Original Article

HMEM: A Hybrid Meta-Ensemble Model for Early Prediction of Student Dropout

Thi-Van Nguyen[1*], Thi-Ngoc-Anh Vu[1], Van-Binh Ngo[2]

[1] *Phenikaa School of Computing, PHENIKAA University. Nguyen Trac Street, Duong Noi Ward, Hanoi 12116, Vietnam*

[2]*FPT University. Education and Training Area, Hoa Lac High-Tech Park, Km29 Thang Long Boulevard, Hoa Lac Commune, Hanoi, Vietnam*

Abstract: Student dropout is a significant concern in higher education, posing challenges for institutional performance and long-term learner success. While various machine learning models have been applied to address this issue, many approaches face limitations in handling feature heterogeneity, class imbalance, and model fusion. In this study, we present the Hybrid Meta-Ensemble Model (HMEM), a modular predictive pipeline that combines three gradient boosting learners (CatBoost, LightGBM, XGBoost), probabilistic output enrichment using statistical descriptors (mean, standard deviation, entropy), SMOTE-based meta-level resampling, and a final classification layer using Tab-Transformer.

We evaluate HMEM on the UCI Student Performance dataset under two scenarios—with and without SMOTE—to examine the impact of meta-level balancing. Experimental results show that the full HMEM pipeline with SMOTE achieves consistent improvements across key metrics: Accuracy (0.9384), Precision (0.9412), Recall (0.9412), F1-score (0.9412), and AUC (0.9539). Compared to both the base learners and the meta-ensemble without SMOTE, the proposed approach demonstrates moderate but systematic gains, particularly in detecting minority-class instances. Ablation studies further indicate that probabilistic enrichment, feature partitioning, and SMOTE each contribute meaningfully to performance, and additional baselines (Logistic Regression and simple stacking) confirm that HMEM offers a more favourable balance between discrimination ability and minority-class coverage. Visual analysis of ROC curves and enriched feature distributions corroborates these findings.

Keywords: Student Dropout Prediction, Ensemble Learning, TabTransformer, SMOTE, Educational Data Mining.

---

*Corresponding author.
E-mail address: van.nguyenthi1@phenikaa-uni.edu.vn

1. Introduction

Student dropout in higher education is a persistent global challenge, with wide-ranging consequences for institutional performance, public investment, and social equity. According to the OECD, only 43% of students across member countries complete their undergraduate degrees on time, though this figure rises to 70% within an additional three years[1]. In many cases, early-stage dropout—especially during the first academic year—is associated with factors such as academic underpreparedness, financial difficulties, and insufficient social integration [1].

In Vietnam, the problem is equally concerning. At the Ho Chi Minh City University of Technology and Education, the dropout rate reached 24.45% in 2024, with 2.13% leaving during the first year[2]. Research attributes these outcomes to limited academic engagement, weak institutional support, and challenges in adapting to new learning environments [2, 3].

To address this, data-driven approaches in educational data mining (EDM) and learning analytics have emerged to identify at-risk students early. These models leverage behavioral logs, academic records, and demographic data to support timely interventions. While notable progress has been made using machine learning (ML) and deep learning (DL), existing methods still face challenges such as fragmented feature representation, class imbalance, and limited model fusion.

This study introduces a modular predictive framework that integrates diverse learning strategies with advanced feature enrichment and attention-based modeling. The goal is to improve the reliability and interpretability of dropout prediction systems, particularly within real-world, imbalanced, and heterogeneous educational datasets.

In this study, we introduce a modular predictive framework that integrates diverse learning strategies with advanced feature enrichment and attention-based modeling. The goal is to improve the reliability and interpretability of dropout prediction systems, particularly within real-world, imbalanced, and heterogeneous educational datasets.

In summary, this paper makes the following contributions:

- We propose the Hybrid Meta-Ensemble Model (HMEM), which combines feature-specialised gradient boosting learners, probabilistic meta-feature enrichment, SMOTE-based balancing at the meta level, and an attention-based TabTransformer classifier for early dropout prediction.

- We provide a detailed empirical evaluation on the UCI Student Performance dataset, including comparisons between HMEM and its non-SMOTE variant, three strong boosting baselines (CatBoost, LightGBM, XGBoost), and simple baselines based on Logistic Regression and stacking.

- We conduct ablation studies to quantify the contribution of probabilistic enrichment, feature partitioning, and meta-level resampling, and we analyse ROC curves and meta-feature distributions to offer additional insight into the behaviour of the proposed architecture.

2. Related Work

The literature on student dropout prediction spans a wide range of modeling strategies, from statistical methods to advanced ensemble and deep learning techniques. Among these, gradient boosting algorithms—such as **XGBoost**, **LightGBM**, and **CatBoost**—have consistently

---

[1] https://www.oecd.org/en/publications/2025/09/education-at-a-glance-2025$_c$58$fc9ae.html$

[2] https://vietnamnet.vn/en/low-on-time-graduation-rates-raise-concerns-for-vietnamese-universities-2450731.html

delivered high performance on educational datasets, due to their ability to manage mixed feature types and capture non-linear dependencies [4, 5]. However, most existing applications utilize these models independently or through simple ensembles, lacking structured meta-level integration.

In recent years, deep learning has enabled more expressive modeling of structured and sequential data. The **TabTransformer**, for instance, leverages attention mechanisms to encode categorical features and model their interaction with numerical variables, significantly improving performance on tabular datasets [6, 7]. Similarly, temporal models such as LSTMs [15] and graph-based approaches using GNNs(e.g., GCN, GAT) [8–10] have shown promise in capturing behavioral dynamics and peer influence in learning environments.

Another major challenge in dropout prediction is class imbalance, where dropout instances are often underrepresented. Techniques like **SMOTE** [5] and cost-sensitive learning [11, 12] have been proposed to address this, but are seldom combined with ensemble outputs or used at intermediate levels of model representation.

Recent works have also emphasized the role of contextual and psychological variables—including burnout, psychological capital, and time pressure—in mediating dropout behaviors [13]. Despite their importance, such dimensions are often excluded from operational predictive systems. Additionally, ensemble outputs are typically treated as final decisions rather than as sources of valuable statistical signals (e.g., entropy, variance) that could enrich downstream learning [14].

Overall, prior research has advanced several important directions in dropout prediction, including gradient boosting on tabular data, attention-based tabular deep learning, and class-imbalance handling. However, relatively few studies jointly exploit (i) feature partitioning into semantically coherent groups (academic, behavioural, demographic), (ii) uncertainty-aware meta-features derived from base-model probabilities, and (iii) resampling and attention-based fusion performed at the meta-learning level. To address these gaps, the next section introduces the proposed Hybrid Meta-Ensemble Model (HMEM), which combines feature-specialised boosting learners, probabilistic meta-feature enrichment, meta-level SMOTE, and a TabTransformer meta-classifier for early dropout prediction.

3. Proposed Model

This section presents the architecture of the **Hybrid Meta-Ensemble Model (HMEM)**, as illustrated in Figure 1 , a multi-stage learning framework designed to address feature heterogeneity, class imbalance, and limited model fusion in student dropout prediction. HMEM integrates the strengths of boosting-based models and deep tabular learning through a structured pipeline comprising three components: (i) base learners trained on specialized feature subsets, (ii) probabilistic meta-feature enrichment with SMOTE-based resampling, and (iii) final fusion using a TabTransformer classifier.

3.1. Stage 1: Specialized Base Learners

In designing the base-learner layer, we deliberately combine CatBoost, LightGBM, and XGBoost and train them on semantically distinct feature subsets (academic, behavioral, and demographic). Gradient boosting methods are well known for their strong performance on structured tabular data and their ability to handle heterogeneous feature types. CatBoost is particularly robust to categorical and ordinal variables, while XGBoost and LightGBM are highly effective at modeling non-linear interactions with competitive training efficiency.

Our choice of tree-based ensembles as base learners is also motivated by prior
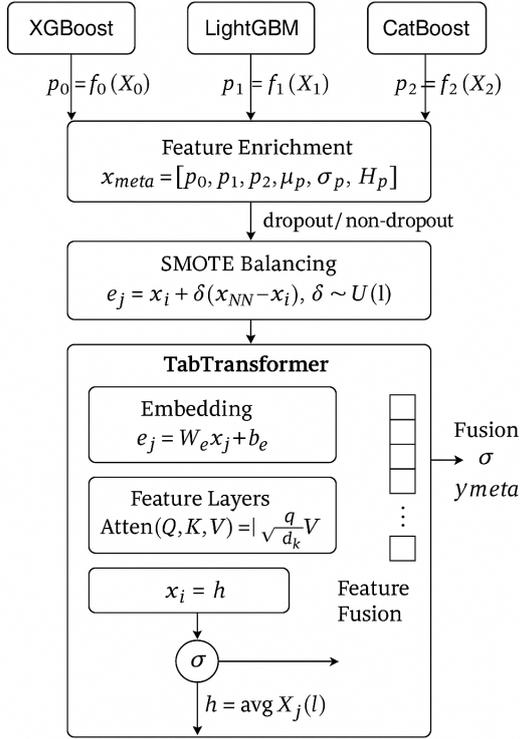
Figure 1. Architectural diagram Hybrid
Meta-Ensemble Model.

empirical evidence on tabular classification outside the educational domain. In a recently accepted conference paper on churn prediction for telecommunications services [16], we systematically compared a broad range of models, including linear classifiers, support vector machines, probabilistic methods, and various tree ensembles. The results showed that tree-based ensembles such as ExtraTrees, RandomForest, XGBoost, and LightGBM consistently outperformed simpler linear and kernel-based methods across Accuracy, Balanced Accuracy, ROC AUC, and F1-score. Although the application domain differs from student dropout prediction, the data characteristics (mixed numerical and categorical attributes, tabular structure, and class imbalance) are similar, providing additional justification for adopting gradient boosting ensembles as the

backbone of our base-learner layer in HMEM.

Conceptually, alternative configurations are also possible, such as using a single boosting model on the full feature set or incorporating additional tree ensembles. In this work, however, we focus on the three-model configuration for two main reasons: (i) it offers a good balance between model diversity and architectural simplicity, and (ii) it aligns with both our prior empirical findings on tabular data and the semantic partitioning of features into academic, behavioral, and demographic groups.

We begin by partitioning the input data into three semantically distinct feature subsets: *academic performance*, *behavioral engagement*, and *demographic attributes*. Each subset is used to train an independent gradient boosting model:

- **XGBoost** on behavioral indicators (e.g., login frequency, forum participation),

- **LightGBM** on academic history (e.g., grades, assignment delays),

- **CatBoost** on categorical/demographic data (e.g., gender, major, socioeconomic status).

Each learner outputs a probability score representing the estimated dropout likelihood from its respective feature group:

$$p_0 = f_0(X_0), \quad p_1 = f_1(X_1), \quad p_2 = f_2(X_2).$$

3.2. Stage 2: Probabilistic Meta-Feature Enrichment

Let $p_0, p_1, p_2 \in [0, 1]$ denote the dropout probabilities predicted by the three base learners (XGBoost, LightGBM, and CatBoost) for a given student. To integrate and enrich these outputs, we construct a meta-feature vector that combines the raw probabilities with higher-order statistical descriptors:

$$\mu_p = \frac{1}{3} \sum_{i=0}^{2} p_i, \quad \sigma_p = \sqrt{\frac{1}{3} \sum_{i=0}^{2} (p_i - \mu_p)^2},$$

$$H_p = -\sum_{i=0}^{2} \tilde{p}_i \log \tilde{p}_i,$$

where $\tilde{p}_i = \frac{p_i}{\sum_{j=0}^{2} p_j + \epsilon}$ are normalized probabilities and ϵ is a small constant to avoid division by zero. The resulting meta-feature vector is defined as

$$x_{\text{meta}} = [p_0, p_1, p_2, \mu_p, \sigma_p, H_p].$$

Intuitively, these descriptors play complementary roles:

- μ_p encodes the overall confidence of the ensemble in predicting dropout for a student;

- σ_p measures disagreement between base learners, where higher variance indicates more conflicting predictions;

- H_p summarizes predictive uncertainty in a single scalar, with higher entropy corresponding to less decisive combined predictions.

By explicitly modeling both consensus and disagreement among base learners, the enriched vector x_{meta} provides the TabTransformer with richer, uncertainty-aware information than using $[p_0, p_1, p_2]$ alone.

Since dropout is often a minority class, we then apply **SMOTE** (Synthetic Minority Oversampling Technique) on the set of meta-features x_{meta} to generate synthetic minority samples and mitigate class imbalance:

$$e_j = x_i + \delta(x_{NN} - x_i), \quad \delta \sim U(0,1),$$

where x_i is a minority meta-feature vector and x_{NN} is one of its nearest neighbors in the meta-feature space.

3.3. Stage 3: Attention-Based Fusion with TabTransformer

The final classification is performed using a **TabTransformer** architecture, which is well suited for structured tabular inputs. At the meta-learning stage, all features $(p_0, p_1, p_2, \mu_p, \sigma_p, H_p)$ are continuous; they are first projected into an embedding space via a shared linear layer and then processed by a stack of transformer encoder blocks.

Concretely, the TabTransformer meta-classifier is implemented with L transformer encoder layers (we use $L = 2$ in our experiments), each with 4 attention heads and an embedding dimension of 64. All meta-features $(p_0, p_1, p_2, \mu_p, \sigma_p, H_p)$ are treated as numerical tokens in this embedding space. We train the TabTransformer using the Adam optimizer with a learning rate of 10^{-3}, a batch size of 32, and early stopping on the validation F1-score with a patience of 20 epochs. Since the meta-feature space at this stage is purely numerical, no additional categorical encodings are required within the attention mechanism.

From a structural perspective, the TabTransformer consists of:

- a linear embedding layer that maps each meta-feature into a shared embedding space,

- a stack of multi-head self-attention layers that model contextual interactions among the embedded meta-features,

- a **fusion layer** (e.g., average pooling followed by fully connected layers) that aggregates the learned representations into a final prediction.

The final output is computed as:

$$y_{\text{meta}} = \sigma(h), \quad h = \text{avg}(X_j^{(l)}),$$

where $\sigma(\cdot)$ is the sigmoid function for binary classification, $X_j^{(l)}$ denotes the intermediate representations from the last transformer layer, and $\text{avg}(\cdot)$ denotes an averaging operation over the token dimension.

3.4. Summary of Design Benefits

The HMEM architecture is designed to:

1. Leverage diverse perspectives via feature-specialized base learners.
2. Quantify prediction uncertainty through statistical enrichment.
3. Improve generalization in imbalanced settings using SMOTE at the meta level.
4. Capture non-linear feature dependencies using attention-based fusion.

Overall, HMEM offers a modular and interpretable framework that balances model complexity with practical deployability in educational settings.

4. Experiments and Results

4.1. Evaluated Metrics

We evaluated the performance of the HMEM on the test data using the following metrics: Accuracy, Precision, Recall,F1-score, ROC Curve and AUC.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

The ROC curve of the models can be drawn by calculating the values of the False Positive Rate (FPR) and True Positive Rate (TPR) of the classification model and taking them as the horizontal and vertical axes, respectively. FPR and TPR are calculated as follows:

$$\text{FPR} = \frac{FP}{FP + TN} \quad (5)$$

$$\text{TPR} = \frac{TP}{TP + FN} \quad (6)$$

Area Under the Curve (AUC) is the area under the ROC curve. Similarly, a larger value of AUC is better. where TP is True Positive, FP is False Positive, TN is True Negative, FN is False Negative.

4.2. Dataset Description

This study employs the publicly available Student Performance dataset from the UCI Machine Learning Repository[3]. The dataset comprises academic and socio-demographic information from students enrolled in two Portuguese secondary schools. It includes 33 variables spanning student grades, study habits, parental education, and family background—providing a comprehensive and multidimensional profile of each learner.

The dataset was selected based on three key considerations. First, its rich attribute space aligns well with the tripartite segmentation of academic, behavioral, and demographic features utilized in the proposed architecture. Second, the inclusion of final grade outcomes enables the construction of dropout risk proxies through academic performance thresholds. Third, its public accessibility promotes reproducibility and comparative evaluation across studies, consistent with prior work in early dropout detection [17].

In the original UCI Student Performance dataset, there is no explicit dropout label. To operationalise a binary outcome compatible with early-warning settings, we derive a proxy label from the final grade $G3 \in [0, 20]$, which represents the overall course performance. Following the conventional pass/fail threshold for this dataset, we define the target variable y as

$$y = \begin{cases} 1, & \text{if } G3 < \tau, \\ 0, & \text{otherwise,} \end{cases}$$

where $\tau = 10$ denotes the minimum passing grade. In this formulation, $y = 1$ is interpreted

---

as "at risk of dropout or severe academic failure", while $y = 0$ corresponds to students who are not at immediate risk. We adopt this risk-oriented proxy label to approximate the dropout phenomenon in the absence of explicit withdrawal records, and later reflect on this choice in the *Limitations* section.

Although the data originates from a European context, the underlying factors of academic underachievement and socioeconomic vulnerability are highly transferable to educational systems such as Vietnam's, where similar patterns of student attrition have been documented [2]. Thus, the dataset remains suitable for evaluating predictive models intended for early warning and targeted intervention.

4.3. Experimental Design

To evaluate the effectiveness of the proposed HMEM, we first conduct experiments under two primary scenarios: (i) using the **SMOTE** algorithm for class balancing at the meta-learning stage (*HMEM (with SMOTE)*), and (ii) without applying any resampling technique (*HMEM (no SMOTE)*). This design allows us to isolate and assess the impact of SMOTE on predictive performance, particularly in addressing class imbalance—a common issue in dropout datasets where the positive class (i.e., dropout cases) is typically underrepresented.

In both scenarios, the HMEM architecture remains unchanged in structure: it integrates three gradient boosting classifiers (XGBoost, LightGBM, and CatBoost) trained on heterogeneous feature groups, corresponding to academic, behavioural, and demographic attributes. The outputs from these base learners are then transformed into enriched meta-features $[p_0, p_1, p_2, \mu_p, \sigma_p, H_p]$, which capture both predictive consensus and uncertainty, and are passed to a TabTransformer-based meta-classifier for final prediction.

When enabled, SMOTE is applied only on the training partition of this enriched

meta-feature space, after the construction of $[p_0, p_1, p_2, \mu_p, \sigma_p, H_p]$ and before fitting the TabTransformer. We employ the standard SMOTE implementation from the `imbalanced-learn` library with $k = 5$ nearest neighbours and the default sampling strategy, which oversamples the minority (dropout) class towards an approximately 1:1 ratio with the majority class. No resampling is performed on the held-out test set, and all base learners are trained on the original, imbalanced data.

Regarding the stacking protocol, the three base learners are first fitted on the training partition. Their predicted dropout probabilities on these training samples are then used to construct the enriched meta-features $[p_0, p_1, p_2, \mu_p, \sigma_p, H_p]$, which serve as input to the TabTransformer meta-classifier. At test time, the same base learners, trained only on the training set, are used to generate probabilities for the held-out test instances; these are transformed into meta-features and passed through the TabTransformer to obtain final predictions. In this way, the test set remains strictly unseen during both base- and meta-level training. In the subsequent *Ablation Studies* subsection, we further vary the enrichment and partitioning components to quantify their individual contributions in addition to the SMOTE versus non-SMOTE comparison.

All experiments are conducted on a single stratified train–test split, which is kept fixed across all model variants to ensure a fair comparison.

4.4. Results and Evaluation

To assess the effectiveness of the proposed model, we conducted experiments under two distinct configurations of the meta-ensemble: one with **SMOTE** applied at the meta-learning stage (*HMEM (with SMOTE)*), and one without (*HMEM (no SMOTE)*). These were evaluated alongside three base learners—**XGBoost**,

LightGBM, and **CatBoost**—to facilitate comparative analysis.

Table 1. Performance metrics of base models and HMEM variants

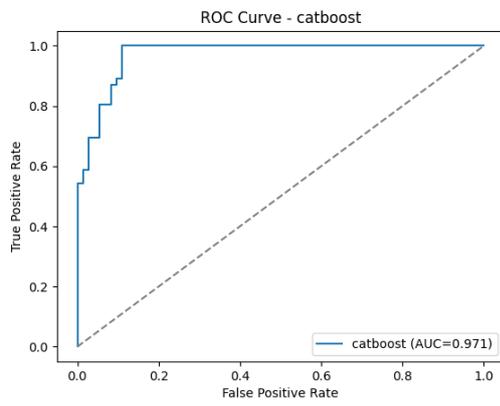| Model	Accuracy	Precision	Recall	F1	AUC
CatBoost	0.8992	0.8696	0.8696	0.8696	0.9711
LightGBM	0.8824	0.8478	0.8478	0.8478	0.9726
XGBoost	0.8824	0.8478	0.8478	0.8478	0.9699
HMEM (no SMOTE)	0.9187	0.9215	0.9215	0.9215	0.9495
HMEM (with SMOTE)	**0.9384**	**0.9412**	**0.9412**	**0.9412**	**0.9539**

Figure 2. ROC curve for CatBoost (AUC = 0.9711).

The experimental results in Table 1 show that the proposed **HMEM** achieves a more balanced performance profile than the individual base learners. Among the base models, **CatBoost** and **LightGBM** obtain slightly higher AUC values than XGBoost (all above 0.969; see Figures 2–3), indicating strong discrimination ability on this dataset. However, all base learners exhibit comparatively lower F1-scores (ranging from 0.8478 to 0.8696), reflecting a limited balance between Precision and Recall—an important consideration in imbalanced dropout classification tasks where correctly identifying at-risk students is critical.

Overall, the HMEM configurations improve Accuracy, Precision, Recall, and F1-score compared with the base models. Both meta-ensembles achieve F1-scores above 0.92, with the SMOTE-enabled variant *HMEM (with SMOTE)*
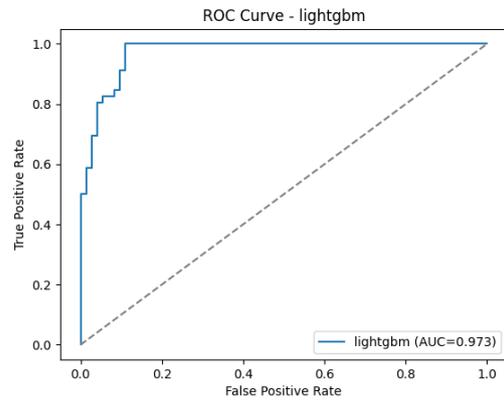
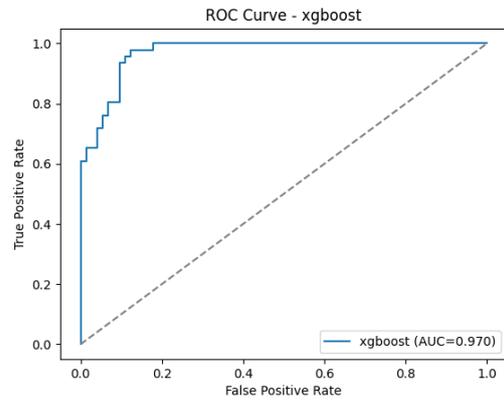

Figure 3. ROC curve for LightGBM (AUC = 0.9726).

Figure 4. ROC curve for XGBoost (AUC = 0.9699).

attaining the best overall balance: Accuracy of 93.84%, Precision and Recall of 94.12%, and an F1-score of 94.12%. Its AUC (0.9539) remains high and competitive, though slightly lower than the strongest base learners. In this sense, the meta-ensemble does not strictly dominate the base models on every metric; instead, it offers a more favourable trade-off by improving Recall and F1 for the minority (dropout) class while maintaining a strong level of discriminative performance.

The comparison between *HMEM (with SMOTE)* and *HMEM (no SMOTE)* further illustrates the contribution of meta-level
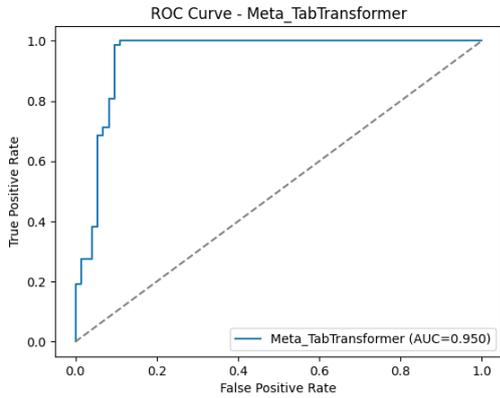
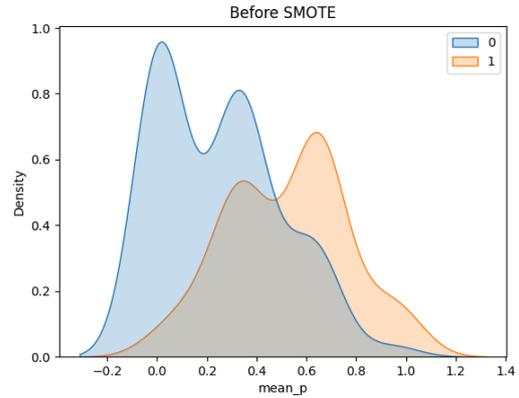Figure 5. ROC curve for HMEM (without SMOTE)
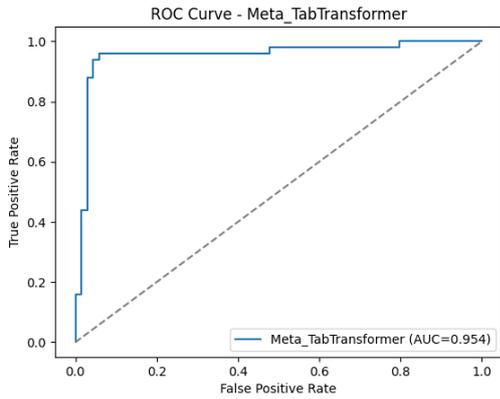(AUC = 0.9495).

Figure 7. Distribution of μ_p before SMOTE.

Figure 6. ROC curve for HMEM (with SMOTE)
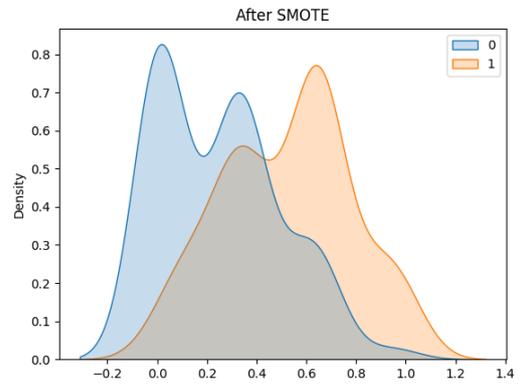(AUC = 0.9539).

Figure 8. Distribution of μ_p after SMOTE.

resampling. The SMOTE-based configuration yields higher Accuracy and F1-score (0.9384 and 0.9412 versus 0.9187 and 0.9215, respectively), and a modest increase in AUC (from 0.9495 to 0.9539). This suggests that synthetic balancing applied to the enriched meta-feature space contributes to better generalisation, particularly for minority-class detection, without substantially degrading the global ranking performance captured by AUC.

Additional insight into the effect of SMOTE is provided by the meta-feature distribution plots in Figures 7 and 8. Before SMOTE,

the distribution of the meta-feature μ_p (the average predicted dropout probability across base learners) is skewed between classes, which can bias the learning process at the meta level. After applying SMOTE, the class-conditional distributions of μ_p become more balanced and exhibit greater overlap in the meta-feature space, making it easier for the TabTransformer to learn a robust decision boundary between dropout and non-dropout instances.

The ROC curves for the two meta-models (Figures 5–6) also align with this observation. Both curves show high true positive rates across a wide range of thresholds, but the SMOTE-

enhanced model presents a slightly more convex shape and a marginally larger area under the curve, indicating a modest gain in robustness against threshold variation.

In summary, the proposed architecture effectively leverages:

- the predictive strength of gradient boosting learners (XGBoost, LightGBM, CatBoost),

- the discriminative power of meta-feature enrichment and probabilistic fusion,

- and the structural learning capability of the TabTransformer.

Rather than strictly outperforming the base learners on every individual metric, HMEM—especially with SMOTE at the meta-learning stage—provides a more balanced and practically useful performance profile for early-stage dropout prediction in educational settings, with improved Recall and F1-score for at-risk students while preserving competitive AUC.

4.5. Ablation Studies

To better understand the contribution of each component in HMEM, we conducted an ablation study over four variants: (i) the full HMEM with feature partitioning, probabilistic enrichment and SMOTE (*HMEM (with SMOTE)*); (ii) a variant without enrichment, where the meta-learner only receives the raw base probabilities $[p_0, p_1, p_2]$ (*NoEnrichment*); (iii) a variant without feature partitioning, in which all three base learners are trained on the full feature space (*NoPartition*); and (iv) a variant without SMOTE at the meta level (*HMEM (no SMOTE)*). The results are summarised in Table 2.

Overall, the *HMEM (with SMOTE)* configuration achieves the most balanced performance, with Accuracy, Precision, Recall and F1-score all around 0.94 and an AUC of 0.9539. Removing the probabilistic enrichment (*NoEnrichment*) leads to a clear degradation

Table 2. Ablation study on the components of HMEM

Variant	Accuracy	Precision	Recall	F1	AUC
HMEM (with SMOTE)	0.9384	0.9412	0.9412	0.9412	0.9539
NoEnrichment (p_0, p_1, p_2 only)	0.9241	0.8846	0.8846	0.8846	0.9507
NoPartition (all features per base)	0.9114	0.8800	0.8462	0.8627	0.9652
HMEM (no SMOTE)	0.9187	0.9215	0.9215	0.9215	0.9495

in all threshold-based metrics: Accuracy drops from 0.9384 to 0.9241 and F1-score from 0.9412 to 0.8846, while AUC decreases slightly from 0.9539 to 0.9507. This suggests that the statistical descriptors (μ_p, σ_p, H_p) provide useful uncertainty-aware information beyond the raw base probabilities, enabling the TabTransformer to better distinguish between dropout and non-dropout profiles.

Disabling feature partitioning (*NoPartition*) has an even stronger impact on the meta-level behaviour. When all three base learners are trained on the full feature space, Accuracy and F1-score decrease to 0.9114 and 0.8627, and Recall falls to 0.8462. Interestingly, the AUC in this variant is the highest among the four (0.9652), indicating that the non-partitioned ensemble can still rank instances well across decision thresholds, but is less well calibrated at the default operating point, leading to a less favourable Precision–Recall trade-off. This observation is consistent with the behaviour of the base learners under *NoPartition*: CatBoost, LightGBM and XGBoost reach AUC values of 0.9695, 0.9666 and 0.9507, respectively, while their F1-scores remain below 0.89 (0.8846, 0.8627 and 0.8627). These trends support our design choice of assigning semantically coherent feature groups (academic, behavioural, demographic) to different boosting models, which appears to stabilise the decision boundary and improve the balance between Precision and Recall at the meta level.

Finally, the comparison between *HMEM (with SMOTE)* and *HMEM (no SMOTE)*

highlights the effect of meta-level resampling. Omitting SMOTE results in lower Accuracy (0.9187 vs. 0.9384) and a reduction in F1-score and Recall (0.9215 vs. 0.9412), together with a slight decrease in AUC (0.9495 vs. 0.9539). In other words, applying SMOTE on the enriched meta-feature space allows HMEM to recover additional at-risk students and to improve overall classification quality, with only a modest change in discriminative performance. Given the high cost of missing dropout cases, this shift towards higher Recall and F1-score is desirable in many early-warning scenarios, and justifies the use of SMOTE at the meta-learning stage of the proposed framework.

4.6. Additional Baseline Models

To further contextualise the performance of HMEM, we additionally compare it with a set of simple yet widely used baseline models trained directly on the raw feature space and on the base-learner outputs. All baselines use the same stratified train-test split as the main experiments. Table 3 reports the performance of additional

Table 3. Additional baseline models on the UCI Student Performance dataset

Model	Accuracy	Precision	Recall	F1	AUC
Logistic Regression (raw features)	0.9241	0.9545	0.8077	0.8750	0.9710
Simple stacking (LR on $[p_0, p_1, p_2]$)	0.8861	0.9048	0.7308	0.8085	0.9528
HMEM (with SMOTE)	0.9384	0.9412	0.9412	0.9412	0.9539

baseline models. Logistic Regression trained on the full (one-hot encoded) feature space attains a competitive AUC of 0.9710, which is comparable to the boosting-based base learners, and achieves high Precision (0.9545). However, its Recall (0.8077) and F1-score (0.8750) remain below those of *HMEM (with SMOTE)*, indicating that the linear model tends to prioritise conservative positive predictions at the cost of missing a non-negligible proportion of at-risk students. The simple stacking baseline, which fits a

Logistic Regression meta-learner on the raw base probabilities $[p_0, p_1, p_2]$ produced by the three partitioned base learners (LightGBM, XGBoost, and CatBoost), obtains moderate performance (Accuracy = 0.8861, F1 = 0.8085, AUC = 0.9528), but clearly lags behind HMEM in both Recall (0.7308 vs. 0.9412) and F1-score. In this baseline, no probabilistic enrichment or meta-level resampling is applied; the meta-learner operates solely on the three base probabilities.

Overall, these results suggest that the proposed HMEM does not merely outperform the individual boosting models, but also offers clear advantages over standard linear baselines and a simple stacking scheme. By combining feature-specialised boosting learners, probabilistic enrichment (μ_p, σ_p, H_p), meta-level SMOTE, and an attention-based TabTransformer, HMEM achieves a more favourable balance between discrimination ability (AUC) and minority-class performance (Recall and F1-score). In the context of early-warning systems, where the cost of failing to identify students at risk is particularly high, this shift towards higher Recall and F1-score at a competitive AUC level supports the practical value of the HMEM architecture.

4.7. Comparison with Other Studies

Table 4. Performance Comparison with Related Studies

Study / Model	Acc.	Prec.	Recall	F1	AUC
Almeida et al. (2025) / GraphSAGE [8]	0.9100	-	-	0.8100	-
Elbouknify et al. (2025) / LightGBM [9]	0.8800	-	-	0.8600	0.8700
Wang et al. (2022) / LSTM [15]	-	-	-	0.8900	-
HMEM (Proposed)	**0.9384**	**0.9412**	**0.9412**	**0.9412**	**0.9539**

As presented in Table 4, recent studies have made valuable contributions to the task of student dropout prediction using a variety of modeling approaches. For instance, Almeida et al. [8] employed GraphSAGE to capture social interaction structures, while Elbouknify et al. [9]

utilized LightGBM to build a robust feature-based classifier. Likewise, recurrent architectures such as LSTM [15] have shown potential in modeling temporal engagement patterns.

Within this landscape, the proposed **HMEM** offers a complementary direction by integrating multiple ensemble learners and attention-based fusion through TabTransformer. Despite differences in datasets and modeling assumptions across studies, HMEM achieves comparatively high scores across all metrics—particularly with an F1-score and AUC exceeding 0.94 and 0.95, respectively. These results suggest that enriching model outputs with probabilistic descriptors and addressing class imbalance at the meta-level can provide practical benefits in real-world educational settings.

While not intended as a replacement for other modeling paradigms, HMEM demonstrates that hybrid ensemble strategies can contribute meaningfully to improving early-warning systems, especially in contexts where data are heterogeneous and class imbalance is present.

5. Conclusion

This study proposes the Hybrid Meta-Ensemble Model (HMEM), an integrated architecture that combines multiple gradient boosting learners, probabilistic output enrichment, SMOTE-based meta-level balancing, and attention-based fusion via a TabTransformer. The empirical results show that HMEM (with SMOTE) achieves consistently higher Accuracy, Recall and F1-score than the individual base learners and its non-SMOTE counterpart, while maintaining competitive AUC. Ablation studies and additional baseline experiments with Logistic Regression and simple stacking further confirm that probabilistic enrichment, feature partitioning, and meta-level resampling each contribute meaningfully to the overall performance and yield clear advantages over simpler linear or stacking alternatives.

Despite these promising results, the work has several limitations. The UCI Student Performance dataset does not contain explicit dropout labels; instead, we derive a binary risk indicator from the final grade G3 using a pass/fail threshold of $\tau = 10$, which only approximates true dropout behaviour. Moreover, the experiments are conducted on a single dataset from Portuguese secondary education, whereas the intended deployment context is Vietnamese higher education, so further validation on local institutional data is required to assess external validity. Methodologically, HMEM combines existing techniques (gradient boosting, SMOTE and a TabTransformer meta-learner) into a coherent framework rather than introducing a fundamentally new learning paradigm.

Future work will therefore focus on evaluating HMEM across multiple educational datasets and regions, as well as exploring extensions such as integrated explainability (e.g., SHAP-based analyses), alternative fusion strategies, and uncertainty-aware or temporal transformer variants to better capture complex patterns of student engagement over time.

References

[1] D.-M. Córdova-Esparza, J. R. Terven, J.-A. Romero-González, K.-E. Córdova-Esparza, R.-E. López-Martínez, T. García-Ramírez, R. Chaparro-Sánchez, "Predicting and Preventing School Dropout with Business Intelligence: Insights from a Systematic Review," *Information*, 2025. https://doi.org/10.3390/info16040326.

[2] L. T. M. Lê, T. T. A. Trn, T. A. Tran, "The Overview of the Reality and the Factors Affecting University Dropouts," *Science Technology Development Journal - Social Science and Humanities*, 2024. https://doi.org/10.32508/stdjssh.v8i3.1033.

[3] C. B. Mai, N. G. Tran, N. H. Nguyen, T. T. N. Nguyen, N. P. H. Bao, "Factors Influencing Students' Dropout Intentions in Ho Chi Minh City, Vietnam," *International Journal of Cognitive Research in Science, Engineering and Education (IJCRSEE)*, 2023.

[4] S. Lotfi, J. Lahnin, S. Atifi, "Optimizing Retention in Professional Training: A Hybrid Artificial Intelligence

Approach for Predicting and Characterizing Dropouts," unpublished.

[5] S. Rizwan, C. K. Nee, S. Garfan, "Identifying the Factors Affecting Student Academic Performance and Engagement Prediction in MOOC Using Deep Learning: A Systematic Literature Review," *IEEE Access*, vol. 13, pp. 18952–18982, 2025. https://doi.org/10.1109/ACCESS.2025.3533915.

[6] X. Zhang, Y. Zhang, A. L. Chen, M. Yu, L. Zhang, "Optimizing Multi Label Student Performance Prediction with GNN-TINet: A Contextual Multidimensional Deep Learning Framework," *PLOS ONE*, vol. 20, 2025. https://doi.org/10.1371/journal.pone.0314823.

[7] A. Patel, J. Kim, S. Yoon, "Student Retention Prediction Using TabTransformer: A Novel Architecture for Tabular Data," *Knowledge-Based Systems*, 2025.

[8] P. G. Almeida, G. Silva, V. C. Santos, G. J. P. Moreira, P. H. L. Silva, E. Luz, "Deep Learning for School Dropout Detection: A Comparison of Tabular and Graph-Based Models for Predicting At-Risk Students," *ArXiv*, vol. abs/2508.14057, 2025. https://doi.org/10.48550/arXiv.2508.14057.

[9] I. Elbouknify, I. Berrada, L. Mekouar, Y. Iraqi, E. H. Bergou, H. Belhabib, Y. Nail, S. Wardi, "AI-Based Identification and Support of At-Risk Students: A Case Study of the Moroccan Education System," *ArXiv*, vol. abs/2504.07160, 2025. https://doi.org/10.48550/arXiv.2504.07160.

[10] D. Park, J. Lee, "Using Social Network Analysis and GNNs to Predict Student Dropout in Blended Learning Environments," *Computers & Education Open*, 2025.

[11] H. Huang, L. Jew, D. Qi, "Take a MOOC and Then Drop: A Systematic Review of MOOC Engagement Pattern and Dropout Factor," *Heliyon*, vol. 9, 2023. https://doi.org/10.1016/j.heliyon.2023.e15220.

[12] R. Roy, P. Singh, "A Reinforcement Learning Agent for Adaptive Intervention of At-Risk Students," *Artificial Intelligence in Education Journal*, 2025.

[13] A. Bettahi, F.-Z. Beloudha, H. Harroud, "Continuous-Time Modeling in Educational Data Mining and Learning Analytics: A Literature Review on Methods, Ethics, and Emerging AI Trends," *IEEE Access*, vol. 13, pp. 181443–181464, 2025. https://doi.org/10.1109/ACCESS.2025.3622103.

[14] J. García-Sánchez *et al.*, "Explainable AI for Student Dropout Prediction: SHAP and LIME in Higher Education," *Education and Information Technologies*, 2023.

[15] S. Mao, C. Zhang, Y. Song, J. Wang, X.-J. Zeng, Z. Xu, Q. Wen, "Time Series Analysis for Education: Methods, Applications, and Future Directions," *ArXiv*, vol. abs/2408.13960, 2024. https://doi.org/10.48550/arXiv.2408.13960.

[16] Thi-Van Nguyen, Van-Binh Ngo, "Optimizing Stacking Models for Accurate Customer Churn Prediction in Telecommunications," in *Proceedings of the International Conference on Advances in Information and Communication Technology (ICTA)*, Ha Noi, Vietnam, 2025, accepted and presented, to appear.

[17] A. Goren, L. Cohen, D. Rubinstein, "Early Prediction of Student Dropout in Higher Education Using Machine Learning Models," in *Proceedings of EDM Conference*, 2024. https://doi.org/10.5281/zenodo.12729834.