structure plays an essential role in the performance of the model. Data preprocessing helps improve performance and reduce computational complexity.

3.3 Exploratory Data Analysis

It helps in understanding structure, pattern and quality of data. Hence it shows the relationship between variables, anamolies detection such as missing values, null values and outliers, correlation, trends also identified. Visualization is significant property of EDA which simply presents the details of data.

3.4 Feature Extraction

It is important to analyze the features and extract the required data. By extracting only the important features, the size and computational complexity of the model can be reduced. For this process, DNN is used. It only extracts the important features from the dataset. The model aims to find the relation between different variables to evaluate the overall performance of the student. Figure 2 show sthe overall flow of the model.

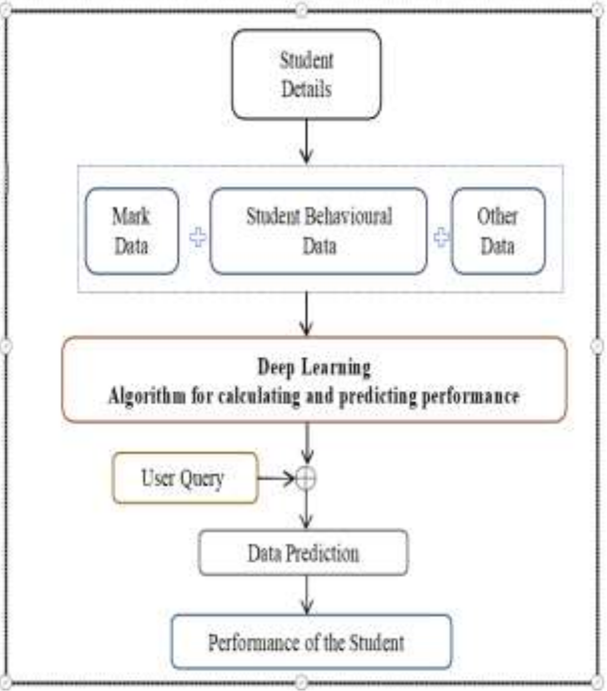


***Figure 2**. Overall workflow*

Before finding the performance of the proposed model, the total number of data collected from the student dataset is classified into two phases training and testing. For training, 80% of the data are used, and the reaming 20% of the data are used for testing. The final prediction result is evaluated using the result obtained from various stages. The main goal of this research work is to visualize the students' performance effectively. And it is more beneficial to the parents, teachers, colleges, and other education units to analyze the problem among the students during classes and find the optimal solution to those issues.

3.5 Deep Neural Network

This model combines the property of DNN and RF to create the hybrid model which utilized for students' performance prediction. It belongs to the type of artificial neural network and consist of input layer, hidden layer and output layer. Every node act as neuron and input layer receives the input and hidden layer process the input and output layer produce the prediction and each layer performs the weight, bias with the help of activation function, Adam optimizer. These predictions were evaluated by evaluation metrics. Figure 3 presents the architecture of DNN

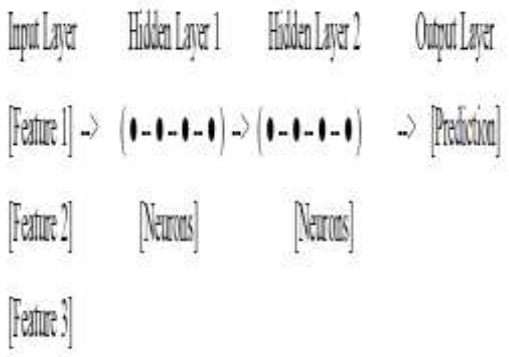


Figure 3. *Architecture of DNN*

3.6 Random Forest

Random Forest works based on the concept of decision tree, which utilized multiple decision trees and each tree is trained by training data and features. It produces good result even for high dimensional data and avoid the Overfitting. It also produced feature importance score for predictions. Figure 4 label, the architecture of RF.

3.7 Light GBM

It works on the concept of decision trees, It is a framework of gradient boosting and ach tree rectify the errors created by previous one. Hence it improves the performance of the model. It carryout operations in leaf wise and choose the leaf which minimize the error highly and splits it. It also used histogram-based binning to capture continues features and gradient based one-time sampling reduce the computation time. Table 1 show sthe parameters in light GBM. Figure 5 shows the architecture of Light GBM.

Steps involved in Deep Neural Network- Light GBM

Step 1: Data Preprocessing

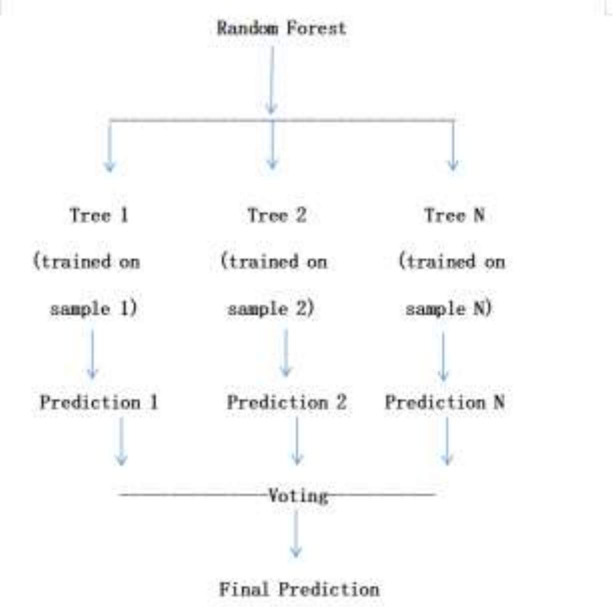


Figure 4*. RF architecture*

Table 1. S*the parameters in light GBM.*

S.No	Parameters	Activity
1.	num_leaves	Maximum number of leaves controlled to avoid Overfitting
2.	max_depth	Controls the depth of each tree
3.	learning_rate	Fix the size of step for boosting iteration
4.	Nestimators	Ensemble tree numbers
5.	Min_data_in_leaf	Fits minimum number of samples
6.	Feature_fraction	Selects features for iteration

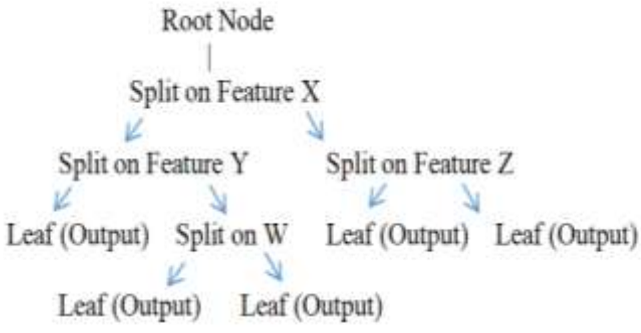
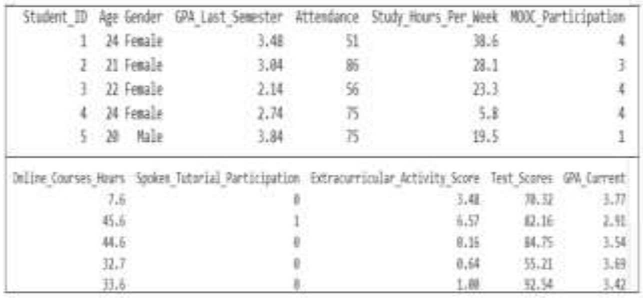


Figure 5. *Architecture of Light GBM*

Step 2: Split the dataset into training and testing
Step 3: Train the DNN model
Step 4: Train the Light GBM model
Step 5: Train the meta model by combining outcome of DNN and Light GBM model for prediction

Step6: Outcome is evaluated by evaluation metrics
Steps involved in Deep Neural Network- Random Forets
Step 1: Data Preprocessing
Step 2: Split the dataset into training and testing
Step 3: feature extraction by DNN
Step 4: RF get input from the last layer of DNN
Step 5: Predication performed by DNN-RF
Step6 : Outcome is evaluated by evaluation metrics

4. Result and Discussion

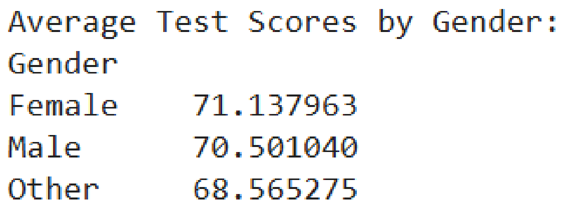


```
   Student_ID  Age Gender  GPA_Last_Semester  Attendance  Study_Hours_Per_Week  MOOC_Participation
0           1   24 Female               3.48          51                  38.6                   4
1           2   21 Female               3.04          86                  28.1                   3
2           3   22 Female               2.14          56                  23.3                   4
3           4   24 Female               2.74          75                   5.8                   4
4           5   20   Male               3.84          75                  19.5                   1

   Online_Courses_Hours  Spoken_Tutorial_Participation  Extracurricular_Activity_Score  Test_Scores  GPA_Current
                    7.6                              0                            3.48        70.32         3.77
                   45.6                              1                            6.57        82.16         2.91
                   44.6                              0                            0.16        84.75         3.54
                   32.7                              0                            0.64        55.21         3.69
                   33.6                              0                            1.00        92.54         3.42
```

Figure 6. *Data Description*

Figure 6 presents the head features of data.
Figure 6 shows the number of features in the dataset. Figure 7 depicts the average test score of female and male.



```
Average Test Scores by Gender:
Gender
Female     71.137963
Male       70.501040
Other      68.565275
```

Figure 7. *Average Test Score by Gender*

This outcome shows that female students got 1.14% higher score than male students.
Gender with the highest average test score: Female with a score of 71.13796277145812
Figure 8 shows that there is no correlation between attendance and test score.The correlation value 0.03 is very close to 0, so there is no linear relationship.
The figure 9 presents the correlation between study hours and current study. From the figure 9 its observed that there is no strong relationship between study hours and GPA.The figure 10 shows that there is strong relationship between GPA_Current, MOOC_Participation, online_Courses_Hours, and Spoken_Tutorial _Participation. EDA analysis of dataset.
The figure 11 depicts that current GPA improved 0.40 than GPA_Last_Semester.
Findings from EDA analysis
1. Online and MOOC, spoken tutorial supports the improvement of current GP.

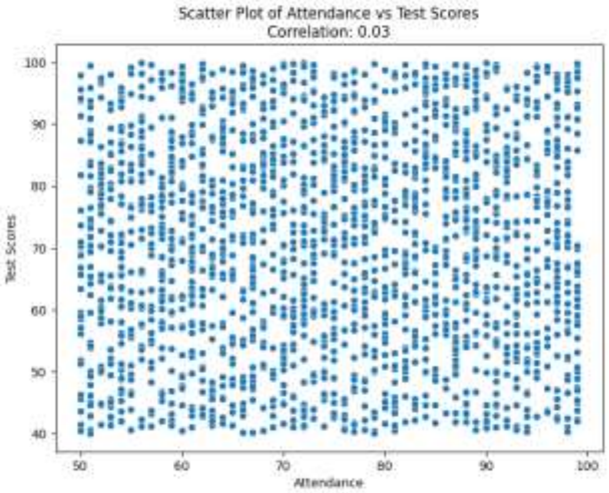


Figure 8. *Correlation between attendance and test score*

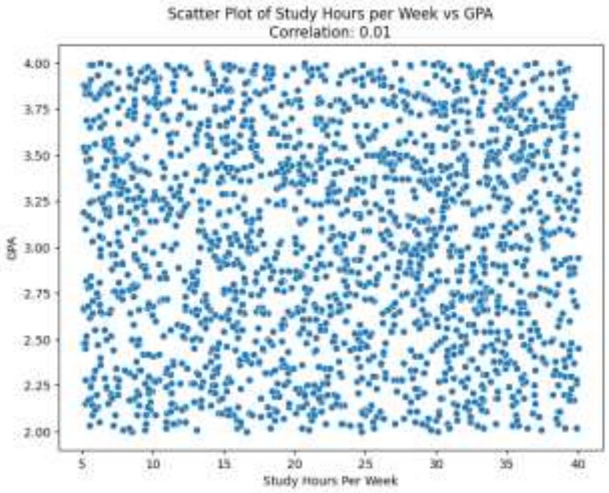
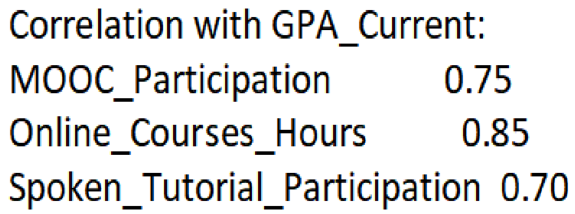
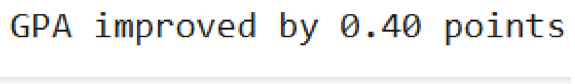


Figure 9. *Correlation between study hours and current study*



```
Correlation with GPA_Current:
MOOC_Participation          0.75
Online_Courses_Hours        0.85
Spoken_Tutorial_Participation  0.70
```

Figure 10. *Dependency between features*



```
GPA improved by 0.40 points
```

Figure 11. *Comparison between GPA_Last_Semester and Current GPA*

2. There is no strong relationship between study hours and GPA current.
3. There is no linear relationship between study hours and GPA current.
4. There is 0.40 improve between last and current GPA which show sthe importance of Online and MOOC, Spoken tutorial.
EDA analysis helps to identify the dominant features and patterns of the data.

Following evaluation metrics were used to measure the experimented model outcome.

4.1 Accuracy

This metrics is used to evaluate the overall accuracy and efficiency of the proposed model. It is calculated by the ratio of positively predicted value to the total number of prediction values. It is evaluated using the expression

$$Accuracy = \frac{number\ of\ correct\ prediction}{Total\ number\ of\ prediction} \times 100 \qquad (1)$$

4.2 Precision

This metric is used to evaluate the actually predicted positive value from the total number of positively predicted values. it can be calculated.

$$Precision = \frac{TP}{TP+FP} \qquad (2)$$

4.3 Recall

This metric evaluates the total number of incorrectly predicted positive values. It is clearly defined in the expression (3). The ratio of predicted true positive value to the sum of true positive and false negative value observes it.

$$Recall = \frac{TP}{TP+FN} \qquad (3)$$

4.4 F1 Score

To evaluate the prediction accuracy of the model, this metrics is used. It produces the final result by computing the precision and recall value. The following formula is performed to find the F1-score value.

$$F1Score = 2 \times \frac{Preciison \times Recall}{Preciison + Recall} \qquad (4)$$

Where TP-True Positive, TN-True Negative, FP-False Positive, FN-False Negative

Table 2 and figure 12 shows the model performance of DNN-RF and DNN-Light GBM in percentage.

Table 2. *DNN-RF and DNN-Light GBM model Evaluation*

Models	Accuracy	Precision	Recall	F1-Score
DNN-RF	97.21	96.82	96.73	96.95
DNN-Light GBM	96.53	95.13	95.34	95.93

DNN-RF achieved high accuracy, precision, recall and f1-score than model DNN-Light GBM. Hence it is more reliable method for students' performance prediction. Deep learning is applied in different fields and reported [8-16].

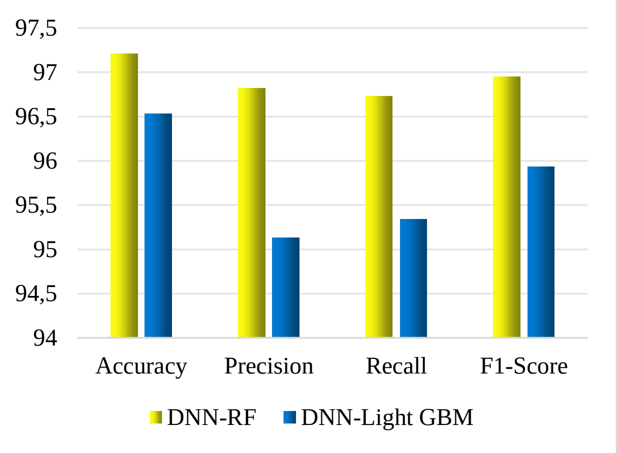


Figure 12. *DNN-RF and DNN-Light GBM model evaluation*

5. Conclusion

The proposed student performance analysis model, which used a decision DNN-RF, DNN-Light GBM algorithm to predict the student performance, has been discussed in detail. The preprocessing of the data helped in faster processing and efficient prediction results. An realtime dataset is considered for evaluating the performance of the model. The model's performance is compared to hybrid deep learning algorithms. Most ML models provide faster prediction but are less accurate, and the correlation between the features can only partially be achieved. In this proposed model, EDA analysis performed to identify correlation between features and its importance in students' performance prediction. EDA analysis finds that Online and MOOC, spoken tutorial strongly supports the improvement of Current GPA and there is improvement of 0.40 over last semester GPA. The DNN-RF, DNN-Light GBM correlate the features, and a collective prediction of the student's academic performance is made. The research shows that the proposed model DNN-RF achieves an accuracy of 97.21% which provides an extensive and accurate analysis of student performance and prediction process.

Author Statements:

- **Ethical approval:** The conducted research is not related to either human or animal use.
- **Conflict of interest:** The authors declare that they have no known competing financial interests or personal relationships that could

have appeared to influence the work reported in this paper

- **Acknowledgement:** The authors declare that they have nobody or no-company to acknowledge.
- **Author contributions:** The authors declare that they have equal right on this paper.
- **Funding information:** The authors declare that there is no funding to be acknowledged.
- **Data availability statement:** The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

References

[1] Sarker, I.H. (2021). Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. *SN COMPUT. SCI.* 2(420). https://doi.org/10.1007/s42979-021-00815-1

[2] Alshamaila, Y., Alsawalqah, H., Aljarah, I. et al. (2024). An automatic prediction of students' performance to support the university education system: a deep learning approach. *Multimed Tools Appl.* 83;46369–46396. https://doi.org/10.1007/s11042-024-18262-4

[3] Alshamaila et al.,(2024) investigated the data from Jordan UG students and anlaysed the features demographic information, students majors, faculty, high school average and four semester marks. Author suggest gmean for students performance analysis.

[4] Alnasyan, B., Basheri, M., & Alassafi, M. (2024). Deep Learning Techniques for Predicting Student's Academic Performance on Virtual Learning Environments: A Review. https://doi.org/10.21203/rs.3.rs-3888441/v1

[5] Kannan, K.R., Abarna, K.T.M., & Vairachilai, S. (2023). Graph Neural Networks for Predicting Student Performance: A Deep Learning Approach for Academic Success Forecasting. *International Journal of Intelligent Systems and Applications in Engineering.* 12(1s);228–232. https://ijisae.org/index.php/IJISAE/article/download/3410/1997/8338

[6] Liu, H., Zhu, Y., Zang, T., Xu, Y., Yu, J., & Tang, F. (2021). Jointly Modeling Heterogeneous Student Behaviors and Interactions among Multiple Prediction Tasks. *ACM Transactions on Knowledge Discovery from Data.* 16;1-24. https://doi.org/10.1145/3458023

[7] Khan, B., Afzal, S., Rahman, T., Khan, I., Ullah, I., Rehman, A., Baz, M., Hamam, H., & Cheikhrouhou, O. (2021). Student-Performulator: Student Academic Performance Using Hybrid Deep Neural Network. *Sustainability.* 13(9775). https://doi.org/10.3390/su13179775

[8] Hafez, I. Y., & El-Mageed, A. A. A. (2025). Enhancing Digital Finance Security: AI-Based Approaches for Credit Card and Cryptocurrency Fraud Detection. *International Journal of Applied Sciences and Radiation Research*, 2(1). https://doi.org/10.22399/ijasrar.21

[9] Ibeh, C. V., & Adegbola, A. (2025). AI and Machine Learning for Sustainable Energy: Predictive Modelling, Optimization and Socioeconomic Impact In The USA. *International Journal of Applied Sciences and Radiation Research*, 2(1). https://doi.org/10.22399/ijasrar.19

[10] Goverdhan Reddy Jidiga, P. Karunakar Reddy, Arick M. Lakhani, Vasavi Bande, Mallareddy Adudhodla, & Lendale Venkateswarlu. (2025). Blockchain and Deep Learning for Secure IoT: A Hybrid Cryptographic Approach. *International Journal of Computational and Experimental Science and Engineering,* 11(1). https://doi.org/10.22399/ijcesen.1132

[11] Johnsymol Joy, & Mercy Paul Selvan. (2025). An efficient hybrid Deep Learning-Machine Learning method for diagnosing neurodegenerative disorders. *International Journal of Computational and Experimental Science and Engineering,* 11(1). https://doi.org/10.22399/ijcesen.701

[12] Sivananda Hanumanthu, & Gaddikoppula Anil Kumar. (2025). Deep Learning Models with Transfer Learning and Ensemble for Enhancing Cybersecurity in IoT Use Cases. *International Journal of Computational and Experimental Science and Engineering,* 11(1). https://doi.org/10.22399/ijcesen.1037

[13] S. Leelavathy, S. Balakrishnan, M. Manikandan, J. Palanimeera, K. Mohana Prabha, & R. Vidhya. (2024). Deep Learning Algorithm Design for Discovery and Dysfunction of Landmines. *International Journal of Computational and Experimental Science and Engineering,* 10(4). https://doi.org/10.22399/ijcesen.686

[14] N.B. Mahesh Kumar, T. Chithrakumar, T. Thangarasan, J. Dhanasekar, & P. Logamurthy. (2025). AI-Powered Early Detection and Prevention System for Student Dropout Risk. *International Journal of Computational and Experimental Science and Engineering,* 11(1). https://doi.org/10.22399/ijcesen.839

[15] Rajitha Kotoju, B.N.V. Uma Shankar, Ravinder Reddy Baireddy, M. Aruna, Mohammed Abdullah Mohammed Alnaser, & Imad Hammood Sharqi. (2025). A Deep auto encoder based Framework for efficient weather forecasting. *International Journal of Computational and Experimental Science and Engineering,* 11(1). https://doi.org/10.22399/ijcesen.429

[16] Olola, T. M., & Olatunde, T. I. (2025). Artificial Intelligence in Financial and Supply Chain Optimization: Predictive Analytics for Business Growth and Market Stability in The USA. *International Journal of Applied Sciences and Radiation Research,* 2(1). https://doi.org/10.22399/ijasrar.18