International Journal of Scientific Research and Engineering Development-– Volume 9 Issue 2, Mar-Apr 2026
Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

EduGuard: An AI-Powered Dropout Early Warning System for Rural Students Using Predictive Analytics and NLP

Agesta Jenifer.A*, S. Sakthi Eswar**
*Student, Department of Computer Science, Holy Cross Engineering College
Tuticorin, Tamil Nadu, India
Email: agestajenifer@gmail.com
**Student, Department of Computer Science, Sir Issac Newton College of Engineering and Technology
Paapakovil, Nagapattinam-611102., Tamil Nadu, India
Email: sakthie631@gmail.com

----------**********----------

Abstract:

Student dropout in rural India remains a critical socio-educational crisis, with over 3.22 crore children out-of-school according to ASER 2023 estimates. This paper presents EduGuard, an AI-powered early warning system designed to predict and prevent student dropout in rural government schools using a multi-modal approach combining Predictive Analytics, Natural Language Processing (NLP), and behavioral pattern recognition. The proposed system processes attendance records, academic performance indicators, teacher-reported observations, and SMS-based student feedback through an ensemble classification model achieving a dropout prediction accuracy of 94.7%. A 5-tier risk stratification engine triggers contextual interventions ranging from automated parental alerts to direct NGO and block education officer notifications. EduGuard is deployable on low-bandwidth infrastructure, supports regional language input (Tamil, Hindi), and integrates with existing government school management systems. Pilot evaluation across 12 rural schools in Tamil Nadu demonstrated a 38% reduction in dropout incidents over a single academic year. EduGuard represents a scalable, government-ready solution to bridge the educational equity gap in rural India.

Keywords — **Student Dropout Prediction, Rural Education, Predictive Analytics, Natural Language Processing, AI Early Warning System, Educational Data Mining, Risk Stratification, Tamil Nadu Schools**

----------**********----------

I. INTRODUCTION

India's education system faces a persistent and deeply entrenched challenge: student dropout, particularly in rural areas. Despite significant policy interventions including the Right to Education (RTE) Act 2009 and the National Education Policy (NEP) 2020, millions of children discontinue schooling before completing secondary education. The Annual Status of Education Report (ASER) 2023 highlights that dropout rates remain disproportionately high among rural, tribal, and economically disadvantaged communities.

The causes of dropout are multifactorial: poverty, child labour, early marriage, lack of transportation, parental illiteracy, and language barriers collectively undermine retention. Critically, warning signals preceding dropout — such as irregular attendance, declining performance, and disengagement — are

rarely detected in time for meaningful intervention. Existing monitoring mechanisms rely on manual record-keeping and teacher intuition, both of which are insufficient in resource-constrained rural settings.

Artificial Intelligence and machine learning offer transformative potential for education systems. AI-driven predictive models can identify at-risk students weeks before formal dropout, enabling proactive, targeted interventions. This paper presents EduGuard, a comprehensive AI-powered dropout early warning system specifically designed for deployment in rural Indian government schools. The system integrates structured student data analysis with NLP-based sentiment processing of teacher and parent feedback, producing a real-time risk stratification output accessible via a lightweight web and SMS interface.

The remainder of this paper is organized as follows: Section II reviews related work; Section III describes the system architecture; Section IV details the prediction models; Section V presents experimental results; Section VI outlines future enhancements; and Section VII concludes the paper.

II. LITERATURE REVIEW

Early research in educational dropout prediction focused on statistical methods applied to enrollment data. Rumberger and Lim [1] conducted a comprehensive review establishing correlations between socioeconomic indicators and dropout likelihood in U.S. school systems. Their findings underscored the multidimensional nature of dropout causation, informing later ML-based approaches.

Pal [2] applied decision tree algorithms to predict student performance in Indian higher education, demonstrating that attendance, assignment completion, and socioeconomic status were the most predictive features. Subsequent studies by Fernandes et al. [3] employed ensemble methods — Random Forest and Gradient Boosting — achieving prediction accuracies above 85% on Portuguese secondary school datasets. Deep learning approaches have further advanced the field. Hussain et al. [4] utilized LSTM networks to model temporal attendance patterns, capturing sequential deterioration signals invisible to static classifiers. Transformer-based models[5] have been applied to educational NLP tasks, including analysis of student-written text and teacher observation logs to identify disengagement signals. Existing systems, however, share critical limitations for Indian deployment: they assume high-connectivity infrastructure, operate in English only, lack SMS-based low-tech interfaces, and are not calibrated for the specific socioeconomic and cultural context of rural Indian schools[6]. EduGuard directly addresses each of these gaps through an offline-capable, multilingual, government-integrated architecture.

III. SYSTEM ARCHITECTURE AND METHODOLOGY

A. System Overview

EduGuard is architected as a four-layer processing pipeline. Student data traverses the following sequential modules: (1) Multi-Modal Data Ingestion, (2) Feature Engineering and NLP Processing, (3) Ensemble Prediction Engine, and (4) Intervention Dispatch System. The architecture is designed for deployment on government school servers with minimum 2G network connectivity.

B. Multi-Modal Data Ingestion

EduGuard processes four categories of student data:

• **Structured Academic Data:** Daily attendance records, unit test scores, assignment submission rates, and co-curricular participation logs extracted from school management systems.

• **Socioeconomic Indicators:** Family income bracket, parental occupation, distance from school, and sibling enrollment status sourced from student registration profiles

. • **NLP-Processed Observations:** Teacher daily diary entries and parent feedback SMSes processed through a regional language NLP module to extract risk-relevant sentiment signals.

• **Behavioral Signals:** Library visit frequency, mid-day meal participation, and school event attendance as proxy indicators of engagement.

C. Feature Engineering and NLP

Module The NLP module processes Tamil and Hindi text inputs using a lightweight transliteration-aware tokenizer. Key operations include:

• **Tokenization:** Rule-based segmentation handling code-switched Tamil-English teacher diary entries.

• **Sentiment Classification:** A fine-tuned mBERT model classifies teacher observations into five sentiment categories: Concern, Neutral, Positive, Absenteeism, and Family Crisis.

• **Risk Keyword Extraction:** A curated dropout lexicon of 340 terms across Tamil, Hindi, and English flags high-risk phrases (e.g., 'needs to help family,' 'stopped coming,' 'married soon').

• **Temporal Trend Analysis**: Sliding 30-day windows compute attendance velocity and performance trajectory features.

D. Risk Stratification Engine The Risk Stratification Engine computes a composite dropout probability score in the range [0%, 100%] and maps each student to one of five risk tiers as shown in

Table 1: Risk Stratification Tiers and System Actions

Risk Tier	Score Range	Automated Action
Watch	0% – 25%	Log entry; no alert
Low Risk	26% – 45%	Teacher dashboard flag
Moderate Risk	46% – 65%	SMS to parent + counselor referral
High Risk	66% – 85%	NGO alert + home visit trigger
Critical	86% – 100%	BEO notification + emergency intervention

IV. ENSEMBLE PREDICTION MODEL

A. Model Architecture

EduGuard employs a stacked ensemble classifier combining three base learners: a Gradient Boosting Classifier (GBC) for structured tabular features, a Bidirectional LSTM (BiLSTM) for temporal attendance sequences, and a fine-tuned mBERT model for NLP-derived sentiment features. A logistic regression meta-learner combines the probabilistic outputs of all three models into a final dropout probability score.

Training was performed on a dataset of 8,420 student records spanning five academic years (2018–2023) from 24 government schools in the Tiruvannamalai and Villupuram districts of Tamil Nadu, collected in collaboration with the Tamil Nadu School Education Department. The dataset was balanced using SMOTE oversampling to address the class imbalance between dropout (18%) and non-dropout (82%) instances.

B. Feature Importance

SHAP (SHapley Additive exPlanations) analysis identified the ten most predictive features. The top five were: (1) 30-day attendance velocity, (2) consecutive absence streaks exceeding 5 days, (3) teacher NLP

concern score, (4) family income category, and (5) performance decline gradient. Together, these five features account for 71.3% of total model predictive power.

V. RESULTS AND PERFORMANCE EVALUATION

Metric	EduGuard Performance
Dropout Prediction Accuracy	94.7%
Precision (At-Risk Class)	91.3%
Recall (At-Risk Class)	93.8%
F1-Score	92.5%
Average Alert Response Time	< 4 seconds
Dropout Reduction (Pilot)	38% (vs. control schools)
System Uptime	99.6% (24/7)
False Positive Rate	4.2%
Teacher Adoption Rate	87%

Comparative analysis against conventional manual tracking systems reveals EduGuard's significant advantages. While teacher-based manual monitoring detects at-risk students on average 6–8 weeks after warning signals emerge, EduGuard identifies risk within 3–5 school days of pattern onset. Across pilot schools, 312 high-risk students were identified; 89% received timely interventions, and 76% of these students continued enrollment through the academic year — compared to a 41% retention rate in equivalent control school populations.

Deployment Contexts:

• Rural Government Schools: Primary deployment target with offline-capable sync for low-bandwidth environments.

• District Education Offices: Aggregate risk dashboards enabling block-level resource allocation.

• NGO and CSR Programmes: API access for dropout-prevention NGOs to receive targeted referrals.

• State Education Departments: Integration with UDISE+ and Samagra Shiksha portals for national-scale monitoring.

EduGuard was evaluated through a 12-month pilot deployment across 12 rural government schools in Tamil Nadu, covering 3,840 students. System performance metrics are presented in Table 2.

Table 2: EduGuard System Performance Metrics

VI. FUTURE SCOPE AND ENHANCEMENTS

The current EduGuard implementation establishes a strong foundational platform. Planned development across four phases includes:

• Phase 1 — Language Expansion: Extension of NLP support to 12 additional Indian regional languages including Telugu, Kannada, Malayalam, and Marathi using IndicBERT multilingual embeddings.

• Phase 2 — Multimodal Sensing: Integration of IoT-based school gate RFID attendance automation and voice-note teacher diary input via WhatsApp-based interfaces.

• Phase 3 — Causal Intervention Engine: Personalized AI-generated intervention recommendations tailored to each student's dropout risk causation profile, including scholarship matching and bridge course suggestions.

• Phase 4 — National Deployment: Integration with the Ministry of Education's PM-POSHAN and NIPUN Bharat frameworks for nationwide early warning coverage across 1.1 million government schools.

VII. CONCLUSION

This paper presented EduGuard, a comprehensive AI-powered student dropout early warning system combining ensemble machine learning, temporal sequence modeling, and multilingual NLP to identify at-risk rural students with 94.7% accuracy. The system addresses critical limitations of existing monitoring frameworks by providing real-time risk stratification, automated multi-channel interventions, and low-bandwidth deployability tailored for rural Indian school infrastructure. Pilot deployment across

12 Tamil Nadu schools demonstrated a 38% reduction in dropout incidents, validating EduGuard's real-world impact potential. By combining predictive intelligence with empathetic, community-aware intervention design, EduGuard represents a significant step toward achieving the Universal Elementary Education mandate of NEP 2020. Technology, when designed with an understanding of ground realities, has the power to transform educational equity — EduGuard demonstrates this potential at scale.

REFERENCES

[1] R. W. Rumberger and S. A. Lim, "Why Students Drop Out of School: A Review of 25 Years of Research," Policy Brief, UC Santa Barbara California Dropout Research Project, 2008.

[2] S. Pal, "Mining Educational Data to Reduce Dropout Rates of Engineering Students," International Journal of Information Engineering and Electronic Business, vol. 4, no. 2, pp. 1–7, 2012.

[3] E. Fernandes, M. Holanda, M. Victorino, V. Borges, R. Carvalho, and G. Van Erven, "Educational Data Mining: Predictive Analysis of Academic Performance in Public Schools," Expert Systems with Applications, vol. 135, pp. 19–30, 2019.

[4] S. Hussain, N. A. Dahan, F. A. Ba-Alawi, and N. A. Ribata, "Educational Data Mining and Analysis of Students' Academic Performance Using WEKA," Indonesian Journal of Electrical Engineering and Computer Science, vol. 9, no. 2, pp. 447–459, 2018.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv:1810.04805, 2018.

[6] Annual Status of Education Report (ASER) 2023, ASER Centre, New Delhi, India, 2023.

[7] Ministry of Education, Government of India, "National Education Policy 2020," MoE, New Delhi, 2020.

[8] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in Proc. NeurIPS, 2017.

[9] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002.

[10] UDISE+ 2022-23 Report, Ministry of Education, Government of India. Available: https://udiseplus.gov.in

