Literature Survey on Student Dropout Analysis in School Education

Psyridou et al., Scientific Reports (2024, Finland)

Outcome: Upper-secondary school dropout, measured from official educational records.

Techniques/Model: Tree-based ensembles (gradient boosting) for classification.

Dataset: Longitudinal Finnish cohort from kindergarten through Grade 9 (13 years).

Challenges: Modest AUC at early grades (≈0.61 in Grade 6); limited generalizability.

Result: AUC improved to ≈0.65 by Grade 9; early detection useful for triage, not decisions.

Christie & Jarratt, ERIC White Paper (2019, USA)

Outcome: Dropout risk scores for Grades 6–9, updated twice yearly.

Techniques/Model: Regularized logistic regression and tree-based models.

Dataset: State information systems (attendance, grades, discipline).

Challenges: Severe class imbalance, interpretability, and deployment across districts.

Result: Statewide EWS operational, delivering usable risk scores for interventions.

Lee et al., Applied Sciences (2019, Korea)

Outcome: Binary prediction of student dropout.

Techniques/Model: Random Forest, XGBoost, SVM, Logistic Regression; imbalance handled using SMOTE.

Dataset: K-12 administrative records (demographics, attendance, academic).

Challenges: Severe class imbalance, fairness concerns across subgroups.

Result: Tree ensembles outperformed linear models on F1 and AUC.

Uekawa et al., REL Mid-Atlantic (2010, Delaware, USA)

Outcome: High school dropout among Grades 9–12.

Techniques/Model: Logistic regression with cut-point analysis.

Dataset: Delaware state longitudinal student records.

Challenges: Limited scope (no socio-emotional factors), difficulty tuning thresholds.

Result: Attendance, math, and English grades were strongest predictors; simple interpretable rules.

Bulut (2024, USA HSLS:09)

Outcome: High school dropout by end of Grade 12.

Techniques/Model: Random Forest vs Deep Learning; human–machine collaboration focus.

Dataset: HSLS:09 national longitudinal cohort (survey + transcripts).

Challenges: Need for actionable features; balancing accuracy with usability.

Result: Random Forest outperformed deep learning; emphasized interpretable, actionable risk factors.

de Vasconcelos et al., Frontiers in Psychology (2023, Brazil)

Outcome: Multi-dimensional dropout risk index (relational & psychological).

Techniques/Model: Psychometric validation, factor analysis.

Dataset: Brazilian school survey samples.

Challenges: Capturing relational constructs reliably, integrating soft skills into EWS.

Result: Validated relational-risk scale complements academic indicators.

Vaarma et al., IJER (2024, Finland, higher-ed context)

Outcome: Course/program dropout (higher education).

Techniques/Model: Logistic Regression, Random Forest, Gradient Boosting.

Dataset: Demographics, transcripts, LMS activity logs.

Challenges: Temporal drift across cohorts, complexity of data fusion.

Result: ML improved predictions over baselines; data fusion critical for sustainable EWS.

Venkatesan et al., PLOS One (2023, India)

Outcome: District-level dropout hotspots (especially secondary).

Techniques/Model: Spatial autocorrelation (Moran's I, LISA).

Dataset: UDISE+ 2020 nationwide Indian data.

Challenges: Measurement errors in large datasets, regional heterogeneity.

Result: Identified dropout hotspots, guiding state/district-level interventions.

Hassan et al., Applied Sciences (2024, Somaliland)

Outcome: Student dropout rate measured from the 2022 National Education Accessibility Survey (NEAS).

Techniques/Model: Logistic Regression, Probit Regression, Naïve Bayes, Decision Tree, Random Forest, SVM, and K-Nearest Neighbors.

Dataset: NEAS 2022, ~1,957 households with demographic, educational, and socioeconomic variables.

Challenges: Limited sample size; balancing across diverse groups; interpreting results in Somaliland's context.

Result: Random Forest achieved ~95% accuracy; key predictors included student's grade, age, household income, and housing type.

Elbouknify et al., arXiv preprint (2025, Morocco)

Outcome: Identification of at-risk students across education levels in Morocco.

Techniques/Model: Advanced ML with SHAP (Shapley Additive Explanations) for interpretability.

Dataset: Moroccan Ministry of National Education administrative data (multiple grades and regions).

Challenges: Data quality issues; heterogeneity across educational settings; need for actionable interpretation.

Result: Achieved 88% accuracy, 88% recall, 86% precision, AUC 87%; SHAP identified key predictors guiding interventions.