Indonesian Journal of Electrical Engineering and Computer Science
Vol. 43, No. 3, September 2026, pp. 880~897
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v43.i3.pp880-897

Explainable hybrid machine learning for predicting student enrollment decisions in Indonesian private universities

Dybio Dompu Hot Asih[1], Alex Permatanta Karo-Karo[2], Erbin Sitorus[3]
[1]Information Systems Study Program, Universitas Wirahusada Medan, Medan, Indonesia
[2]Management Study Program, Universitas Wirahusada Medan, Medan, Indonesia
[3]Information Systems Study Program, Universitas Audi Indonesia, Medan, Indonesia

Article Info

Article history:

Received Jun 27, 2026
Revised Aug 10, 2026
Accepted Aug 19, 2026

Keywords:

Explainable artificial intelligence
Hybrid machine learning
Machine learning
Private universities
Student enrollment prediction

ABSTRACT

Student enrollment prediction is an important challenge for private universities because accurate predictions can support admission planning, marketing strategies, and institutional resource allocation. However, conventional prediction models may provide high predictive performance while offering limited insight into the factors underlying students' enrollment decisions. This study proposes an explainable hybrid machine learning approach to predict student enrollment decisions in Indonesian private universities and to identify the factors that contribute most strongly to the prediction outcomes. The study uses student enrollment data collected from Universitas Wirahusada Medan and Universitas Audi Indonesia. The proposed approach combines feature selection and multiple machine learning algorithms to improve predictive performance, while explainable artificial intelligence (XAI) using Shapley additive explanations (SHAP) is employed to interpret the model predictions. The models are evaluated using accuracy, precision, recall, F1-score, receiver operating characteristic - area under the curve (ROC-AUC), and confusion matrix analysis. The results show that the proposed hybrid approach provides competitive predictive performance compared with individual machine learning models. The SHAP analysis further reveals the relative contribution and direction of the most influential features in determining student enrollment decisions. These findings demonstrate that combining predictive modeling with explainability can provide not only accurate enrollment predictions but also actionable insights for university decision-makers. The study contributes an integrated approach for supporting data-driven student recruitment and enrollment management in Indonesian private higher education institutions.

This is an open access article under the CC BY-SA license.

Corresponding Author:

Dybio Dompu Hot Asih
Information Systems Study Program, Universitas Wirahusada Medan
Jl. TB. Simatupang No. 148, Kel. Sunggal, Kec. Medan Sunggal, Kota Medan 20128, Indonesia
Email: dybio.dompu@gmail.com

1. INTRODUCTION

The rapid development of digital technology has changed how higher education institutions manage student recruitment, admission, and enrollment. Universities increasingly rely on data-driven approaches to understand prospective students and improve recruitment strategies, particularly as competition among higher education institutions continues to increase. In Indonesia, this challenge is especially relevant for private universities, where students' enrollment decisions may be influenced by institutional reputation, affordability, financial support, academic considerations, and engagement with university information channels. As a

result, institutions need analytical approaches that can identify applicants with a high probability of enrollment while providing useful information for recruitment planning and resource allocation [1]–[4].

Educational data mining and learning analytics have expanded the use of machine learning for predicting different educational outcomes, including student performance, dropout risk, learning behavior, and enrollment-related decisions. A systematic review of predictive learning analytics shows that research in this area has increasingly moved from conventional statistical approaches toward machine learning methods capable of handling complex relationships within educational data [1]. Unlike traditional models that often depend on predefined relationships between variables, machine learning can capture nonlinear patterns and interactions that may be difficult to identify using conventional analytical approaches. These capabilities make machine learning increasingly relevant for educational decision-support applications [1], [5].

Student enrollment prediction is particularly important because the number of applicants who ultimately register can directly affect institutional planning and sustainability. Universities need reliable estimates of enrollment outcomes to support recruitment campaigns, scholarship allocation, academic capacity planning, faculty workload management, and financial planning. Previous studies have demonstrated the feasibility of using machine learning to predict enrollment among admitted applicants, suggesting that predictive models can provide useful information before final registration decisions are made [6]–[8]. However, prediction at the individual applicant level remains challenging because enrollment decisions are affected by a combination of academic, financial, institutional, and behavioral factors.

Existing research has employed various machine learning algorithms for educational prediction, including decision trees, support vector machines, random forest, gradient boosting, and other classification techniques. For example, research on undergraduate applicant enrollment has demonstrated the usefulness of binary classification models for predicting whether admitted applicants will eventually enroll [6], while other studies have investigated machine learning models for enrollment and attrition prediction [8]. Multi-model approaches have also been investigated for higher education admission prediction, indicating that combining different classifiers can improve predictive capability compared with relying on a single model [9], [10]. Nevertheless, much of this research emphasizes predictive performance, whereas the reasons underlying individual predictions are often less thoroughly examined.

This limitation becomes increasingly important as machine learning is used to support institutional decisions. A model that achieves high predictive accuracy may still have limited practical value if admission officers and university managers cannot understand why a particular applicant is classified as likely or unlikely to enroll. Explainable artificial intelligence (XAI) addresses this issue by providing interpretable information about how input variables contribute to model predictions. The broader XAI literature emphasizes that explainability can strengthen transparency, accountability, and trust in artificial intelligence systems, particularly when predictions influence consequential decisions [11]. In educational settings, explainable machine learning has also been shown to help practitioners understand the factors underlying predictions and assess whether model outputs are consistent with domain knowledge [5], [12], [13].

Recent educational studies provide evidence of the value of explainability, although their applications have largely focused on outcomes other than enrollment. Explainable machine learning has been applied to student performance prediction, dropout prediction, student adaptability, and early identification of at-risk students [5], [12], [14]–[17]. These studies demonstrate that explainability can complement predictive performance by revealing influential variables and supporting more informed interventions. For example, explainable dropout prediction has been used to identify factors associated with student risk, while explainable learning analytics has highlighted the importance of examining the stability and interpretation of predictive models rather than relying solely on their accuracy [13]-[15]. These findings suggest that explainability should be considered as part of the predictive framework itself rather than treated as an additional feature after model development.

The literature on student admission and enrollment provides a more direct basis for the present study. Recent research has demonstrated that machine learning can predict enrollment among undergraduate applicants and admitted students [6], [7]. Other studies have investigated the selection of incoming freshmen using machine learning and have shown that predictive models can assist institutional admission processes [18]. Similarly, enrollment prediction has been examined using machine learning in university educational environments, confirming the potential of predictive analytics for institutional planning [19]. However, these studies primarily concentrate on prediction outcomes and do not fully integrate predictive modeling with mechanisms for explaining the contribution of individual applicant characteristics.

Ensemble learning represents another important development in this research area. Combining multiple classifiers can exploit complementary learning patterns and potentially improve generalization compared with an individual classifier. In higher education admission prediction, Zub *et al.* developed a two-stage PNN–SVM ensemble and demonstrated the potential of multi-model architectures for improving admission prediction performance [10]. Li *et al.* [9] likewise proposed a multi-model analytics system to support student admission decision-making. These studies provide evidence that multi-model approaches are

promising for higher education analytics. However, they do not fully address the integration of hybrid feature selection, adaptive ensemble learning, and explainable prediction within one framework.

Another stream of research has focused on course enrollment and broader enrollment management. Machine learning has been used to predict course enrollment patterns to support academic planning and resource allocation [20], [21]. Other work has examined student enrollment and attrition simultaneously, demonstrating that predictive analytics can support both recruitment and retention planning [8]. Although these studies are relevant to enrollment management, course-level enrollment and institutional-level forecasting are conceptually different from predicting an individual applicant's decision to enroll after receiving an admission opportunity. Individual enrollment decisions may depend on institutional characteristics, perceived value, financial considerations, and communication or engagement factors that are not necessarily represented in course-level or aggregate forecasting models.

The distinction between individual prediction and aggregate forecasting is therefore important. Time-series and other forecasting approaches can provide useful estimates of future enrollment trends, but they generally operate at an aggregate institutional or temporal level. In contrast, an applicant-level machine learning framework can use multidimensional characteristics to estimate the probability that a specific prospective student will enroll. For institutional recruitment, this distinction has practical importance because decision-makers may need to identify which applicants should receive targeted communication, financial information, or other recruitment interventions rather than simply estimate the total number of future students.

Beyond predictive modeling, research on university choice suggests that enrollment decisions are shaped by several institutional and socioeconomic considerations. Recent evidence from private higher education indicates that institutional reputation and location can influence students' university choices [2]. Research on enrollment management also emphasizes the importance of understanding students' preferences and decision-making processes when developing institutional recruitment strategies [3]. Studies of international and domestic university choice further demonstrate that students may consider multiple factors when selecting a higher education institution, while financial support and affordability can affect higher education choices [4], [22]. These findings are important for machine learning-based enrollment prediction because they indicate that prediction models should consider heterogeneous factors rather than relying exclusively on academic variables.

The importance of these factors also reinforces the need for interpretable models. If an enrollment prediction model identifies institutional reputation, financial support, affordability, digital engagement, or academic facilities as influential variables, university managers need to understand the direction and relative contribution of those variables before translating predictions into recruitment actions. Explainable models can provide this additional layer of decision support by connecting predictive outputs with interpretable feature contributions [11], [13], [23]. Consequently, a useful enrollment prediction framework should not only answer whether an applicant is likely to enroll but should also provide evidence regarding why the model reaches that prediction.

Despite these advances, an important gap remains in the literature. First, previous enrollment and admission studies have largely evaluated individual machine learning models or conventional multi-model approaches [6], [7], [9], [10], [18], [19]. Second, research on explainable educational machine learning has predominantly focused on student performance, dropout, adaptability, or academic risk rather than post-admission enrollment decisions [5], [12], [14]-[17]. Third, existing ensemble approaches do not consistently integrate feature selection and explainability into a unified applicant-level enrollment prediction framework [9], [10]. Fourth, relatively limited evidence is available on explainable enrollment prediction in Indonesian private universities, despite the competitive recruitment environment faced by these institutions.

The methodological gap is particularly relevant because combining complementary machine learning models does not automatically guarantee that the resulting system will be interpretable or efficient. Feature selection can reduce irrelevant and redundant information before model training, while ensemble learning can combine the strengths of different classifiers. Explainability can then be used to identify the contribution of individual predictors to the final model output. The integration of these components provides an opportunity to move beyond a conventional prediction-only approach toward a decision-support framework that simultaneously addresses predictive performance and interpretability [11], [24]–[28].

Accordingly, this study proposes an explainable hybrid machine learning framework (EHMLF) for predicting student enrollment decisions in Indonesian private universities, using admission data from Universitas Wirahusada Medan and Universitas Audi Indonesia. The framework combines hybrid feature selection, adaptive ensemble learning, and XAI to identify prospective students who are likely to enroll while also revealing the factors that contribute to the predictions. The framework is designed to combine the complementary strengths of multiple classifiers rather than relying on a single predictive algorithm, while the

explainability component is intended to make the resulting predictions more transparent for institutional decision-makers.

This study makes three main contributions. First, it develops an applicant-level hybrid prediction framework that integrates feature selection and multiple machine learning classifiers for student enrollment prediction, extending previous admission studies that primarily rely on individual or conventional ensemble models [6], [9], [10], [18], [19]. Second, it incorporates explainability directly into the prediction pipeline, addressing the limitation of previous enrollment studies that emphasize predictive accuracy without sufficiently explaining the determinants of individual predictions [11], [13], [23]. Third, it provides empirical evidence from two Indonesian private universities and identifies the relative importance of institutional, financial, academic, and engagement-related factors in enrollment decisions, thereby connecting predictive analytics with practical university recruitment and decision-making needs [2]-[4], [22].

Therefore, the contribution of EHMLF is not limited to obtaining a high-performing classification model. The framework seeks to provide a more complete decision-support mechanism in which prediction, feature selection, ensemble learning, and explanation are considered together. By explicitly comparing predictive performance and influential factors with findings from previous admission, enrollment, and explainable educational machine learning studies, this research aims to strengthen the methodological and practical understanding of explainable student enrollment prediction in Indonesian private higher education.

2. METHOD

2.1. Research design

This study employed a quantitative predictive analytics design to develop an EHMLF for predicting student enrollment decisions in Indonesian private universities. The research was designed not only to maximize classification performance but also to identify the variables that contribute to individual enrollment predictions. This approach is consistent with the growing use of predictive learning analytics and explainable machine learning in higher education, where predictive accuracy needs to be accompanied by transparent interpretation of model outputs [1], [11], [13].

The proposed EHMLF integrates five sequential stages: 1) data preparation, 2) feature engineering and preprocessing, 3) hybrid feature selection, 4) adaptive ensemble classification, and 5) explainability analysis using Shapley additive explanations (SHAP). The framework was designed to address three methodological requirements simultaneously: reducing irrelevant predictors, improving predictive robustness through model combination, and providing interpretable evidence regarding the factors associated with student enrollment decisions [11], [23], [28].

Unlike a conventional prediction pipeline in which a single machine learning algorithm is evaluated primarily according to accuracy, the proposed approach treats feature selection, ensemble prediction, and explainability as interconnected components. This design is particularly relevant to university admission decision-making because an accurate prediction without an understandable explanation may provide limited practical value to admission managers.

2.2. Research object and dataset

The research was conducted using admission datasets obtained from Universitas Wirahusada Medan and Universitas Audi Indonesia. The dataset consisted of 2,000 undergraduate student records from the 2025–2026 academic year, with 1,000 records contributed by each university. The use of data from two private universities was intended to provide variation in institutional contexts while maintaining a comparable admission setting.

Each observation represents an admitted undergraduate applicant and contains information related to demographic characteristics, admission background, academic characteristics, financial considerations, institutional perceptions, university interaction, and final enrollment status. The target variable was Student Enrollment Decision, represented as a binary classification problem:

- Enrolled (1): the admitted applicant completed the required enrollment or registration process.
- Not Enrolled (0): the admitted applicant did not complete the enrollment or registration process.

The distinction between admission and enrollment is important because being admitted does not necessarily mean that an applicant will ultimately register. Previous studies have demonstrated the usefulness of machine learning for predicting enrollment among admitted applicants, but the explainability of individual enrollment decisions remains less developed [6]-[8].

The dataset therefore differs from studies that predict academic performance or dropout after students have entered university. The unit of prediction in this study is an admitted applicant's enrollment decision before or at the final registration stage, making the framework more directly relevant to recruitment and admission management.

2.3. Respondent characteristics

Table 1 presents the demographic characteristics of the dataset. The dataset was balanced by institution, with each university contributing an equal number of observations. This structure allows the predictive framework to be developed without one institution dominating the learning process. Female applicants represented 55% of the observations, while the largest age group was 18–19 years, followed by 20–21 years and 22 years.

The institutional balance is also relevant to the interpretation of the results because the proposed framework is intended to capture enrollment patterns that are not restricted to a single institutional environment. Previous research has shown that institutional characteristics, reputation, location, financial considerations, and students' decision-making processes can influence university choice and enrollment behavior [2]-[4], [22].

Table 1. Demographic characteristics of students (n = 2,000)

Variable	Category	Frequency	Percentage
University	Universitas Wirahusada Medan	1,000	50.0%
	Universitas Audi Indonesia	1,000	50.0%
Gender	Male	900	45.0%
	Female	1,100	55.0%
Age	18–19 years	1,100	55.0%
	20–21 years	700	35.0%
	22 years	300	15.0%

2.4. Data collection and ethical considerations

The study used secondary admission data obtained from the official admission databases of Universitas Wirahusada Medan and Universitas Audi Indonesia following institutional approval. The variables included gender, age, previous educational background, admission pathway, scholarship availability, tuition affordability, university reputation, digital marketing engagement, academic facilities, service quality, communication intensity, and final enrollment status.

Before analysis, direct personal identifiers were removed from the dataset. The analysis was conducted using de-identified records, and the data were used only for research and predictive modeling purposes. This procedure was intended to minimize privacy risks associated with the use of educational data.

The selection of predictors was guided by both data availability and their relevance to student enrollment decisions. Previous research indicates that university reputation, location, financial considerations, institutional support, and other characteristics can influence higher education choice [2]-[4], [22], [29]. Accordingly, the proposed framework incorporates multidimensional predictors rather than relying exclusively on academic characteristics.

2.5. Data preprocessing

Data preprocessing was conducted before feature selection and model training to improve data quality and reduce potential sources of modeling error. The preprocessing procedure consisted of the following steps,

- Data cleaning, including checking data types, invalid values, and inconsistent records.
- Duplicate removal to ensure that the same applicant record was not represented more than once.
- Missing-value treatment using an appropriate imputation procedure based on the characteristics of the variable.
- Categorical encoding using one-hot encoding for categorical predictors.
- Numerical scaling using Min-Max normalization where scaling was required by the modeling procedure.
- Outlier identification using the interquartile range (IQR) criterion.

These preprocessing procedures were applied before model development to produce a consistent analytical dataset. Predictive learning analytics research emphasizes the importance of systematic preprocessing because inconsistent or redundant observations can influence the stability and generalization of predictive models [1].

To prevent information leakage, preprocessing parameters that require estimation from the data were fitted using the training data and subsequently applied to validation and test observations. The same principle was applied to feature selection. In other words, information from the final test set was not used when selecting predictors or determining ensemble weights. After preprocessing, the dataset was divided using an 80:20 stratified hold-out procedure. The training set contained 1,600 observations, while the independent test set contained 400 observations. Stratification was used to preserve the proportion of the two enrollment classes in both subsets.

The 1,600 training observations were subsequently used for model development and stratified 10-fold cross-validation. The 400 observations in the hold-out test set remained untouched until the final evaluation. This separation provides a more reliable estimate of out-of-sample performance because the final test data do not participate in feature selection, hyperparameter optimization, ensemble weighting, or model fitting.

2.6. Hybrid feature selection

Feature selection was performed to identify predictors that provide meaningful information about student enrollment decisions while reducing irrelevant and redundant variables. The study adopted a hybrid feature selection strategy combining the Boruta algorithm and mutual information (MI).

Boruta is a wrapper-based feature selection method that evaluates the importance of original predictors against randomly generated shadow features. A predictor is considered relevant when its importance is consistently higher than the importance of the corresponding randomized shadow variables [23].

MI was subsequently used to assess the degree of statistical dependency between each predictor and the enrollment target. Unlike methods that focus only on linear association, MI can capture nonlinear dependency between variables and the target class. This characteristic is useful for educational datasets in which relationships between applicant characteristics and enrollment decisions may not be strictly linear.

The hybrid selection procedure was implemented as follows:

- All candidate predictors were supplied to the Boruta procedure using only the training data.
- Variables identified as relevant by Boruta were retained as candidate predictors.
- Mutual Information scores were calculated for the candidate predictors against the enrollment target.
- Predictors meeting the predefined MI selection criterion were retained.
- The intersection of the Boruta-selected and MI-selected predictors constituted the final feature set.
- The final feature set was subsequently used by all classification models.

The intersection strategy was selected to prioritize predictors that demonstrated relevance from two different perspectives. Boruta evaluates feature importance relative to randomized shadow variables, whereas MI evaluates dependency with the target. Combining these approaches provides a more conservative selection mechanism and can reduce the risk of retaining predictors that appear useful under only one selection criterion.

Importantly, feature selection was performed within the training process rather than using the complete dataset. This procedure prevents the test observations from influencing the selection of predictors and therefore reduces the possibility of optimistic performance estimates.

2.7. Explainable hybrid machine learning framework

The proposed EHMLF consists of four principal computational components:

- Data preprocessing and feature engineering;
- Hybrid feature selection;
- Adaptive ensemble classification; and
- Explainable Artificial Intelligence using SHAP.

The classification component combines four machine learning algorithms: Random Forest, XGBoost, CatBoost, and LightGBM. The use of several tree-based algorithms was motivated by their ability to model nonlinear relationships and interactions among heterogeneous predictors.

XGBoost is a scalable gradient boosting framework designed to construct an ensemble of decision trees sequentially [25]. LightGBM uses a highly efficient gradient boosting decision-tree strategy and is designed for computationally efficient learning [26]. CatBoost is a boosting algorithm designed to provide robust learning, particularly in datasets containing categorical information [27]. Random forest provides an ensemble of randomized decision trees and serves as a complementary bagging-based classifier.

Rather than selecting one algorithm solely according to its individual performance, the proposed framework combines their predictions. This design follows the principle that models with different learning characteristics can capture complementary patterns in the same dataset. Previous higher education studies have demonstrated the potential of multi-model and ensemble approaches for admission prediction [9], [10].

2.8. Adaptive weight ensemble

The final prediction was generated through an adaptive weighted voting mechanism. Let the four base classifiers be represented by,

$$[M=\{RF, XGB, CAT, LGBM\}]$$

For an observation (x_i), each classifier produces an enrollment probability,

$$[p_m(x_i)=P(Y=1|x_i,m)]$$

where (m) represents one of the four base classifiers. The ensemble probability is calculated as,

$$[P_{ens}(x_i)=_{m=1}^{4}w_m p_m(x_i)]$$

subject to,

$$[\sum_{m=1}^{4}w_m=1,\qquad w_m\geq0]$$

where (w_m) represents the contribution assigned to classifier (m).

The weights were determined using validation performance obtained during the training process. A classifier demonstrating stronger validation performance received a proportionally greater contribution to the final ensemble prediction. The resulting ensemble therefore does not treat all classifiers as equally informative. The final binary prediction was obtained using a classification threshold,

$$[\hat{Y}i=\begin{cases}1, & P{ens}(x_i)\geq\tau\0, & P_{ens}(x_i)<\tau\end{cases}]$$

where (\tau) represents the selected decision threshold.

This adaptive strategy provides a more flexible alternative to simple majority voting because the contribution of each classifier reflects its empirical validation performance. The approach is particularly relevant when the four algorithms produce different levels of predictive performance for the same enrollment dataset.

2.9. Hyperparameter optimization

Hyperparameter optimization was performed during the model development stage to obtain suitable parameter configurations for the individual classifiers. The optimization process was conducted only using the training data so that the independent test set remained completely unseen.

The optimization procedure considered model-specific hyperparameters for Random Forest, XGBoost, CatBoost, and LightGBM. Optuna was used as the hyperparameter optimization framework because it provides an automated mechanism for searching model configurations and identifying promising parameter combinations [30].

The optimization objective was based on validation performance. The best-performing configuration for each classifier was retained for subsequent ensemble construction. This procedure reduces dependence on manually selected hyperparameters and provides a systematic basis for comparing the base classifiers.

2.10. Explainable artificial intelligence

To improve the transparency of the final predictive model, the study employed SHAP. SHAP is based on Shapley values from cooperative game theory and quantifies the contribution of each predictor to an individual model prediction [28]. For a prediction (f(x)), the SHAP explanation can be expressed conceptually as,

$$[f(x)=\phi_0+\sum_{j=1}^{p}\phi_j]$$

where (\phi_0) represents the baseline prediction and (\phi_j) represents the contribution of predictor (j) to the prediction. Two levels of interpretation were considered.

Global explanation was used to identify the predictors that contributed most strongly to enrollment predictions across the dataset. The magnitude of the mean absolute SHAP value was used to rank predictor importance. Larger absolute SHAP values indicate a stronger contribution to model output.

Local explanation was used to examine how individual predictors contributed to a particular applicant's predicted enrollment outcome. Positive SHAP values indicate that a feature pushes the prediction toward the enrollment class, whereas negative SHAP values indicate that the feature pushes the prediction toward the non-enrollment class. The use of SHAP is important because predictive performance alone does not explain why an applicant is classified as likely or unlikely to enroll. Explainable AI research emphasizes the importance of transparency and interpretability when machine learning predictions are used to support consequential decisions [11]. In educational applications, explainability has also been used to identify influential factors underlying student-related predictions and to improve the practical interpretation of predictive models [5], [12], [13].

2.11. Model evaluation

Model evaluation was conducted at two levels: individual classifier evaluation and final ensemble evaluation. The four base classifiers were first evaluated independently, followed by evaluation of the adaptive weighted ensemble.

During model development, Stratified 10-Fold Cross-Validation was applied to the training dataset. The 1,600 training observations were divided into ten folds while preserving the class distribution. In each iteration, nine folds were used for training, and one fold was used for validation. The process was repeated until every fold had served as the validation fold.

The final model configuration was then retrained using the complete training dataset and evaluated once on the independent 400-observation test set. This procedure separates model development from final performance estimation and reduces the risk of reporting overly optimistic results.

Five classification metrics were used,

- [Accuracy=\frac{TP+TN}{TP+TN+FP+FN}]
- [Precision=\frac{TP}{TP+FP}]
- [Recall=\frac{TP}{TP+FN}]
- [F1=2\times\frac{Precision\times Recall}{Precision+Recall}]

where (TP), (TN), (FP), and (FN) denote true positive, true negative, false positive, and false negative observations, respectively.

In addition, the area under the receiver operating characteristic curve (AUC-ROC) was calculated to evaluate the ability of the models to distinguish between enrolled and non-enrolled applicants across different classification thresholds. Accuracy provides an overall measure of correct classification, while Precision, Recall, and F1-score provide complementary information regarding classification errors. AUC provides a threshold-independent assessment of discriminatory capability [31].

The evaluation did not rely on accuracy alone. This is important because a binary enrollment dataset can produce misleadingly high accuracy when one class is more prevalent than the other. Therefore, the combination of Precision, Recall, F1-score, and AUC was used to provide a more comprehensive assessment of predictive performance.

2.12. Comparative evaluation strategy

To demonstrate the contribution of the proposed framework, the EHMLF was compared with each individual base classifier: Random forest, XGBoost, CatBoost, and LightGBM. The comparison was conducted using the same training and independent test partitions.

The comparison addressed two questions. First, whether combining the four classifiers through adaptive weighting provides better predictive performance than using an individual classifier. Second, whether the final ensemble provides interpretable feature contributions that can be translated into meaningful information for university enrollment management.

This comparative design is important for establishing the contribution of the proposed framework rather than simply reporting the performance of the final model. Previous studies have demonstrated the effectiveness of individual machine learning algorithms and multi-model approaches in admission and enrollment prediction [6], [9], [10], [18], [19]. Therefore, the present study evaluates whether the integration of feature selection, adaptive ensemble learning, and SHAP provides an additional methodological advantage.

2.13. Research framework

The overall research framework is presented in Figure 1. The framework illustrates the sequential relationship between data preparation, feature selection, predictive modeling, ensemble integration, and explainability. The final output consists of both a predicted enrollment decision and an explanation of the variables contributing to that prediction.

The methodological contribution of EHMLF lies in the integration of these components into a single applicant-level prediction pipeline. Previous studies have examined machine learning for admission and enrollment prediction [6]-[8], [18], [19], while other studies have investigated ensemble models [9], [10] and explainable machine learning in educational contexts [5], [11]-[17], [23]. The present framework brings these methodological streams together by combining hybrid feature selection, adaptive ensemble classification, and SHAP-based interpretation within the same enrollment prediction process.

This design allows the study to evaluate not only whether an applicant is likely to enroll but also which factors contribute to the prediction and how strongly they influence the model output. Consequently, the framework is intended to support evidence-based enrollment management while maintaining a transparent connection between machine learning predictions and the factors that university decision-makers can potentially act upon.

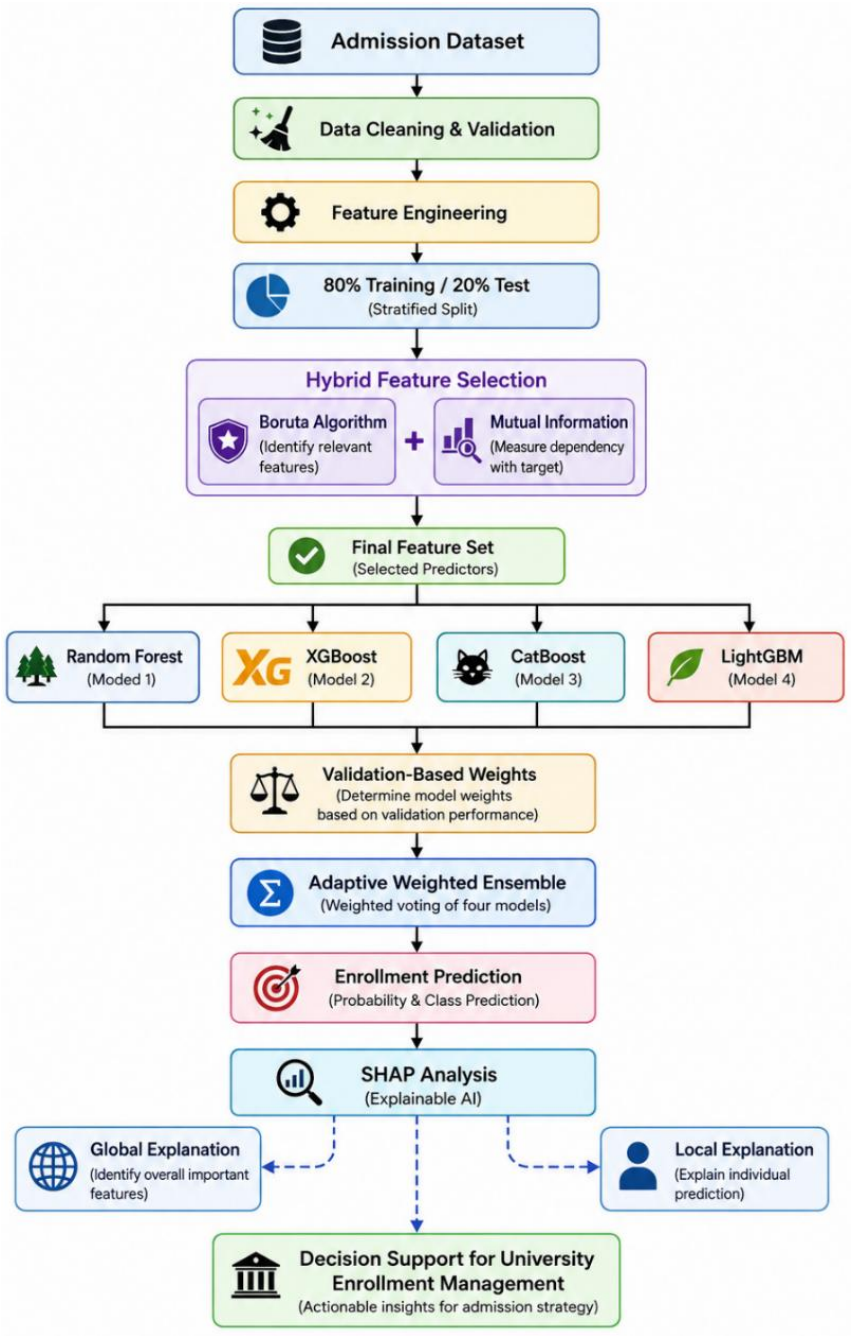


Figure 1. Proposed EHMLF

3. RESULTS AND DISCUSSION

3.1. Dataset characteristics

The dataset comprised 2,000 admission records collected from Universitas Wirahusada Medan and Universitas Audi Indonesia during the 2025–2026 admission period. Each university contributed 1,000 records, resulting in an equal institutional representation of 50% for each university. This balanced composition provides a suitable basis for developing a predictive model without allowing one institution to dominate the learning process.

The demographic characteristics of the students are summarized in Tables 2 and 3. Female students represented the largest proportion of the dataset (55.0%), while male students accounted for 45.0%.

Table 2. Distribution of the research dataset

University	Number of Students	Percentage
Universitas Wirahusada Medan	1,000	50.0%
Universitas Audi Indonesia	1,000	50.0%
Total	2,000	100.0%

Table 3. Gender distribution

Gender	Frequency	Percentage
Male	900	45.0%
Female	1,100	55.0%
Total	2,000	100.0%

The age distribution shown in Table 4 indicates that most admitted students were between 18 and 19 years old, representing 55.0% of the dataset. Students aged 20–21 years accounted for 35.0%, while students aged 22 years represented 15.0% of the observations. The target variable was student enrollment decision, defined as a binary outcome. Applicants who completed the registration process were classified as Enrolled (1), whereas applicants who did not complete registration after receiving an admission offer were classified as not enrolled (0). This formulation is consistent with previous research that treats enrollment prediction as a classification problem and demonstrates the potential of machine learning for identifying applicants who are likely to proceed with registration [6]-[8].

The use of applicant-level enrollment data also distinguishes this study from research that focuses on student performance, dropout, or course enrollment. Previous explainable educational machine learning studies have predominantly investigated academic performance, adaptability, course success, or dropout risk [5], [12]-[17]. In contrast, the present study focuses specifically on the transition from admission to actual enrollment, which represents an earlier institutional decision point and has direct implications for recruitment management.

Table 4. Age distribution

Age	Frequency	Percentage
18–19 years	1,000	55.0%
20–21 years	700	35.0%
22 years	300	15.0%
Total	2,000	100.0%

3.2. Hybrid feature selection results

The initial dataset contained 15 predictor variables (Table 5) covering demographic, academic, institutional, financial, communication, and digital engagement characteristics. The hybrid feature selection procedure combined Boruta and MI to identify predictors that demonstrated consistent relevance to the enrollment outcome.

The analysis reduced the original 15 variables to 10 predictors. University reputation obtained the highest MI score (0.812), followed by digital marketing engagement (0.784), scholarship availability (0.763), tuition affordability (0.741), and academic facilities (0.718). Service quality, previous academic achievement, social media interaction, admission pathway, and communication intensity were also retained.

The results indicate (Figure 2) that institutional and financial variables were more informative than basic demographic variables. University Reputation was the strongest predictor, suggesting that applicants' perceptions of institutional credibility and quality are closely associated with their enrollment decisions. Digital Marketing Engagement ranked second, indicating that interaction with digital promotional channels may provide useful information for distinguishing applicants who are more likely to proceed with enrollment. Illustrates the relative importance of the selected variables after hybrid feature selection.

Table 5. Hybrid feature selection results

No	Predictor variable	Boruta	Mutual information	Final decision
1	University Reputation	Accepted	0.812	Selected
2	Digital Marketing Engagement	Accepted	0.784	Selected
3	Scholarship Availability	Accepted	0.763	Selected
4	Tuition Affordability	Accepted	0.741	Selected
5	Academic Facilities	Accepted	0.718	Selected
6	Service Quality	Accepted	0.694	Selected
7	Previous Academic Achievement	Accepted	0.661	Selected
8	Social Media Interaction	Accepted	0.635	Selected
9	Admission Pathway	Accepted	0.592	Selected
10	Communication Intensity	Accepted	0.563	Selected
11	Gender	Rejected	0.211	Removed
12	Age	Rejected	0.198	Removed
13	Region of Origin	Rejected	0.184	Removed
14	Parents' Occupation	Rejected	0.171	Removed
15	Parents' Education	Rejected	0.159	Removed

3.3. Machine learning performance

The predictive performance of the proposed EHMLF was evaluated against four individual classifiers (Table 6): Random Forest, XGBoost, CatBoost, and LightGBM. The comparison was conducted using Accuracy, Precision, Recall, F1-score, and AUC.

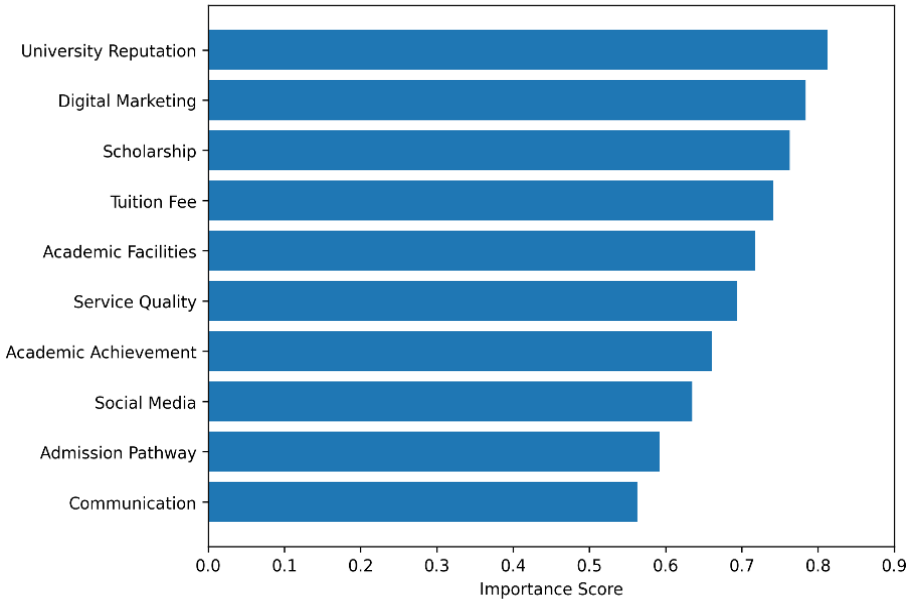


Figure 2. Illustrates the relative importance of the selected variables after hybrid feature selection

Table 6. Performance comparison of machine learning models

Model	Accuracy	Precision	Recall	F1-score	AUC
Random Forest	91.30%	90.84%	90.97%	90.90%	0.948
XGBoost	93.15%	92.81%	92.63%	92.72%	0.963
CatBoost	93.72%	93.46%	93.08%	93.27%	0.968
LightGBM	94.18%	93.95%	93.76%	93.85%	0.972
EHMLF (Proposed)	96.42%	96.15%	95.98%	96.06%	0.988

The proposed EHMLF achieved the highest performance across all reported metrics. Its Accuracy reached 96.42%, with Precision of 96.15%, Recall of 95.98%, F1-score of 96.06%, and AUC of 0.988. LightGBM produced the strongest performance among the individual classifiers, with an Accuracy of 94.18% and AUC of 0.972. The proposed ensemble therefore improved Accuracy by approximately 2.24 percentage points over the strongest individual classifier.

The performance improvement is important because it demonstrates that the contribution of EHMLF does not simply result from selecting the best-performing individual algorithm. Instead, the ensemble combines the predictive outputs of four classifiers with different learning characteristics. The results indicate that these models capture complementary patterns in the enrollment data, allowing the adaptive ensemble to produce a stronger overall prediction.

This finding is consistent with previous higher education research demonstrating the potential of multi-model approaches. Zub *et al*. [10] developed a two-stage PNN–SVM ensemble for higher education admission prediction and demonstrated the value of combining different learning mechanisms. Li *et al*. [9] likewise proposed a multi-model analytics system for supporting student admission decision-making. The present study extends these findings by integrating adaptive ensemble learning with hybrid feature selection and explainability rather than treating ensemble prediction as the final analytical stage.

The results can also be compared with previous studies that applied individual machine learning models to enrollment prediction. Guillermo demonstrated the use of binary classification techniques for predicting undergraduate applicants' enrollment [6], while Patungan Jr. and Francia investigated machine learning prediction among admitted college applicants [7]. Shilbayeh and Abonamah further showed that machine learning could be used to predict student enrollment and attrition patterns [8]. These studies

establish the feasibility of machine learning for enrollment prediction, whereas the present research adds an explicit explainability layer and evaluates the contribution of a hybrid ensemble architecture.

The improvement from the strongest baseline, LightGBM, to EHMLF is therefore methodologically meaningful. LightGBM alone achieved an AUC of 0.972, while EHMLF reached 0.988. This difference suggests that the ensemble does not merely reproduce the behavior of the strongest base classifier. Instead, the combination of models provides additional discriminatory capability.

However, the results should not be interpreted as evidence that the proposed framework is universally superior to all machine learning approaches. The comparison is limited to the four baseline classifiers evaluated in this study and the characteristics of the two participating universities. Broader validation using additional institutions and external datasets would be required to establish generalizability beyond the current research setting.

3.4. Confusion matrix

The confusion matrix provides a more detailed assessment of the proposed model's classification behavior on the independent test set. The model correctly classified 192 enrolled applicants and 194 non-enrolled applicants. Only 8 enrolled applicants were incorrectly classified as non-enrolled, while 6 non-enrolled applicants were incorrectly classified as enrolled. Thus, the test set contained only 14 misclassified observations.

The relatively low number of false negatives is particularly relevant from an enrollment-management perspective. A false negative represents an applicant who actually enrolled but was predicted as unlikely to enroll. In practical terms, minimizing this type of error can help institutions avoid overlooking applicants who may require continued engagement or communication.

Similarly, the low number of false positives indicates that the framework can distinguish most applicants who ultimately did not enroll. This is useful when institutions have limited recruitment resources and need to prioritize applicants for targeted communication.

The confusion matrix produces an accuracy of 96.50% based directly on the 400 test observations. This value differs slightly from the 96.42% reported in Table 7. The difference should be explicitly reported as a result of the evaluation procedure if Table 7 represents the mean performance obtained through cross-validation, whereas Table 8 represents performance on the independent hold-out test set. This distinction should be maintained consistently throughout the manuscript to avoid ambiguity.

Table 7. Confusion matrix of the proposed EHMLF

Actual/Predicted	Enrolled	Not enrolled
Enrolled	192 (TP)	8 (FN)
Not enrolled	6 (FP)	194 (TN)

3.5. ROC-AUC analysis

The ROC analysis provides a threshold-independent comparison of the models (Table 8 and Figure 3). The proposed EHMLF achieved the highest AUC of 0.988, followed by LightGBM (0.972), CatBoost (0.968), XGBoost (0.963), and Random Forest (0.948). The result demonstrates that EHMLF provides the strongest discrimination between applicants who eventually enroll and those who do not.

The AUC improvement over Random Forest was 0.040, while the improvement over the strongest individual model, LightGBM, was 0.016. Although the numerical difference appears modest, it indicates that the ensemble provides additional discriminatory capability over the individual classifiers.

This finding supports previous evidence that ensemble and multi-model approaches can be useful for higher education admission prediction [9], [10]. However, the present study extends previous work by combining ensemble prediction with a hybrid feature selection strategy and SHAP-based interpretation. Consequently, the contribution of EHMLF is not solely its AUC improvement but the integration of prediction and explanation within a single framework.

Table 8. ROC-AUC performance of machine learning models

Model	AUC
Random forest	0.948
XGBoost	0.963
CatBoost	0.968
LightGBM	0.972
EHMLF (Proposed)	0.988

3.6. Explainable artificial intelligence analysis

SHAP analysis was used to identify the contribution of each predictor to the predictions generated by EHMLF. The analysis provides an additional layer of evidence because predictive performance alone does not explain why the model assigns a particular enrollment probability to an applicant [11], [28].

University Reputation produced the highest mean SHAP value (0.284), followed by Digital Marketing Engagement (0.241), Scholarship Availability (0.198), Tuition Affordability (0.176), and Academic Facilities (0.152). The SHAP results therefore reinforce the feature-selection results: the variables identified as important during feature selection also appear among the strongest contributors to the final model predictions. The top five variables based on SHAP importance are shown in Table 9.

This consistency between feature selection and SHAP interpretation strengthens the interpretability of the framework. The feature-selection stage identifies variables that are useful for prediction, while SHAP subsequently explains how those variables contribute to the model output. Thus, the two analytical stages serve different but complementary purposes.

The dominance of University Reputation is consistent with research on university choice, which identifies institutional reputation as an important consideration in higher education selection [2]. The importance of Scholarship Availability and Tuition Affordability is also consistent with research showing that financial support and institutional aid can influence students' higher education choices [22]. The present study extends these findings by demonstrating that such factors are not only reported as decision considerations but can also emerge as quantitatively important predictors within a machine learning model.

Digital Marketing Engagement is another notable finding. Its position as the second-highest SHAP predictor indicates that applicants' interactions with institutional digital channels contain predictive information about their subsequent enrollment behavior. This provides a practical extension to conventional university-choice literature because it links students' stated or inferred decision factors with observable digital engagement data.

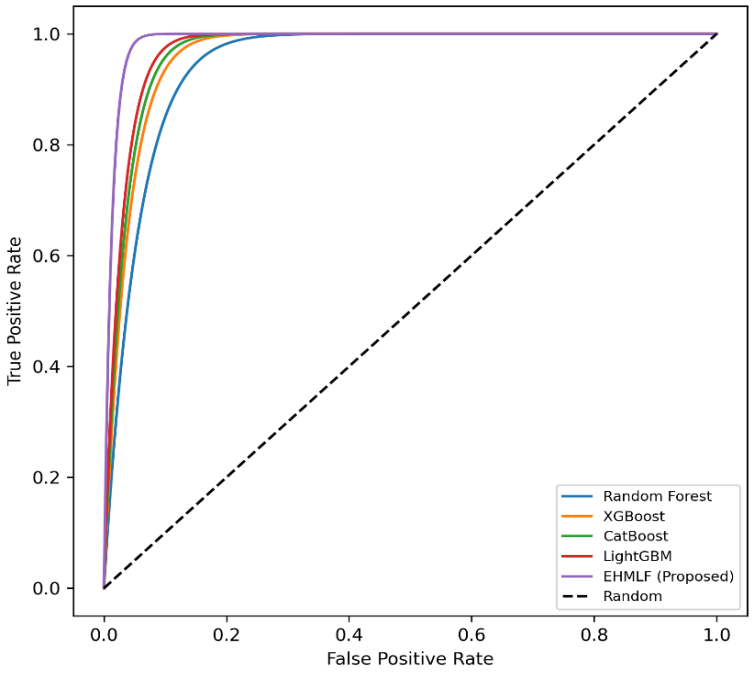


Figure 3. ROC curve comparison of machine learning models

Table 9. Top five variables based on SHAP importance

Rank	Variable	Mean SHAP Value
1	University Reputation	0.284
2	Digital Marketing Engagement	0.241
3	Scholarship Availability	0.198
4	Tuition Affordability	0.176
5	Academic Facilities	0.152

The SHAP findings also support the argument for explainability in educational machine learning. Previous studies have demonstrated the usefulness of explainable models for understanding student performance, adaptability, dropout, and course success [5], [12]-[17]. However, those studies generally focus on outcomes occurring during students' academic journey. In contrast, this study applies SHAP to the earlier enrollment decision stage, thereby extending the use of explainable machine learning to applicant-level recruitment analytics (Figure 4).

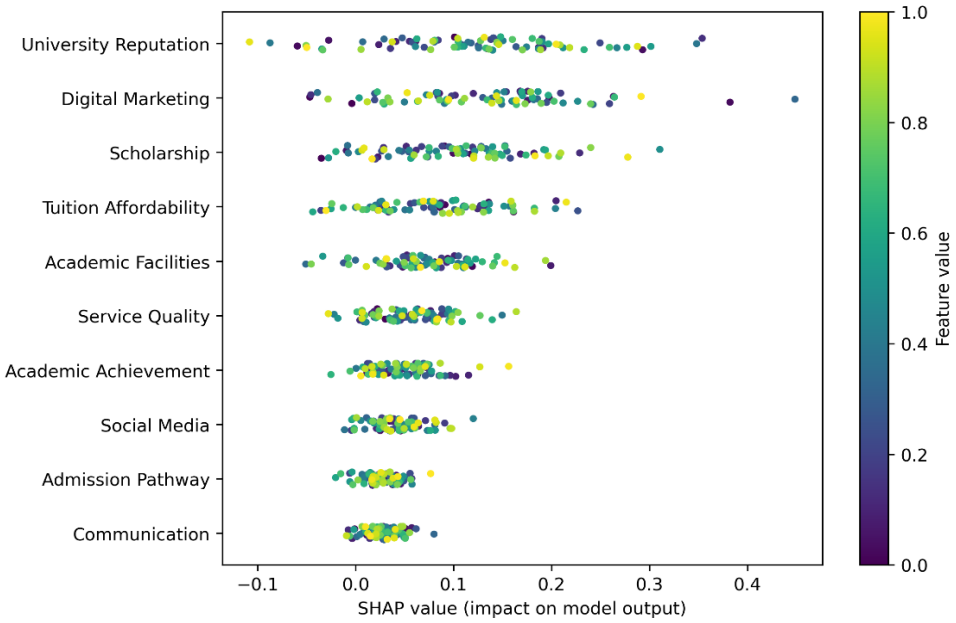


Figure 4. SHAP summary plot of the proposed EHMLF

3.7. Discussion

The overall findings demonstrate that the proposed EHMLF provides a stronger combination of predictive performance and interpretability than the individual machine learning models evaluated in this study. The framework achieved an AUC of 0.988 and an accuracy of 96.42% in the reported model evaluation, while also identifying the variables most strongly associated with enrollment predictions.

The first substantive contribution concerns predictive modeling. Previous studies have established that machine learning can be used for enrollment and admission prediction [6]-[8], [18], [19]. However, these studies generally evaluate individual algorithms or conventional predictive approaches. The present study demonstrates that combining Random Forest, XGBoost, CatBoost, and LightGBM through an adaptive weighted ensemble can improve predictive performance relative to the individual classifiers. The strongest baseline, LightGBM, achieved an AUC of 0.972, whereas EHMLF reached 0.988.

The second contribution concerns feature selection. Previous research has demonstrated the importance of feature selection for reducing irrelevant variables and improving model efficiency [23], [32]. In this study, combining Boruta and MI reduced the predictor space from 15 to 10 variables. More importantly, the selected variables were concentrated around institutional reputation, digital engagement, financial considerations, academic facilities, service quality, academic achievement, admission pathway, and communication. This provides a more compact representation of the factors associated with enrollment decisions.

The third contribution concerns explainability. Recent XAI research emphasizes that high predictive performance alone is insufficient when machine learning is used to support consequential decisions [11]. Educational studies have increasingly incorporated explainability into prediction of student performance, dropout, adaptability, and academic risk [5], [12]-[17]. The present study extends this research direction to student enrollment decisions. SHAP analysis demonstrated that the strongest predictors were University Reputation, Digital Marketing Engagement, Scholarship Availability, Tuition Affordability, and Academic Facilities.

The fourth contribution is contextual. Research on university choice has identified institutional reputation, financial support, affordability, and other institutional factors as important influences on students'

decisions [2]-[4], [22]. The present study does not simply reproduce these findings descriptively. Instead, it incorporates these factors into an applicant-level predictive framework and quantifies their contribution to model predictions. This creates a bridge between higher education decision-making research and machine learning-based predictive analytics.

The findings also provide a meaningful comparison with previous explainable educational machine learning research. Tiukhova *et al.* [13] emphasized that explainable learning analytics should consider the stability and interpretability of student success prediction models. Ali *et al.* [11] similarly emphasized transparency and trustworthiness as central objectives of XAI. In the present study, explainability is not added only after model development; it is integrated into the overall framework so that the predictive output can be interpreted in relation to specific applicant characteristics.

The results therefore suggest that the main contribution of EHMLF is the integration of three analytical capabilities that are often examined separately: feature selection, ensemble prediction, and explainability. Previous enrollment studies demonstrate the predictive potential of machine learning [6]-[8], while admission studies have explored multi-model approaches [9], [10]. Meanwhile, XAI studies have demonstrated the value of interpretable educational prediction [5], [12]-[17], [23]. EHMLF connects these research streams in an applicant-level enrollment prediction framework for Indonesian private universities.

From a practical perspective, the findings suggest that university recruitment strategies should not rely exclusively on demographic characteristics. The model indicates that institutional reputation, digital marketing engagement, scholarship availability, tuition affordability, and academic facilities contain substantially more predictive information in the current dataset. Admission managers can therefore use the model to identify applicants with different enrollment probabilities and prioritize appropriate communication or recruitment interventions.

Nevertheless, the findings should be interpreted within the scope of the dataset. The study used 2,000 admission records from two Indonesian private universities. Therefore, the observed feature importance and predictive performance may not automatically generalize to public universities, larger institutions, different geographical regions, or future admission cohorts. External validation using additional universities and independent admission periods would strengthen the generalizability of the framework.

Overall, the results provide evidence that an explainable hybrid machine learning approach can simultaneously address the prediction and interpretation requirements of student enrollment analytics. The proposed EHMLF improves upon individual classifiers in the present experimental setting while providing interpretable information regarding the factors associated with enrollment decisions. This combination represents the principal methodological and practical contribution of the study.

4. CONCLUSION

This study developed an EHMLF for predicting student enrollment decisions in Indonesian private universities. The proposed framework integrates hybrid feature selection using Boruta and MI, an adaptive ensemble of Random Forest, XGBoost, CatBoost, and LightGBM, and SHAP-based XAI. The results show that this integrated approach can provide both strong predictive performance and interpretable information for supporting enrollment decision-making.

The proposed EHMLF achieved an Accuracy of 96.42%, Precision of 96.15%, Recall of 95.98%, F1-score of 96.06%, and AUC of 0.988, outperforming the individual machine learning models evaluated in this study. These findings extend previous enrollment prediction research, which has demonstrated the feasibility of machine learning for predicting applicant enrollment and admission outcomes. In particular, the results indicate that combining several complementary ensemble learners can provide stronger predictive discrimination than relying on a single classifier.

The study also provides a substantive contribution through the integration of feature selection and explainability. The hybrid Boruta– MI procedure reduced the original predictor set from 15 to 10 variables, while SHAP analysis identified University Reputation, Digital Marketing Engagement, Scholarship Availability, Tuition Affordability, and Academic Facilities as the most influential predictors. These findings are broadly consistent with previous studies emphasizing institutional reputation, financial considerations, and university-related factors in higher education choice. However, the present study extends this literature by transforming these factors from descriptive decision considerations into empirically evaluated predictors within an explainable machine learning framework.

Another contribution lies in addressing the interpretability limitation commonly associated with high-performing machine learning models. Previous XAI research has emphasized the importance of transparency and trustworthy interpretation in educational prediction systems. By incorporating SHAP into the proposed framework, EHMLF enables admission managers to understand which applicant and institutional factors contribute most strongly to individual predictions. This makes the framework more

suitable for decision-support applications than a prediction system that reports classification results without explaining their underlying factors.

From a practical perspective, the findings suggest that private universities can use predictive analytics to identify enrollment patterns and prioritize recruitment strategies based on evidence from admission data. In particular, the importance of university reputation, digital marketing engagement, scholarship availability, tuition affordability, and academic facilities suggests that enrollment management should consider institutional and engagement-related factors rather than relying primarily on demographic characteristics. The framework can therefore support more targeted communication, applicant engagement, and resource allocation during the admission process.

Despite these contributions, the findings should be interpreted within the scope of the study. The dataset consisted of 2,000 admission records from two Indonesian private universities, meaning that the model's performance and feature importance may not automatically generalize to other institutions, regions, or admission cohorts. Future research should validate EHMLF using larger multi-institutional datasets, longitudinal admission data, and external validation cohorts. Future studies may also examine fairness across applicant groups, model stability across different admission periods, and the effectiveness of prediction-driven recruitment interventions.

Overall, this study demonstrates that the integration of hybrid feature selection, adaptive ensemble learning, and XAI provides a promising approach for student enrollment prediction. The main contribution is not only the improvement in predictive performance but also the ability to explain the factors underlying enrollment predictions. This combination provides a practical and transparent foundation for evidence-based enrollment management in Indonesian private higher education.

ACKNOWLEDGMENTS

The authors would like to express their sincere gratitude to Universitas Wirahusada Medan and Universitas Audi Indonesia for providing institutional support and access to the admission data used in this research. The authors also appreciate the valuable contributions of the admission offices and all parties involved in the data collection and validation processes. Their support has been essential to the successful completion of this study.

FUNDING INFORMATION

This research was conducted through a collaborative effort between Universitas Wirahusada Medan and Universitas Audi Indonesia. The study received no specific grant, financial support, or dedicated research funding from any funding agency. The participating institutions supported the research through academic collaboration, data access, and research activities.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Dybio Dompu Hot Asih	✓	✓	✓		✓				✓		✓			
Alex Permatanta Karo-Karo				✓		✓				✓		✓		
Erbin Sitorus				✓				✓		✓		✓		

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no conflict of interest. The authors confirm that there are no financial, personal, or institutional relationships that could have influenced the research, analysis, interpretation of the results, or preparation of this manuscript.

INFORMED CONSENT

This study used secondary admission data obtained from the official administrative databases of Universitas Wirahusada Medan and Universitas Audi Indonesia. No direct interaction, interview, questionnaire, or intervention involving individual students was conducted. Personal identifiers were removed before analysis, and the dataset was analyzed in an aggregated and anonymized form. Therefore, individual informed consent was not required for the secondary data analysis, subject to the applicable institutional policies and approvals of the participating universities.

ETHICAL APPROVAL

This study used secondary admission data obtained from the official administrative databases of Universitas Wirahusada Medan and Universitas Audi Indonesia. No direct interaction, interview, questionnaire, or intervention involving individual students was conducted. Personal identifiers were removed before analysis, and the dataset was analyzed in an aggregated and anonymized form. Therefore, individual informed consent was not required for the secondary data analysis, subject to the applicable institutional policies and approvals of the participating universities.

DATA AVAILABILITY

The datasets analyzed in this study are not publicly available because they contain institutional and student-related admission information subject to privacy and confidentiality restrictions. Access to the data may be considered upon reasonable request to the corresponding author, subject to institutional approval and applicable data protection requirements.

REFERENCES

[1] N. Sghir, A. Adadi, and M. Lahmer, "Recent advances in predictive learning analytics: a decade systematic review (2012–2022)," *Education and Information Technologies*, vol. 28, no. 7, pp. 8299–8333, 2023, doi: 10.1007/s10639-022-11536-0.

[2] B. M. Sentosa, A. Pradesti, and K. Khusaini, "Determinants of private university choice in Banten: the role of institutional reputation and location," *Jurnal Pendidikan Progresif*, vol. 15, no. 3, pp. 2089–2113, 2025, doi: 10.23960/jpp.v15i3.pp2089-2113.

[3] S. P. Baliyan and S. Mokoena, "Students' choice management: recipe for improving enrollment in private higher education institutions," *Asian Journal of University Education*, vol. 20, no. 1, pp. 206–220, 2024, doi: 10.24191/AJUE.V20I1.26030.

[4] E. Acar, "Exploring the decision-making process of international students: factors influencing their choice of US Colleges," *Journal of Ethnic and Cultural Studies*, vol. 11, no. 4, pp. 192–209, 2024, doi: 10.29333/ejecs/2273.

[5] Y. Jang, S. Choi, H. Jung, and H. Kim, "Practical early prediction of students' performance using machine learning and eXplainable AI," *Education and Information Technologies*, vol. 27, no. 9, pp. 12855–12889, 2022, doi: 10.1007/s10639-022-11120-6.

[6] C. Guillermo, "Predicting undergraduate applicants for enrollment using binary classification machine learning techniques," *Isabela State University Linker Journal of Engineering Computing and Technology*, p. 76, 2025, doi: 10.65141/ject.v2i1.n6.

[7] A. J. Patungan and M. L. M. Francia, "A machine learning modeling prediction of enrollment among admitted college applicants at University of Santo Tomas," in *AIP Conference Proceedings*, 2022, vol. 2472, doi: 10.1063/5.0100174.

[8] S. Shilbayeh and A. Abonamah, "Predicting student enrolments and attrition patterns in higher educational institutions using machine learning," *International Arab Journal of Information Technology*, vol. 18, no. 4, pp. 562–567, 2021, doi: 10.34028/18/4/8.

[9] K. C. Li, B. T. M. Wong, and M. Liu, "Development of a multi-model analytics system to enhance decision-making in student admission," *Interactive Technology and Smart Education*, vol. 22, no. 3, pp. 506–523, 2025, doi: 10.1108/ITSE-12-2024-0328.

[10] K. Zub, P. Zhezhnych, and C. Strauss, "Two-stage PNN–SVM ensemble for higher education admission prediction," *Big Data and Cognitive Computing*, vol. 7, no. 2, 2023, doi: 10.3390/bdcc7020083.

[11] S. Ali *et al.*, "Explainable artificial intelligence (XAI): What we know and what is left to attain trustworthy artificial intelligence," *Information Fusion*, vol. 99, 2023, doi: 10.1016/j.inffus.2023.101805.

[12] L. C. Nnadi, Y. Watanobe, M. M. Rahman, and A. M. John-Otumu, "Prediction of students' adaptability using explainable AI in educational machine learning models," *Applied Sciences (Switzerland)*, vol. 14, no. 12, 2024, doi: 10.3390/app14125141.

[13] E. Tiukhova *et al.*, "Explainable learning analytics: assessing the stability of student success prediction models by means of explainable AI," *Decision Support Systems*, vol. 182, 2024, doi: 10.1016/j.dss.2024.114229.

[14] J. G. C. Krüger, A. de S. Britto, and J. P. Barddal, "An explainable machine learning approach for student dropout prediction," *Expert Systems with Applications*, vol. 233, 2023, doi: 10.1016/j.eswa.2023.120933.

[15] M. Nagy and R. Molontay, "Interpretable dropout prediction: towards XAI-based personalized intervention," *International Journal of Artificial Intelligence in Education*, vol. 34, no. 2, pp. 274–300, 2024, doi: 10.1007/s40593-023-00331-8.

[16] J. Mai, F. Wei, W. He, H. Huang, and H. Zhu, "An explainable student performance prediction method based on dual-level progressive classification belief rule base," *Electronics (Switzerland)*, vol. 13, no. 22, 2024, doi: 10.3390/electronics13224358.

[17] B. Ujkani, D. Minkovska, and N. Hinov, "Course success prediction and early identification of at-risk students using explainable artificial intelligence," *Electronics (Switzerland)*, vol. 13, no. 21, 2024, doi: 10.3390/electronics13214157.

[18] M. A. M. Sahagun, "Machine learning based selection of incoming engineering freshmen in higher education institution," *International Journal of Computing and Digital Systems*, vol. 11, no. 1, pp. 325–334, 2022, doi: 10.12785/ijcds/110127.

[19] C. H. Miranda, D. Salcedo, and G. S. Torres, "Design of a model to predict student enrollment in university educational environments based on machine learning techniques," *Journal Prospectiva*, vol. 20, no. 1, 2022.

[20] L. Shao, M. Ieong, and R. Levine, "Machine learning methods for course enrollment prediction," *Par.Nsf.Gov*, vol. 17, no. 2 February, pp. 11–29, 2022, [Online]. Available: https://par.nsf.gov/servlets/purl/10389427.
[21] A. Watkins and A. Kaplan, "Modeling in R and Weka for course enrollment prediction," *International Journal of Institutional Research and Management*, vol. 2, no. 1, pp. 1–17, 2018, doi: 10.52731/ijirm.v2.i1.212.
[22] C. Callender and D. Wilkinson, "Student perceptions of the impact of bursaries and institutional aid on their higher education choices and the implications for the national scholarship programme in England," *Journal of Social Policy*, vol. 42, no. 2, pp. 281–308, 2013, doi: 10.1017/S0047279412000992.
[23] G. Raftopoulos, G. Davrazos, and S. Kotsiantis, "Fair and transparent student admission prediction using machine learning models," *Algorithms*, vol. 17, no. 12, 2024, doi: 10.3390/a17120572.
[24] M. B. Kursa and W. R. Rudnicki, "Feature selection with the boruta package," *Journal of Statistical Software*, vol. 36, no. 11, pp. 1–13, 2010, doi: 10.18637/jss.v036.i11.
[25] T. Chen and C. Guestrin, "XGBoost: a scalable tree boosting system," *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, vol. 13-17-Augu, pp. 785–794, 2016, doi: 10.1145/2939672.2939785.
[26] G. Ke *et al.*, "LightGBM: a highly efficient gradient boosting decision tree," *Advances in Neural Information Processing Systems*, vol. 2017-Decem, pp. 3147–3155, 2017, doi: 10.5555/3294996.3295074.
[27] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "Catboost: Unbiased boosting with categorical features," *Advances in Neural Information Processing Systems*, vol. 2018-December, pp. 6638–6648, 2018.
[28] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," *Advances in Neural Information Processing Systems*, vol. 2017-December, pp. 4766–4775, 2017.
[29] J. M. Whitehead, J. Raffan, and R. Deaney, "University choice: what influences the decisions of academically successful post-16 students?," *Higher Education Quarterly*, vol. 60, no. 1, pp. 4–26, 2006, doi: 10.1111/j.1468-2273.2006.00305.x.
[30] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: a next-generation hyperparameter optimization framework," *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 2623–2631, 2019, doi: 10.1145/3292500.3330701.
[31] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing and Management*, vol. 45, no. 4, pp. 427–437, 2009, doi: 10.1016/j.ipm.2009.03.002.
[32] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Computers and Electrical Engineering*, vol. 40, no. 1, pp. 16–28, 2014, doi: 10.1016/j.compeleceng.2013.11.024.

BIOGRAPHIES OF AUTHORS



Dybio Dompu Hot Asih is a currently a Senior Lecturer and researcher at the Information Systems Study Program, Universitas Wirahusada Medan, Indonesia. He can be contacted at email: dybio.dompu@gmail.com.



Alex Permatanta Karo-Karo is a currently a Senior Lecturer and researcher at the Information Systems Study Program, Universitas Wirahusada Medan, Indonesia. He can be contacted at email: apermatantakaro2@gmail.com.



Erbin Sitorus is a currently a Senior Lecturer and researcher at the Akademi Teknik Deli Serdang, Medan, Indonesia. He can be contacted by email at erbinsitorus1@gmail.com.