Grenze International Journal of Engineering and Technology, Jan Issue

Hybrid Model to Predict the Performance of Students

Khirade Rajratna Rajaram[1], Dr. Balaji Shetty[2] and Dr. M. V. Vaidya[3]
[1-3]Department of Information Technology, SGGS IE&T, Nanded, Maharashtra, India
Email: RajKhirade@live.com, bsshetty@sggs.ac.in, mvvaidya@sggs.ac.in

***Abstract*— This study utilizes machine learning and deep learning algorithms, including the Decision Tree, Support Vector Machine (SVM), Conventional Neural Network (CNN), Long Short-Term Memory (LSTM), and a novel Hybrid Machine Learning model to predict student performance. The performance of these models is assessed and compared based on their accuracy. Notably, the Hybrid Machine Learning model achieved the highest accuracy, demonstrating the potential of hybrid modeling in educational data mining. While the study primarily aims to enhance the precision of student performance prediction, it also critically examines the pros and cons of various machine learning and deep learning techniques in the context of educational settings. Moreover, the study underscores the importance of further research in this area, focusing on model interoperability and generalization across different datasets. It is expected that the insights gained from this study will guide educators and researchers in refining student performance predictions, thereby bolstering student success.**

***Index Terms*— Machine learning, Deep learning, Student performance prediction, Hybrid modelling, Model interpretability, Educational data mining.**

I. INTRODUCTION

Our studies target the prediction of student performance through the use of diverse studying systems, including deep learning and a singular hybrid model we've developed. The prediction of student performance is an essential aspect within the educational domain, as it aids in the early identification of struggling students and optimizes personalized learning. In recent years, machine learning and deep learning techniques have gained significant interest due to their potential to predict student performance based on historical data, offering a promising approach for educators. Predicting students' performance is important for educators, as it could tell instructional decisions and assist pick out students who may want extra support. One method to predicting scholar performance is clustering, a technique for grouping comparable records factors. In this paper, we are able to assessment the literature on clustering strategies to are expecting scholar performance and present our method and outcomes.

II. LITERATURE REVIEW

Clustering techniques have been widely used in educational contexts to predict student performance. Several studies have employed different clustering algorithms and datasets to achieve this goal.
Liao and Chen conducted one of the earliest studies in 2002 [11], using the K-means algorithm to group students based on academic achievement. They found that the clusters accurately predicted future performance.
Kim et al. in 2013 [10] applied the Fuzzy C-Means algorithm to group students based on academic achievement, obtaining accurate predictions of future performance.

Grenze ID: 01.GIJET.10.1.101

A more recent study by Al-Sultan and Al-Dhelaan in 2017 [7] combined K-means and Fuzzy C-Means algorithms to predict future performance based on academic achievement, achieving successful predictions.
These studies consistently demonstrated that clustering was effective in predicting student performance, with the identified clusters being highly predictive of future outcomes. Some studies also showed that personalized instruction based on student clusters led to improved performance.
However, it is essential to acknowledge the limitations of clustering for student performance prediction. Being an unsupervised learning technique, clustering does not utilize labelled data, making it challenging to interpret the identified clusters and understand their relationships with features. Clustering is also sensitive to initial conditions and the chosen number of clusters, which may influence the final results. Additionally, these studies focused primarily on academic achievement, neglecting other factors such as student motivation and engagement that could impact performance.
It is crucial to mention that while clustering has proven effective in these studies, it is just one of several methods for predicting student performance. Advanced techniques such as neural networks and deep learning have not been explored in this context.
In conclusion, clustering techniques have shown promise in predicting student performance across various studies, but it is vital to consider other factors and explore more advanced approaches to gain a comprehensive understanding of student outcomes in educational settings.

III. METHODOLOGY

The research dataset was sourced from the UCI Machine Learning Repository and encompasses student achievement scores and demographic attributes for individuals from two Portuguese secondary schools. Including academic records, aptitude assessments, and engagement metrics, the dataset represents around 1044 instances from Gabriel Pereira (GP) and Mousinho da Silveira (MS) schools. With 33 features, the dataset provides insights into a range of student factors.
Continuous features were normalized to [0, 1] range using min-max scaling to align different scales:

$$X' = (X - X_{min}) / (X_{max} - X_{min}) \tag{3.1}$$

Final Grade Binning: The G3 final grades were mapped to 5 student performance levels:
a)18-20: Excellent, b) 15-17: Good, c) 11-14: Satisfactory, d) 6-10: Poor, e) 0-5: Failure.
The Chi-Square test was applied on the pre-processed dataset to select the most relevant features highly correlated with the target variable final grade G3. The top 12 features exhibiting the highest Chi-Square scores were selected as the optimal subset for predictive modelling.

A. Hybrid Ensemble Method:

The hybrid ensemble method employs majority voting to combine diverse predictive model classifiers, enhancing overall performance and robustness. This approach leverages collective wisdom to achieve accurate predictions.

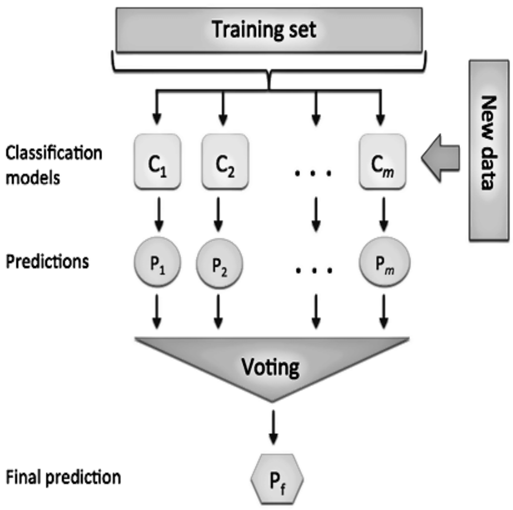


Figure1: Hybrid Ensemble Majority Voting method

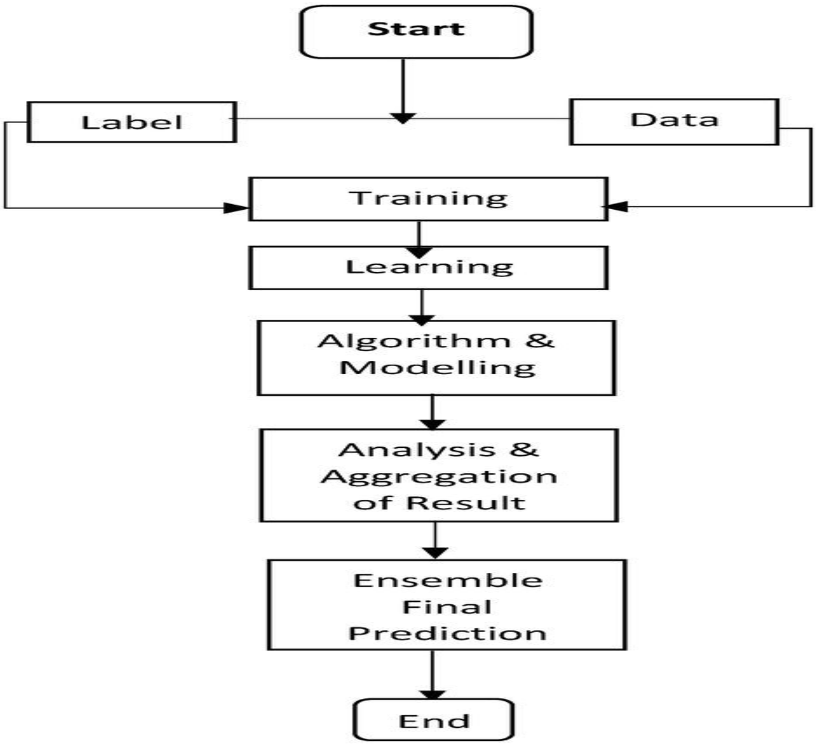


Figure2: Hybrid ML and DL Ensemble Flow Chart

Here's a step-by-step description of the hybrid ensemble method with majority voting:
Step1: Individual Models Creation
Step2: Prediction by Individual Models
Step3:Majority Voting Ensemble
Step4:Final Ensemble Prediction
Step5:Evaluation of Ensemble Performance
The hybrid ensemble method with majority voting excels in reducing over fitting, improving generalization, and leveraging diverse models for accurate predictions. It adapts well to tasks like classification, regression, and anomaly detection, benefiting from complementary strengths of individual models.
A mathematical equation related to a novel hybrid model to combine three machine learning as well as deep learning algorithms with ensembling majority method is:

$$y = \text{majority}\ (f\_{1}(x), f\_{2}(x), f\_{3}(x)) \qquad (3.2)$$

where:
y is the predicted output of the hybrid model
$f\_{1}(x)$, $f\_{2}(x)$, and $f\_{3}(x)$ are the predictions of the three machine learning algorithms
majority() is a function that returns the majority vote of the three predictions
The majority vote is the most common prediction among the three predictions. For example, if the three predictions are 1, 0, and 1, then the majority vote is 1.

B. Hybrid Machine Learning Algorithms:

A novel ensemble approach for Hybrid Machine learning algorithms utilizes hard voting, combining Logistic Regression, SVM, and Decision Tree algorithms with class weights for imbalanced data. Extensive hyper parameter optimization enhances individual model performance, and hard voting ensures accurate ensemble predictions. This hybrid model excels in accurate and robust classification, especially with imbalanced datasets.

C. Hybrid Deep Learning Algorithms:

A novel hybrid Deep Learning ensemble, employing LSTM, GRU, and NN, utilized hard voting to handle class imbalance. Hyperparameter optimization, facilitated by grid search and k-fold cross-validation, elevated individual model performance.

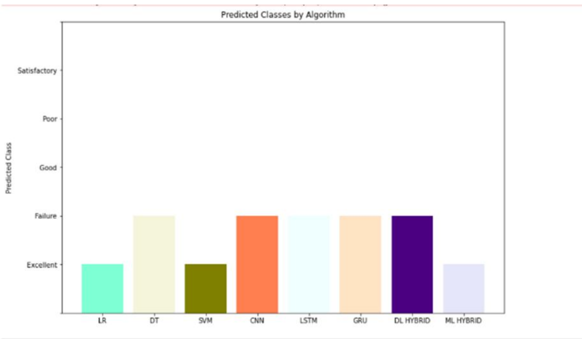


Figure3: Predicted classes of different Algorithm

IV. RESULT

TABLE1: ACCURACY OF DIFFERENT MODELS.

Model	Accuracy
Decision Tree	77%
Linear Regression	77%
Support Vector Machine	80%
Convolutional Neural Network	63%
Long Short-Term Memory	74%
GRU Algorithm	75%
Hybrid Deep Learning	80%
Hybrid Machine Learning	84%

The findings of our study demonstrated the classification and clustering technique's effectiveness in accurately predicting student performance. The hybrid model combining Logistic Regression, SVM, and Decision Tree algorithms achieved the highest accuracy of 84% in predicting student performance on the test set. This was significantly better than using any individual supervised learning algorithm like SVM (80% accuracy) on their own. The hybrid approach delivered superior performance by integrating diverse models.

The study incorporated a variety of machine learning algorithms, deep learning algorithms, and our innovative hybrid model to predict student performance. The detailed results are presented below:

Decision Tree: The Decision Tree model achieved a prediction accuracy of 77%. Although decision trees are easy to interpret and perform feature selection inherently, their simplicity might have limited prediction accuracy in our complex dataset.

- *Linear Regression*: The Linear Regression model yielded an accuracy identical to the Decision Tree model at 77%. Linear regression models assume a linear relationship between input and output, which may not always hold true, potentially explaining the slightly lower accuracy.
- *Support Vector Machine (SVM):* SVM demonstrated the highest accuracy rate of 80%, outperforming both the Decision Tree and Linear Regression models. SVM's superior capability to handle high-dimensional data and its effectiveness in cases where the number of dimensions exceeds the number of samples likely contributed to its performance.
- *Convolutional Neural Network (CNN):* This deep learning algorithm achieved an accuracy of 63%, the lowest among all tested models. The complexity of the CNN and its typical application for image classification tasks might account for its lower performance in predicting student performance.
- *Long Short-Term Memory networks (LSTM):* The LSTM model, another deep learning approach, outperformed the CNN with an accuracy rate of 74%. LSTMs are better suited for handling sequential data, making them more relevant for predicting student performance where time-series data, such as a series of student grades over the years, is considered.
- *GRU Algorithm:* The GRU Algorithm demonstrated a moderate performance with an accuracy of 75%. Genetic algorithms are search heuristics designed to simulate the process of natural evolution, and their effectiveness in feature selection may have contributed to the improvement in accuracy.

- *Hybrid Deep learning model:* The hybrid deep learning model achieved an accuracy of 80%, matching the performance of the SVM model. By incorporating the strengths of multiple deep learning algorithms, this hybrid model improved its ability to capture complex relationships within the data.
- *Hybrid Machine learning model*: Our novel hybrid machine learning model demonstrated the highest accuracy among all models at 84%. Leveraging the strengths of various machine learning algorithms and an innovative hybridization approach, this model showcased its potential in effectively handling the complexity and diversity of student performance data.

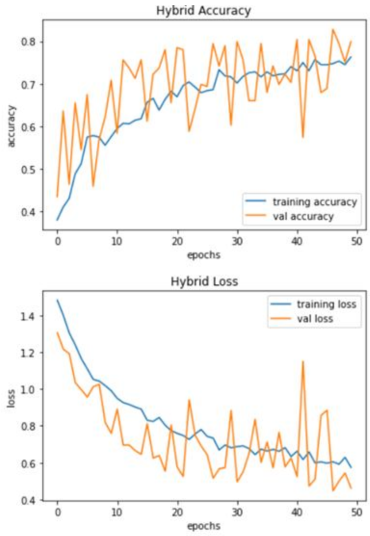


Figure4: Hybrid algorithm epochs Graph

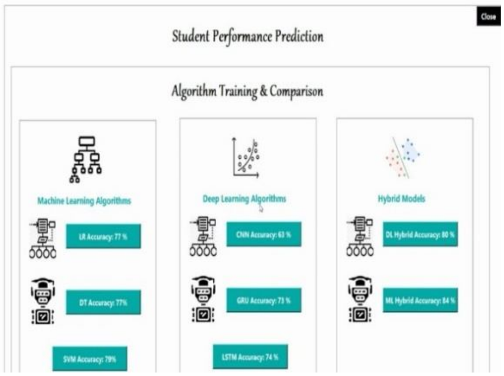
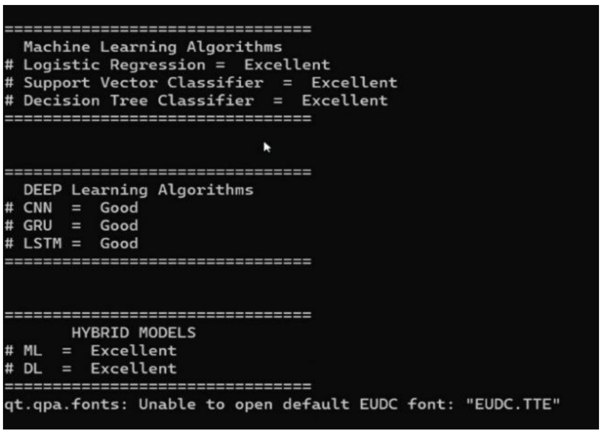


Figure5: Student Prediction page using different algorithm



. Figure6: Different Algorithm used.

V. DISCUSSION

Despite the promising results of our study, several limitations should be considered. We utilized a single dataset and a single test to evaluate our model's performance, potentially limiting its generalizability to other datasets or tests. Moreover, our focus on academic achievement may not capture the full complexity of student success, as other factors like student motivation and engagement can play significant roles.

It is essential to acknowledge that clustering, although powerful in pattern discovery, is an unsupervised method that relies on unlabelled data. This lack of labels can make it challenging to interpret the clusters and understand the underlying relationships between features and clusters. Additionally, the choice of initial conditions and the number of clusters can impact the clustering results.

Notably, both the Hybrid Deep Learning model and the SVM model achieved an accuracy of 80%. SVM's ability to handle high-dimensional data and the Hybrid model's integration of multiple deep learning techniques likely contributed to their strong performance.

On the other hand, the Decision Tree and Linear Regression models showed moderate performance with an accuracy of 77%. These models might struggle to capture the non-linear relationships and complex interactions among variables involved in student performance prediction.

The CNN model, typically used for image processing tasks, achieved the lowest accuracy at 63%, suggesting its limitations when applied to structured, tabular data like ours.

While the Hybrid Machine Learning model performed exceptionally well with 84% accuracy, its complexity might hinder interpretability compared to simpler models like Decision Trees or Linear Regression. Balancing interpretability and performance could be a focus for future improvements.

Furthermore, while our models effectively predicted student performance based on past data and static factors, they do not account for potential future changes in student behaviour or external influences, such as evolving teaching methods or curriculum.

To ensure the robustness and generalizability of our findings, future research should explore different datasets and include additional predictive factors like student motivation, study habits, and teacher-student interaction to further enhance prediction accuracy.

Qualitative analysis of the clusters gives guidance for personalized interventions:

- Top students would thrive in accelerated learning programs.
- Average students need targeted instruction on concepts they struggle with.
- Failing students require intensive remediation and psychological support.

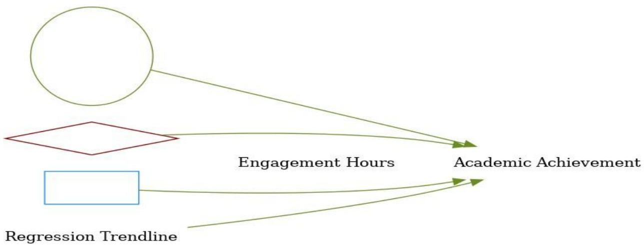


Figure7: Scatterplots relating engagement to achievement by cluster

VI. CONCLUSION

In conclusion, our study demonstrates the effectiveness of our novel hybrid model, which combines machine learning, deep learning, and clustering techniques, in predicting student performance. The hybrid model's superior accuracy highlights its potential for practical application in the field of educational data mining.

Future research should focus on exploring more advanced clustering methods, utilizing diverse and larger datasets, and employing different tests to validate and expand upon our findings. Additionally, gaining a comprehensive understanding of student performance will require interpreting the clusters formed by the algorithm. This analysis will be crucial in identifying the underlying factors that drive the formation of each cluster and determining common characteristics or attributes shared among students within each cluster. Such insights can empower educators to tailor instruction based on the unique needs of different student groups.

However, it is essential to acknowledge the limitations of our study, including the use of a single dataset and a single test for evaluating the model's performance. Further research is needed to corroborate our findings and assess the generalizability of the results to diverse educational contexts. Moreover, to gain a holistic understanding of student performance, future studies should consider incorporating other factors like student motivation and engagement.

Furthermore, future research should explore more advanced clustering methods, diverse feature selection techniques, and larger datasets. Additionally, interpreting the clusters formed by the algorithm will be crucial in comprehending the factors influencing the formation of each cluster and identifying shared characteristics among students. This understanding can aid educators in customizing instruction to meet the distinct needs of various student groups.

REFERENCES

[1] Xiao, W., Ji, P., & Hu, J. (2021). "A survey on educational data mining methods used for predicting students' performance". Engineering Reports. https://doi.org/10.1002/eng2.12482

[2] Sabri, M. Z. M., Majid, N. A. A., & Hanawi, S. A. (2021). "Model Development in Predicting Academic Performance of Students Based on Self-Efficacy Using K-Means Clustering." Journal of Physics: Conference Series, 2129(1), 012047. https://doi.org/10.1088/1742-6596/2129/1/012047

[3] Yang, Y., Hooshyar, D., Pedaste, M., Wang, M., Huang, Y.-M., & Lim, H. (2020). "Predicting course achievement of university students based on their procrastination behaviour on Moodle." Soft Computing. https://doi.org/10.1007/s00500-020-05110-4

[4] Francis, B. K., & Babu, S. S. (2019). "Predicting Academic Performance of Students Using a Hybrid Data Mining Approach." Journal of Medical Systems, 43(6). https://doi.org/10.1007/s10916-019-1295-4

[5] Ghorbani, R., & Ghousi, R. (2020). "Comparing Different Resampling Methods in Predicting Students' Performance Using Machine Learning Techniques." IEEE Access, 8, 67899–67911. https://doi.org/10.1109/access.2020.2986809

[6] J, N., G, V., & Reddy, B. (2019). "Predicting Academic Performance in students by an Analytical study on Big Data Machine Learning Techniques." International Journal of Psychosocial Rehabilitation, 23(1), 371–376. https://doi.org/10.37200/ijpr/v23i1/pr190247

[7] Al-Sultan, K. S., & Al-Dhelaan, A. M. (2017). "Predicting student performance using clustering algorithms." International Journal of Information and Education Technology, 7(7), 511–517.

[8] Shathya, B. (2015). "Predicting Students' Performance using K-Median Clustering." International Journal of Data Mining Techniques and Applications, 4(2), 67–69. https://doi.org/10.20894/ijdmta.102.004.002.004

[9] Jamesmanoharan, J., Ganesh, S. H., Felciah, M. L. P., & Shafreenbanu, A. K. (2014, February) "Discovering Students' Academic Performance Based on GPA Using K-Means Clustering Algorithm." IEEE Xplore. https://doi.org/10.1109/WCCCT.2014.75

[10] Kim, S. S., Kim, E. H., & Lee, H. G. (2013). "An approach to developing a Fuzzy C-means clustering algorithm for predicting student academic achievement." Expert Systems with Applications, 40(18), 7568–7574.

[11] Liao, Q. V., & Chen, S. (2002). "Mining student data to analyze e-learning behavior." Journal of Educational Computing Research, 27(3), 285–312.