Calibrated Probabilistic Stacking with Linear Meta-Learning for Admission Outcome Prediction

Khrystyna Zub[1*], Oksana Mulesa[2]
[1] Department of Social Communication and Information Activities, Institute of the Humanities and Social Sciences, Lviv Polytechnic National University, Lviv, Ukraine
[2] Department of Physics, Mathematics, and Technologies, Faculty of Humanities and Natural Sciences, University of Presov, Presov, Slovakia
E-mail: khrystyna.v.zub@lpnu.ua, oksana.mulesa@unipo.sk
*Corresponding author

Keywords: Artificial intelligence, machine learning, ensemble learning, probabilistic stacking, meta-learning, calibrated probabilities, probabilistic classification, intelligent data analysis, predictive modeling, decision support systems, educational data mining, admission prediction, higher education analytics, classification.

Received: April 10, 2026

This paper presents a two-stage probabilistic stacking ensemble for admission outcome prediction based on calibrated heterogeneous classifiers and linear decision fusion in probability space. The proposed approach combines HistGradientBoosting, ExtraTrees, and RandomForest models to generate posterior class probabilities, while probability calibration is applied to improve the consistency and comparability of probabilistic estimates. To prevent information leakage, the meta-level training space is formed exclusively from calibrated out-of-fold probability representations. At the second stage, linear meta-models are used to aggregate probabilistic outputs and produce the final decision. The method was evaluated on a real-world dataset collected from the admission campaign of Lviv Polytechnic National University. Experimental studies using holdout validation and stratified cross-validation demonstrate that the proposed ensemble achieves high and stable predictive performance while preserving the quality of probabilistic estimates. In particular, the method reached F1-scores up to 0.990 and MCC values up to 0.979 on the holdout test set, together with low LogLoss values. Comparative analysis with baseline classifiers and standard stacking approaches confirms that calibrated probabilistic fusion improves both classification quality and the reliability of posterior probability estimates in practical decision-support tasks.

Povzetek: Članek predstavlja zanesljiv ansambelski pristop za napovedovanje izidov sprejema, ki z umerjenim združevanjem več klasifikacijskih modelov izboljša natančnost in kakovost verjetnostnih ocen.

1 Introduction

In the modern competitive academic environment, prospective students often face difficulties in realistically assessing their chances of admission due to heterogeneous selection criteria, program-specific requirements, and fluctuating admission thresholds across universities. At the same time, traditional approaches, such as paid educational consultancy services or simple rule-based online calculators, are often expensive, subjective, and insufficient for capturing complex relationships among applicant characteristics. Consequently, with the growing competition for university places, predicting admission outcomes has become an important decision-support task for prospective students. This creates an increasing demand for intelligent systems capable of supporting applicants in making informed decisions about their higher education opportunities [1], [2].

With the rapid development of machine learning (ML) and data analytics, recent studies have proposed various ML-based admission prediction models and decision-support systems that estimate the probability of acceptance using historical applicant data and academic performance indicators [3], [4]. Such models enable applicants to better evaluate their chances of enrollment and support more informed application strategies. In this context, probabilistic prediction models are particularly important, as they provide not only categorical decisions but also quantitative estimates of admission likelihood, allowing applicants to assess risk and compare alternative application strategies.

However, the complexity and heterogeneity of admission data often limit the predictive performance of individual models [5]. Admission datasets typically include multiple correlated attributes, non-linear relationships, and significant variability between institutions and programs. As a result, single classifiers may fail to capture all relevant patterns in the data. It is due to the use of ensemble learning approaches that combine multiple models to improve prediction accuracy and robustness [6], [7], [8].

Ensemble learning techniques integrate the outputs of several base classifiers in order to reduce prediction variance and enhance generalization performance [9], [10]. At the same time, reliable probabilistic prediction requires well-calibrated classifier outputs, since poorly calibrated probabilities may lead to misleading estimates of admission chances even when classification accuracy is high [11], [12]. In probabilistic ensemble frameworks, decision fusion in the probability space allows combining calibrated posterior probabilities from multiple classifiers, enabling more consistent and interpretable final predictions [13].

Improving the accuracy and reliability of data-driven predictive models remains a fundamental challenge in many applied domains [13].

Traditional ML methods do not always provide sufficient classification accuracy for practical decision-support applications. Therefore, hybrid ensemble methods have attracted increasing attention in recent years.

Despite the broad use of stacking architectures, several issues remain insufficiently addressed in applied probabilistic classification tasks [10]. In particular, standard stacking does not always ensure comparability of posterior estimates produced by heterogeneous base classifiers, while the use of complex meta-models may increase the risk of overfitting [14], [15]. In addition, relatively few studies explicitly analyze stacking schemes in which the secondary learning stage is constructed only from calibrated out-of-fold probabilities and the final decision is obtained by dedicated linear fusion in the probability space.

This study examines the role of calibrated probabilistic stacking in improving both predictive performance and probabilistic reliability in admission prediction tasks. Particular attention is paid to the role of probability calibration, out-of-fold probability generation, and constrained meta-learning within a unified ensemble framework.

The following research questions are considered in this study:

- Can calibrated posterior probability representations improve the consistency of probabilistic outputs generated by heterogeneous classifiers?
- Does the use of strict out-of-fold probability generation help reduce the risk of information leakage during meta-learning?
- Can linear aggregation in probability space provide stable generalization performance while controlling the effective complexity of the meta-model?
- Does the proposed probabilistic stacking framework improve predictive performance compared to conventional single models and standard stacking approaches?

To address these questions, a structurally constrained probabilistic stacking ensemble for admission outcome prediction is proposed. Its main idea is to construct the secondary feature space only from calibrated out-of-fold posterior probabilities produced by heterogeneous base classifiers and to perform the final decision fusion by means of a linear meta-model. Such an organization is aimed at improving the consistency of probabilistic outputs, reducing the effective complexity of the meta-level model, and providing a more reliable decision rule for admission prediction.

Compared with standard stacking, the proposed approach differs in three aspects:

- only calibrated posterior probabilities are transferred to the meta-level;
- only out-of-fold estimates are used to prevent information leakage;
- the second stage uses a constrained linear aggregation model to control the effective complexity of the meta-learning procedure.

Accordingly, the objectives of the study are:

- to develop a two-stage probabilistic stacking ensemble based on calibrated heterogeneous classifiers;
- to construct a leakage-free meta-learning procedure using out-of-fold probability estimates;
- to investigate the effect of calibrated probability fusion and linear aggregation on predictive performance and probabilistic reliability;
- to compare the proposed method with baseline classifiers and standard stacking approaches using a real-world admission prediction dataset.

The main contribution of this paper is as follows:

1. The study proposes a structured calibrated probabilistic stacking ensemble for predictive modeling in decision-support tasks.
2. The proposed ensemble performs second-stage learning in the space of calibrated out-of-fold posterior probabilities, which ensures that the meta-level model operates on mutually consistent probabilistic inputs.
3. A constrained linear aggregation model is used at the meta-level to reduce the effective complexity of second-stage learning and to improve generalization.
4. The effectiveness of the proposed ensemble is demonstrated on a real-world admission campaign dataset, where it achieves competitive predictive performance while providing more reliable probabilistic outputs for practical decision-making.

Unlike conventional stacking schemes, where outputs of base learners are combined without explicit control of their probabilistic consistency, the proposed ensemble uses calibrated out-of-fold posterior probabilities as inputs to the second-stage model. This design ensures that the meta-level learner aggregates comparable probabilistic estimates rather than heterogeneous raw outputs. In addition, the second-stage model is intentionally restricted to linear aggregation, which reduces its effective complexity and limits the risk of overfitting. Therefore, the proposed approach should not be interpreted merely as another application of stacking to a real dataset. Instead, it represents a methodologically grounded probabilistic

ensemble framework designed for more reliable predictive decision support.

2 Related work

Ensemble learning and probabilistic prediction have been widely studied in ML as effective approaches for improving predictive performance and robustness [10]. In particular, stacking-based ensemble architectures and probability calibration techniques have attracted significant attention because they enable the integration of heterogeneous classifiers while preserving the reliability of probabilistic outputs [11], [16]. This section reviews the most relevant research related to stacked ensembles, probabilistic prediction, and probability calibration methods that form the methodological basis for the proposed approach.

The concept of stacked generalization, commonly referred to as stacking, was originally introduced by Wolpert as a ensemble for combining multiple predictive models through a meta-learning layer. In this approach, the outputs of base classifiers serve as input features for a secondary model that learns how to optimally combine their predictions [17]. A crucial aspect of stacking is the use of predictions generated on validation folds of the training data, often referred to as OOF outputs. This strategy reduces bias associated with in-sample predictions and enables the meta-model to capture systematic errors of individual classifiers. As a result, stacking provides a flexible mechanism for constructing hierarchical predictive systems in which the final decision is obtained through a learned combination of multiple predictors [17].

The stacking paradigm was further formalized within a statistical ensemble through the Super Learner method proposed by van der Laan, Polley, and Hubbard. In this approach, ensemble construction is formulated as an optimization problem where the optimal combination of candidate algorithms is selected based on cross-validation risk minimization. The Super Learner ensemble provides theoretical and empirical justification for constructing meta-predictions using cross-validated outputs of base models. Importantly, this methodology emphasizes that the meta-level model should rely exclusively on OOF predictions to ensure unbiased estimation of generalization performance and prevent information leakage from the training data [18].

Since stacking methods can operate in the space of posterior probabilities, the reliability of probability estimates produced by base classifiers becomes a critical factor. One of the classical approaches to probability calibration is Platt scaling [19], which transforms classifier scores into posterior probability estimates by fitting a sigmoid function. This method was originally introduced for Support Vector Machines by Platt. This technique learns a logistic transformation that maps classifier scores to calibrated probability estimates. Later work by Lin et al. improved the numerical stability and practical implementation of this calibration procedure, making it applicable to probabilistic outputs of various classifiers [20].

Empirical research by Niculescu-Mizil and Caruana [21] demonstrated that many commonly used ML algorithms produce systematically distorted probability estimates. Some models tend to generate overconfident predictions, while others produce overly conservative probability distributions. Their results highlight the importance of evaluating probabilistic predictions using proper loss functions and calibration diagnostics, such as log-loss and reliability diagrams, particularly when probability estimates are used as inputs for subsequent decision-making models [21].

The problem of improving probability estimates has also been investigated for specific classes of ML algorithms. Zadrozny and Elkan [11] analyzed calibration techniques for Decision Trees and naive Bayes classifiers and demonstrated that relatively simple post-processing procedures can significantly improve the quality of predicted probabilities. Their findings indicate that even models with strong classification performance may produce poorly calibrated probability estimates, and that calibration methods can effectively correct these distortions [11].

In subsequent work, the same authors explored methods for producing consistent probability estimates in multiclass classification problems. Their approach converts score-based outputs into calibrated multiclass probabilities through the combination of calibrated binary models. The study [22] established the general principle that probability calibration can be applied independently of the underlying classifier and becomes particularly important when probabilistic outputs are further combined by other models or decision rules.

More recent studies have also emphasized the importance of calibration for modern ML models and classifier ensembles. For instance, Mortier et al. in [16] investigated the calibration properties of probabilistic classifier sets and proposed statistical methods for evaluating whether ensembles provide reliable probability estimates. Their findings show that even complex ensembles of modern models may produce poorly calibrated outputs, highlighting the need for explicit calibration procedures in probabilistic prediction systems.

Recent research has also explored probabilistic ensemble architectures designed to improve uncertainty estimation and robustness of ML models [13]. Some studies propose ensemble frameworks based on probabilistic model averaging or neural architecture ensembles that combine predictions generated by different model configurations. These approaches demonstrate that probabilistic ensembles can improve predictive performance while simultaneously providing more reliable uncertainty estimates.

Another line of work investigates ensemble approaches that produce probabilistic prediction intervals or predictive distributions using conformal prediction techniques [23]. Such methods combine ensemble learning with distribution-free statistical guarantees, enabling uncertainty-aware predictions that adapt to local variability in the data while maintaining theoretical reliability.

In addition, existing studies have explored ensemble models that combine predictions directly in probability space rather than at the level of class labels [24]. These approaches integrate posterior probability estimates from multiple classifiers to construct more informative and interpretable ensemble decisions. Operating in probability space allows ensemble models to preserve uncertainty information and exploit complementary probabilistic patterns across base classifiers.

The table summarizes the methodological characteristics of the most relevant studies discussed in this section.

Table 1: Comparative analysis of existing stacking-based approaches and the methodological positioning of the proposed framework

Reference	Main methodological focus	Ensemble or probabilistic strategy	Probability calibration discussed	Meta-level learning characteristics	Relation to the proposed approach
[17]	Introduction of stacked generalization	Two-level stacking architecture	Not a central focus	General-purpose meta-learning	Provides the conceptual foundation for stacking ensembles
[18]	Cross-validated Super Learner ensemble	Risk-minimization ensemble learning	Not emphasized as a primary component	Flexible meta-learning framework	Supports the use of strict out-of-sample meta-learning
[20]	Calibration of classifier probability estimates	Sigmoid-based probability transformation	Explicitly addressed	Not a stacking-oriented framework	Provides the calibration basis used in the proposed method
[11]	Improvement of probabilistic outputs for classifiers	Classifier probability calibration	Explicitly addressed	Not focused on meta-learning	Supports probabilistic consistency across heterogeneous classifiers
[21]	Analysis of probabilistic reliability in ML models	Comparative probabilistic evaluation	Strongly emphasized	Not a stacking-based framework	Motivates the importance of calibrated probability estimates
[16]	Calibration analysis for probabilistic classifier sets	Probabilistic ensemble evaluation	Explicitly addressed	Ensemble-level probabilistic analysis	Supports reliability analysis of probabilistic ensembles
[13]	Two-stage PNN–SVM ensemble	Hybrid ensemble architecture	Partially discussed	Nonlinear secondary model	Conceptually related two-stage ensemble organization
Proposed method	Calibrated probabilistic stacking with constrained fusion	Two-stage probability-space ensemble	Explicit calibration before aggregation	Linear aggregation in probability space	Integrates calibration, OOF learning, and constrained meta-learning

The conducted comparative analysis demonstrates that existing stacking and probabilistic ensemble approaches mainly address individual components of ensemble learning separately. Some studies focus on stacking architectures and meta-learning strategies, while others investigate probability calibration or probabilistic reliability of classifiers. Although previous studies have demonstrated the usefulness of stacking, probability calibration, and probabilistic classifier fusion, limited attention has been paid to their joint integration within a unified two-stage framework. In particular, the literature still lacks sufficient empirical evidence on whether calibrated out-of-fold probability representations and linear fusion in probability space can simultaneously improve discriminative performance and probabilistic reliability in practical admission prediction tasks. In many conventional stacking approaches, posterior probabilities produced by heterogeneous classifiers are aggregated without explicit calibration, which may reduce the consistency and comparability of probabilistic estimates. In addition, the use of complex nonlinear meta-models may increase the effective hypothesis space and lead to a higher risk of overfitting. Furthermore, the absence of strict out-of-fold probability generation mechanisms in some ensemble schemes creates a potential risk of information leakage during second-stage learning. The proposed approach addresses these limitations by combining calibrated posterior probability estimation, leakage-free out-of-fold meta-feature construction, and structurally constrained linear aggregation in probability

space. Such an organization allows the ensemble to preserve probabilistic consistency, control the effective complexity of the meta-level model, and improve the reliability and generalization ability of the final decision-making process.

3 Methodology

This section describes the proposed two-stage probabilistic ensemble method.

The proposed approach belongs to the class of ensemble ML algorithms and implements a two-stage decision ensemble in which the secondary model is trained exclusively in the space of the calibrated posterior probabilities generated by several heterogeneous base classifiers.

The key characteristics of the method include:

- the use of calibrated ensemble models as probability generators;
- construction of a secondary training space based on OOF probability estimates;
- application of a linear meta-model that performs alignment and aggregation of posterior probability estimates.

Such a design reduces the effective complexity of the decision function and improves the generalization capability of the model without directly using the original feature space at the second stage.

3.1 Formal problem statement

Let a training dataset be given as

$$D = \{(x_i, y_i)\}_{i=1}^{N},$$

where:

$$x_i \in R^d,$$

is a vector of input features describing the i-th observation, and:

$$y_i \in \{1, \dots, C\},$$

denotes the corresponding class label, where C is the number of classes.

The goal is to construct a classifier:

$$\hat{y} = F(x),$$

that maps an input feature vector x to a predicted class label and maximizes the generalization performance according to evaluation metrics such as F1-score, Matthews Correlation Coefficient (MCC), or Cohen's Kappa.

3.2 Stage 1: Generation of calibrated posterior probabilities

This section describes the base classifiers, the probability calibration procedure, and the OOF strategy used to generate unbiased posterior probability estimates.

At the first stage, a set of M heterogeneous classifiers is trained:

$$H = \{h^1, h^2, \dots, h^M\}$$

In this study, the following heterogeneous classifiers are used at the first stage of the ensemble:

- HistGradientBoosting with default learning rate and tree depth parameters;
- ExtraTrees with probability calibration, using 200 trees and the Gini criterion;
- RandomForest with probability calibration, using 200 trees and the Gini criterion.

Probability calibration for ExtraTrees and RandomForest models is performed using Platt scaling implemented through a sigmoid-based calibration procedure. All models are implemented using the Scikit-learn library with fixed random seeds to ensure reproducibility of the experimental results.

Each classifier produces a vector of posterior class probabilities:

$$p^{(m)}(x) = (p_1^{(m)}(x), \dots, p_c^{(m)}(x)), m = 1, \dots, M,$$

where

$$p_c^{(m)}(x) \approx P(Y = c|X = x)$$

Thus, each model estimates the probability that the input observation x belongs to class c.

For tree-based ensemble models (ExtraTrees and RandomForest), Platt scaling is applied to improve the reliability of predicted probabilities. In this study, Platt scaling was selected as the calibration method for tree-based ensemble models. Isotonic regression was considered as an alternative non-parametric calibration approach. However, it was not used in the final experimental pipeline because the main objective of this work was not to compare calibration methods, but to evaluate the effect of calibrated probability representations in a two-stage probabilistic stacking framework. In addition, isotonic regression is more flexible and may require larger calibration subsets to avoid overfitting, whereas Platt scaling provides a simpler parametric calibration procedure based on a sigmoid transformation. This makes it more suitable for the proposed framework, where calibrated probabilities are further used as inputs for a linear meta-model. Therefore, Platt scaling was chosen as a computationally simple and methodologically consistent calibration strategy.

The calibrated probability estimate is defined as

$$\tilde{p}_c^{(m)}(x) = \sigma\left(a_c^{(m)} p_c^{(m)}(x) + b_c^{(m)}\right),$$

where:

$$\sigma(t) = \frac{1}{1+e^{-t}},$$

is the logistic function, and

$$a_c^{(m)}, b_c^{(m)},$$

are calibration parameters estimated on held-out validation data.

This calibration step improves the comparability and reliability of probability estimates across different models, which is essential for subsequent probability-based aggregation.

To prevent information leakage, stratified K-fold cross-validation with K = 5 is used during out-of-fold probability generation. For each fold, the base classifiers are trained on K−1 folds and generate probability estimates for the held-out subset. This procedure ensures that all probability representations used for meta-learning remain strictly out-of-sample. For each observation x_i, an OOF probability representation is constructed as:

$$z_i = [\tilde{p}^{(1)}(x_i)\tilde{p}^{(2)}(x_i), \dots, \tilde{p}^{(M)}(x_i)] \in R^{M \cdot C}.$$

Thus, the vector z_i contains calibrated class probabilities produced by all base models. All vectors z_i form a new training matrix

$$Z = \begin{bmatrix} z_1^T \\ z_2^T \\ \vdots \\ z_N^T \end{bmatrix}$$

This matrix represents the secondary feature space composed exclusively of posterior probability estimates.

3.3 Stage 2: Linear aggregation in probability space

This section presents the meta-model operating in the probability space and the corresponding final decision rule used for prediction.

At the second stage, a linear classifier is trained in the probability space:

$$g: R^{M \cdot C} \to R^C.$$

The use of a linear meta-model is motivated by its ability to ensure stable aggregation of calibrated probabilities, reduce the risk of overfitting, and preserve interpretability of the contribution of individual base classifiers. Logistic Regression and Linear SVC with probability calibration are used as meta-learners at the second stage of the ensemble. The choice of linear meta-models is intentional. In the proposed framework, the second-stage model operates only on calibrated posterior probability representations produced by the base classifiers. Therefore, its main role is to align and aggregate probabilistic estimates rather than to learn a new complex nonlinear feature representation. More flexible non-linear meta-models, such as multilayer perceptrons or tree-based meta-learners, could increase the effective hypothesis space of the ensemble and may lead to a higher risk of overfitting at the meta-learning stage. For this reason, Logistic Regression and calibrated Linear SVC were selected as controlled and interpretable aggregation models. This choice is consistent with the main methodological aim of the study, which is to evaluate the effect of calibrated probability fusion under a constrained probabilistic stacking design.

The Logistic Regression model uses L2 regularization with default optimization parameters, while Linear SVC is calibrated using sigmoid-based probability estimation. For Logistic Regression, the class probability is defined by the softmax function:

$$g_c(z) = \frac{\exp(w_c^T z + b_c)}{\sum_{k=1}^{C} \exp(w_k^T z + b_k)},$$

where:

$$w_c \in R^{M \cdot C},$$

is the weight vector associated with class c, and

$$b_c \in R,$$

is the corresponding bias term.

The final class prediction is determined as:

$$\hat{y} = \arg\max_c g_c(z)$$

Thus, the secondary model does not operate on the original feature space. Instead, it performs linear aggregation and alignment of posterior probability estimates produced by the base classifiers.

3.4 Structural constraints and algorithmic overview

The proposed approach incorporates several structural constraints that define the functional capacity of the model architecture:

- the meta-model operates exclusively in the probability space generated by base classifiers;
- all probability estimates are calibrated prior to aggregation;
- a linear hypothesis class is used at the second stage;
- a strict OOF mechanism is employed for constructing the meta-level training data;
- the architecture is restricted to a two-level ensemble structure.

These constraints significantly reduce the effective hypothesis space of the model and consequently decrease the risk of overfitting while preserving the expressive power of the ensemble.

The two-stage calibrated probabilistic ensemble can be summarized as follows:

1. Split the training dataset into K stratified folds.
2. For each fold: train the base models on $K-1$ folds, and generate calibrated probability estimates for the held-out fold.
3. Construct the probability feature matrix from the OOF predictions.
4. Train the linear meta-model using matrix Z.
5. For a new observation x: obtain calibrated probability estimates from the base models and aggregate them using the trained meta-model to obtain the final prediction.

The proposed design is motivated by three methodological considerations. First, the use of out-of-fold probability estimates prevents information leakage when constructing the secondary training space. Second, probability calibration makes the outputs of heterogeneous base classifiers more comparable before aggregation. Third, restricting the meta-model to a linear hypothesis class reduces the effective complexity of the second-stage decision rule and makes the contribution of each base classifier easier to interpret. Thus, the proposed architecture aims to improve not only predictive accuracy, but also the consistency and transparency of the final probabilistic decision.

4 Experimental design and evaluation procedure

The experimental study was conducted to objectively evaluate the generalization ability of the proposed two-step ensemble method and to compare it with basic ML algorithms. The modeling procedure was designed to ensure the validity of statistical estimates, the absence of information leakage, and the reproducibility of results.

4.1 Dataset and preprocessing

Experimental studies were performed on the collected and cleaned dataset of archived results of admission campaign of Lviv Polytechnic National University (Lviv, Ukraine).
The input data were represented as a feature matrix:

$$X \in R^{N*d},$$

and a corresponding vector of class labels:

$$y \in \{1, \dots, C\}^N.$$

Categorical variables were encoded using one-hot encoding, ensuring that all features were transformed into a numerical representation. Missing values in numerical attributes were imputed using median values, which reduces the influence of outliers and prevents distortion of feature distributions. The target variable was encoded into a compact numerical label space.
No explicit feature selection was performed in this study in order to preserve the original information structure of the admission dataset. Stratified sampling was applied during both holdout and cross-validation procedures to preserve the empirical class distribution across training and validation subsets.
The admission dataset was not characterized by extreme class imbalance. Nevertheless, stratified sampling was consistently used during both holdout validation and cross-validation in order to preserve the empirical class distribution across all training and evaluation subsets. No additional resampling procedures or explicit class-weighting strategies were applied because the observed class proportions allowed the models to achieve stable predictive performance without a pronounced bias toward the majority class.
For the primary evaluation of model performance, a holdout validation strategy was employed. The dataset was divided into training and testing subsets in an 80/20 ratio. Thus, each of the studied methods received about 12736 observations for training and 3185 observations for model application at each run.
The split was performed using stratified sampling, ensuring that the empirical class distribution was preserved in both subsets. A fixed random seed was used to guarantee reproducibility of the experiments. The dataset contained approximately 15,900 admission records represented by 12 original attributes and the corresponding admission outcome. The features included both numerical and categorical variables related to applicant scores, priorities, quotas, and specialization information.

4.2 Experimental design

The proposed method was compared with several baseline classifiers, including both single models and ensemble algorithms: Hist Gradient Boosting, Random Forest, Extra Trees, Logistic Regression, Linear SVC with probability calibration.
All baseline models were trained exclusively on the training subset and evaluated on the held-out test data without additional hyperparameter tuning. Default hyperparameter settings from the Scikit-learn framework were used unless explicitly specified otherwise. This setup allows the study to focus primarily on the structural properties of the proposed probabilistic stacking framework rather than on extensive model-specific hyperparameter optimization.
The proposed approach implements a two-stage decision procedure. At the first stage, a set of heterogeneous base classifiers (Hist Gradient Boosting, Extra Trees, and Random Forest) is trained to generate posterior probability estimates of class membership.
For tree-based ensemble models (Extra Trees and Random Forest), probability calibration is applied using Platt scaling to ensure consistency of probability estimates across different classifiers.
To construct the training dataset for the meta-level model, an OOF strategy is employed. The training data are divided into K stratified folds. For each fold, base models are trained on $K-1$ folds and produce probability estimates for the held-out fold. This process produces the secondary feature matrix

$$Z \in R^{N*(M*C)},$$

which contains only out-of-sample posterior probability estimates and therefore eliminates information leakage.
At the second stage, a meta-model (Logistic Regression or calibrated Linear SVC) is trained exclusively in the probability space represented by matrix Z. The original feature space X is not used at this stage, which substantially reduces the hypothesis space and mitigates the risk of overfitting.
For inference on test data, base models are retrained on the entire training dataset. Their calibrated probability outputs are then aggregated by the meta-model to produce the final prediction.
To evaluate the generalization performance of the models, stratified K-fold cross-validation was applied.
The outer evaluation stage uses stratified 5-fold cross-validation, while the inner stage is responsible for generating out-of-fold probability estimates for meta-learning. At the outer level, model performance was evaluated on validation folds, while the inner level was used to generate OOF probability estimates for training the meta-model.
This approach provides unbiased performance estimates and enables a reliable comparison between the proposed ensemble and baseline models.
Holdout metrics correspond to the results obtained from a fixed stratified split of the dataset into training (80%) and testing (20%) subsets. The metrics reported for the Holdout Test set reflect model behavior on independent data that were not used during any stage of training or model construction.
Cross-validation estimates were obtained using stratified K-fold cross-validation and are reported as mean ± standard deviation across validation folds.
For the proposed two-stage ensemble method, the cross-validation procedure incorporates the OOF probability generation mechanism, ensuring that probability estimates used to train the meta-model remain strictly out-of-sample. The cross-validation mean serves as the primary criterion for comparing model performance, while the standard deviation reflects the stability of the models with respect to variations in the training data.

Comparing results obtained under the Holdout and cross-validation schemes allows assessing the consistency and generalization capability of the models. Similar metric values across both evaluation protocols, together with small cross-validation standard deviations, indicate stable model behavior and absence of significant overfitting.
All experiments were implemented using fixed random seeds and a unified computational pipeline. The full sequence of preprocessing, out-of-fold probability generation, calibration, meta-level training, and evaluation is described in sufficient detail to support reproducibility of the proposed ensemble on comparable data.
The experimental framework was implemented in Python 3 using the Scikit-learn machine learning library. Standard scientific computing libraries, including NumPy and Pandas, were additionally used for data processing and experimental evaluation.
All experiments were performed within a consistent software environment to ensure reproducibility of the obtained results.

4.3 Evaluation metrics and computational cost

Model performance was evaluated using several complementary metrics: Accuracy, F1-score, MCC, Cohen's Kappa
These metrics provide a comprehensive assessment of both predictive accuracy and classification balance.
Additionally, the probabilistic quality of predictions was evaluated using the LogLoss function.
For each model, the mean and standard deviation of the metrics were reported based on cross-validation results. The reported standard deviations were additionally used to assess the stability of model behavior across different cross-validation folds. Computational time corresponding to the full experimental pipeline was additionally recorded to assess the overall complexity of the proposed framework.
All experiments were conducted using fixed random seeds, and the entire modeling procedure was implemented as a unified reproducible computational pipeline. This design ensures that the results can be independently replicated and verified.

5 Results and discussion

This section presents the experimental results and provides a comparative analysis of the proposed method against baseline models for the considered admission prediction task (table 2).
The proposed two-step method, based on the aggregation of calibrated posterior probabilities from three heterogeneous ensemble classifiers, demonstrates high and stable generalization ability. Both meta-level implementations (Logistic Regression and calibrated Linear SVC) achieve F1 scores exceeding 0.986 and MCC scores exceeding 0.97 in cross-validation results, confirming the absence of overfitting and the correctness of the OOF strategy.
The model with the Logistic Regression meta-level demonstrates the best performance on the holdout set, while Linear SVC (calibrated) provides slightly better generalization stability in cross-validation and minimal LogLoss values. The computational complexity of the method is higher compared to single-stage ensembles, which is the expected trade-off for using multi-model OOF aggregation.

Table 2: Generalization performance and computational complexity of the proposed method.

Meta-model	Holdout F1	Holdout MCC	Holdout Kappa	Holdout LogLoss	CV F1 (mean ± std)	CV MCC (mean ± std)	CV Kappa (mean ± std)	CV LogLoss (mean ± std)	CV time (s)
Logistic Regression	0.990	0.979	0.979	0.033	0.99 ± 0.002	0.97 ± 0.004	0.97 ± 0.004	0.05 ± 0.005	60.20
LinearSVC (calibrated)	0.989	0.978	0.978	0.033	0.99 ± 0.002	0.97 ± 0.003	0.97 ± 0.003	0.05 ± 0.005	57.31

An analysis of the cross-validation results shows that the improved two-step method, based on the aggregation of calibrated posterior probabilities from three heterogeneous ensemble classifiers, Hist Gradient Boosting, Extra Trees, and Random Forest, yields the highest or second-highest values for the F1 score, MCC, and Cohen's Kappa. This indicates the model's high discriminative power and its ability to perform correctly under conditions of potential class imbalance.
It is fundamentally important that the improvement in classification quality is not accompanied by a degradation of probabilistic estimates. The values of the LogLoss function for the two-step method remain at the same level or are better compared to most baseline models. This behaviour confirms the feasibility of forced calibration of ensemble tree algorithms prior to their subsequent aggregation at the meta-level

5.1 Comparison with baseline models and holdout evaluation

To provide a clear comparison of the performance of the improved two-step method and the baseline models, Tables 3 and 4 present summary metrics of classification quality based on the results of cross-validation and evaluation on an independent test set.
Table 3 reflects the generalization ability of the models based on the results of cross-validation in the format mean

± std. The time shown corresponds to the total execution time of a full cross-validation cycle, including model training, probability estimation, and, in the case of the two-step method, the formation of OOF representations and meta-model training.

Table 3: Cross-validation comparison of baseline models and the proposed method

Method	F1 (mean ± std)	MCC (mean ± std)	Kappa (mean ± std)	LogLoss (mean ± std)	CV time (s)
HistGradientBoosting	0.983 ± 0.002	0.964 ± 0.005	0.964 ± 0.005	0.060 ± 0.006	18
Random Forest	0.984 ± 0.003	0.967 ± 0.006	0.967 ± 0.006	0.066 ± 0.009	25
Extra Trees	0.983 ± 0.002	0.965 ± 0.004	0.965 ± 0.004	0.053 ± 0.005	22
Standard stacking (Meta: LogReg, raw probabilities)	0.985 ± 0.003	0.969 ± 0.005	0.969 ± 0.005	0.055 ± 0.006	54
Standard stacking (Meta: LinearSVC, raw probabilities)	0.985 ± 0.003	0.97 0± 0.004	0.969 ± 0.004	0.054 ± 0.006	52
Two-stage (Meta: LogReg)	0.986 ± 0.002	0.971 ± 0.004	0.971 ± 0.004	0.050 ± 0.005	60
Two-stage (Meta: LinearSVC)	0.988 ± 0.002	0.972 ± 0.003	0.972 ± 0.003	0.050 ± 0.005	57

The summary results show that all the basic ensemble methods considered achieve high classification accuracy (F1 > 0.98), but differ significantly in terms of stability and the quality of their probability estimates. Random Forest achieves slightly higher average F1 and MCC values, but is characterized by the largest standard deviations and the worst LogLoss metrics, indicating increased variability and insufficient calibration of probabilities. Hist Gradient Boosting demonstrates more stable behavior, but its probability estimates are smoothed, resulting in higher LogLoss values. Extra Trees, thanks to additional randomization, provides the best probabilistic accuracy among the baseline models and smaller standard deviations, although it is slightly inferior in discriminative metrics.

The comparison with standard stacking based on raw posterior probabilities shows that the proposed two-stage method provides slightly better discriminative performance and lower LogLoss for the considered admission prediction task. This indicates that the improvement is associated not only with the stacking architecture itself, but also with the use of calibrated probability representations at the meta-learning stage.

The improved two-step method consistently outperforms the baseline approaches on all key generalization performance metrics. The increase in the mean F1, MCC, and Cohen's Kappa values is accompanied by a decrease in standard deviations, indicating more stable model behavior across different data partitions. At the same time, the LogLoss values are the smallest or among the smallest among all the methods considered, confirming the effectiveness of aggregating calibrated posterior probabilities.

The obtained estimates confirm the conclusions drawn from cross-validation and indicate the absence of overfitting. The two-step method consistently outperforms baseline models on key discriminative metrics while maintaining high-quality probabilistic estimates.

Results obtained on an independent test set (table 4) demonstrate a clear hierarchy of methods in terms of discriminative power and generalization quality. All ensemble methods noticeably better than Hist Gradient Boosting and Random Forest, indicating the nonlinear nature of the relationships in the data and the limitations of linear models in this task.

Table 4: Comparison of methods based on holdout test results

Method	F1	MCC	Cohen's Kappa	LogLoss	Time, sec
Improved: (HGB + ET(cal) + RF(cal)) - Meta: LogReg	**0.990**	**0.979**	**0.979**	0.033	68.37
Improved: (HGB + ET(cal) + RF(cal)) - Meta: LinearSVC(cal)	0.989	0.978	0.978	0.033	62.77
Standard stacking: (HGB + ET + RF) - Meta: LogReg	0.989	0.977	0.977	0.038	61.24

Standard stacking: (HGB + ET + RF) - Meta: LinearSVC	0.988	0.976	0.976	0.037	58.96
HistGradientBoosting	0.988	0.975	0.975	0.048	3.60
RandomForest	0.987	0.974	0.974	0.046	1.73
ExtraTrees	0.987	0.973	0.973	0.032	3.39
Logistic Regression	0.961	0.918	0.917	0.124	1.19
LinearSVC (calibrated)	0.96	0.917	0.916	0.121	1.50

Among the basic ensembles, Hist Gradient Boosting, Random Forest, and Extra Trees demonstrate similar values for F1, MCC, and Cohen's Kappa (≈ 0.987–0.988 for F1), indicating a high level of classification accuracy. At the same time, there are significant differences in the quality of probability estimates among them. In particular, Extra Trees provides the lowest LogLoss value among the base models, confirming its tendency to generate smoother and better-calibrated probabilities. In contrast, Random Forest and Hist Gradient Boosting exhibit worse LogLoss values, which is expected given their decision-making mechanisms and the absence of explicit calibration optimization.

The comparison with standard stacking based on raw posterior probabilities indicates that the observed gain is associated not only with ensemble combination itself, but also with the use of calibrated probability representations at the meta-learning stage. This supports the methodological role of probability calibration in the proposed framework.

The obtained improvements in F1-score, MCC, and LogLoss appear to be associated with several complementary factors. First, the use of calibrated posterior probabilities makes probabilistic outputs produced by heterogeneous classifiers more comparable before aggregation at the meta-learning stage. Second, strict out-of-fold probability generation reduces the risk of information leakage and allows the meta-model to learn more reliable decision boundaries. Third, the use of a constrained linear meta-model limits the effective complexity of second-stage learning and decreases the tendency to overfit compared to more flexible nonlinear aggregation schemes.

The reduction in LogLoss values is particularly important because it indicates not only improved classification accuracy, but also better probabilistic consistency of the final predictions. This behavior suggests that the proposed framework produces probability estimates that are more aligned with the true class distribution, which is especially relevant for admission prediction tasks where predicted probabilities are directly interpreted as estimates of admission likelihood.

Although formal statistical hypothesis testing was not the primary focus of the present study, the observed improvements remain consistent across both holdout and cross-validation evaluation protocols. In addition, the relatively small standard deviations obtained during cross-validation indicate stable model behavior and suggest that the performance gains are not associated with random variations of individual data splits.

The improved method consistently outperforms all baseline approaches on key discriminative metrics. Improvements in F1, MCC, and Cohen's Kappa are consistent for both meta-level variants (Logistic Regression and calibrated Linear SVC), indicating effective compensation for the errors of baseline models in the posterior probability space. Importantly, this improvement is achieved without degrading probabilistic correctness: the LogLoss values for the two-step method are comparable to the best baseline model (Extra Trees) and significantly better than Hist Gradient Boosting and Random Forest.

Runtime analysis shows that the two-step method is computationally more expensive, due to the generation of OOF probabilities and the additional meta-model training step. However, this computational cost is methodologically justified, as it allows for higher accuracy, better generalization ability, and more balanced behavior between discriminative and probabilistic metrics. For tasks where the reliability and accuracy of predictions are critical, such a trade-off is acceptable.

The present study is focused on a single real-world admission dataset corresponding to the target application for which the ensemble was developed. Therefore, the conclusions of the paper should be interpreted primarily within this admission prediction setting. Broader cross-domain validation may be considered in future research, but it does not constitute the main objective of the present study.

5.2 Discussion

The proposed ensemble was designed for a specific real-world admission prediction task rather than as a universal benchmark method for heterogeneous public datasets. Therefore, its value should be interpreted primarily in terms of its suitability for the target decision-support setting. In this context, the use of a real institutional dataset is an advantage, since it allows the method to be evaluated under practically meaningful conditions that directly correspond to the intended deployment scenario.

Unlike purely benchmark-oriented studies, the present work emphasizes ecological validity, since the ensemble is developed and evaluated on real data from the decision environment for which it is intended.

A comparison of the results obtained using holdout and cross-validation schemes demonstrates a high degree of

consistency in the quality assessments. Similar metric values on the independent test set and cross-validation averages, as well as small standard deviations, indicate the absence of significant overfitting and the stability of the proposed approach with respect to variations in the training set.

Unlike one-stage ensemble models, in which improved accuracy is often achieved by increasing the complexity or number of base components, the proposed method achieves improvement through structurally constrained information aggregation. The meta-model operates exclusively in the space of calibrated posterior probabilities, which significantly reduces the effective hypothesis space and, as a result, enhances the model's generalization ability.

The comparison with the studies summarized in Table 1 shows that existing ensemble approaches usually focus on separate methodological components. Some methods mainly investigate stacking architectures, while others focus on probability calibration or probabilistic reliability. At the same time, many conventional stacking schemes aggregate posterior probabilities without explicit calibration and without strict control of out-of-sample probability generation. Because of this, probabilistic outputs produced by heterogeneous classifiers may remain insufficiently consistent at the meta-learning stage. In addition, complex nonlinear meta-models may increase the effective hypothesis space and lead to a higher risk of overfitting. In contrast, the proposed approach combines calibrated posterior probability fusion, strict out-of-fold probability generation, and constrained linear aggregation within a unified framework. Such an organization improves the consistency of probabilistic estimates and supports stable generalization performance.

However, the use of a two-step scheme involving the calculation of out-of-fold probability representations leads to increased computational complexity compared to conventional single-stage ensemble models. The additional computational cost is mainly associated with repeated model training during out-of-fold probability generation, probability calibration, and second-stage meta-learning. At the same time, the effective complexity of the meta-level model remains limited because the second stage operates on compact probability representations rather than on the original high-dimensional feature space. As a result, the memory requirements of the meta-model remain relatively moderate despite the multi-stage organization of the ensemble.

For substantially larger datasets, the computational burden associated with repeated out-of-fold probability generation may become more significant. In such scenarios, parallel implementation strategies or distributed model training may be required to improve scalability. Nevertheless, the obtained experimental results indicate that the proposed framework remains computationally feasible for practical admission prediction tasks while providing improved probabilistic reliability and stable generalization performance.

During inference, the proposed framework requires sequential probability estimation by the base classifiers followed by linear aggregation at the meta-learning stage. As a result, the inference complexity is higher than that of a single classifier because predictions from multiple models must be generated before the final decision is obtained. At the same time, the second-stage model operates on compact probability representations and uses a linear aggregation scheme, which keeps the additional inference overhead relatively moderate. The main computational burden of the proposed framework is therefore associated primarily with training and out-of-fold probability generation rather than with the inference stage itself.

The results confirm that simply aggregating the outputs of individual models is not sufficient to improve classification performance. A key role in the proposed approach is played by the combination of heterogeneous sources of probabilistic information, which are characterized by different inductive biases and tend to make errors in different regions of the feature space. This diversity creates the necessary conditions for effective error compensation at the second stage.

An important component of the proposed architecture is the enforced calibration of ensemble tree-based models. Without calibration, the posterior probability estimates produced by Random Forest and Extra Trees are not directly comparable, which hinders their correct linear aggregation. Calibration enables all components of the secondary feature space to be interpreted as consistent estimates of the conditional probability

$$P(Y|X),$$

which is critical for ensuring the stable and reliable operation of the meta-model.

The obtained results indicate that the improvement is related not only to ensemble aggregation itself, but also to the probabilistic organization of the meta-learning stage. Unlike standard stacking approaches based on raw posterior outputs, the proposed method operates on calibrated probability representations that are more comparable across heterogeneous classifiers. This is particularly important for admission prediction tasks, where predicted probabilities are directly interpreted as estimates of admission likelihood. Therefore, probabilistic reliability becomes important together with classification accuracy.

The use of linear meta-models (Logistic Regression and Linear SVC) is fundamental in terms of generalization ability. Unlike more complex nonlinear meta-classifiers, linear aggregation does not lead to a sharp increase in variability and does not impair stability, as evidenced by the small standard deviations of metrics in cross-validation. Thus, the improvement is achieved not by increasing complexity, but through a structurally constrained transformation of the feature space.

Although non-linear meta-models may be useful in some stacking architectures, they were not the main focus of this study. The proposed method was designed to investigate whether calibrated out-of-fold probability representations can be effectively combined using a simple and controlled linear decision rule. Such an organization reduces the effective complexity of the second-stage model and helps separate the influence of probability calibration and OOF probability generation from the influence of a more

flexible meta-learner. Therefore, the use of Logistic Regression and calibrated Linear SVC is consistent with the main methodological objective of the proposed framework. The comparative analysis indicates that even under these structural constraints, the ensemble achieves stable predictive performance together with improved probabilistic reliability.

This result also distinguishes the proposed approach from many existing stacking architectures discussed in the literature. In several related studies, performance improvement is achieved by increasing the complexity of the meta-model. In contrast, the proposed method improves predictive performance through a more consistent organization of probabilistic information transferred between ensemble stages. The results suggest that constrained linear aggregation in probability space provides a balanced trade-off between predictive accuracy, probabilistic reliability, and generalization ability.

The practical usefulness of the proposed framework should be understood in the context of decision support rather than automatic decision-making. In a real admission environment, the model could be used to support applicants in estimating their admission chances, to help university admission offices analyze application flows, or to provide analytical dashboards during admission campaigns. In these scenarios, the predicted probabilities may help compare alternative application strategies and identify cases that require additional human attention.

At the same time, the proposed method should not be considered a universal solution for all admission or educational prediction tasks. Its performance may depend on the structure of institutional data, admission rules or applicant behavior. The model may require recalibration or retraining when applied to another university, another admission year, or a substantially different educational system. Possible failure cases include distribution shifts between admission campaigns, changes in admission policy, incomplete or biased applicant records, and poorly calibrated outputs of the base classifiers. Therefore, practical deployment should include periodic validation, monitoring of probabilistic outputs, and human oversight in decision-sensitive use cases.

The present study has several limitations that should be considered in the context of modern ensemble learning approaches. First, the proposed framework is computationally more expensive than conventional single-stage ensembles because it requires the generation of calibrated out-of-fold probability representations and an additional meta-learning stage. Second, the current implementation investigates a limited set of heterogeneous base classifiers and only one calibration strategy based on Platt scaling. Other calibration methods and ensemble configurations may also influence the probabilistic behavior of the model. Third, the experimental study is focused on a real-world institutional admission dataset corresponding to the intended application setting of the ensemble. Although this improves the practical relevance of the study, additional validation on datasets from other institutions would be useful for assessing the transferability of the proposed probabilistic stacking strategy.

In addition, the obtained results may be influenced by dataset-specific characteristics, including institutional admission rules, applicant demographics, feature distributions, and temporal properties of a particular admission campaign. Such factors may affect the probabilistic structure of the data and influence the behavior of the ensemble when applied to other institutions or educational systems. Therefore, practical deployment in substantially different environments may require recalibration, retraining, or adaptation of the proposed framework.

Overall, the obtained results show that the proposed approach should not be interpreted only as another application of stacking to an admission prediction problem. The main contribution of the study lies in the joint integration of calibrated posterior probability fusion, leakage-free out-of-fold probability generation, and constrained linear meta-learning within a unified probabilistic ensemble framework. The comparative analysis with related studies indicates that these components are often investigated separately in the existing literature. Their combined use allows the proposed method to achieve improved probabilistic consistency together with stable predictive performance and reliable generalization behavior.

6 Conclusion

In this paper, a two-stage probabilistic stacking ensemble based on calibrated heterogeneous classifiers and linear decision fusion in probability space was proposed for admission outcome prediction. The developed framework combines calibrated posterior probability estimation, out-of-fold probability generation, and constrained meta-learning within a unified probabilistic ensemble architecture. The obtained results demonstrate that the proposed organization improves both predictive performance and the consistency of probabilistic estimates compared to conventional single models and standard stacking approaches.

The experimental study performed on a real-world admission campaign dataset from Lviv Polytechnic National University showed that the proposed framework achieves stable classification performance across both holdout and cross-validation evaluation protocols. The use of calibrated posterior probabilities and leakage-free out-of-fold probability representations contributed to improved LogLoss values and more reliable probability estimates, while the use of linear meta-models helped limit the effective complexity of second-stage learning and reduce the tendency to overfit.

The conducted analysis additionally demonstrated that the proposed probabilistic stacking strategy may be practically useful for decision-support scenarios related to admission analytics and applicant assessment. At the same time, the obtained findings should be interpreted in the context of the considered admission prediction task and the specific institutional dataset used in the study. Additional validation on datasets from other institutions and application domains is required to further assess the

transferability and generalizability of the proposed framework.

Data availability statement

The dataset used in this study has been deposited in a publicly available open-access repository and can be accessed at: https://www.researchgate.net/publication/403689404_Enrollment_Data_LPNU.

References

[1] B. G. Banik and A. B. Syed, 'Predicting University Admission Chances Using Machine Learning', *Next-Generation Computing Systems and Technologies*, vol. 2, no. 1, pp. 1–9, Mar. 2026, doi: 10.62762/NGCST.2026.766610.

[2] K. K. Reddy, 'AI-Based University Admission Prediction System Using Random Forest Regression', *IJRASET*, vol. 13, no. 12, pp. 3269–3272, Dec. 2025, doi: 10.22214/ijraset.2025.76688.

[3] K. Zub and M. Gregus, 'Machine Learning-based Classification of Higher Education Admission Success for Informed Decision-Making', *Procedia Computer Science*, vol. 272, pp. 534–539, 2025, doi: 10.1016/j.procs.2025.10.243.

[4] P. Golden, K. Mojesh, L. M. Devarapalli, P. N. S. Reddy, S. Rajesh, and A. Chawla, 'A Comparative Study on University Admission Predictions Using Machine Learning Techniques', *IJSRCSEIT*, pp. 537–548, Apr. 2021, doi: 10.32628/CSEIT2172107.

[5] J.-P. Wu, M.-S. Lin, and C.-L. Tsai, 'A Predictive Model That Aligns Admission Offers with Student Enrollment Probability', *Education Sciences*, vol. 13, no. 5, p. 440, Apr. 2023, doi: 10.3390/educsci13050440.

[6] I. Izonin, R. Muzyka, R. Tkachenko, M. Gregus, R. Korzh, and K. Yemets, 'An enhanced cascade ensemble method for big data analysis', *IJ-AI*, vol. 14, no. 2, p. 963, Apr. 2025, doi: 10.11591/ijai.v14.i2.pp963-974.

[7] C. Rokde, J. Chakole, and A. Ukey, 'Financial Forecasting with Deep Learning Models Based Ensemble Technique in Stock Market Analysis', *IJIEEB*, vol. 17, no. 4, pp. 1–13, Aug. 2025, doi: 10.5815/ijieeb.2025.04.01.

[8] S. A. Hamim, R. S. Aftab, M. Ahmed, F. Faiza, and M. F. Mridha, 'Advanced Heart Attack Prediction Using a Stacked Ensemble Machine Learning Model and Diverse Data Integration', *IJISA*, vol. 17, no. 5, pp. 49–67, Oct. 2025, doi: 10.5815/ijisa.2025.05.04.

[9] K. V. Zub, 'The evaluation of the hei`s entrants admission chances based on the stacking model of the support vectors machine', *Sci. Pap. UAP*, vol. 2, no. 63, pp. 168–176, 2021, doi: 10.32403/1998-6912-2021-2-63-168-176.

[10] D. R. Patil, T. M. Pattewar, T. S. Shinde, K. S. Kumavat, and S. N. Deshpande, 'Stacking and Voting-Based Boosting Ensembles for Robust Malicious URL Classification', *IJCAI*, vol. 49, no. 35, Dec. 2025, doi: 10.31449/inf.v49i35.7762.

[11] B. Zadrozny and C. Elkan, 'Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers', in *Proceedings of the Eighteenth International Conference on Machine Learning*, in ICML '01. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., Jun. 2001, pp. 609–616.

[12] M. A. Voronenko *et al.*, 'Using Bayesian methods in the task of modeling the patients' pharmacoresistance development', *IAPGOS*, vol. 12, no. 2, pp. 77–82, Jun. 2022, doi: 10.35784/iapgos.2968.

[13] K. Zub, P. Zhezhnych, and C. Strauss, 'Two-Stage PNN–SVM Ensemble for Higher Education Admission Prediction', *BDCC*, vol. 7, no. 2, p. 83, Apr. 2023, doi: 10.3390/bdcc7020083.

[14] A. Kowshir Bitto, Md. H. Imam Bijoy, A. Das, J. Ferdousi, A. Begum, and I. Mahmud, 'A Novel CatML Stacking Classifier Based Intelligent System for Predicting Postgraduate Admission Chances: A Study on Bangladesh', *IJMECS*, vol. 17, no. 4, pp. 82–100, Aug. 2025, doi: 10.5815/ijmecs.2025.04.06.

[15] N. S. K. M. K. Tirumanadham and T. S., 'Enhancing Student Performance Prediction in ELearning Environments: Advanced EnsembleTechniques and Robust Feature Selection', *IJMECS*, vol. 17, no. 2, pp. 67–86, Apr. 2025, doi: 10.5815/ijmecs.2025.02.03.

[16] T. Mortier, V. Bengs, E. Hüllermeier, S. Luca, and W. Waegeman, 'On the Calibration of Probabilistic Classifier Sets', in *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, PMLR, Apr. 2023, pp. 8857–8870. Accessed: Apr. 07, 2026. [Online]. Available: https://proceedings.mlr.press/v206/mortier23a.html

[17] D. H. Wolpert, 'Stacked generalization', *Neural Networks*, vol. 5, no. 2, pp. 241–259, Jan. 1992, doi: 10.1016/S0893-6080(05)80023-1.

[18] M. J. Van Der Laan, E. C. Polley, and A. E. Hubbard, 'Super Learner', *Statistical Applications in Genetics and Molecular Biology*, vol. 6, no. 1, Sep. 2007, doi: 10.2202/1544-6115.1309.

[19] J. Platt, 'Probabilistic Outputs for Support vector Machines and Comparisons to Regularized Likelihood Methods', 1999. Accessed: Apr. 07, 2026. [Online]. Available: https://www.semanticscholar.org/paper/Probabilistic-Outputs-for-Support-vector-Machines-Platt/42e5ed832d4310ce4378c44d05570439df28a393

[20] H.-T. Lin, C.-J. Lin, and R. C. Weng, 'A note on Platt's probabilistic outputs for support vector machines', *Mach Learn*, vol. 68, no. 3, pp. 267–276, Aug. 2007, doi: 10.1007/s10994-007-5018-6.

[21] A. Niculescu-Mizil and R. Caruana, 'Predicting good probabilities with supervised learning', in *Proceedings of the 22nd international conference on Machine learning - ICML '05*, Bonn, Germany: ACM Press, 2005, pp. 625–632. doi: 10.1145/1102351.1102430.

[22] B. Zadrozny and C. Elkan, 'Transforming classifier scores into accurate multiclass probability estimates', in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, Edmonton Alberta Canada: ACM, Jul. 2002, pp. 694–699. doi: 10.1145/775047.775151.

[23] V. Jensen, F. M. Bianchi, and S. N. Anfinsen, 'Ensemble Conformalized Quantile Regression for Probabilistic Time Series Forecasting', *IEEE Trans. Neural Netw. Learning Syst.*, vol. 35, no. 7, pp. 9014–9025, Jul. 2024, doi: 10.1109/TNNLS.2022.3217694.

[24] F. Ren, Y. Li, and M. Hu, 'Multi-classifier ensemble based on dynamic weights', *Multimed Tools Appl*, vol. 77, no. 16, pp. 21083–21107, Aug. 2018, doi: 10.1007/s11042-017-5480-5.