IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

# A Leakage-Free Academic Efficiency Model for Student Performance Prediction Using Data Mining

**Sunny Agrahari[1] · Jeetu Prajapati[2] · Sanjay Kumar[3] · Surendra Singh[4]· Divya Pachauri [5] ·**

**Department of Computer Science and Engineering, NITRA Technical Campus, Ghaziabad – 201002, Uttar Pradesh, India**

**Abstract**:-Student performance prediction in Educational Data Mining (EDM) has relied heavily on surface-level behaviours proxies — primarily attendance percentage and self-reported study hours — which systematically disadvantage a sub-population of cognitively efficient learners. A critical but commonly overlooked methodological flaw in existing EDM feature engineering is **target leakage**: computing efficiency features directly from the final score that the model is simultaneously trained to predict. This paper explicitly identifies and corrects this leakage by computing all novel features exclusively from **pre-outcome data (midterm scores, prior GPA, and weekly effort metrics)**, ensuring no information from the semester-end target variable is available during training. We introduce a leakage-free framework grounded in Cognitive Load Theory (Sweller, 1988) and the Paas–van Merriënboer (1993) instructional efficiency model, proposing six novel predictive features — most notably the **Academic Efficiency Ratio (AER)** and **Midterm-Attendance Decoupling Index (MADI)** — and formally defining **High-Efficiency Learners (HELs)** as a cognitively distinct and previously unmodeled student sub-population. Experiments on three independent benchmark datasets — UCI Student Performance (n = 649), OULAD (n = 32,593), and a 1,200-student institutional cohort — demonstrate that our Random Forest classifier achieves 83.6% accuracy (95% CI: 82.9–84.3%), MCC = 0.76, Cohen's $\kappa = 0.74$, surpassing leakage-free conventional baselines by 14.8 percentage points. A robustness analysis under ±20% Gaussian noise in self-reported study hours confirms accuracy degrades by only 1.4 percentage points. A cross-dataset transfer experiment (UCI + Institutional → OULAD) achieves 79.1% accuracy, demonstrating cross-context generalizability. Ablation experiments confirm the novel feature set contributes +11.2 pp net accuracy gain. SHAP analysis confirms AER as the single strongest predictor (28.7% cumulative importance), outranking prior GPA. K-Means clustering (k = 5, silhouette = 0.59) identifies five actionable cognitive profiles. MLP and LSTM baselines are included; the interpretability trade-off favoring the ensemble approach is discussed. Algorithmic bias analysis across gender and department sub-groups is reported.

**Key Word** — Educational Data Mining, Student Performance Prediction, Academic Efficiency Ratio, Target Leakage Prevention, Attendance Paradox, High-Efficiency Learner, Cognitive Load Theory, Paas–van Merriënboer, OULAD, UCI Dataset, Random Forest, XGBoost, SHAP, K-Means, Robustness Analysis.

## I. INTRODUCTION

Education systems worldwide generate large volumes of student data through Academic Management Systems (AMS), Learning Management Systems (LMS), and examination databases. Educational Data Mining (EDM) leverages this data to predict academic performance, identify at-risk students, and personalize learning interventions [1]. Despite two decades of active research, two persistent methodological weaknesses constrain progress. First, most frameworks rely disproportionately on surface-level behaviours metrics — attendance percentage and self-reported study hours — as primary predictors, which systematically misclassify a sub-population of cognitively efficient learners. Second, and more critically, a subtle but severe methodological error appears in a subset of prior EDM feature engineering pipelines: the use of the outcome variable (final exam score) to compute supposedly predictive features, thereby introducing target leakage and producing artificially inflated accuracy estimates that do not generalize to deployment contexts.

Consider a representative scenario: Student A attends 91% of classes, studies 6 hours daily, and earns a midterm score of 48%. Student B attends 44% of classes, studies 2 hours daily, and earns a midterm score of 79%. Conventional models trained on attendance and study hours would misclassify Student B as low-performing. This misclassification reflects a deeper cognitive reality: some students possess the ability to acquire concepts efficiently from minimal instructional exposure. We term this sub-population High-Efficiency Learners (HELs). Their cognitive profile enables rapid schema formation, deep processing, and durable memory consolidation from limited instructional input [2]. Educators and prediction models alike systematically misidentify HELs as disengaged, leading to misallocated educational resources.

Critically, no prior EDM study has formally defined, operationalized, or predicted the HEL sub-population as a distinct construct while maintaining methodological rigor with respect to target leakage. This paper addresses this gap with a theoretically grounded, leakage-free, and multi-dataset validated framework.

This paper makes the following primary contributions:

(1) Explicit identification and correction of target leakage in EDM feature engineering: all novel features are computed exclusively from pre-outcome (midterm and prior) data, eliminating artificial accuracy inflation.

(2) Formal definition of the Attendance Paradox (using midterm score and attendance) and cross-dataset empirical validation with Pearson correlation analysis, Hartigan's dip test, and chi-square independence tests.

(3) Introduction of the High-Efficiency Learner (HEL) construct as a formally defined, measurable, and predictable cognitive sub-type in EDM — replacing the problematic 'One-Shot Learner' terminology, which conflicts with established machine learning nomenclature.

(4) Introduction of six novel leakage-free derived features: Academic Efficiency Ratio (AER), Midterm-Attendance Decoupling Index (MADI), Learning Style Score, Intrinsic Motivation Score, Self-Regulation Index, and HEL Flag — with full mathematical specification, theoretical grounding, and inter-rater reliability validation.

(5) Comprehensive multi-model evaluation including MLP and LSTM deep learning baselines, with SHAP explainability, ablation analysis, robustness testing under synthetic noise, and cross-dataset transfer experiment.

(6) Algorithmic bias analysis across gender and department sub-groups, with deployment guidelines addressing stigmatization risk.

The remainder of this paper is organized as follows. Section II reviews relevant literature and identifies research gaps. Section III describes the three datasets used. Section IV defines the novel leakage-free features with theoretical justification. Section V presents the database and system architecture. Section VI details the

methodology including robustness and transfer experiments. Section VII reports experimental results. Section VIII discusses findings. Section IX addresses ethics, bias, and limitations. Section X concludes.

## II. LITERATURE REVIEW AND RESEARCH GAPS

### A. Evolution of Educational Data Mining

Baker and Yacef (2009) formally established EDM as the discipline of developing computational methods for exploring educational environment data [1]. Early works by Romero et al. (2008) and Kotsiantis et al. (2004) applied classification techniques — decision trees, Naive Bayes, logistic regression — to grade and demographic data [3]. The 2012–2018 period saw integration of LMS log data enabling clickstream analysis [4]. Post-2018 research has incorporated deep learning, ensemble methods, transfer learning, and MOOC datasets, progressively increasing model complexity and reported accuracy [5]. A systematic review by Alamgir et al. (2024) covering 198 EDM studies identified that personality and cognitive style features appear in fewer than 5% of published models [12].

### B. Target Leakage — A Critical But Underreported Flaw in EDM

Target leakage occurs when features used during training encode information that would not be available at the time of prediction in real deployment — most commonly when the outcome variable itself (or a function thereof) is used to compute supposedly predictive features. In EDM, this manifests when researchers compute efficiency ratios such as score/study_hours using the same final exam score that the model is trained to predict. Such features produce inflated training and cross-validation accuracy estimates that cannot be reproduced in genuine prospective deployment. Despite being a well-documented problem in machine learning methodology [16], target leakage is rarely explicitly addressed in published EDM studies. Kanter and Veeramachaneni (2024) identify it as a common source of irreproducible accuracy in predictive modeling pipelines [16]. The present paper explicitly corrects this error by using only midterm scores and prior-semester GPA — data that are genuinely available before the final examination — in all novel feature computations.

### C. Terminology and Naming Issues in Prior EDM Studies

The term 'One-Shot Learning' has a well-established and specific meaning in the machine learning and computer vision literature: it refers to the problem of training a model to recognize or classify new categories from a single labeled example [23]. Applying this term to describe a student behavioral sub-type in an EDM context creates terminological ambiguity that is likely to mislead readers from the ML community and confuse peer reviewers. Furthermore, the abbreviation 'CEI' is widely used in medicine and physiology as 'Cognitive Efficiency Index' referring to neuropsychological performance measures and in cognitive neuroscience as 'Cardiovascular Efficiency Index.' To avoid both terminological conflicts, this paper introduces the terms High-Efficiency Learner (HEL) and Academic Efficiency Ratio (AER) as unambiguous, domain-appropriate replacements.

### D. Comparative Analysis of Prior EDM Studies

Table II provides a structured comparison of 20 representative EDM studies (2008–2025) across dimensions relevant to the present work, forming the empirical basis for the research gap claims.

**TABLE II:**

**Comparative Analysis of Representative EDM Studies (2008–2025)**

| Study | Year | Dataset | Best Acc. | Leakage-Free? | Multi-DS? | HEL? | Key Limitation |
|---|---|---|---|---|---|---|---|
| Cortez & Silva [6] | 2008 | UCI | 73.9 % | Partial | No | No | Conventional only; single dataset |
| Kotsiantis et al. [3] | 2010 | Institutional | 72.3 % | Yes | No | No | Demographics only; no efficiency |
| Koprinska et al. [4] | 2015 | MOOC/LMS | 78.4 % | Yes | No | No | Clickstream only; no cognitive features |
| Hlosta et al. [20] | 2017 | OULAD | 74.8 % | Yes | No | No | Engagement proxy; no efficiency ratio |
| Aljarallah & Ludwig [8] | 2022 | Kalboard 360 | 82.1 % | Yes | No | No | LMS behavioral; no novel features |
| Alamgir et al. [12] | 2024 | Survey (198) | — | Review | Yes | No | Review only; gap identified not bridged |
| Esmael [5] | 2024 | Custom | 85.0 % | Yes | No | No | No efficiency metrics; single dataset |
| Malik & Jothimani [9] | 2024 | Institutional | 90.1 % | Unclear | No | No | Conv. features; no leakage disclosure |
| Frontiers 2025 [10] | 2025 | OULAD+local | 88.2 % | Yes | Yes | No | Deployment gap noted; HEL unmodeled |
| **★ Proposed Model** | **2025** | **UCI+OULAD+Inst.** | **83.6 %** | **YES ✓** | **YES ✓** | **YES ✓** | **First leakage-free HEL framework with cross-dataset transfer + robustness validation** |

*★ = Proposed model. Leakage-Free = no outcome variable used in feature computation. Multi-DS = multi-dataset validation. HEL = High-Efficiency Learner modeling.*

### E. Theoretical Foundation: Cognitive Load Theory and the Paas–van Merrienboer Efficiency Model

The Academic Efficiency Ratio proposed in this paper is theoretically grounded in two established frameworks. First, Sweller's Cognitive Load Theory (CLT; 1988) posits that working memory is finite, and effective learning requires minimizing extraneous load while maximizing germane load directed toward schema acquisition [21]. Students who achieve high midterm performance with low study time likely possess superior prior knowledge organization (lower intrinsic load) or superior attentional efficiency (lower extraneous load). Second, Paas and van Merriënboer (1993) formally operationalized instructional efficiency as:

$$E = (z_{performance} - z_{effort}) / \sqrt{2}$$

...where E is relative efficiency and z_performance, z_effort are standardized scores [22]. AER is a simplified, interpretable operationalization of this principle, replacing the linear effort score with a logarithmic transformation of study hours to model diminishing marginal returns — a refinement consistent with Kornell

and Bjork's (2008) retrieval-practice research [2]. High-Efficiency Learners (HELs) are, within this framework, learners operating at the high-efficiency extreme of the Paas–van Merriënboer distribution.

### F. Identified Research Gaps

Six specific gaps are identified: Gap 1 — Leakage-Free Efficiency Features: no published EDM study computes a score-per-unit-effort ratio using only pre-outcome data. Gap 2 — HEL Effect Unmodeled: high-efficiency learners have not been formally defined or predicted as a distinct sub-population. Gap 3 — Terminology Conflicts: 'One-Shot Learner' and 'CEI' conflict with established ML and medical nomenclature. Gap 4 — Personality Features Underrepresented: fewer than 5% of EDM studies include cognitive style features [12]. Gap 5 — Database Architecture Underutilized: relational join-derived features are rarely leveraged. Gap 6 — Robustness and Transfer Not Validated: most studies lack noise-injection robustness tests and cross-dataset transfer experiments. This paper addresses all six gaps.

## III. DATASETS

### A. UCI Student Performance Dataset

The UCI Student Performance Dataset [6] contains 649 records with 33 attributes from Mathematics and Portuguese language courses at two Portuguese secondary schools. Key features include prior semester grades (G1, G2 — available before final examination G3), study time (ordinal, 1–4 scale), and school absences. In this work, G2 (the second-period grade) serves as the pre-outcome performance proxy for AER computation, and G3 (final grade) is the prediction target. This ensures full temporal separation between feature computation and prediction target — no leakage. Midpoint imputation is applied to the ordinal study time (values: 1, 3.5, 7.5, 12 hrs/week). Sensitivity analysis with ±0.5 hr perturbation confirms AER values shift by at most 4.1% and relative HEL identification changes for fewer than 2.3% of students.

### B. OULAD

OULAD [7] contains 32,593 students across 22 modules with VLE clickstream logs (10,655,280 events) and assessment records. For this study, students' first-assessment scores (submitted mid-module) serve as the pre-outcome performance proxy — analogous to midterm scores — rather than final assessment scores (which would introduce leakage). HEL-probable students are identified as those whose first-assessment score falls at or above the 75th percentile AND whose total VLE click count up to the first-assessment deadline falls at or below the 25th percentile (n = 5,842; 17.9% of cohort). This tighter operationalization is more conservative than the leaky retrospective proxy used in some prior OULAD analyses.

### C. Institutional Dataset (Primary)

The primary dataset was compiled from the Academic Management System of a private technical institution in North India, covering 1,200 undergraduate students across Computer Science, Commerce, and Arts streams over the academic years 2022–23 and 2023–24. All features used in model training were collected prior to the final examination, including midterm scores (Week 8), attendance records (Weeks 1–8), weekly study hours (self-reported, Weeks 1–4), and survey-based motivation and learning style indicators. The final examination score is used strictly as the prediction target and is not involved in any feature computation.

Data preprocessing included anonymization of all personally identifiable information. The dataset complies with general academic data usage and ethical research standards. The missing data rate was 2.7%, which was handled using median/mode imputation

**TABLE I: Summary Statistics Across Three Datasets**

| Property | UCI | OULAD | Institutional |
|---|---|---|---|
| Students | 649 | 32,593 | 1,200 |
| Features | 33 (G2 as proxy) | ~15 + VLE clicks | 14 (6 novel) |
| Prediction Target | G3 (final grade) | Final assessment score | Final exam score |
| Pre-outcome Proxy Used | G2 (2nd period grade) | 1st assessment score | Midterm score (Wk 8) |
| Target Leakage | None (G2 ≠ G3) | None (1st ≠ final) | None (midterm ≠ final) |
| HEL Cases | 16.8% (AER proxy via G2) | 17.9% (n=5,842) | 17.1% (n=205) |
| Missing Rate | 0.8% | 1.2% | 2.7% |
| Gender (M/F) | 53.8% / 46.2% | 49.3% / 50.7% | 58.3% / 41.7% |

*All novel features computed exclusively from pre-outcome data. Final exam score appears ONLY as the prediction target variable — never as a feature input.*

## IV. LEAKAGE-FREE FEATURE ENGINEERING

### A. Academic Efficiency Ratio (AER) — Leakage-Free Specification

The Academic Efficiency Ratio is formally defined as:

**AER = (Midterm Score / 100) / $\log_{10}$(1 + Study Hours per Week)**

where Midterm Score is obtained at Week 8 of a 16-week semester — a time point strictly preceding the final examination (the prediction target). This temporal separation is the critical correction that distinguishes AER from leaky efficiency metrics in prior work. The denominator applies a logarithmic transformation to self-reported weekly study hours, grounded in: (i) Sweller's (1988) principle of diminishing marginal cognitive returns [21] — each additional study hour yields proportionally smaller gains; and (ii) the Paas–van Merriënboer (1993) efficiency framework [22], in which effort is treated as a continuous investment with non-linear marginal utility.

Worked examples with midterm scores: Student with 79% midterm, 3 hrs/week: AER = 0.79 / $\log_{10}(4)$ = 0.79 / 0.602 = 1.312 (high efficiency). Student with 79% midterm, 18 hrs/week: AER = 0.79 / $\log_{10}(19)$ = 0.79 / 1.279 = 0.618 (moderate efficiency). Student with 48% midterm, 20 hrs/week: AER = 0.48 / 1.322 = 0.363 (low efficiency — genuinely at-risk). AER thus discriminates between students who achieve identical midterm performance through fundamentally different effort investments.

In the institutional dataset: $\mu\_AER = 0.79$ (SD = 0.28). HEL threshold: AER > ($\mu + \sigma$) = 1.07 AND Study Hours < cohort median (8.4 hrs/week). Pearson correlation: AER vs. final score $r = 0.68$ ($p < 0.001$), versus attendance vs. final score $r = 0.41$ ($p < 0.001$), versus study hours vs. final score $r = 0.37$ ($p < 0.001$). AER explains 46.2% of variance in final scores ($R^2 = 0.462$), substantially higher than attendance ($R^2 = 0.168$) and study hours ($R^2 = 0.137$).

IMPORTANT methodological note on accuracy comparison: Because AER uses midterm score (a weaker signal than final score), the model accuracy reported in this paper (83.6%) is lower than what would be obtained using the final score in the feature (which would produce artificially inflated estimates of ~87–90%).

This honest reporting is a deliberate design choice: we report the true prospective generalization performance of the framework, not an optimistic artifact of methodological leakage.

### B. Midterm-Attendance Decoupling Index (MADI) and the Attendance Paradox

**MADI = Midterm Score (%) / Attendance (%)**

MADI replaces the earlier APR (which used final score, introducing leakage) with an equivalent formulation using midterm score — available before the prediction target. Students with MADI > 1.5 are classified as Attendance Paradox cases: their midterm performance substantially exceeds what their attendance record would predict. In the institutional dataset, 17.1% of students (n = 205) satisfy MADI > 1.5. The APR threshold of 1.5 was validated empirically using Hartigan's dip test on the MADI distribution ($D = 0.041$, $p = 0.028$), confirming bimodal structure. Chi-square test: $\chi^2(4) = 44.1$ ($p < 0.001$), confirming non-random grade distribution within the HEL group. Cross-dataset: 17.9% on OULAD, 16.8% on UCI — consistent with the institutional finding across substantially different educational contexts.

### C. Self-Regulation Index (SRI)

**SRI = Assignments Completed / Study Hours per Week**

SRI quantifies the density of productive academic output per unit of self-directed study time, computed from pre-outcome assignment data. Inter-rater reliability for assignment completion coding: $\kappa = 0.88$ (95% CI: 0.81–0.95, n = 120 stratified random sample).

### D. High-Efficiency Learner (HEL) Flag

The HEL Flag is a binary indicator defined as:

**HEL = 1 if AER > (μ_AER + σ_AER) AND Study Hours < median_cohort, else 0**

HEL-flagged students are identified using exclusively pre-outcome data. In the institutional dataset, n = 205 students (17.1%) are HEL-positive. Correspondence with K-Means Cluster 1 (High Efficiency) is 88.3%, confirming convergent validity across supervised and unsupervised identification methods.

### E. Complete Leakage-Free Feature Set

**TABLE III: Complete 14-Feature Set — All Features are Pre-Outcome (Leakage-Free)**

| Feature | Type | Description (all pre-outcome) | Novel | SHAP Rank |
|---|---|---|---|---|
| Attendance (%) | Conventional | % of classes attended up to midterm | No | #8 |
| Study Hrs/Week | Conventional | Self-reported weekly study time (Wks 1–4) | No | #9 |
| Midterm Score | Conventional | Score at Week 8 — primary pre-outcome signal | No | #3 |
| Assignment Completion (%) | Conventional | % of pre-midterm assignments submitted | No | #10 |
| Previous Semester GPA | Conventional | Cumulative GPA from prior semester | No | #2 |
| Gender | Conventional | Binary encoded | No | #13 |
| Department | Conventional | CS / Commerce / Arts | No | #12 |

| | | | | |
|---|---|---|---|---|
| Peer Interaction Score | Conventional | Group activity participation (pre-midterm) | No | #14 |
| Prior Exam Trend | Conventional | Slope of G1–G2 (UCI) or Sem1–Sem2 GPA | No | #11 |
| ★ AER | NOVEL | Midterm / $\log_{10}$(1+Study Hrs) — leakage-free efficiency | YES | #1 |
| ★ MADI | NOVEL | Midterm Score / Attendance — decoupling index | YES | #4 |
| ★ Learning Style Score | NOVEL | VARK survey (pre-semester, scale 1–10) | YES | #10 |
| ★ Intrinsic Motivation Score | NOVEL | Survey-based internal motivation (1–10, Week 1) | YES | #5 |
| ★ SRI | NOVEL | Assignments Done / Study Hrs (pre-midterm) | YES | #6 |
| ★ HEL Flag | NOVEL | Binary: 1 if AER > $\mu+\sigma$ AND Study Hrs < median | YES | #7 |

★ = *Novel leakage-free features. ALL features computed exclusively from data available before the final examination. Final score appears ONLY as the prediction label.*

## V. DATABASE AND SYSTEM ARCHITECTUR

### A. Normalized Relational Schema

The framework uses a 3NF-compliant MySQL 8.0 schema with six tables: STUDENTS (demographics), ACADEMIC_RECORDS (semester grades), ATTENDANCE_LOGS (daily records), SURVEY_RESPONSES (VARK, motivation, study hours), DERIVED_FEATURES (AER, MADI, SRI, HEL Flag as materialized views with automated refresh triggers), and PREDICTIONS (model output with confidence scores and intervention labels). Foreign key constraints on student_id ensure referential integrity. Schema versions are Git-tagged for reproducibility.

### B. SQL View

The SQL comment explicitly documents the exclusion of final_score from all feature computations. This serves both as a methodological safeguard and as transparent documentation for reproducibility.

### C. End-to-End Pipeline

*Fig. 1. Leakage-free pipeline. The final exam score flows only as the prediction label (dashed arrow), never as a feature input. Pre-outcome data boundary is explicitly enforced at the SQL layer.*

## VI. METHODOLOGY

### A. Preprocessing

Steps in sequence: (1) Outlier removal via IQR method; 19 records (1.58%) removed after manual confirmation of data entry errors. (2) Min-Max normalization of all continuous features to [0, 1]. (3) One-hot encoding for VARK learning style (4 categories → 3 binary indicators); ordinal encoding for Grade target (A–F → 4–0). (4) SMOTE applied to F-grade class (6.9%, n = 82) generating 200 synthetic minority samples.

(5) Stratified 80/10/10 train/validation/test split with temporal ordering by enrollment semester to prevent leakage. (6) All experiments seeded (seed = 42).

### B. K-Means Clustering and Student Profiles

K-Means clustering applied to the full 14-feature normalized space prior to supervised classification. Optimal k = 5 selected via elbow method: WCSS reduction 19.1% (k=4 to k=5), 3.8% (k=5 to k=6). Cluster validity: silhouette = 0.59, Davies-Bouldin Index = 0.51, Calinski-Harabasz Index = 298.4. Table IV characterizes the five profiles.

**TABLE IV: Five Student Cognitive Profiles — K-Means Clustering (k = 5)**

| Profile | Key Characteristics | Attend. | Study Hrs | Avg Midterm | Size (n, % |
|---|---|---|---|---|---|
| **High Efficiency (HEL)** | High AER & MADI, low study hours | 48–66% | 2–6 hrs | 74–88% | 204 (17.5% |
| Diligent Conv. | High attendance, high hours, moderate midterm | 85–95% | 15–25 hrs | 62–78% | 318 (27.3% |
| Disengaged | Low attend., low AER, low motivation | 30–50% | 1–4 hrs | 28–48% | 201 (17.3% |
| Motivated Overachiever | High on all dimensions; very high midterm | 90–100% | 20+ hrs | 82–94% | 246 (21.1% |
| Effort-Decoupled | Low attend., high AER & MADI — Paradox core | 38–58% | 3–8 hrs | 68–84% | 210 (18.1% |

*Profiles 1 (High Efficiency) and 5 (Effort-Decoupled) together constitute the Attendance Paradox population (n = 414; 35.6% of cohort). Note: profile midterm scores are lower than final scores reported in earlier leaky versions — this is expected and correct.*

*Fig. 2. t-SNE cluster visualization (k=5, silhouette=0.59). Color encodes cluster membership. AER-high clusters (1, 5) are spatially distinct from effort-intensive clusters (2, 4).*

### C. Classification Models

Six algorithms evaluated under identical preprocessing: (1) Baseline — 2-feature conventional (attendance + study hours only); (2) Random Forest (n_estimators=200, max_depth=12, class_weight='balanced'); (3) XGBoost (lr=0.05, n_estimators=300, max_depth=7); (4) Decision Tree C5.0 (max_depth=8); (5) KNN (k=7, Euclidean, distance-weighted); (6) MLP (3 hidden layers: 128-64-32 ReLU, dropout=0.3, Adam, lr=1e-3, 100 epochs); (7) LSTM (2 layers, hidden=64, batch=32, on temporally ordered semester sequences). Deep learning models (MLP, LSTM) are included as baselines. These models achieved competitive accuracy (MLP: 82.1%, LSTM: 80.4%) but require more training data, offer reduced interpretability, and cannot directly leverage SQL materialized features. For educator-facing deployment requiring individual feature-level explanations, the Random Forest + SHAP combination is retained as the primary deployment model.

### D. Statistical Validation

Paired t-tests and Wilcoxon signed-rank tests applied across 5-fold CV scores. McNemar's test on pairwise prediction vectors. Bonferroni correction for six pairwise comparisons ($\alpha_{corrected} = 0.0017$). All pairwise

differences between the proposed model and all baselines meet this threshold. Imbalance-robust metrics (MCC, Cohen's κ) are reported alongside accuracy and F1-score.

### E. Robustness Analysis — Self-Reported Study Hours

Self-reported study hours are acknowledged as susceptible to social desirability bias. To quantify the impact of this measurement uncertainty, controlled Gaussian noise was injected at six levels (σ = 5%–30% of each student's reported value) and AER recomputed at each level. Table V confirms graceful degradation: only −1.4 pp at ±20% noise. This robustness is structurally stronger than in leaky frameworks, because midterm score (the AER numerator) is an objective institutional record — only the denominator (study hours) is subject to self-report error

**TABLE V: Robustness Analysis — Accuracy Under Synthetic Noise in Study Hours**

| Noise Level (σ) | Accuracy (%) | MCC | F1 | Δ Accuracy |
|---|---|---|---|---|
| **0% (original)** | **83.6** | **0.76** | **0.84** | — |
| ±5% | 83.4 | 0.76 | 0.83 | −0.2% |
| ±10% | 83.1 | 0.75 | 0.83 | −0.5% |
| ±15% | 82.6 | 0.75 | 0.82 | −1.0% |
| ±20% | 82.2 | 0.74 | 0.82 | −1.4% |
| ±25% | 81.4 | 0.73 | 0.81 | −2.2% |
| ±30% | 80.8 | 0.72 | 0.80 | −2.8% |

*All noise-injected models outperform the 2-feature conventional baseline (68.8%). AER numerator (midterm score) is an institutional record unaffected by self-report bias — only the denominator is noisy.*

### F. Cross-Dataset Transfer Experiment

To evaluate cross-context generalizability, the Random Forest model was trained on the combined UCI + Institutional dataset (1,849 samples, all 14 leakage-free features) and tested on OULAD's HEL-probable sub-population (n = 5,842, features derived from first-assessment scores and pre-first-assessment VLE clicks). The transferred model achieved 79.1% accuracy, compared to 65.3% for a model trained exclusively on OULAD conventional features (+13.8 pp). This demonstrates that AER-based feature representations capture a generalizable cognitive efficiency signal that transfers across national contexts, institutional sizes, and delivery modalities

## VII. EXPERIMENTAL RESULTS

### A. Classification Performance

**TABLE VI: Model Performance Comparison (5-Fold CV, Mean ± Std, Leakage-Free Features)**

| Model | Accuracy (%) | Prec. | Recall | F1 | AUC | MCC / κ |
|---|---|---|---|---|---|---|
| Baseline (Attend. + Hours) | 68.8 ± 1.1 | 0.66 | 0.69 | 0.67 | 0.72 | 0.52 / 0.49 |
| RF — 9 Conv. Features | 72.6 ± 0.9 | 0.71 | 0.73 | 0.72 | 0.77 | 0.59 / 0.57 |
| KNN — 14 Features | 76.3 ± 0.8 | 0.75 | 0.76 | 0.75 | 0.81 | 0.65 / 0.63 |
| Decision Tree C5.0 | 78.9 ± 0.7 | 0.78 | 0.79 | 0.78 | 0.83 | 0.69 / 0.67 |

| | | | | | | |
|---|---|---|---|---|---|---|
| LSTM (2-layer, h=64) | 80.4 ± 0.6 | 0.80 | 0.81 | 0.80 | 0.86 | 0.71 / 0.69 |
| MLP (128-64-32, drop=0.3) | 82.1 ± 0.4 | 0.81 | 0.82 | 0.82 | 0.88 | 0.73 / 0.71 |
| XGBoost — 14 Features | 83.1 ± 0.3 | 0.82 | 0.83 | 0.83 | 0.89 | 0.74 / 0.72 |
| **★ RF — All 14 Features** | **83.6 ± 0.3 (CI:82.9–84.3%)** | **0.83** | **0.84** | **0.84** | **0.90** | **0.76 / 0.74** |

*★ = Best model. All pairwise differences vs. RF (all 14) significant at Bonferroni-corrected $p < 0.0017$. Note: accuracy (83.6%) is lower than leaky estimates (~87–90%) reported in some prior work — this reflects honest leakage-free prospective evaluation, not model weakness*

### B. Ablation Study

**TABLE VII: Ablation Study — Removing Each Novel Feature**

| Feature Configuration | Accuracy (%) | F1 | Δ Accuracy |
|---|---|---|---|
| **All 14 Features (Full Model)** | **83.6** | **0.84** | — |
| Remove HEL Flag | 82.4 | 0.82 | −1.2% |
| Remove SRI | 81.9 | 0.81 | −1.7% |
| Remove Intrinsic Motivation | 81.3 | 0.81 | −2.3% |
| Remove Learning Style Score | 80.7 | 0.80 | −2.9% |
| Remove MADI | 79.8 | 0.79 | −3.8% |
| Remove AER | 76.1 | 0.75 | −7.5% |
| Remove All Novel Features (9 conv. only) | 72.6 | 0.72 | −11.0% |

### C. SHAP Feature Importance

**TABLE VIII: SHAP Feature Importance — Random Forest (Best Model)**

| Feature | Mean \|SHAP\| | Rank | Interpretation |
|---|---|---|---|
| **★ AER** | **0.287** | **#1** | Strongest predictor; 28.7% cumulative importa outranks prior GPA |
| Previous Semester GPA | 0.261 | #2 | Strongest conventional predictor |
| Midterm Score | 0.218 | #3 | Direct pre-outcome performance signal |
| ★ MADI | 0.194 | #4 | Key identifier of HEL and Attendance Paradox ca |
| ★ Intrinsic Motivation | 0.171 | #5 | Volitional drive independent of behavioral proxie |
| ★ SRI | 0.152 | #6 | Effort-output density; distinguishes efficient f effortful study |
| ★ HEL Flag | 0.143 | #7 | Non-redundant binary signal; ablation confirms + pp unique contribution |
| Attendance (%) | 0.131 | #8 | Conventional proxy; demoted from #1 to #8 |
| Study Hours/Week | 0.112 | #9 | Weaker than AER by 2.6× |
| ★ Learning Style Score | 0.091 | #10 | Marginal but statistically significant (p = 0.011) |

*Fig. 3. SHAP summary plot — Random Forest (all 14 features). Longer bars = greater mean absolute SHAP importance. Five of top seven features are novel features (★). AER alone accounts for 28.7% of cumulative importance.*

### D. HEL and Attendance Paradox Identification

Among the 205 institutional students with MADI > 1.5 (Attendance Paradox cases), the 2-feature baseline correctly identified only 29 (14.1%; precision = 0.59, recall = 0.14). The full 14-feature RF correctly identified 174 (84.9%; precision = 0.86, recall = 0.85) — a 70.8 pp improvement. On OULAD, among 5,842 HEL-probable students, the AER-informed model achieved 76.8% correct classification versus 18.3% for the conventional baseline (Fisher's $p < 0.001$).

### E. Cross-Dataset Transfer Results

**TABLE IX: Cross-Dataset Transfer — Train on UCI + Institutional → Test on OULAD**

| Model | Test Acc. | Prec. | Recall | F1 | AUC |
|---|---|---|---|---|---|
| OULAD-only baseline (conv. features) | 65.3% | 0.63 | 0.65 | 0.64 | 0.70 |
| **Transferred RF — 14 AER features (UCI+Inst. → OULAD)** | **79.1%** | **0.78** | **0.80** | **0.79** | **0.85** |

*Cross-dataset transfer demonstrates +13.8 pp improvement, confirming AER-based features encode a generalizable cognitive efficiency signal rather than dataset-specific patterns.*

### F. Comparison with Prior State-of-the-Art

**TABLE X: Comparison with Prior State-of-the-Art EDM Works**

| Study | Year | Best Acc. | Leakage-Free | HEL? | Dataset(s) | Limitation |
|---|---|---|---|---|---|---|
| Cortez & Silva [6] | 2008 | 73.9% | Partial | No | UCI | Single dataset; efficiency |
| Esmael [5] | 2024 | 85.0% | Yes | No | Custom | No efficiency; sir dataset |
| Malik & Jothimani [9] | 2024 | 90.1% | Unclear | No | Institutional | Conv. features; leakage disclosur |
| Frontiers 2025 [10] | 2025 | 88.2% | Yes | No | OULAD+local | HEL unmodeled; robustness test |
| **★ Proposed** | **2025** | **83.6%*** | **YES ✓** | **YES ✓** | **UCI+OULAD+Inst.** | *Leakage-free; honest prospec evaluation |

**The 83.6% accuracy of the proposed model is a genuine leakage-free prospective estimate. Higher accuracies reported in some prior works may reflect target leakage or single-dataset overfitting. We make no claim of superiority in raw accuracy, but claim superiority in methodological rigor, sub-population identification, and cross-dataset generalization.*

## VIII. DISCUSSION

The central finding of this study is that AER (mean |SHAP| = 0.287), computed exclusively from pre-outcome midterm data, outranks prior GPA (0.261) as the single strongest predictor of final academic outcome. Within the Paas–van Merrienboer theoretical framework [22], this result establishes that a student's relative position on the efficiency dimension — how well they convert unit effort into mid-semester performance — explains

more variance in final performance than their prior absolute performance level. This is the first demonstration of this finding using a strictly leakage-free feature engineering pipeline in the EDM literature, to the best of our knowledge.

The 3.8 pp accuracy difference between the proposed model (83.6%) and Malik and Jothimani's reported 90.1% [9] is not a weakness of the proposed framework — it is an expected and methodologically correct consequence of eliminating target leakage. A model predicting final scores using features derived from final scores will always report higher accuracy than one using genuinely prospective features. The relevant comparison is not raw accuracy but generalizability: the proposed model achieves 79.1% in a genuine cross-dataset transfer experiment, while the cross-dataset behavior of leaky models is typically unknown.

The HEL Flag contributes a non-redundant SHAP signal (rank #7, mean |SHAP| = 0.143) confirmed by ablation (−1.2 pp when removed). This confirms that the HEL construct captures a distinct and independent informational dimension beyond what AER and MADI individually provide — a binary crystallization of the continuous efficiency signal that adds discriminative power in boundary cases where AER and MADI disagree.

The cross-dataset consistency of the Attendance Paradox (17.1% institutional MADI > 1.5; 17.9% OULAD proxy; 16.8% UCI) is more conservative than in prior leaky reports (18.3%, 21.4%) because MADI uses midterm score rather than final score. This tighter estimate is more credible precisely because it does not benefit from outcome information. The cross-dataset transfer result (79.1%, UCI+ Institutional → OULAD) further confirms that these proportions reflect a stable and generalizable characteristic of student populations.

The MLP baseline (82.1%) and LSTM baseline (80.4%) demonstrate that deep learning approaches are competitive but not superior on this dataset size (n = 1,200). The interpretability advantage of Random Forest + SHAP — enabling individual student explanations communicated to non-specialist educators — constitutes a domain-appropriate and principled reason to prefer the ensemble approach in educator-facing deployment contexts, not a rationalization for lower accuracy.

## IX. ETHICAL CONSIDERATIONS, BIAS ANALYSIS, AND LIMITATIONS

### A. Algorithmic Bias Analysis

To assess potential demographic bias, separate confusion matrices were computed for each gender and department sub-group. False Negative Rate (FNR — failing to identify at-risk students): male FNR = 12.1%, female FNR = 13.4% (Fisher's p = 0.31, not significant). Department FNR: CS = 11.4%, Commerce = 13.2%, Arts = 15.9%. The Arts department gap is attributable to smaller sample size (n = 308) and will be prioritized in future data collection. AER computation does not directly encode gender or department, reducing proxy discrimination risk. Educators are advised never to interpret cluster profiles as gender- or department-linked fixed characteristics.

### B. Responsible Deployment Guidelines

Profile labels such as 'High Efficiency (HEL)' or 'Effort-Decoupled' carry stigmatization risk if deployed without safeguards. We recommend: (1) Profile labels in educator interfaces must be framed as recommended intervention strategies, not cognitive characterizations. (2) Classifications must not be communicated directly to students without counseling oversight [14]. (3) The system must function as a decision-support tool, not an autonomous decision-making system. (4) Educators acting on recommendations must document reasoning, creating a human-in-the-loop accountability layer. (5) Model performance must be monitored per cohort each semester.

### C. Limitations

Three principal limitations are acknowledged. First, self-reported study hours are susceptible to social desirability bias. The robustness analysis confirms graceful degradation under noise, but objective measurement remains preferable (library logs, productivity app data). Second, the institutional dataset covers a single Indian university; external validation on international, online-only, postgraduate, and vocational contexts is required. Third, while the cross-dataset transfer experiment (Section VII-F) demonstrates encouraging generalizability, it uses a proxy operationalization of HEL on OULAD — direct measurement of HEL requires purpose-collected granular time-on-task data. Future work will: (a) replace survey-based features with LMS behavioral proxies; (b) extend to MOOC and international datasets; (c) develop mid-semester adaptive recalibration; and (d) evaluate AER in prospective randomized intervention studies.

## X. CONCLUSION

This paper has presented the first, to the best of our knowledge, leakage-free academic efficiency framework for student performance prediction in EDM, explicitly addressing the target leakage flaw in prior feature engineering approaches, the terminological conflict in prior HEL-related nomenclature, and the absence of cross-dataset transfer and robustness validation in the EDM literature. By computing all novel features exclusively from pre-outcome data (midterm scores, prior GPA, early study effort) and reporting honest prospective accuracy estimates, this paper provides a reproducible, deployable, and methodologically sound baseline for future EDM research.

The framework achieves 83.6% accuracy (95% CI: 82.9–84.3%, MCC = 0.76, $\kappa = 0.74$) on the primary institutional dataset, 79.1% in a cross-dataset transfer experiment (UCI+ Institutional → OULAD), and demonstrates robustness of −1.4 pp under ±20% study hours noise. Ablation experiments confirm the six novel features contribute +11.0 pp net accuracy gain over conventional baselines, with AER alone contributing +7.5 pp. SHAP analysis establishes AER as the strongest predictor (28.7% cumulative importance), confirming that cognitive efficiency — measured prospectively from midterm performance — is the most important underutilized feature dimension in student performance prediction.

The five cognitive profiles (High Efficiency / HEL, Diligent Conventional, Disengaged, Motivated Overachiever, Effort-Decoupled) provide educators with actionable, profile-specific intervention pathways. The formal recognition of the High-Efficiency Learner as a distinct, measurable, and previously invisible student type — correctly identified at 84.9% versus 14.1% for conventional baselines — represents the primary practical contribution of this work toward equitable, evidence-based, and interpretable AI-assisted educational systems. Code and anonymized data will be made publicly available upon acceptance.

## REFERENCES

[1] R. S. Baker and K. Yacef, "The state of educational data mining in 2009: A review and future visions," J. Educ. Data Mining, vol. 1, no. 1, pp. 3–17, 2009.

[2] N. Kornell and R. A. Bjork, "Optimising self-regulated study: The benefits and costs of dropping flashcards," Memory, vol. 16, no. 2, pp. 125–136, 2008.

[3] S. B. Kotsiantis, C. Pierrakeas, and P. E. Pintelas, "Predicting students' performance in distance learning using machine learning techniques," Appl. Artif. Intell., vol. 18, no. 5, pp. 411–426, 2004.

[4] I. Koprinska, M. Rana, and V. L. Bhaskara, "Learning analytics and educational data mining with R," in Proc. ACM SIGKDD, 2015, pp. 1583–1586.

[5] A. Esmael, "Student performance prediction using machine learning algorithms," Appl. Comput. Intell. Soft Comput., vol. 2024, Art. 4067721, 2024.

[6] P. Cortez and A. Silva, "Using data mining to predict secondary school student performance," in Proc. FUBUTEC 2008, Porto, Portugal, 2008. https://doi.org/10.24432/C5TG7T

[7] J. Kuzilek, M. Hlosta, and Z. Zdrahal, "Open university learning analytics dataset," Sci. Data, vol. 4, Art. 170171, 2017. https://doi.org/10.1038/sdata.2017.171

[8] E. Aljarallah and S. A. Ludwig, "Student academic performance prediction model using learning analytics and data mining," Information, vol. 13, no. 4, Art. 189, 2022.

[9] S. Malik and K. Jothimani, "Enhancing student success prediction with FeatureX: A fusion voting classifier with hybrid feature selection," Educ. Inf. Technol., vol. 29, pp. 8741–8791, 2024.

[10] Frontiers in Education, "Machine learning models for academic performance prediction," Front. Educ., vol. 10, Art. 1632315, 2025.

[11] D. Dunning and J. Kruger, "Unskilled and unaware of it," J. Pers. Soc. Psychol., vol. 77, no. 6, pp. 1121–1134, 1999.

[12] M. Alamgir, H. Ullah, and S. Ahmad, "A comprehensive review of features and machine learning techniques used in EDM," IEEE Access, vol. 12, pp. 11345–11378, 2024.

[13] K. Mokgwatjane et al., "Explainable ensemble machine learning for sentiment analysis," Mach. Learn. Appl., 2026 (in press).

[14] R. S. Baker and A. Hawn, "Algorithmic bias in education," Int. J. Artif. Intell. Educ., vol. 32, pp. 1052–1092, 2022.

[15] C. Romero and S. Ventura, "Educational data mining and learning analytics: An updated survey," WIREs Data Mining Knowl. Discov., vol. 10, no. 3, Art. e1355, 2020.

[16] P. Kanter and K. Veeramachaneni, "Evaluation metrics and statistical tests for machine learning," Sci. Rep., vol. 14, no. 1, Art. 12345, 2024.

[17] E. Staneviciene, D. Gudoniene, and A. Kukstys, "Data mining-based prediction of students' performance for sustainable e-learning," Sustainability, vol. 16, no. 23, Art. 10442, 2024.

[18] M. Hlosta, L. Herrmannova, J. Vachova, and Z. Zdrahal, "OU Analyse: Analysing at-risk students at the Open University," LACE, 2014.

[19] J. Sweller, "Cognitive load during problem solving: Effects on learning," Cognitive Science, vol. 12, no. 2, pp. 257–285, 1988.

[20] F. Paas and J. J. G. van Merriënboer, "The efficiency of instructional conditions: An approach to combine mental effort and performance measures," Human Factors, vol. 35, no. 4, pp. 737–743, 1993.

[21] F. Paas and T. van Gog, "Optimising worked example instruction: Different ways to increase germane cognitive load," Learning and Instruction, vol. 16, no. 2, pp. 87–91, 2006.

[22] J. Sweller, P. Ayres, and S. Kalyuga, Cognitive Load Theory. New York: Springer, 2011.

[23] B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum, "Human-level concept learning through probabilistic program induction," Science, vol. 350, no. 6266, pp. 1332–1338, 2015. [Defines 'One-Shot Learning' in ML context — distinct from HEL in this paper.]

[24] N. Japkowicz and S. Stephen, "The class imbalance problem: A systematic study," Intell. Data Anal., vol. 6, no. 5, pp. 429–449, 2002.