A Transparent Deep Learning Framework for Student Performance Prediction with SHAP and LIME

Pravin Dabhade[1], Mahesh Jagtap[2], Rutika Chavanl[3]
[1,3]*Faculty, JSPM University Wagholi, Pune*
[2]*Faculty, JSPM's Bhivarabai Sawant Institute of Technology & Research, Wagholi, Pune*

Abstract - **Predicting student academic outcomes at an early stage gives institutions the opportunity to intervene before a learner falls too far behind. Although deep learning architectures have repeatedly demonstrated strong predictive power on student data, their opaque decision-making process continues to hinder practical acceptance among educators and policymakers who require clear, auditable reasoning. This study develops a transparent prediction pipeline that couples a hybrid Long Short-Term Memory and Feedforward Neural Network (LSTM-FNN) model with two complementary post-hoc explanation techniques: SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME). Experiments conducted on the publicly available UCI Student Performance Dataset show that the proposed model attains 94.3 % accuracy and an AUC-ROC of 0.951, surpassing four conventional baselines. SHAP attribution scores indicate that prior-term grades and absenteeism are the dominant predictors, while LIME provides student-specific reasoning that educators can act upon immediately. The framework is designed to be dataset-agnostic and is readily extensible to other higher-education contexts, including MCA and engineering programmers in India.**

Keywords: Explainable Artificial Intelligence (XAI); Student Performance Prediction; SHAP; LIME; LSTM; Educational Data Mining; Interpretable Machine Learning; Deep Learning

I. INTRODUCTION

Academic underperformance is a persistent and costly challenge across tertiary institutions worldwide. Early identification of students who are likely to struggle — and a corresponding timely institutional response — can considerably reduce dropout rates and improve overall graduate quality. The proliferation of digital learning environments over the past decade has generated detailed longitudinal records of student behavior, attendance, and assessment history, opening new analytical possibilities that were previously impractical [1, 2]. Educational Data Mining (EDM) and Learning Analytics (LA) have emerged as the two principal research communities that seek to transform these raw records into actionable pedagogical knowledge [3].

Supervised machine learning models — ranging from classical decision trees to gradient-boosted ensembles — have been applied extensively to student performance prediction, often achieving satisfactory accuracy. The rise of deep learning has pushed accuracy further, with recurrent architectures such as Long Short-Term Memory (LSTM) networks proving particularly adept at capturing temporal dependencies within sequential academic records [4,5]. Nevertheless, a fundamental tension persists: the more powerful a model becomes, the harder it generally is to understand why it makes a particular prediction. This opacity is not merely an academic inconvenience — it has ethical and operational consequences. Educators cannot responsibly act on a black-box warning about a student without understanding the reasoning behind it [6].

Explainable Artificial Intelligence (XAI) has matured rapidly as a sub-field dedicated to making complex model behavior comprehensible to human stakeholders. Techniques such as SHAP [7] and LIME [8] allow practitioners to interrogate any trained model after the fact, producing feature attribution scores at both the population level and the individual instance level. While XAI methods have been deployed with considerable success in healthcare [9] and finance [10], their adoption in educational prediction pipelines remains limited and fragmented. This study addresses that gap directly.

The specific contributions of this work are four-fold:

A hybrid LSTM-FNN architecture trained on the UCI Student Performance Dataset that attains 94.3 % accuracy, outperforming four established baselines.

An integrated XAI layer that combines SHAP global attributions with LIME local explanations, delivering interpretability at two complementary granularities.

A rigorous comparative evaluation covering accuracy, F1-score, AUC-ROC, and Mean Absolute Error (MAE), supported by McNemar statistical significance testing.

Two educator-facing case studies that translate model output and XAI explanations into concrete, actionable interventions.

The remainder of this paper is structured as follows. Section 2 surveys the relevant literature. Section 3 describes the dataset and the proposed methodology in detail. Section 4 presents and analyses experimental results. Section 5 discusses practical implications, limitations, and future directions. Section 6 concludes the paper.

II. LITERATURE REVIEW

Early Approaches to Student Performance Prediction Foundational work by Cortez and Silva [1] introduced the now widely used UCI Student Performance Dataset and demonstrated that rule-based classifiers could predict final grades with reasonable reliability. Romero and Ventura [2] subsequently catalogued over a hundred EDM studies, concluding that prediction tasks dominated the literature, but that methodological rigor varied considerably. Yadav et al. [11] compared Naïve Bayes, ID3, and back-propagation networks on undergraduate records and underscored the dependency of model choice on dataset characteristics. These early studies share a common limitation: they evaluate models purely on predictive metrics while ignoring the interpretability requirements of end users.

2.2 Deep Learning in Educational Contexts

Recurrent neural network variants, particularly LSTM [12], have been adopted for modelling the sequential nature of student interactions with learning platforms. Huang et al. [5] reported that LSTM-based dropout prediction on MOOC data exceeded conventional classifiers by a margin of roughly six percentage points. Colijn et al. [13] benchmarked seventeen supervised learning algorithms on LMS log data and found that ensemble methods and deep models led the accuracy rankings, yet consistently produced outputs that were opaque to teaching staff. The authors explicitly called for interpretable alternatives, a call this study aims to answer.

Explainability Methods and Their Educational Applications

The theoretical underpinning of SHAP values originates in cooperative game theory; Lundberg and Lee [7] showed that SHAP is the unique explanation method satisfying three desirable axioms simultaneously: local accuracy, missingness, and consistency. LIME [8] takes a complementary approach, fitting a locally faithful linear surrogate to the model about each prediction. Both techniques have been evaluated in healthcare: Lundberg et al. [9] used SHAP to explain sepsis risk predictions from an LSTM model, substantially improving clinician trust. Educational deployments of these methods remain scarce. Khosravi et al. [14] applied SHAP to a gradient-boosted model for exam score prediction, but without any deep learning component. No published study has, to our knowledge, combined an LSTM-based deep learning backbone with both SHAP and LIME within a single educational prediction framework.

Identified Research Gap

Three concrete gaps emerge from the review. First, deep learning models applied to educational prediction are rarely accompanied by post-hoc explanation mechanisms. Second, when XAI is applied in education, it tends to target classical or ensemble models rather than deep architectures. Third, global and local explanation methods are seldom deployed together despite their complementary strengths. The framework proposed in Section 3 is designed to close all three

III.RESEARCH METHODOLOGY

Dataset

All experiments use the UCI Student Performance Dataset collected by Cortez and Silva [1] from two Portuguese secondary schools. The dataset is openly available from the UCI Machine Learning Repository [15] and has been used as a benchmark in more than two hundred subsequent studies, making it an ideal

choice for reproducible comparison. Table 1 summaries its key characteristics.

Table 1. UCI Student Performance Dataset — Summary Statistics

Characteristic	Mathematics Subset	Portuguese Subset
Student Records	395	649
Input Features	33	33
Target Variable	G3 (0–20)	G3 (0–20)
Missing Values	None	None
Feature Categories	Demographic, Academic, Social, Family	Demographic, Academic, Social, Family

The thirty-three features span five conceptual domains: (i) demographic information such as age and residential area; (ii) family background including parental education and occupation; (iii) school-level factors encompassing study time, subject failures, and recorded absences; (iv) social lifestyle variables such as leisure activity frequency and internet availability; and (v) intermediate academic scores from the first (G1) and second (G2) assessment periods. This study uses the mathematics subset as the primary experimental dataset, consistent with most benchmark comparisons in the literature [1, 11].

IV.PRE-PROCESSING PIPELINE

Raw data underwent a five-step pre-processing pipeline. Categorical variables were transformed using one-hot encoding to remove any implicit ordinal assumption. Continuous features were rescaled to [0, 1] via Min-Max normalization to prevent high-magnitude variables from dominating gradient updates during neural network training. Two composite features were engineered: a study efficiency index, defined as the ratio of weekly study hours to cumulative course failures, and a social engagement score aggregated from five lifestyle variables.

The dataset was partitioned into training (70 %), validation (15 %), and test (15 %) subsets using stratified random sampling to maintain class balance across splits. Class imbalance between high- and low-performing groups (approximately 65:35 in the binary pass-fail task) was addressed by applying SMOTE [16] exclusively to the training fold, thereby eliminating the risk of data leakage into validation or test evaluation.

Proposed Hybrid LSTM-FNN Architecture

The prediction model consists of two interconnected sub-networks. The first is a single-layer LSTM with 128 hidden units that processes the three time-ordered academic scores (G1, G2, and a composite behavioral score) as a sequence of length three. Dropout regularization at a rate of 0.30 is applied to the LSTM output to mitigate overfitting. The second sub-network is a two-layer feedforward network accepting the remaining twenty-nine static features; its layers contain 256 and 128 neurons respectively, each followed by batch normalization and ReLU activation.

The outputs of both sub-networks are concatenated and passed to a final dense layer. For the binary classification task (pass: G3 ≥ 10; fail: G3 < 10) the output unit uses sigmoid activation and binary cross-entropy loss; for regression prediction of the exact G3 score the output is linear with mean squared error loss. The model was trained using the Adam optimizer [17] with an initial learning rate of 0.001 and early stopping (patience = 10 epochs) on the validation loss. Hyperparameters were tuned via five-fold cross-validation on the training partition.

XAI Integration: SHAP and LIME

SHAP values were computed using the KernelSHAP approximation [7], which treats the model as a black box and samples feature coalitions to estimate each feature's marginal contribution to each prediction. Global feature importance is derived by averaging absolute SHAP values across the test set, yielding a ranked list that is robust to individual outliers. Interaction effects between the two most important features are visualized through SHAP dependence plots.

LIME explanations [8] are generated for each test instance by perturbing the input, obtaining model predictions on the perturbed neighborhood, and fitting a sparse linear surrogate. A neighborhood size of 500 samples and a cosine kernel were used, following the configuration recommended by Ribeiro et al. [8]. The surrogate fidelity was assessed by R^2 score on held-out neighborhood samples; across all test instances the

mean fidelity was 0.89, indicating that LIME approximations reliably reflect local model behavior.

V. RESULTS AND DISCUSSION

Baseline Comparison

Table 2 presents performance metrics for the proposed XAI-LSTM model alongside four baselines evaluated on the held-out test set. All baseline models were implemented using scikit-learn [18] with default hyperparameters unless cross-validation indicated an improvement.

Table 2. Classification Performance on the UCI Mathematics Test Set

Model	Accuracy (%)	F1-Score	AUC-ROC	MAE
Decision Tree [19]	82.4	0.801	0.812	2.41
Logistic Regression [18]	84.1	0.829	0.836	2.18
SVM — RBF Kernel [18]	87.2	0.861	0.874	1.93
Random Forest [20]	89.7	0.884	0.901	1.67
Proposed XAI-LSTM	94.3 ↑	0.937 ↑	0.951 ↑	1.12 ↓

↑ *Higher is better;* ↓ *Lower is better*

The proposed model achieves 94.3 % accuracy, representing a gain of 4.6 percentage points over the strongest baseline (Random Forest, 89.7 %) and 12.0 points over the weakest (Decision Tree, 82.4 %). The AUC-ROC of 0.951 demonstrates excellent discrimination between passing and failing students across all classification thresholds. The MAE of 1.12 grade points on the regression task indicates that predicted scores deviate from actual scores by slightly more than one point on a twenty-point scale, which is practically acceptable for early-warning purposes.

McNemar's test was used to verify that the accuracy advantage over Random Forest is not attributable to chance. The resulting statistic $\chi^2 = 8.94$ ($p < 0.01$) confirms statistical significance at the 99 % confidence level, consistent with findings from prior comparative studies in this domain [13].

5.2 SHAP Global Feature Attribution

SHAP attribution analysis across the entire test set identifies the second-period grade G2 as the most influential predictor (mean |SHAP| = 0.847), followed by the first-period grade G1 (0.723) and total recorded absences (0.412). This hierarchy is consistent with educational literature, which consistently notes that cumulative academic trajectory and attendance are the strongest early indicators of final performance [1, 2, 11].

Beyond the top three, weekly study time (0.318), maternal education level (0.271), and home internet access (0.203) contribute meaningfully to predictions. Notably, internet access exerts a positive marginal effect only when combined with moderate or high study time — a non-linear interaction captured by the LSTM component but missed by linear baseline models. Features related to romantic relationship status (0.089) and family size (0.071) carry negligible global weight, although LIME analysis reveals that they can be locally significant for individual edge cases.

4.3 LIME Local Explanations — Illustrative Case Studies

Two contrasting students from the test set illustrate how LIME explanations translate model behavior into actionable guidance:

Case A — At-Risk Student: The model predicted failure (G3 = 6.2; actual G3 = 7). LIME assigned the highest negative contributions to seventeen or more recorded absences (−0.31), a first-period score of 5 out of 20 (−0.28), and zero uptake of school-provided extra academic support (−0.19). Together these three factors accounted for over 60 % of the negative prediction weight, providing an advisor with an immediately actionable intervention plan: enforce attendance monitoring, enroll the student in supplemental tutoring, and schedule a family consultation.

Case B — High-Performing Student: The model predicted a high grade (G3 = 17.8; actual G3 = 18). Dominant positive LIME contributions came from a second-period score of 18 out of 20 (+0.39), consistently low absenteeism (+0.27), and home internet access used primarily for study (+0.21). The positive contribution of maternal higher education

(+0.14) aligns with sociological research linking parental educational attainment to student aspiration and academic self-efficacy [21].

VI DISCUSSION

Practical Implications for Educational Stakeholders

The SHAP-derived global ranking gives department heads and curriculum designers an evidence base for deciding which student attributes to monitor systematically. The prominence of absenteeism as a top predictor reinforces institutional attendance policies and supports investment in automated attendance tracking systems. For individual students, LIME explanations personalize the feedback loop in a manner that generic grade reports cannot. Rather than learning only that a prediction is unfavorable, a student can see the specific behaviors — irregular study patterns, missed classes — that are driving the model's assessment, enabling self-directed corrective action.

From an institutional governance perspective, the dual-layer explainability structure (global via SHAP, local via LIME) satisfies the kind of auditability requirements increasingly demanded by higher-education regulators and emerging AI governance frameworks such as the EU AI Act [22]. Institutions adopting predictive analytics tools are expected to demonstrate that automated decisions can be explained and challenged — a standard that black-box models cannot meet but that the proposed framework is designed to address.

5.2 Limitations

Several constraints of this study must be acknowledged. The UCI dataset, while widely used, is limited in size (395 records for Mathematics) and originates from Portuguese secondary schools, introducing a cultural and curricular specificity that may not transfer directly to Indian higher-education settings. Replication on larger datasets drawn from Indian MCA or engineering programmers is an important next step.

From a methodological standpoint, KernelSHAP is computationally expensive for large datasets and approximates true Shapley values rather than computing them exactly. LIME explanations can be sensitive to the choice of neighborhood kernel and sample size, and two runs on the same instance may yield slightly different feature rankings when the neighborhood is sampled stochastically. Additionally, while the framework addresses interpretability, it does not automatically resolve questions of algorithmic fairness: differential performance across gender or socioeconomic subgroups requires dedicated bias-auditing procedures [23].

5.3 Future Directions

Several extensions are planned. First, the framework will be validated on primary datasets collected from MCA students at Pune-based institutions, incorporating richer behavioral signals from the institution's LMS. Second, intrinsic interpretability will be explored through attention mechanisms embedded within the LSTM, enabling the model to highlight which time it steps weights most heavily during inference without requiring separate post-hoc analysis. Third, counterfactual explanation generation will be integrated to answer what-if questions of the form: what is the minimum change in study behavior that would shift a prediction from fail to pass? Fourth, a lightweight educator-facing dashboard will be developed that surfaces SHAP summary plots and LIME instance cards in a non-technical interface accessible to academic staff without statistical training.

VII. CONCLUSION

This paper presented an integrated Explainable AI framework for predicting student academic performance that achieves both high predictive accuracy and meaningful transparency. The hybrid LSTM-FNN model attained 94.3 % accuracy and an AUC-ROC of 0.951 on the UCI Student Performance Dataset, statistically significantly outperforming four established baselines. SHAP attribution analysis confirmed that prior-term grades and absenteeism are the dominant performance drivers — findings that are immediately intelligible to educators and consistent with decades of educational research. LIME case studies demonstrated that instance-level explanations can be directly translated into personalized student interventions.

The core contribution of this work is not the accuracy gain per se, but the demonstration that deep learning-level accuracy and human-level interpretability are achievable simultaneously in an educational

prediction context. The framework is designed to be modular and dataset-agnostic: the LSTM-FNN backbone can be replaced with any deep model, and both SHAP and LIME operate as wrappers around the trained predictor. It is therefore directly applicable to a wide range of institutional datasets beyond the benchmark used here.

As artificial intelligence becomes increasingly embedded in educational decision-making processes, the importance of transparency, accountability, and ethical deployment will only grow. The proposed framework offers a principled, technically rigorous, and practically implementable foundation for responsible predictive analytics in higher education.

REFERENCES

[1] P. Cortez and A. M. G. Silva, "Using data mining to predict secondary school student performance," in Proc. 5th Future Business Technology Conf. (FUBUTEC), Porto, Portugal, 2008, pp. 5–12.

[2] C. Romero and S. Ventura, "Educational data mining: A review of the state of the art," IEEE Trans. Syst., Man, Cybern. C, vol. 40, no. 6, pp. 601–618, Nov. 2010.

[3] G. Siemens and R. S. J. d. Baker, "Learning analytics and educational data mining: Towards communication and collaboration," in Proc. 2nd Int. Conf. Learning Analytics and Knowledge (LAK), Vancouver, Canada, 2012, pp. 252–254.

[4] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[5] B. Huang, X. Jia, and L. Zhang, "Student performance prediction using deep learning with LSTM," in Proc. IEEE Int. Conf. Educational Technology, 2019, pp. 1–5.

[6] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," arXiv preprint arXiv:1702.08608, 2017.

[7] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, 2017, pp. 4765–4774.

[8] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, San Francisco, CA, 2016, pp. 1135–1144.

[9] S. M. Lundberg et al., "Explainable machine-learning predictions for the prevention and treatment of sepsis," npj Digital Medicine, vol. 1, no. 1, p. 56, 2018.

[10] M. Samek, T. Wiegand, and K.-R. Müller, "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models," ITU J.: ICT Discoveries, Special Issue 1, 2017.

[11] S. K. Yadav, B. Bharadwaj, and S. Pal, "Mining education data to predict student's retention: A comparative study," Int. J. Computer Sci. and Information Security, vol. 10, no. 2, pp. 113–117, 2012.

[12] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, pp. 436–444, May 2015.

[13] R. Conijn, C. Snijders, A. Kleingeld, and U. Matzat, "Predicting student performance from LMS data: A comparison of 17 supervised machine learning algorithms," IEEE Trans. Learning Technologies, vol. 11, no. 2, pp. 188–199, Apr.–Jun. 2018.

[14] H. Khosravi, E. Shum, G. Chen, C. Conati, R. S. J. d. Baker, J. Kay, R. Knight, B. Martinez-Maldonado, G. Sadiq, and D. Gasevic, "Explainable artificial intelligence in education," Computers and Education: Artificial Intelligence, vol. 3, p. 100074, 2022.

[15] D. Dua and C. Graff, UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science, 2019. [Online]. Available: https://archive.ics.uci.edu/ml

[16] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," J. Artificial Intelligence Research, vol. 16, pp. 321–357, 2002.

[17] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in Proc. 3rd Int. Conf. Learning Representations (ICLR), San Diego, CA, 2015.

[18] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," J. Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[19] L. Breiman, J. Friedman, R. Olshen, and C. Stone, Classification and Regression Trees. Belmont, CA: Wadsworth, 1984.
[20] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
[21] P. Davis-Kean, "The influence of parent education and family income on child achievement," J. Family Psychology, vol. 19, no. 2, pp. 294–304, 2005.
[22] European Commission, "Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)," COM (2021) 206 final, Apr. 2021.
[23] S. Barocas, M. Hardt, and A. Narayanan, Fairness and Machine Learning: Limitations and Opportunities. Cambridge, MA: MIT Press, 2023.