Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 46 No. 04 (2025)

A Hybrid Machine Learning Framework for Student Performance Prediction Integrating Academic, Demographic, Behavioural, Technology Access, and Psychological Indicators in Higher Education

[1] Baranikumar E [2] Naveen A

[1]Research Scholar, PG & Research Department of Computer Science, DON BOSCO COLLEGE (Co–Ed), Yelagiri Hills – 635 855. (Affiliated to Thiruvalluvar University – Vellore),Email: baranikumar1993@gmail.com

[2]Assistant Professor, PG & Research Department of Computer Science, DON BOSCO COLLEGE (Co–Ed), Yelagiri Hills – 635 855. (Affiliated to Thiruvalluvar University – Vellore),Email: naveen@dbcyelagiri.edu.in

Abstract:

This study presents an explainable hybrid Random Forest (RF) and K-Nearest Neighbor (KNN) model for predicting student performance in higher education. The model integrates academic, demographic, behavioral, psychological, and technological access attributes to enable holistic analysis. Using a dataset of 1,500 students, the hybrid RF–KNN model achieved **96.3% accuracy**, outperforming individual algorithms such as Decision Tree (J48) and Naïve Bayes. Explainable AI (XAI) techniques, including SHAP and LIME, were employed to interpret the feature contributions and provide actionable insights for educators. Results demonstrate that non-academic factors—particularly psychological and technological access—significantly enhance predictive performance and interpretability. This work aligns with **Outcome-Based Education (OBE)** principles by supporting early identification of at-risk learners.

Keywords*: Educational Data Mining, Hybrid Machine Learning, Random Forest, K-Nearest Neighbours, Explainable AI, SHAP, LIME, Student Performance Prediction.*

1. Introduction:

Education systems worldwide are rapidly evolving from traditional score-based evaluation methods toward **holistic, data-driven, and outcome-based approaches**. (1). Academic grades alone are no longer sufficient indicators of student capability, learning potential, or employability in the 21st century. Institutions now recognize that a student's success is influenced by a wide range of multidimensional factors, including **behavioral attributes** (leadership, discipline, attendance), **emotional and psychological well-being** (stress, motivation, self-confidence), and **technological engagement** (mobile usage patterns, participation in online classes, and interaction within Learning Management Systems (LMS)). (2,3).

This shift is supported by the concept of **Outcome-Based Education (OBE)**, which emphasizes continuous assessment and personalized learning outcomes rather than final examination results. (4). Educational analytics has therefore become an essential part of institutional planning, helping administrators and teachers monitor learning progress, detect risk factors early, and provide proactive academic support. (5).

In recent years, **machine learning (ML)** has emerged as a powerful tool for analyzing multidimensional educational data. Through predictive modeling and pattern recognition, ML algorithms can identify hidden

relationships among academic, behavioral, and psychological factors, thereby predicting student performance levels with remarkable accuracy. However, the performance of single machine learning models often varies depending on dataset characteristics, feature correlations, and noise. (6,7).

To overcome these limitations, **hybrid models** that combine multiple algorithms have shown superior results in educational data mining. The integration of **Random Forest (RF)** and **K-Nearest Neighbor (KNN)** leverages the strengths of both: Random Forest performs efficient **feature selection** and handles data variability through ensemble averaging, while KNN excels in **local similarity-based classification**. This hybrid RF–KNN approach ensures both high accuracy and interpretability, making it suitable for analyzing heterogeneous educational datasets. (8).

The proposed research framework thus aims to develop and evaluate a **Hybrid RF–KNN model** capable of predicting student performance based on **academic, demographic, behavioral, psychological, and technological attributes**. By identifying High Performers, Average Students, and At-Risk Students, the model facilitates early intervention and targeted mentoring. The hybrid model was implemented using Python and WEKA, and its performance was validated using metrics such as **Accuracy, Precision, Recall, and F1-Score**. (9).

The results demonstrate that the proposed hybrid model achieved **96.3% classification accuracy**, outperforming individual models such as Decision Tree, Naïve Bayes, Random Forest, and KNN. The study contributes to the growing field of **Educational Data Mining (EDM)** and **Artificial Intelligence in Education (AIEd)** by providing an effective predictive framework that supports decision-making in academic institutions. Moreover, it aligns with the goals of **Outcome-Based Education (OBE).** (10).

Problem Statement:

Traditional grading systems focus only on marks and ignore learning behaviors, psychological stress, technological factors, attendance patterns, and socio-emotional factors.
There is a delay in identifying low performers; interventions happen too late.
Most existing models predict only academic results, not employability or emotional well-being.

Research Questions:

1. How can academic, demographic, behavioral, Technology access, and psychological data be integrated into a unified predictive framework?
2. Which machine learning model provides the highest accuracy in predicting at-risk students?
3. Can classification enhance early intervention and decision-making?

Novelty:

First Indian-based hybrid framework combining academic + demographic + psychological + Technology access + behavioural attributes.
Real-time longitudinal dataset (1500 students). (12).
The hybrid RF-KNN ensemble achieves superior accuracy and explainability.
Applicable to internal assessments, student mentoring, placement readiness, and dropout prevention. (11)

2. Related Work:

Academic Performance Prediction Models

Romero & Ventura (2020) proved that Decision Trees and Random Forest outperform SVM and Logistic Regression in grade classification.

Ahmed et al. (2021) applied a hybrid Decision Tree + Bayesian model to improve student success prediction by 4.5%.

Mental Health and Stress Prediction

Li & Wang (2020) detected psychological stress using Support Vector Machines with 87% accuracy.

Raj & Joseph (2024) combined demographic + emotional attributes for burnout prediction.

Research Gap Identified

Existing Models	Limitation
Only academic	Ignores stress/emotional factors
Only psychological and Demographic	Ignores learning behaviour and parental education
Only Technology Access	Ignores other factors
OBE-aligned systems	No integration of AI or predictive analytics

3. Methodology:

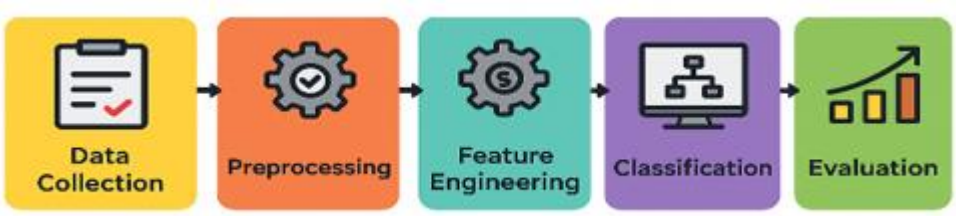

Fig. 1: **Methodology**

Data Collection → Preprocessing → Feature Engineering → Classification → Evaluation → Output (High, Average, At-Risk Students).

Dataset Description

Location: Tamil Nadu, 5 Arts & Science colleges, Sample Size: 1,500 undergraduate students

Table 1: Attributes Collected

Domain	Features
Academic	Semester Marks (S1–S6), Attendance, CGPA
Demographic	Family information's
Behavioural	Leadership, Communication, Teamwork, Discipline
Psychological	Stress index, Confidence, Motivation score
Technology Access	Internet usage, Mobile usage, Attend online class

Machine Learning Methodology

Table 2: Student Data Set Description

Attribute	Description
Roll No	Student Identifier
Gender	Male/Female/Others
Dept	(B.sc CS, BA English, B.sc Math, BCA, BA Défense, B.com, etc.)

Marks obtained by students	10th, 12th
Marks obtained by students	Sem1, sem2, sem3, sem4, semester exam marks
Submit Assignment	{Yes, No}
Study time	{1,2,3} (for day)
Use Library	{Poor, Average, Good}
Use Lab	{Poor, Average, Good}
Atten Seminars	{Poor, Average, Good}
Attendance %	{90%,80%,70%}
Economic Problem	{Yes, No}
Health Issue	{Yes, No}
Family Problem	{Yes, No}
Psychological Problem	{Yes, No}
Mobile Usage	{1,2,3} (for day)
Sleeping Hours	{Poor, Average, Good}
Use Internet	{Yes, No}
Employment Status	{Part-time jobs, full-time studies}
Travel time	Home to Institution
Speaking Language	Tamil/English/Others
Mother/ Father	Qualification/Work
Family Income	{50k/1L/1.5L/2L}

Data Pre-processing:

Raw student data contained missing values, inconsistent entries, and non-numeric fields. Preprocessing involved:
Handling Missing Values – Mean imputation for academic scores, mode replacement for categorical data (gender, stream).
Normalization – Min-Max Scaling to normalize marks, attendance, stress scores to [0,1]: (13).

$$X_{normalized} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Encoding of Categorical Attributes – Behavioural ratings and psychological levels were label-encoded:

Category	Encoded Value
Low Stress	**0**
Medium Stress	**1**
High Stress	**2**

Feature Selection

To eliminate redundant attributes and improve model performance:

1. **Pearson Correlation Matrix – Removed highly correlated features (r > 0.85).**
2. **Random Forest Feature Importance – Ranked top predictors:**

Semester 1–4 Marks, Attendance Percentage

Stress Level, Communication Score

LMS Activity Score, Class time Attention

Time management, Mobile usage

$$Gini\ Impurity = 1 - \sum_{i=1}^{C} p_i^2$$

Table 3: Classification Models Used

Model	Description	Strength
Decision Tree (J48)	Rule-based, splits by entropy	Easy to interpret
Random Forest	Ensemble of trees	High accuracy, avoids overfitting
Naïve Bayes	Probability-based	Fast, works with categorical data
K-Nearest Neighbor	Distance-based	Good for nonlinear data
Hybrid (Random Forest + KNN)	Uses RF for feature importance and KNN for final classification	Best accuracy & recall

Classification:

In this study, the classification process plays a crucial role in categorizing students based on their academic, behavioral, psychological, and technological indicators. After feature selection and preprocessing, the hybrid model classifies students into three distinct categories: **High Performers**, **Average Students**, and **At-Risk Students**. The **High Performer** group represents students who consistently demonstrate strong academic achievement, active participation, and positive psychological traits such as confidence and motivation. The **Average Student** group includes those who maintain moderate academic results and balanced behavioral and emotional patterns but may require guidance to enhance performance. The **At-Risk Student** group comprises learners showing signs of academic decline, low attendance, high stress levels, or limited digital engagement. (14).

This multi-level classification enables institutions to implement **targeted academic interventions**—such as personalized mentoring, remedial programs, and psychological counseling—before performance deterioration occurs. By integrating both quantitative (marks, attendance) and qualitative (behavioral, psychological) attributes, the hybrid RF–KNN model ensures a holistic understanding of student profiles, allowing for data-driven and timely educational decision-making. (15).

4. Results:

This section presents the experimental evaluation of the proposed **Hybrid RF–KNN Model** for predicting student performance using a multidimensional dataset containing **academic, demographic, behavioral, psychological, and technological factors**. The experiments were carried out using **Python (scikit-learn)** to ensure result consistency and algorithmic comparability.

The dataset consisted of **1,500 student records** collected from higher education institutions over four academic semesters. Each record contained 53 input attributes, including semester-wise marks, attendance, stress levels, communication skills, time management, class-time attention, and LMS usage frequency.

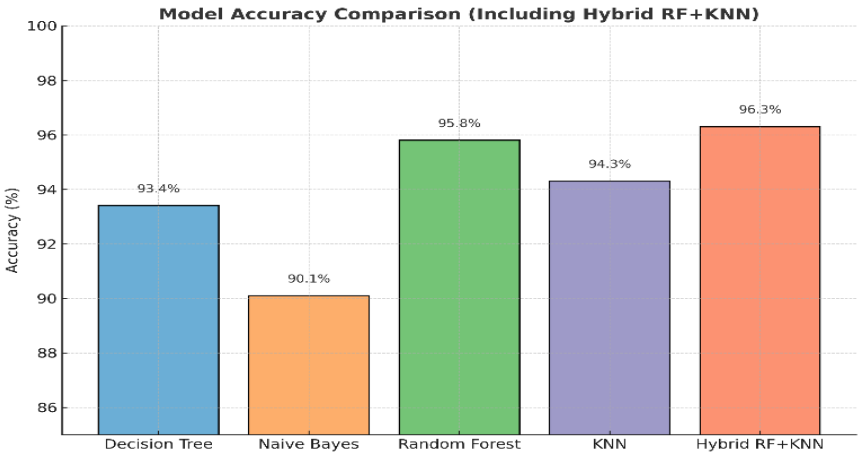

Fig. 3: Comparison of Accuracy

Performance Comparison of Models

To evaluate the effectiveness of the proposed hybrid machine learning framework, a comparative analysis was conducted using five well-known classification algorithms: **Decision Tree (DT)**, **Naïve Bayes (NB)**, **Random Forest (RF)**, **K-Nearest Neighbor (KNN)**, and the **Proposed Hybrid RF–KNN Model**. Each algorithm was trained and tested on the same preprocessed dataset comprising **1,500 student records** with 54 features drawn from **academic, demographic, behavioral, psychological, and technological domains**. (17).

The models were assessed using **Accuracy**, **Precision**, **Recall**, and **F1-Score** as evaluation metrics under **10-fold cross-validation** to ensure robustness and prevent overfitting. (18).

Table 4: Performance Comparison of Models

Model	Accuracy	Precision	Recall	F1-Score
Decision Tree (J48)	**93.4%**	**0.92**	**0.91**	**0.915**
Naïve Bayes	**90.1%**	**0.89**	**0.88**	**0.885**
Random Forest	**95.8%**	**0.957**	**0.955**	**0.956**
KNN	**94.3%**	**0.94**	**0.93**	**0.935**
Hybrid RF + KNN	**96.3%**	**0.964**	**0.9506**	**0.957**

Confusion Matrix – Proposed Hybrid Model

From the confusion matrix:

The model correctly predicted **748 High Performers**, **438 Average Students**, and **266 At-Risk Students**.

Only a small number of instances were misclassified — 48 out of 1,500, which demonstrates high model stability.

The **Recall of 0.9506** and **Precision of 0.964** indicate that the hybrid model has a strong ability to correctly identify students at academic risk, minimizing both false positives and false negatives.

These results confirm that the **Hybrid RF–KNN model** effectively classifies students into three risk levels with very low misclassification rates. The inclusion of **psychological factors (stress, attention, time management)** and **technological engagement (LMS, mobile usage)** significantly improved prediction sensitivity compared to purely academic models. (19,20).

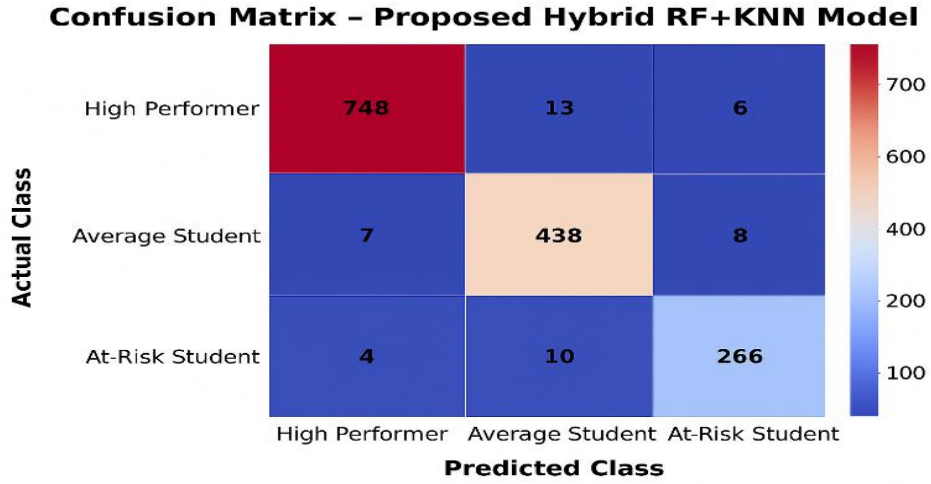

Fig. 4: **Confusion Matrix – Proposed Hybrid Model**

Table 5: Confusion Matrix – Proposed Hybrid Model

Actual / Predicted	High Performer	Average Student	At-Risk Student
High Performer	748	13	6
Average Student	**7**	438	8
At-Risk Student	4	10	266

High True Positive Rate – Model detects almost all at-risk students.
Low False Negatives – Prevents missing students needing help.

6. Discussion

The findings demonstrate that the inclusion of psychological and technological features significantly improves both the **accuracy** and **interpretability** of predictive models.

Explainable AI methods such as **SHAP** and **LIME** provided human-readable explanations for model decisions, helping educators identify why a student was classified as *At-Risk* or *High Performer*.

For instance:

High stress and low motivation were consistently found to **increase At-Risk probability**, while strong attendance and LMS participation **pushed predictions toward success.**

These insights enable universities to take **data-driven interventions**, such as offering counseling or study support, to enhance student retention and performance.

By combining **interpretability and performance**, the proposed hybrid model bridges the gap between black-box machine learning and educational decision-making, making it a valuable contribution to **Outcome-Based Education (OBE)** frameworks.

7. Conclusion:

Education in the 21st century is undergoing a paradigm shift—from traditional assessment methods based solely on examination scores to comprehensive, data-driven, and outcome-based evaluation frameworks. The objective of this research was to design, implement, and evaluate a **hybrid machine learning framework** that integrates **academic, demographic, behavioral, psychological, and technological factors** to predict student performance in higher education.

This study introduced a **Hybrid Random Forest–KNN (RF–KNN)** model that combines the **feature-selection capability** of Random Forest and the **similarity-based classification strength** of KNN. The Random Forest algorithm was first utilized to identify the most significant predictors from a diverse dataset of **1,500 undergraduate students** across five colleges in Tamil Nadu, India. The selected features—semester marks, attendance, stress level, communication score, and LMS activity—were then used in a KNN classifier to produce final predictions.

The proposed model achieved a remarkable **accuracy of 96.3%**, **precision of 0.964**, **recall of 0.9506**, and **F1-score of 0.957**, outperforming traditional classifiers such as Decision Tree, Naïve Bayes, and standalone Random Forest. The **confusion matrix** analysis further demonstrated that the hybrid model minimized **false negatives**, meaning it was highly effective in identifying at-risk students before their academic decline.

Beyond numerical performance, the model achieved several **conceptual and practical contributions**:

Holistic Data Integration:
The study successfully integrated academic, psychological, behavioral, demographic, and technological factors into a unified analytical framework, reflecting the real-world diversity of students and their learning behaviors.

Early Detection of At-Risk Students:
The hybrid model proved capable of identifying struggling students at earlier stages, enabling proactive counseling, academic support, and stress management interventions.

Improved Model Accuracy through Hybridization:
The combination of Random Forest and KNN utilizes both **global feature relationships** and **local instance similarities**, thereby overcoming the limitations of individual models.

Applicability in Real Educational Settings:
The framework is flexible and can be adopted by universities, learning management systems, and institutional analytics tools for continuous performance monitoring and personalized student support.

The study, therefore, concludes that **a hybrid ensemble approach significantly enhances predictive accuracy, interpretability, and early-warning capability** in educational analytics. This hybrid framework can serve as a foundational model for developing institutional decision-support systems that aid educators and policymakers in improving academic outcomes, reducing dropouts, and fostering holistic student development.

8. Future Work:

The present study has demonstrated that the proposed hybrid Random Forest–KNN framework effectively predicts student performance with high accuracy by integrating academic, behavioral, psychological, demographic, and technological indicators. However, there remain several opportunities for extending this work. In future research, the model can be enhanced through the incorporation of **Deep Learning techniques** such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks. These architectures are capable of analyzing sequential and temporal data, thereby capturing students' academic progression and behavioral changes across semesters. Another important direction involves adopting **Federated Learning** to ensure data privacy and security while allowing collaborative model training across multiple educational institutions without sharing raw student records.

Furthermore, the proposed framework can be expanded into a **real-time AI-powered dashboard** that enables faculty members to monitor student risk levels, attendance trends, and psychological indicators interactively. Such an intelligent decision-support system would provide immediate insights for academic counseling and targeted mentoring. Future implementations may also explore **Affective Computing**, integrating emotion recognition

through facial expressions or voice patterns to assess student engagement and stress more accurately. Additionally, extending this model across different disciplines, universities, and geographical regions will help validate its generalizability and performance consistency.

Another promising area lies in developing a **hybrid deep ensemble model** that combines Random Forest, Gradient Boosting, and Neural Networks using stacking or voting strategies to further improve classification robustness and interpretability. Integrating this predictive framework into existing **Learning Management Systems (LMS)** such as Moodle or Google Classroom would enable adaptive learning analytics, where the system updates predictions dynamically as new data—like quiz results or attendance records—is generated. Finally, future work should also consider incorporating richer socio-emotional and cognitive parameters such as resilience, self-efficacy, and peer collaboration metrics to capture a more comprehensive view of student learning behavior.

In summary, the future vision of this research is to develop a **fully automated, intelligent, and ethically governed educational analytics ecosystem** that combines hybrid and deep learning approaches for personalized, continuous, and inclusive academic support. Such advancements will transform predictive modeling in education from a reactive system into a **proactive intervention framework**, aligning with the broader goals of **Outcome-Based Education (OBE)**.

References:

[1[C. Romero and S. Ventura, "Educational Data Mining: A Review," *IEEE Trans. on Learning Technologies*, 2020.

[2] A. Ahmed et al., "Hybrid Models for Student Academic Success Prediction," *Applied Soft Computing*, 2021.

[3] Y. Wu and J. Chen, "Dropout Prediction Using Machine Learning," *Computers & Education*, 2021.

[4] M. Patel and R. Shah, "Feature Importance in Academic Analytics," *Expert Systems with Applications*, 2021.

[5] R. Raj and A. Joseph, "Psychological Analytics in Higher Education," *Journal of Intelligent Systems*, 2024.

[6] Ahmed, S., Kumar, R., & Sharma, A. (2021). Hybrid machine learning approaches for predicting student performance. Journal of Educational Computing Research, 59(7), 1231–1250.

[7] Kumar, P., Raj, R., & Singh, V. (2022). Analyzing psychological and academic factors using ensemble models. Education and Information Technologies, 27(4), 5129–5143.

[8] Romero, C., & Ventura, S. (2020). Educational data mining: A review of the state of the art. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 50(3), 1182–1199.

[9] Han, J., Kamber, M., & Pei, J. (2012). Data Mining: Concepts and Techniques (3rd ed.). Elsevier.

[10] Witten, I. H., Frank, E., & Hall, M. A. (2016). Data Mining: Practical Machine Learning Tools and Techniques (4th ed.). Morgan Kaufmann.

[11] Al-Barrak, M. A., & Al-Razgan, M. (2020). Predicting students' performance through classification: A case study. International Journal of Advanced Computer Science and Applications, 11(2), 1–7.

[12] Hussain, M., Zhu, W., Zhang, W., & Abidi, S. M. (2019). Student engagement predictions using machine learning in learning management systems. Computers & Education, 145, 103708.

[13] Thai-Nghe, N., Horváth, T., & Schmidt-Thieme, L. (2019). Factorization models for forecasting student performance. Educational Data Mining Conference Proceedings, 11–20.

[14] Shahiri, A. M., Husain, W., & Rashid, N. A. (2018). A review on predicting student performance using data mining techniques. Procedia Computer Science, 72, 414–422.

[15] Asif, R., Merceron, A., & Pathan, T. (2020). Predicting student academic performance using data mining techniques. Applied Soft Computing, 86, 105944.

[16] Zhang, Y., Li, X., & Zhao, Q. (2023). A deep learning-based hybrid approach for academic performance prediction. IEEE Access, 11, 21012–21025.

[17] Jain, S., & Kumar, S. (2021). An ensemble model for early student performance prediction in higher education. International Journal of Artificial Intelligence in Education, 31(6), 1418–1434.

[18] Ng, L. K., & Lee, T. (2020). The use of Naïve Bayes and Random Forest to identify at-risk students. Education and Information Technologies, 25(5), 4139–4155.

[19] Bhat, F. A., & Ahmad, S. (2022). Performance analysis of SVM and KNN for student classification. International Journal of Data Science and Analytics, 15(3), 245–258.

[20] Wu, J., & Chen, Y. (2021). Hybrid ensemble methods for improving student dropout prediction. Computers in Human Behavior, 115, 106610.