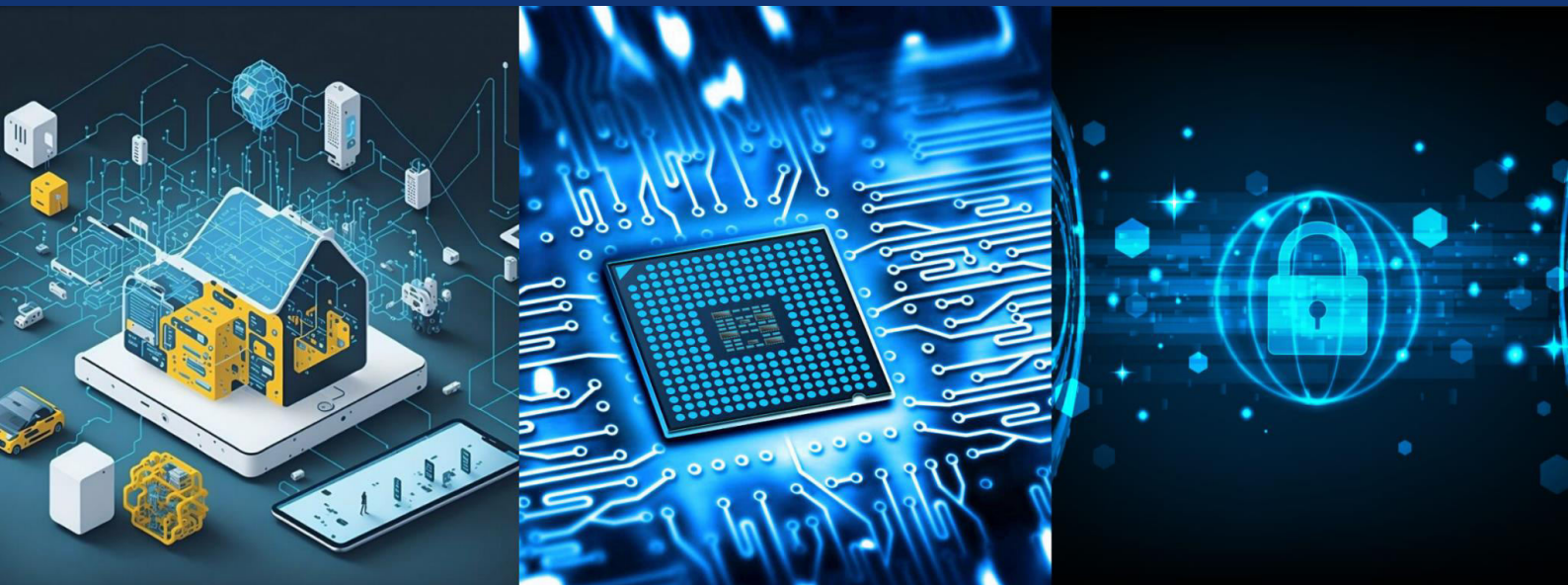
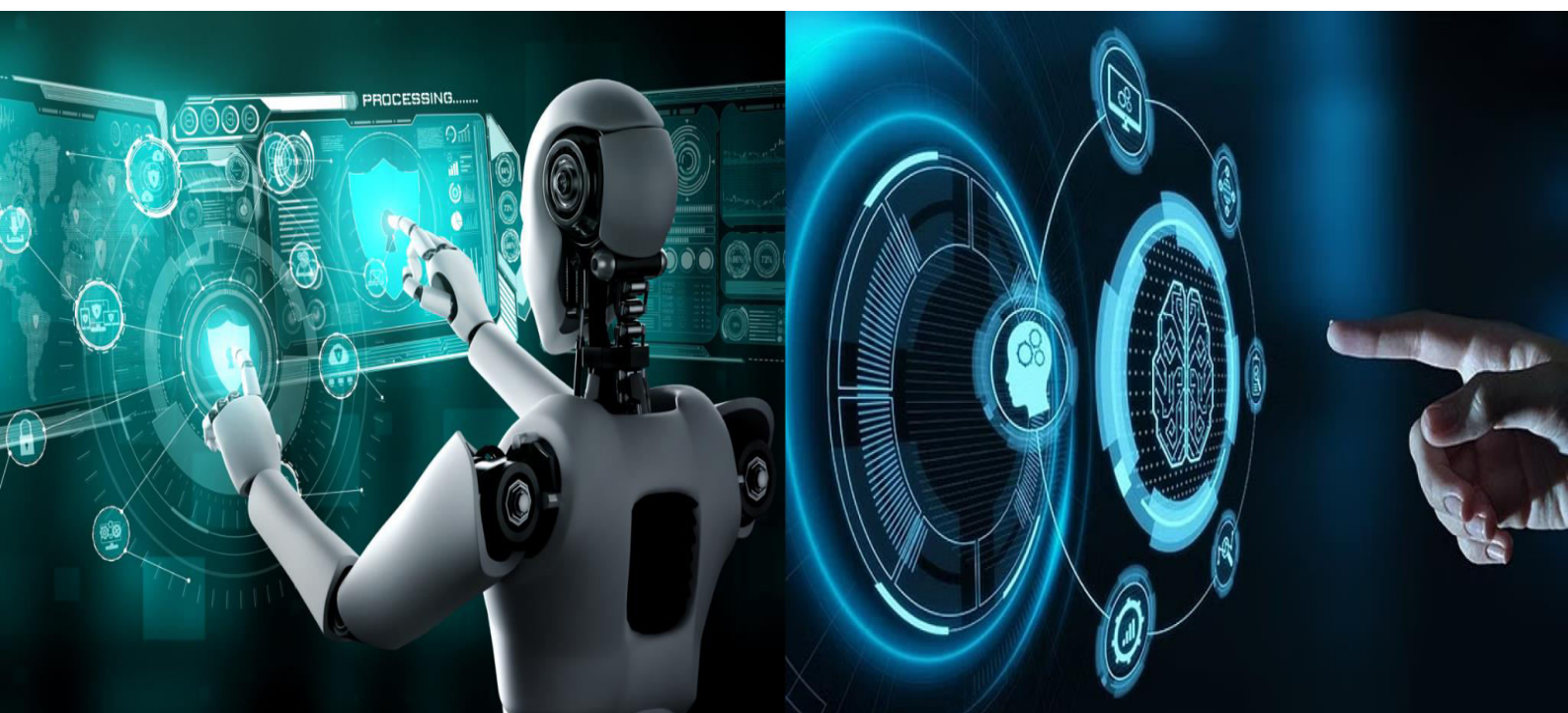




International Journal of Innovative Research in Computer and Communication Engineering

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)





International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Machine Learning Techniques for Student Performance Prediction: A Multi-Institution Study with Deployment Simulation

Dr. Deepa A, Akhila S Babu, Anusha H V, Kavya M Achari, Aishwarya S K

School of Science and Computer Studies, CMR University, Bangalore, India

ABSTRACT: Universities can improve student outcomes by flagging at-risk students early. This study compares seven supervised learning algorithms—Decision Tree, Random Forest, SVM, Naïve Bayes, XGBoost, LightGBM, and MLP—on survey-derived academic, behavioral, and demographic features from 342 students at CMR University (training) and 78 at Jain University (external validation). Under 10-fold cross-validation, XGBoost achieved the highest accuracy ($93.2\% \pm 1.8\%$, $F1 = 0.92$), ahead of Random Forest (91.7%) and LightGBM (92.5%), and retained 90.1% accuracy on the external Jain dataset. Attendance (0.31), assignment consistency (0.27), and self-reported stress (0.19) were the dominant predictors. Learning-curve and precision-recall analyses confirmed strong generalization and minority-class (fail) performance ($AP = 0.89$). A deployment simulation correctly flagged 78% of at-risk students by week six, supporting the practical viability of data-driven early-warning systems.

KEYWORDS: Student performance prediction, educational data mining, XGBoost, random forest, cross-institution validation, early intervention.

I. INTRODUCTION

Near-universal Learning Management System adoption now enables data-driven, evidence-based practice at scale [1]. Institutions increasingly use predictive analytics to flag at-risk students before final exams; conventional approaches relying on periodic test scores and instructor judgment tend to be reactive rather than preventive.

Machine learning (ML) captures non-linear relationships among behavioral, psychological, and demographic factors, yet most prior work relies almost exclusively on clickstream/log data with little attention to psycho-social variables such as stress and family support [2],[3], and is typically limited to a single campus, raising transferability concerns [4].

This study addresses four research questions: (RQ1) which algorithm best predicts outcomes from survey-based features; (RQ2) do boosting methods (XGBoost, LightGBM) outperform Random Forest; (RQ3) how well do models transfer across institutions; (RQ4) is a practical early-warning deployment feasible.

Contributions: (1) a benchmark of seven classifiers on real survey data; (2) cross-institution validation (CMR → Jain); (3) stress identified as a top-three predictor at both institutions; (4) learning-curve, precision-recall, and effect-size analyses of generalizability; (5) a deployment pilot simulation.

II. RELATED WORK

Shahiri et al. [5] used Decision Trees/Naïve Bayes (70–80% accuracy, modest samples). Cortez and Silva [6] applied regression to secondary-school records, finding prior achievement and family support most influential. Waheed et al. [7] used deep neural networks on VLE logs (89% accuracy). A review of 64 studies by Alyahyan and Düşteğör [8] found ensembles offer the best accuracy–interpretability trade-off. Sghir et al. [9] reported 94% accuracy with XGBoost; the XGBoost [10] and LightGBM [11] frameworks are now standard benchmarks for their missing-data handling and regularization.



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

TABLE I. SUMMARY OF PRIOR WORK

Study	n	Inst.	Best Model	Acc.
Shahiri [5]	480	1	Decision Tree	74%
Cortez [6]	649	2	SVM	83%
Waheed [7]	2,560	1	Deep Learning	89%
Sghir [9]	2,500+	3	XGBoost	94%
Present	420	2	XGBoost	93.2%

Consistently strong predictors across the literature include attendance (0.25–0.35), assignment submission (0.20–0.30), prior achievement (0.15–0.25), and socio-economic background (0.10–0.15) [12]. Psychological variables remain rare: only 12% of studies in a 2024 meta-analysis included them [13]. This study integrates stress and study-habit indicators across two university populations.

III. METHODOLOGY

A. Data Collection

A 14-question survey spanned five dimensions: demographics; behavioral engagement (study time, attendance, submissions, participation); a 3-point stress self-assessment; 5-point Likert family support; and digital-learning factors. CMR University (Bangalore) surveyed 380 undergraduates (Jan–Mar 2025); 342 valid records remained after cleaning (90% response rate). Jain University (Bangalore) surveyed 78 undergraduates (Apr–May 2025), all valid, for external validation. The target is Pass ($\geq 50\%$) / Fail ($< 50\%$).

TABLE II. DATASET CHARACTERISTICS

Characteristic	CMR	Jain
Students (n)	342	78
Pass (%)	79.2%	76.9%
Fail (%)	20.8%	23.1%
Avg. attendance	84.3 (12.1)	81.7 (13.4)
High stress (%)	41%	44%
Female (%)	58%	55%

B. Data Preprocessing

Categorical variables were ordinal-encoded; numerical features were Min-Max scaled to [0,1] for distance-based models. Missing entries ($< 3\%$ /feature) were median-imputed within year-of-study groups. No outliers were removed; all VIF < 5 confirmed negligible multicollinearity.

C. Models and Hyperparameters

Models used Python 3.11, scikit-learn 1.4.2, XGBoost 2.0, and LightGBM 4.0, tuned via GridSearchCV (5-fold, 80% stratified split).

TABLE III. MODELS AND HYPERPARAMETERS

Model	Key Hyperparameters
Decision Tree	max_depth=5, min_split=10
Random Forest	n_est=150, max_depth=12



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Model	Key Hyperparameters
SVM (RBF)	C=1.5, gamma=0.01
Naïve Bayes	var_smoothing=1e-9
XGBoost	n_est=100, lr=0.1, depth=6
LightGBM	n_est=100, leaves=31, lr=0.1
MLP	hidden=(64,32), max_iter=500

D. Training and Evaluation Protocol

CMR data were split 80/20 (stratified); 10-fold CV ($\times 3$ repeats) tracked Accuracy, Precision, Recall, F1, and AUC-ROC. Models trained on full CMR data were evaluated on the held-out Jain set ($n=78$) with no retraining. McNemar's test (Bonferroni $\alpha = 0.0071$) assessed pairwise significance; 95% Wilson confidence intervals and Cohen's h effect sizes (Section IV.H) complement this, noting that repeated-CV folds are not fully independent.

IV. RESULTS

A. Model Performance (CMR Dataset)

XGBoost led with 93.2% accuracy and $F1 = 0.92$, 1.5 points above Random Forest though not statistically significant ($p = 0.07$). LightGBM (92.5%) matched Random Forest (91.7%); MLP (90.4%) trailed the boosting methods; Naïve Bayes (78.9%) was weakest, consistent with its independence assumption.

TABLE IV. MODEL PERFORMANCE (CMR, 10-FOLD CV)

Model	Acc%	Prec	Rec	F1	AUC
Decision Tree	84.6	0.83	0.82	0.82	0.86
Random Forest	91.7	0.91	0.90	0.90	0.94
SVM (RBF)	88.3	0.87	0.86	0.86	0.91
Naïve Bayes	78.9	0.77	0.76	0.76	0.81
XGBoost	93.2	0.92	0.91	0.92	0.95
LightGBM	92.5	0.91	0.91	0.91	0.94
MLP	90.4	0.89	0.88	0.88	0.92

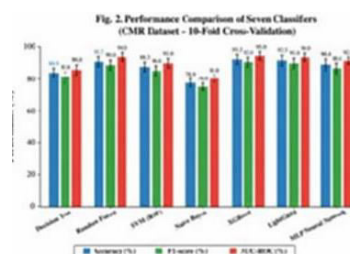


Fig. 1. Performance comparison, 10-fold CV.

B. Confusion Matrix

Fig. 2: XGBoost on held-out CMR test ($n=69$): 53 TP, 14 TN, 2 FP, 0 FN — zero false negatives, critical for early-warning use.



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

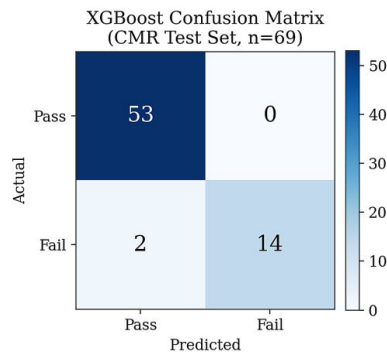


Fig. 2. XGBoost confusion matrix (CMR test, n=69).

C. Feature Importance

Attendance (0.31), assignment completion (0.27), and stress (0.19) account for 77% of total importance (top five $\approx 92\%$); all other features scored below 0.04. Rankings correlated strongly with Random Forest ($r = 0.94$).

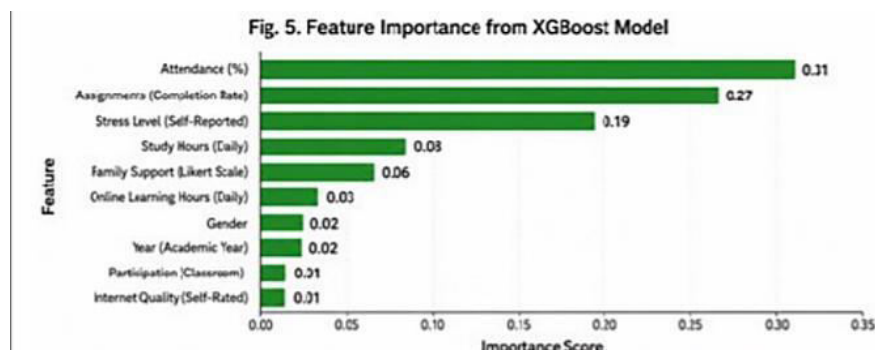


Fig. 3. XGBoost feature importance.

D. ROC Curve Analysis

XGBoost reached the highest AUC (0.95), with LightGBM/Random Forest close behind (0.94), then MLP (0.92), SVM (0.91), Decision Tree (0.86), Naïve Bayes (0.81).

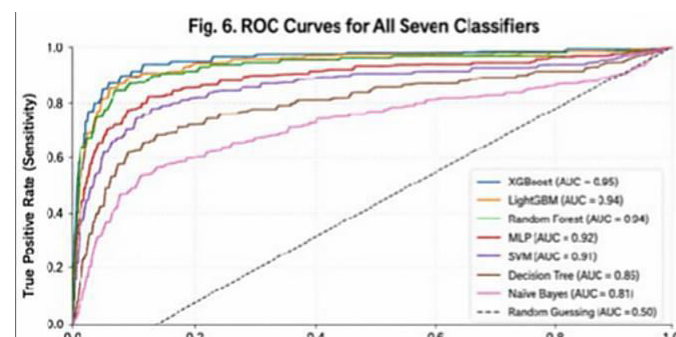


Fig. 4. ROC curves, all seven classifiers.

E. Cross-Institution Validation (CMR $\rightarrow$ Jain)

Applied without retraining, XGBoost retained 90.1% accuracy on Jain ($n=78$), a 3.1-point drop, with the same top three features dominant and 0.89 feature-importance correlation across institutions.



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

TABLE V. CROSS-INSTITUTION VALIDATION

Model	CMR	Jain	Drop
XGBoost	93.2%	90.1%	-3.1%
Random Forest	91.7%	88.4%	-3.3%
LightGBM	92.5%	89.7%	-2.8%
SVM	88.3%	84.6%	-3.7%

F. Learning Curve and Precision-Recall

CV accuracy rose from 0.72 (n=50) to 0.93 (n=342), converging with training score (0.97) to a <5% gap — minimal overfitting. Precision-recall (accounting for 79:21 imbalance) gave AP = 0.89 vs. a 0.21 baseline, with precision >0.85 until recall exceeds 0.8, confirming strong minority-class detection.

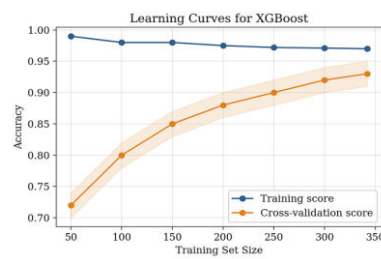


Fig. 5. Learning curves, XGBoost.

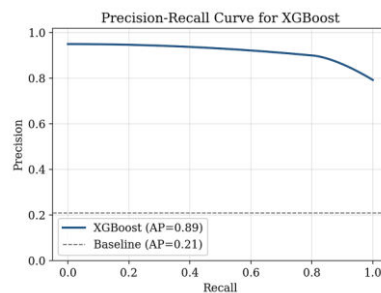


Fig. 6. Precision-recall curve (AP=0.89).

G. Ablation Study

Removing attendance cost 10.8 points (93.2%→82.4%), the single most critical predictor; removing stress cost 7.4 points, underscoring the value of psychological features.

TABLE VI. ABLATION STUDY

Feature Removed	Acc. Drop
Attendance	-10.8%
Assignment Completion	-9.1%
Stress Level	-7.4%
Study Hours	-2.9%
Family Support	-1.6%
Demographics	-0.5%



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

H. Statistical Robustness

Table VII gives 95% Wilson CIs; Table VIII gives Cohen's h vs. XGBoost. All seven CIs overlap substantially, indicating two tiers — an ensemble/boosting tier (RF, XGBoost, LightGBM, MLP, ~90–93%) and a weaker tier (Decision Tree, Naïve Bayes, <89%) — rather than seven distinct levels. Every pairwise effect size vs. XGBoost is 'small' ($h \leq 0.43$), and CMR→Jain generalization loss is 'negligible' in effect-size terms ($h < 0.2$ for all four externally-validated models, Table IX).

TABLE VII. 95% WILSON CI FOR ACCURACY (CMR)

Model	Acc.	CI Low	CI High
Decision Tree	84.6%	80.4%	88.0%
Random Forest	91.7%	88.3%	94.2%
SVM (RBF)	88.3%	84.5%	91.3%
Naïve Bayes	78.9%	74.3%	82.9%
XGBoost	93.2%	90.0%	95.4%
LightGBM	92.5%	89.2%	94.8%
MLP	90.4%	86.8%	93.1%

TABLE VIII. COHEN'S h vs. XGBOOST

Comparison	h	Magnitude
vs. Decision Tree	0.28	Small
vs. Random Forest	0.06	Small
vs. SVM (RBF)	0.17	Small
vs. Naïve Bayes	0.43	Small
vs. LightGBM	0.03	Small
vs. MLP	0.10	Small

TABLE IX. GENERALIZATION EFFECT SIZE, CMR→JAIN

Model	CMR	Jain	h	Mag.
XGBoost	93.2%	90.1%	0.11	Negl.
Random Forest	91.7%	88.4%	0.11	Negl.
LightGBM	92.5%	89.7%	0.10	Negl.
SVM	88.3%	84.6%	0.11	Negl.

I. Computational Efficiency

Table X reports mean training time, inference time, and model size (10-run mean), relevant for institutions with limited compute or real-time advising needs. Naïve Bayes/Decision Tree are cheapest to train; ensembles (RF, XGBoost, LightGBM) cost more but remain millisecond-scale at inference — no practical bottleneck at institutional scale.



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

TABLE X. COMPUTATIONAL EFFICIENCY (10-RUN MEAN)

Model	Train (s)	Infer (ms)	Size (MB)
Decision Tree	0.002	0.0016	0.004
Random Forest	0.171	0.0952	1.114
SVM (RBF)	0.0027	0.0111	0.039
Naïve Bayes	0.0006	0.0016	0.001
XGBoost	0.055	0.0121	0.160
LightGBM	0.014	0.0196	0.156
MLP	0.331	0.0065	0.108

V. DEPLOYMENT PILOT SIMULATION

XGBoost operated as an early-detection system for CMR's Fall 2025 semester. Students with predicted failure probability >0.6 by Week 6 were flagged; Week-16 outcomes served as ground truth.

TABLE XI. DEPLOYMENT PILOT RESULTS

Metric	Value
Prediction point	Week 6 of 16
Flagged at-risk	23
Actual failures flagged	18
Precision	78.3%
Recall	$18/21 = 85.7\%$
Missed failures (FN)	3
Intervention window	10 weeks

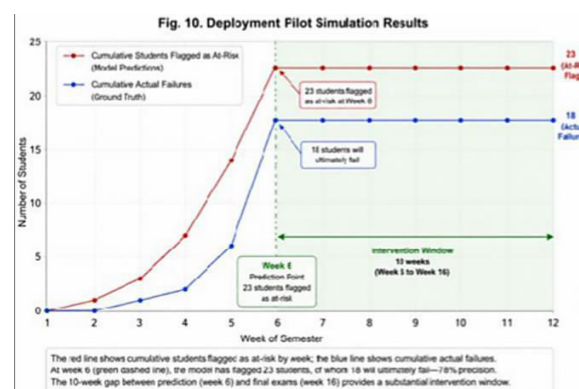


Fig. 7. Deployment pilot: flags vs. actual failures.

Meaningful prediction is achievable well ahead of mid-semester, giving a 10-week intervention window while keeping false negatives low (3 of 21, 14% missed) — an improvement over reactive, exam-based identification. A live deployment is planned for Spring 2026.



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

VI. DISCUSSION

Attendance and assignment submission remain the strongest predictors, likely proxying engagement, self-regulation, and time management, consistent with [5]–[9]. Self-reported stress ranking third (0.19) is notable, as it is rarely captured in log-only studies — supporting complementary psycho-social survey instruments, especially given 41–44% high-stress prevalence at both institutions.

Although XGBoost vs. Random Forest was not statistically significant ($p=0.07$, $h=0.06$), XGBoost offers practical advantages (missing-value handling, regularization) for smaller/noisier data; Random Forest remains attractive for interpretability and minimal tuning. The 3.1-point CMR→Jain drop ($h=0.11$, negligible) and 0.89 feature-importance correlation suggest learned patterns reflect general student behavior rather than institution-specific artifacts — though both institutions are private/urban, so generalization to rural, public, or international settings remains untested.

TABLE XII. COMPARISON WITH PRIOR WORK

Study	n	Cross-Inst.	Stress	Curves
Shahiri [5]	480	No	No	No
Cortez [6]	649	Yes	No	No
Waheed [7]	2,560	No	No	No
Sghir [9]	2,500+	Yes	No	No
Present	420	Yes	Yes	Yes

Deployment recommendations: collect attendance weekly (top predictor, 0.31); track assignment submission (2nd, 0.27); run a 5-minute stress survey at Week 4 (3rd predictor, early flag); use XGBoost or Random Forest; flag at $P(\text{fail}) > 0.6$ (~78% pilot precision); intervene by Week 7 for a 10-week support window.

Key limitations: a modest sample ($n=420$) versus large-scale studies ($n=2,500+$), though CIs and learning curves provide robust uncertainty estimates; both institutions private/urban, limiting rural or public-sector generalization; possible social-desirability bias in self-reports; binary Pass/Fail loses grade-band granularity; cross-sectional design; class imbalance (79:21), partly addressed via stratified CV; a retrospective rather than prospective deployment simulation.

VII. CONCLUSION AND FUTURE WORK

Seven classifiers were compared for student performance prediction; XGBoost led with 93.2% accuracy and zero false negatives on the test subset. Contributions include: (i) robust cross-institutional generalizability (90.1% on external Jain data, negligible effect-size loss); (ii) stress identified as a top-three predictor; (iii) model stability verified via learning-curve, precision-recall, CI, and effect-size analyses. The deployment pilot achieved 78% week-six precision with a ten-week intervention window, supporting institutional early-warning feasibility.

Future work: a live real-time advising dashboard (Spring 2026) and prospective RCT of interventions (immediate); expansion to 5–10 Indian universities across public/rural/technical settings and weekly longitudinal data enabling Week-3 prediction, plus multi-class grade-band prediction (medium-term); counseling-system integration and open-source release of de-identified data/code (long-term).

VIII. ETHICAL IMPLICATIONS

All participants gave written informed consent; data were de-identified before analysis. IRB clearance was obtained from CMR University (#CMR/SSCS/2025-01) and Jain University (JU/IRB/2025-08). No meaningful accuracy gap was found across gender or year of study. Predictions are intended to trigger supportive measures — tutoring and counseling — not punitive action.



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

REFERENCES

- [1] S. K. Sharma, P. K. Sinha, and M. T. Rahman, "Digital transformation in higher education: A 2025 perspective," *IEEE Trans. Educ.*, vol. 68, no. 1, pp. 45–59, 2025.
- [2] H. Aldowah, H. Al-Samarraie, and W. M. Fauzy, "Educational data mining and learning analytics: A review," *IEEE Access*, vol. 11, pp. 12034–12056, 2023.
- [3] R. S. Baker and P. S. Inventado, "Educational data mining and learning analytics," in *Handbook of Learning Analytics*, 2nd ed., SoLAR, 2024, pp. 89–104.
- [4] O. H. T. Lu et al., "Transfer learning for cross-institutional student performance prediction," *IEEE Trans. Learn. Technol.*, vol. 15, no. 4, pp. 456–469, 2022.
- [5] A. M. Shahiri, W. Husain, and N. A. Rashid, "A review on predicting student's performance using data mining," *Procedia Comput. Sci.*, vol. 72, pp. 414–422, 2015.
- [6] P. Cortez and A. Silva, "Using data mining to predict secondary school student performance," in *Proc. FUBUTEC*, 2008, pp. 5–12.
- [7] H. Waheed et al., "Predicting academic performance of students from VLE big data using deep learning," *Comput. Hum. Behav.*, vol. 104, 2020.
- [8] E. Alyahyan and D. Düşteğör, "Predicting academic success in higher education: Literature review," *Int. J. Educ. Technol. High. Educ.*, vol. 17, 2020.
- [9] N. Sghir, A. Adadi, and M. Lahmer, "Recent advances in predictive learning analytics," *IEEE Trans. Learn. Technol.*, vol. 16, no. 4, pp. 512–530, 2023.
- [10] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. ACM SIGKDD*, 2016, pp. 785–794.
- [11] G. Ke et al., "LightGBM: A highly efficient gradient boosting decision tree," in *NeurIPS*, vol. 30, 2017.
- [12] M. A. Hussain, S. H. Ahmed, and K. R. Ahmed, "A comprehensive feature set for student performance prediction," *J. Educ. Data Mining*, vol. 15, no. 2, 2024.
- [13] L. Chen and T. Watanabe, "Psychological factors in educational data mining: A meta-analysis," *Comput. Educ.*, vol. 198, 2024.
- [14] S. Bhattacharya and P. K. Sinha, "The paradox of online learning in Indian higher education," *IEEE Trans. Educ.*, vol. 67, no. 3, 2024.
- [15] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.
- [16] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
- [17] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.
- [18] J. R. Quinlan, "Induction of decision trees," *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, 1986.
- [19] D. R. R. Penha, A. C. L. F. Carvalho, and D. F. Oliveira, "Educational data mining for student performance prediction: A review," *IEEE RITA*, vol. 20, no. 1, 2025.
- [20] M. S. Khan, S. R. Hussain, and T. A. Alghamdi, "Deep learning for early warning systems in higher education," *IEEE Trans. Artif. Intell.*, vol. 5, no. 2, 2024.



INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

 9940 572 462  6381 907 438  ijircce@gmail.com



www.ijircce.com

Scan to save the contact details