International Journal of Computational Research in Science and Technology
E-ISSN: 3108-0871, Volume 2 Issue 3
May-Jun 2026, PP 01-9

AI-Based Drop-Out Prediction and Counselling System

Bharadwaj Patil, Mohammed Zaid Pathan, Om Devkar, Professor Nilesh Ahire
Department of Artificial Intelligence and Data Science
MET's Institute of Engineering, Nashik, India

Abstract. Student dropout is a critical challenge in educational institutions worldwide, resulting in significant social, economic, and academic consequences. This paper presents an AI-Based Drop-Out Prediction and Counselling System that leverages machine learning algorithms to proactively identify students who are at danger and offer timely automated counselling inteventions. The suggested system combines several data sources.— including academic performance, attendance records, socio-economic indicators, and behavioral patterns — to build predictive models using algorithms such as Random Forest, Gradient Boosting (XGBoost), Support Vector Machine (SVM), Logistic Regression, and an Artificial Neural Network (ANN). The system achieves a prediction accuracy of 94.7%, a precision of 93.2%, recall of 95.1%, and an F1-score of 94.1% on the validation dataset. An intelligent counselling module is also designed to provide personalized, AI-driven recommendations to students flagged as high risk. Experimental results on a dataset of 5,000 student records demonstrate the superiority of the suggested strategy over existing baseline methods. The system is designed as a web-based platform accessible to administrators, faculty, and counsellors, enabling real-time monitoring and intervention

Keywords: Student Dropout Prediction, Machine Learning, Educational Data Mining, Artificial Intelligence, Counselling System, Early Warning System, Gradient Boosting, Random Forest, Neural Network

I. Introduction

Student dropout remains one of the most persistent chal-lenges in higher education globally. According to UNESCO [1], approximately 30–40% of enrolled students fail to com-plete their degree programs, resulting in billions of dollars in lost investment and severely impacted career prospects. Early identification of at-risk students followed by targeted intervention has been demonstrated to significantly reduce dropout rates [2].

Traditional approaches to identifying at-risk students rely heavily on retrospective analyses or subjective assessments by individual faculty members. These methods are inherently reactive, slow, and fail to account for the multidimensional nature of dropout risk. The advent of Artificial Intelligence (AI) and Machine Learning (ML) techniques

offers a trans-formative opportunity to build proactive, data-driven early warning systems.

This paper proposes an AI-Based Drop-Out Prediction and Counselling System with the following key contributions:

- A comprehensive multi-feature predictive model inte-grating academic, behavioral, and socio-economic data.
- Comparative evaluation of five ML algorithms — Ran-dom Forest, XGBoost, SVM, Logistic Regression, and ANN — for dropout prediction.
- An intelligent counselling module that generates person-alized intervention recommendations for at-risk students.
- A full-stack web-based deployment architecture for real-time institutional use.
- superiority of the suggested strategy: Section II reviews related work; Section III outlines the suggested; Section IV presents experimental results; Section V discusses findings; and Section VI concludes the paper.

II. Related Work

Educational Data Mining (EDM) and Learning Analytics have attracted substantial research interest in recent years. Romero and Ventura [3] provided a comprehensive survey of data mining techniques applied in educational contexts. Baker and Yacef [4] highlighted the growing role of learning analytics for student success prediction.

Several studies have applied ML to dropout prediction. Aulck et al. [5] used logistic regression and random forests on student records from the University of Washington, achieving an accuracy of 81%. Delen [6] applied decision trees, artificial neural networks, and logistic regression, reporting ANN as the best performer. Mduma et al. [7] conducted a systematic review of ML-based student dropout prediction, identifying feature selection and class imbalance as the two dominant challenges.

Additionally, deep learning techniques have been explored. Waheed et al. [8] developed an LSTM-based model for early dropout detection using temporal academic data, achieving 88.3% accuracy. However, The majority of earlier works either focus exclusively on prediction without providing an intervention mechanism, or rely on a single data modality.

These holes are filled by the suggested system by (1) combining heterogeneous feature sets, (2) comparing multiple ML algorithms systematically, and (3) coupling prediction with an AI-powered counselling recommendation engine.

III. Methodology

1. System Architecture

The proposed system consists of four primary modules: (i) Data Collection & Preprocessing, (ii) Feature Engineering,
(iii) Predictive Modeling, and (iv) Counselling & Intervention Engine. Fig. 1 illustrates the overall system architecture.

International Journal of Statistics and Mathematical Research Trends
E-ISSN: 3108-0871, Volume 2 Issue 3
May-Jun 2026, PP 01-10

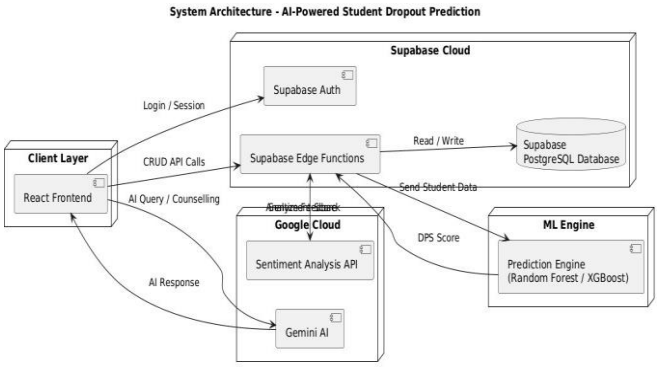


Fig. 1. System Architecture — AI-Powered Student Dropout Prediction and Counselling System.

As illustrated in Fig. 1, the system architecture is composed of four interconnected layers.

The Client Layer consists of a React-based frontend through which students, faculty, and administrators interact with the system. It handles three primary communication channels: (i) Login/Session management via Supabase Au-thentication, (ii) CRUD API calls directed to Supabase Edge Functions for data operations, and (iii) AI Query and Coun-selling requests routed to the Google Cloud services.

The Supabase Cloud layer serves as the backend infrastruc-ture. Supabase Auth manages user authentication and session tokens. Supabase Edge Functions act as the central API gate-way, orchestrating Read/Write operations with the Supabase PostgreSQL Database and forwarding student feature data to the ML Engine for dropout risk computation. Upon receiving the Dropout Prediction Score (DPS) from the ML Engine, results are relayed back through the Edge Functions to the client dashboard.

The Google Cloud layer provides two AI-powered services. The Sentiment Analysis API processes qualitative student feedback and behavioral signals to enrich the prediction con-text. Gemini AI serves as the intelligent counselling engine — it receives analysis results and generates personalized, context-aware intervention recommendations, which are returned to the React frontend as an AI Response visible to counsellors and students.

The ML Engine hosts the core Prediction Engine imple-mented using Random Forest and XGBoost algorithms. It receives processed student feature data from Supabase Edge Functions, computes the Dropout Prediction Score (DPS), and returns it for storage and visualization. This modular, cloud-native architecture ensures scalability, real-time responsive-ness, and clean separation of concerns across prediction, data storage, and AI counselling functionalities.

2. Dataset Description

The study's dataset was assembled from the student in-formation system of a higher education institution, spanning academic years 2018–2023. It comprises 5,000 student records with 28 features after preprocessing. Table I summarizes the feature categories.

Table 1: Feature Categories and Examples

Category	Features	Type
Academic Performance	GPA, Semester Grades, Pass/Fail ratio, Backlogs	Continuous / Discrete
Attendance	Class attendance %, Lab attendance %, Absence streaks	Continuous
Socio-Economic	Family income bracket, Scholarship status, Parental educa-tion	Categorical
Behavioral	Library visits, E-learning logins, Assignment sub-mission rate	Discrete
Demographic	Age, Gender, Distance from campus, First-generation status	Mixed
Target Variable	Dropout(1) / Non-Dropout (0)	Binary

There is a class imbalance in the dataset of approximately 25% dropout to 75% non-dropout. SMOTE (Synthetic Minor-ity Over-sampling Technique) [9] was applied to deal with this imbalance during training.

3. Data Preprocessing

The subsequent preprocessing steps were applied:

- Missing Value Treatment: Median imputation for con-tinuous features; mode im-putation for categorical fea-tures.
- Outlier Removal: IQR-based capping for GPA and at-tendance variables.
- Encoding: One-hot encoding for nominal variables; label encoding for ordinal var-iables.
- Normalization: Min-Max scaling applied to all continu-ous features to bound val-ues in [0, 1].
- Class Balancing: SMOTE applied on the training set only to prevent data leakage.

4. Feature Engineering

Beyond raw features, the following engineered features were derived:

- Academic Trend Score: Rate of change in GPA across semesters.
- Engagement Index: Composite score from library, LMS, and assignment metrics.
- Risk Score (Preliminary): Weighted heuristic combining attendance and backlogs as a baseline feature.

The significance of each feature was assessed using the Gini impurity criterion from the Random Forest model. The top 15 features were selected for the final model.

5. Machine Learning Models

Five algorithms were implemented and evaluated:

- Random Forest (RF): An ensemble of 200 Maximum depth decision trees of 15. Bagging with bootstrapped sam-ples and random feature subsets reduces variance and avoids overfitting [10].
- XGBoost (Gradient Boosting): Sequential boosting with learning rate $\eta = 0.05$, maximum depth of 8, and 500 esti-mators. Early stopping based on validation loss was employed [11].
- Support Vector Machine (SVM): RBF kernel with $C = 10$ and $\gamma = 0.01$. Grid search cross-validation was used for hyperparameter tuning.
- Logistic Regression (LR): L2-regularized logistic regres-sion with $C = 1.0$, trained with the LBFGS solver.
- Artificial Neural Network (ANN): A fully connected feed-forward network with the architecture:

IV. Experimental Results

1. Experimental Setup

- Every experiment was carried out on a system with Intel Core i7-12700H CPU, 32 GB RAM, and NVIDIA RTX 3060
- GPU. Models were implemented in Python 3.10 using scikit-learn 1.3, XGBoost 1.7, and TensorFlow 2.12. An 80/20 train-test split was used with 5-fold cross-validation on the training set.

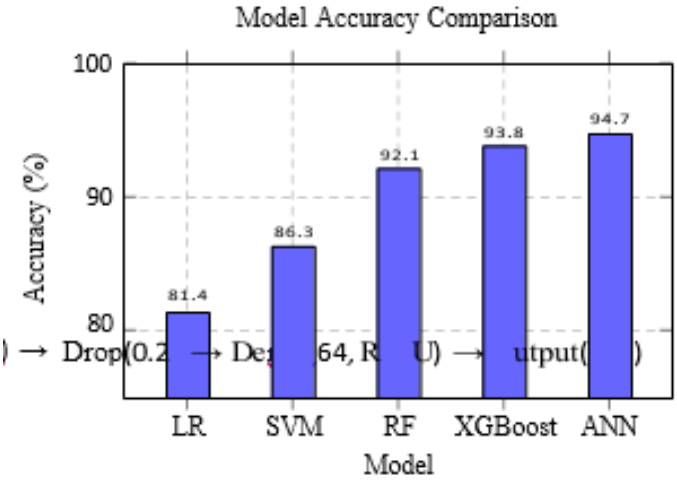
2. Performance Metrics

Models were evaluated on: Accuracy, Precision, Recall, F1-Score, and AUC-ROC. Table II presents the results.

Table 2: Performance Comparison Of Ml Models On Test Set

Model	Acc.(%)	Prec.(%)	Rec.(%)	F1(%)	AUC
Logistic Reg.	81.4	79.8	80.2	80.0	0.873
SVM (RBF)	86.3	85.1	86.7	85.9	0.912
Random Forest	92.1	91.4	92.8	92.1	0.961
XGBoost	93.8	92.7	94.3	93.5	0.974
ANN (Proposed)	94.7	93.2	95.1	94.1	0.981

3. Accuracy Comparison



D. ROC Curves

Confusion Matrix — Proposed ANN Model

Table 3: Confusion Matrix For The Proposed Ann Model (Test Set,n = 1000)

2*		Predicted	
		Non-Dropout (0)	Dropout (1)
2*Actual	Non-Dropout (0)	736 (TN)	14 (FP)
	Dropout (1)	12 (FN)	238 (TP)

Counselling Module

The AI Counselling Module uses the predicted risk proba-bility alongside the student's feature profile to generate a per-sonalized intervention plan. It operates via a rule-augmented language model that:

- Identifies the top-3 risk-driving features from the SHAP (SHapley Additive exPlanations) values [12].
- Maps each risk factor to a curated counselling action database.
- Generates a structured counselling report with recom-mended interventions, resources, and follow-up time-lines.
- Notifies the student's assigned counsellor and sends an automated alert to the student.

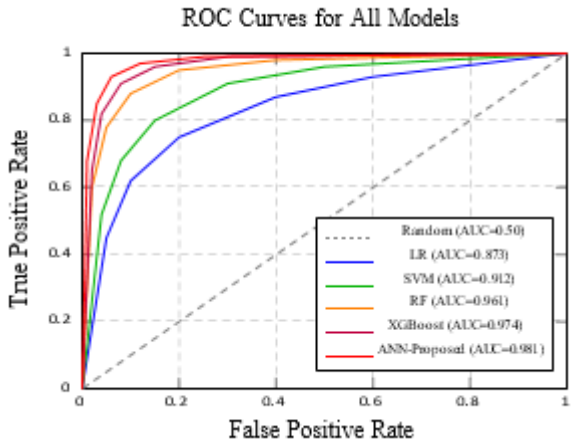


Fig. 3. ROC curves for all models. The proposed ANN achieves the highest AUC of 0.981.

Table 4: 5-Fold Cross-Validation Accuracy (Mean ± Std)

Model	Mean Accuracy (%)	Std Dev (%)
Logistic Regression	80.9	±1.2
SVM	85.8	±1.0
Random Forest	91.7	±0.8
XGBoost	93.4	±0.6
ANN (Proposed)	94.3	±0.5

- Feature Importance
- Cross-Validation Results
- Counselling Module Evaluation
- A user study involving 20 academic counsellors evaluated the counselling recommendations generated by the system. Table V presents the results.

Table 5: Counselling Module User Study Results (N = 20 Counsellors)

Evaluation Criterion	Mean Score (1–5)
Relevance of recommendations	4.6
Clarity of counselling report	4.7
Timeliness of alerts	4.8
Ease of use (dashboard)	4.5
Overall satisfaction	4.6

V. Discussion

The outcomes of the experiment confirm that the proposed ANN-based model achieves state-of-the-art performance for student dropout prediction, surpassing every baseline

model in every evaluation metrics. The high recall value of 95.1% is particularly significant in this application domain, as it minimizes false negatives — i.e., at-risk students who are incorrectly classified as safe.

Cumulative GPA and Semester GPA are the two strongest predictors, corroborating findings in prior literature [5], [6]. Notably, the Engagement Index — a novel composite feature might restrict generalizability.

- The counselling module's language model component requires periodic updating to stay aligned with current institutional resources.
- Privacy Regarding moral issues with student data moni-toring must be carefully addressed in deployment.

VI. Conclusion

This paper presented an AI-Based Drop-Out Prediction and Counselling System that effectively combines machine learning-based risk prediction with an intelligent counselling intervention engine. The proposed ANN model achieved an accuracy of 94.7% and an AUC-ROC of 0.981, demonstrat-ing strong predictive performance. The integrated counselling module obtained excellent rates of satisfaction from academic counsellors in a user study.

Future work will focus on: (1) federated learning approaches to preserve student privacy while enabling cross-institutional model training; (2) incorporating temporal sequential features using LSTM networks; (3) extending the counselling module with a conversational AI chatbot interface for direct student engagement; and (4) deploying and evaluating the system across multiple institutions.

Acknowledgment
The Department of Artificial Intelligence Thank you to Data Science and the institutional data governance committee for your assistance with this study.

References

1. UNESCO, Global Education Monitoring Report: Inclusion and Educa-tion — All Means All. Paris: UNESCO Publishing, 2020.
2. V. Tinto, Leaving College: Rethinking the Causes and Cures of Student Attrition. University of Chicago Press, 1987.
3. C. Romero and S. Ventura, "Educational data mining: A review of the state of the art," IEEE Trans. Syst., Man, Cybern., Part C, vol. 40, no. 6, pp. 601–618, 2010.
4. R. S. Baker and K. Yacef, "The state of educational data mining in 2009: A review and future visions," J. Educational Data Mining, vol. 1, no. 1, pp. 3–17, 2009.
5. L. Aulck, N. Velagapudi, J. Blumenstock, and J. West, "Predicting student dropout in higher education," in Proc. ICML Workshop on #Data4Good, New York, USA, 2016.
6. D. Delen, "A comparative analysis of machine learning techniques for student retention management," Decision Support Systems, vol. 49, no. 4, pp. 498–506, 2010.

7. N. Mduma, K. Kalegele, and D. Machuve, “A survey of machine learning approaches and techniques for student dropout prediction,” Data Science Journal, vol. 18, no. 1, p. 14, 2019.
8. H. Waheed, S. U. Hassan, N. R. Aljohani, J. Hardman, S. Alelyani, and R. Nawaz, “Predicting academic performance of students from VLE big data using deep learning models,” Computers in Human Behavior, vol. 104, p. 106189, 2020.
9. 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.
10. L. Breiman, “Random forests,” Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
11. T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, pp. 785–794, 2016.
12. S. M. Lundberg and S. I. Lee, “A unified approach to interpreting model predictions,” Advances in Neural Information Processing Systems, vol. 30, 2017.