Advances in Artificial Intelligence and Machine Learning; Research (Ahead of Print) Received 23-05-2026; Accepted 01-09-2026; Published online 07-09-2026

Calibration-Aware Hybrid Machine Learning and Deep Learning Ensemble for Cardiovascular Disease Prediction

Marwa Bikai marwahbikai@gmail.com
Department of Computer & Communications Engineering
Lebanese International University
Beirut, Lebanon

Milia Habib milia.habib@liu.edu.lb
Department of Computer & Communications Engineering
Lebanese International University
Beirut, Lebanon

Zaher Merhi zaher.merhi@liu.edu.lb
Department of Computer & Communications Engineering
Lebanese International University
Beirut, Lebanon

Rabih Rammal rabih.rammal@liu.edu.lb
Department of Electrical Engineering
Lebanese International University
Beirut, Lebanon

Ali El Masri ali.masri@liu.edu.lb
Department of Computer & Communications Engineering
Lebanese International University
Beirut, Lebanon

Corresponding Author: Milia Habib

Abstract

Cardiovascular disease (CVD) is a major global cause of morbidity and mortality. This paper proposes a calibration-aware hybrid ensemble framework. It combines traditional machine learning (ML) and deep learning (DL) techniques to predict CVD from tabular clinical datasets of 70,000 records, evaluated with leakage-safe nested cross-validation and calibrated base probabilities, and reports both discrimination and calibration metrics. Six machine learning models, including LR, RF, ET, XGB, LGBM, and KNN, along with three deep learning models, namely MLP, TabNet, and FT-Transformer, are evaluated. The top four ML models and the three DL models are used to construct an ensemble via stacking and soft voting, with both weighted and unweighted variants. The proposed approach integrates hybrid ML/DL stacking, selective probability calibration, leakage-safe nested cross-validation, and feature engineering. The performance is evaluated through nine metrics (ROC-AUC, PR-AUC, accuracy, recall, precision, F1-score, Brier score, log loss, and ECE).

Citation: Marwa Bikai, et al. Calibration-Aware Hybrid Machine Learning and Deep Learning Ensemble for Cardiovascular Disease Prediction. **Advances in Artificial Intelligence and Machine Learning**. 2026. (Ahead of Print) https://dx.doi.org/10.54364/AAIML.2026.65341

The results show that soft voting of XGB and FT-Transformer achieved the highest discrimination performance with an ROC-AUC of 80.16% and the best calibration metrics, including a Brier Score of 0.1803, a Log Loss of 0.505, and ECE of 0.004. While the models demonstrated strong internal performance, further validation on external datasets is needed to establish their potential utility for clinical decision support.

Keywords: Cardiovascular disease, Machine learning, Deep learning, Hybrid ensemble, Calibration, Predictive modeling, Tabular data, Risk prediction, Nested-CV.

1. INTRODUCTION

Cardiovascular Diseases (CVDs), including coronary heart disease, stroke, and related conditions, remain the leading cause of death worldwide. The World Health Organization estimates that around 19.8 million people died from CVDs in 2022, being the reason for almost 32% of all deaths around the world [1]. Furthermore, the cost of heart disease in the US is predicted to rise by 41% between 2010 and 2040 [2]. Heart disease is a complex, multifactorial illness with high mortality, and morbidity, and these depressing figures highlight its significant clinical and public health effects. Therefore, early detection and management are crucial. However, the expense, complexity, and accessibility of current diagnostics limit their use in many circumstances. Tests like stress tests, echocardiograms, and electrocardiograms need specific tools and knowledge, and they may still overlook early or unusual instances [3]. The need for scalable, precise predictive systems that can identify those who are at risk even before symptoms appear is driven by this gap. A viable solution to this problem is machine learning (ML), a branch of artificial intelligence that focuses on data-driven prediction. ML models can forecast the development of cardiac disease and stratify people based on risk by utilizing a variety of health data, including clinical records, imaging, biosignals, and even wearable sensors [2]. On benchmark heart-disease datasets, recent research shows that machine learning classifiers, including logistic regression, decision trees, random forests, support vector machines, and neural networks may reach high accuracy. In cross-validated experiments, hybrid deep learning architectures (such as CNN–LSTM models) have further improved performance, frequently surpassing conventional models [2]. However, careful consideration of best practices is necessary to turn these promising ML models into clinically effective tools. While academic research demonstrates very high accuracies on established datasets, real-world clinical applications are still challenging. Explainability and generalizability are major obstacles. When applied to new populations, models that were trained on a small dataset—typically the UCI Cleveland or Framingham data [4], may perform badly, particularly if patient demographics or data collection methods are different. Naive cross-validation may unintentionally mix test and training samples. Methodological rigor is therefore crucial: outputs should be calibrated so that anticipated risks accurately reflect actual probabilities, and data should be rigorously partitioned (e.g., by patient or facility) to prevent leakage [5]. In light of this, the main contributions of this paper include proposing a calibration-aware framework incorporating feature engineering to support model training, as well as machine learning and deep learning models, including both calibrated and uncalibrated variants, evaluated under a nested cross-validation design to prevent data leakage. This paper also focuses on both discrimination and calibration metrics rather than relying solely on accuracy, which is insufficient for cardiovascular disease risk assessment.

2. RELATED WORK

The domain of cardiovascular disease (CVD) prediction using machine learning has evolved rapidly, leveraging both classical ML and deep learning techniques. Researchers typically rely on structured tabular clinical datasets, which provide demographic, clinical, and lifestyle information relevant to heart disease risk. Among these, the top three datasets most commonly used in CVD prediction research are: Kaggle Cardiovascular Disease Dataset [6], Cleveland Heart Disease Dataset [4], and Framingham Heart Study Dataset [7]. Yasmeen et al.(2025) [8], proposed a stacked ensemble framework that combines TabNet, a deep learning model, with XGBoost, a tree-based models. Their outputs are integrated using a Logistic Regression (LR) or Support Vector Machine (SVM) to improve the predictive performance. Ganie et al. (2025) [9], employed six base models pipelined to develop stacking and voting framework. In addition, they conducted a statistical analysis of the proposed model using Friedman Aligned Ranks test and Holm's post-hoc pairwise comparisons. They used two datasets about heart disease patients, HDDC and UHDD collected from Kaggle. The results showed that the developed ensemble models, particularly stacking, achived higher accuracy for advance heart disease prediction. Ziadi et al. (2025) [10], proposed a novel hybrid ensemble learning framework for CVD risk assessment, integrating voting and stacking techniques, HeartEnsembleNet with Hybrid Random Forest Linear Models (HRFLM), that implements several machine learning classifiers. The model is evaluated against six classical ML classifiers, including support vector machine (SVM), gradient boosting (GB), decision tree (DT), logistic regression (LR), k-nearest neighbor (KNN), and random forest (RF). HeartEnsembleNet outperformed the tested ML classifiers with an accuracy of 92.95%, and a precision of 93.08%. The researchers in [11], investigated the role of hybrid models combining ML techniques, including KNN and XGBoost, with DL models such as CNN and LSTM for cardiovascular disease prediction. Then, they employed majority voting as an ensemble technique to identify the final output class. Furthermore, ensemble optimization deep learning techniques were used by Oyewola et al.(2024) [12], for the early diagnosis of CVD, achieving an accuracy of 98.45%. Setiawan et al.(2025) [13], conducted a comparative study between Multilayer Perceptron (MLP) and Random Forest for CVD prediction on a large dataset with balanced target classes. They evaluated the models using precision, recall, F1-score, AUC-ROC, confusion matrix, and k-fold cross-validation. Random Forest achieved better performance, with an accuracy of 74.23% and an AUC of 0.80. A comparative study was done by Asadi et al. (2024) [14], on 9,499 participants using tree-based machine learning models for CVD prediction. The researchers evaluated Decision Tree, Random Forest, GLMM Tree, and Generalized Mixed Effect Random Forest (GMERF) models using ROC-AUC as the main performance metric. The GMERF model yielded the best AUC value of 0.80. Additionally, they identified the most effective factors of CVD. In this regard, this study proposed a selective calibration prior to ensemble construction using a leakage-safe framework for cardiovascular disease prediction using tabular structured clinical. It integrates traditional machine learning, deep learning, and hybrid ML/DL ensembles with discrimination and calibration reporting to provide a comprehensive evaluation of model performance.

3. DATASET AND RESEARCH METHODOLOGY

3.1 Data and Preprocessing

In this study, the Kaggle Cardiovascular Disease Dataset [6], is used. It consists of 70,000 records and 13 columns, including 12 raw predictor variables and the cardio target variable. Then, several preprocessig steps such as plausibility filtering, duplicate removal, missing-value handling, and feature engineering are applied to reduce the size to 68,606 records while the number of predictors increased to 25 engineered features. TABLE 1, lists the considered features and their descriptions which are processed, cleaned, and engineered to include derived indicators such as BMI, pulse pressure, and mean arterial pressure (MAP) for enhanced model input. The binary columns smoke, alco, active, and cardio are treated as binary indicators. Gender is checked against values 1 and 2. While cholesterol and gluc are checked against values 1, 2, and 3. These checks help ensure data consistency of the modelling pipeline and prevent invalid categories that could arise from corrupted data, wrong file versions, or accidental preprocessing mistakes.

Table 1: Description of Dataset Features

Feature	Description	Type	Notes
age	Age in days	Continuous	Converted to years
gender	1 = female, 2 = male	Categorical	Converted to 0/1
height	Height in cm	Continuous	120–220 cm range
weight	Weight in kg	Continuous	30–200 kg range
ap_hi	Systolic blood pressure	Continuous	60–250 mmHg
ap_lo	Diastolic blood pressure	Continuous	30–150 mmHg
cholesterol	1 = normal, 2 = above normal, 3 = well above normal	Categorical	Risk factor
gluc	Glucose level: 1–3	Categorical	Risk factor
smoke	0 = non-smoker, 1 = smoker	Binary	Lifestyle factor
alco	0 = non-alcoholic, 1 = drinks alcohol	Binary	Lifestyle factor
active	0 = inactive, 1 = active	Binary	Lifestyle factor
cardio	Target variable: 0 = no CVD, 1 = CVD	Binary	Outcome

3.1.1 Clinical plausibility cleaning

A Clinical Plausibility Cleaning is applied after duplicate removal, id removal, numeric conversion, and missing-value removal. The objective is to remove records containing physiologically implausible or invalid measurements. Such records can distort scaling, affect model fitting, and damage the reliability of probability estimates. TABLE 2, presents the applied validity filters rules with their corresponding justification. After applying the plausibility process, the main pipeline reports a final cleaned and feature-engineered shape of 68,606 rows and 25 columns. The class distribution remained approximately balanced. This supports the use of the cleaned data for the main evaluation and stacking experiments.

Table 2: Main pipeline plausibility and validity filters

Variable/group	Applied Rule	Reason
Age	$3650 \leq \text{age} \leq 36500$	Keep plausible age range when age is measured in days
Gender	gender $\in \{1, 2\}$	Preserve valid dataset encoding
Height	$120 \leq \text{height} \leq 220$	Remove implausible height values
Weight	$30 \leq \text{weight} \leq 250$	Remove implausible weight values
ap_hi	$80 \leq \text{ap\_hi} \leq 250$	Keep plausible systolic BP values
ap_lo	$40 \leq \text{ap\_lo} \leq 150$	Keep plausible diastolic BP values
BP relation	ap_hi > ap_lo	Systolic pressure should exceed diastolic pressure
Cholesterol	cholesterol $\in \{1, 2, 3\}$	Preserve ordinal category encoding
Gluc	gluc $\in \{1, 2, 3\}$	Preserve ordinal category encoding
Smoke, alco, active, cardio	$\in \{0, 1\}$	Preserve binary variable encoding

3.1.2 Feature engineering strategy

The feature-engineering step transforms the raw variables into clinically meaningful indicators. The engineered features are designed to represent age interpretation, body composition, blood pressure physiology, binary risk thresholds, lifestyle burden, metabolic burden, and interaction between risk factors. TABLE 3, presents the engineered features with formulas and interpretations.

Table 3: Engineered Features

Engineered feature	Formula or rule	Interpretation
age_years	$\text{age}/365.25$	Age expressed in years
BMI	$\text{weight}/(\text{height}/100)^2$	Body-mass index
pulse_pressure	ap_hi − ap_lo	Difference between systolic and diastolic BP
map_estimate	$(\text{ap\_hi} + 2 \times \text{ap\_lo})/3$	Estimated mean arterial pressure
age_risk	1 if age_years ≥ 50 else 0	Age-related risk indicator
bmi_obese	1 if BMI ≥ 30 else 0	Obesity indicator
bp_high	1 if ap_hi ≥ 140 or ap_lo ≥ 90 else 0	High blood pressure indicator
pulse_pressure_high	1 if pulse_pressure ≥ 60 else 0	Elevated pulse pressure indicator
chol_high	1 if cholesterol ≥ 2 else 0	Above-normal cholesterol indicator
gluc_high	1 if gluc ≥ 2 else 0	Above-normal glucose indicator
lifestyle_risk	smoke + alco + (1 − active)	Count of lifestyle risk components
metabolic_risk	bmi_obese + chol_high + gluc_high	Count of metabolic risk components
age_bp_interaction	age_years $\times$ bp_high	Age modified by high-BP status
bmi_bp_interaction	BMI $\times$ bp_high	BMI modified by high-BP status

3.2 Proposed Methodology

Our proposed framework for cardiovascular disease prediction is stuctured into four main phases as illustated in FIGURE 1. After preparing the data, we integrated classical machine learning, deep learning, and hybrid ensemble strategies, aiming to balance discrimination and calibration while ensuring rigorous evaluation. Hyperparameter optimization is conducted to select the best model parameter configuration, as shown in TABLE 4.

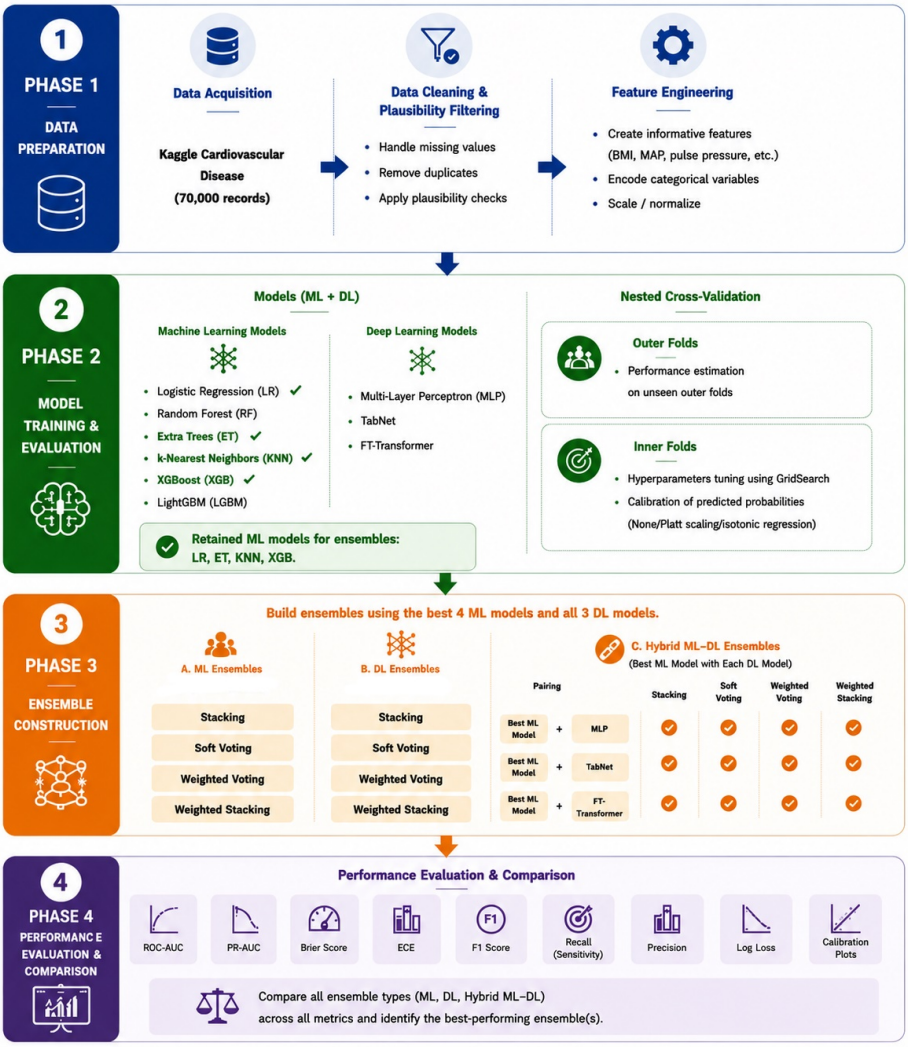


Figure 1: An Overview of the Proposed Methodology for the CVD Prediction Framework.

3.2.1 Machine Learning Models

This study proposes several classical ML models, as follows:

- Logistic Regression (LR): It is a baseline linear classifier for risk assessment [15]. A regularization strength of $C = 10$ with L2 regularization over all outer folds was favored by the hyperparameter optimization.
- An ensemble of decision trees, namely Random Forest (RF) is used to lower variance. With 300 decision trees, a maximum tree depth of 20, and a minimum of two samples needed at each leaf node, the optimal configuration is chosen.
- Extra Trees (ET): It is a randomized tree ensemble for improved diversity [16]. The Random Forest model's optimal hyperparameters were the same as those of ET's.

- k-Nearest Neighbors (KNN): It is a distance-based nonparametric classifier [10]. Hyperparameter tuning was performed over the number of neighbors, weighting strategy, and Minkowski distance parameter. The optimal configuration consistently selected $n\_neighbors = 21$ and `weights = uniform`, while the Minkowski parameter varied between $p = 1$ and $p = 2$ across outer folds.
- XGBoost (XGB): Strong nonlinear learning and ranking capabilities characterize this gradient-boosting ensemble of decision trees [17]. The number of estimators, maximum depth, learning rate, subsample ratio, and column subsampling by tree were all part of the hyperparameter search.
- LightGBM (LGBM): It is an efficient gradient boosting for large-scale datasets [18]. The number of estimators, number of leaves, learning rate, and the minimum number of samples per leaf node were all included in the hyperparameter search.

3.2.2 Deep learning models

The study implements three deep learning models to learn nonlinear representations and complex feature interactions, as follows:

- Multi-Layer Perceptron (MLP): It is a feedforward neural network that can learn nonlinear feature interactions because of its completely linked hidden layers [13]. Hyperparameter optimization is performed using GridSearchCV within the inner cross-validation loop to find the best hidden layer architecture, regularization strength, and initial learning rate. Across the five outer folds, the most frequently selected configuration for the hidden layer sizes was a two-layer architecture with 64 and 32 neurons, followed by a single-layer architecture with 128 neurons.
- TabNet: It is an attentive and interpretable deep learning architecture designed for tabular data [8, 19]. Hyperparameter optimization was performed using a predefined grid-based candidate search within the inner cross-validation loop. The hyperparameter search included the decision feature dimension (n_d), attention feature dimension (n_a), number of decision steps (n_steps), feature reusage coefficient (gamma), sparsity regularization coefficient ($lambda\_sparse$), and learning rate.
- FT-Transformer: It is a transformer-based deep learning architecture specialized for tabular data [19]. Hyperparameter optimization was done using a predefined grid-based candidate search within the inner cross-validation loop. The search included several configurations varying the token embedding dimension, several of transformer blocks, several of attention heads, attention dropout, feed-forward network dropout, learning rate, weight decay, batch size, maximum number of training epochs, and early-stopping patience.

3.2.3 Ensemble learning strategies

The framework highlights the following computational strategies:

- Probability Calibration: Platt scaling or isotonic regression is applied to ensure the predicted probabilities and observed outcomes are in alignment. Each base model (ML and DL) is calibrated independently using Platt scaling or isotonic regression in the inner cross-validation

Table 4: Hyperparameter Search Space and Selected Configurations Across Outer Folds

Model	Hyperparameter Search	Selected Hyperparameters Across Folds
LR	regularization strength C ∈ {0.1, 1, 10}, penalty ∈ {l2}	C = 10, penalty = l2
RF	n_estimators = 300, max_depth ∈ {None, 20}, min_samples_leaf ∈ {1, 2}	max_depth = 20, min_samples_leaf = 2, n_estimators = 300
ET	n_estimators = 300, max_depth ∈ {None, 20}, min_samples_leaf ∈ {1, 2}	max_depth = 20, min_samples_leaf = 2, n_estimators = 300
KNN	n_neighbors ∈ {5, 11, 21}, weights ∈ {uniform, distance}, p ∈ {1, 2}	n_neighbors = 21, weights = uniform, p ∈ {1, 2}
XGB	n_estimators ∈ {200, 300}, max_depth ∈ {3, 5}, learning_rate ∈ {0.05, 0.1}, subsample ∈ {0.8, 1.0}, colsample_bytree ∈ {0.8, 1.0}	colsample_bytree = 0.8, learning_rate = 0.05, max_depth = 3, subsample = 0.8, n_estimators ∈ {200, 300}
LGBM	n_estimators ∈ {200, 300}, num_leaves ∈ {31, 63}, learning_rate ∈ {0.05, 0.1}, min_child_samples ∈ {20, 50}	learning_rate = 0.05, n_estimators = 200, num_leaves = 31, min_child_samples ∈ {20, 50}
MLP	hidden_layer_sizes ∈ {(64,), (128,), (64,32)}, alpha ∈ $\{10^{-4}, 10^{-3}\}$, learning_rate_init ∈ $\{10^{-3}, 5 \times 10^{-4}\}$	hidden_layer_sizes ∈ {(64,32), (128,)}, alpha ∈ $\{10^{-4}, 10^{-3}\}$, learning_rate_init ∈ $\{10^{-3}, 5 \times 10^{-4}\}$
TabNet	n_d ∈ {16, 24, 32}, n_a ∈ {16, 24, 32}, n_steps ∈ {3, 4, 5}, gamma ∈ {1.2, 1.5, 1.8}, lambda_sparse ∈ $\{10^{-5}, 10^{-4}, 10^{-3}\}$, learning rate ∈ $\{10^{-2}, 2 \times 10^{-2}\}$	n_d = 16, n_a = 16, n_steps ∈ {3, 4}, gamma ∈ {1.2, 1.5, 1.8}, lambda_sparse ∈ $\{10^{-5}, 10^{-4}, 10^{-3}\}$, learning rate ∈ $\{10^{-2}, 2 \times 10^{-2}\}$
FT-Transformer	d_token ∈ {32, 64, 128}, n_blocks ∈ {3, 4}, n_heads ∈ {4, 8}, attention_dropout ∈ {0.10, 0.15}, ffn_dropout ∈ {0.10, 0.15}	n_blocks = 3, n_heads = 4, attention_dropout = 0.10, residual_dropout = 0.00, weight_decay = 10^{-4}, batch_size = 512, max_epochs = 80, patience = 10, d_token ∈ {32, 64}, ffn_dropout ∈ {0.10, 0.15}, learning rate ∈ $\{10^{-3}, 7 \times 10^{-4}\}$

folds before building an ensemble. The calibrated probabilities are then used in voting and stacking ensembles. By combining well-calibrated forecasts, the ensembles help ensure more reliable probability estimates while mitigating the impact of biases arising from poorly calibrated base models.

- Nested Cross-Validation: It consists of outer folds for unbiased evaluation, inner folds for hyperparameter tuning, and selective probability calibration.
- Outer folds: They held out for unbiased evaluation. Outer-test folds are not used for hyperparameter tuning and calibration.
- Inner folds: They are only used for hyperparameter tuning and probability calibration of base learners. This procedure ensures that all ensemble predictions are *leakage-free*, and that probability calibration improves reliability without impacting the evaluation set.
- Ensemble Construction: After standalone evaluation, LR, ET, KNN, and XGB were retained for ML ensemble construction, while MLP, TabNet, and FT-Transformer were used for DL and hybrid ensemble experiments, as shown in FIGURE 1. This process consists of:
  - *Stacking*: For each outer-train fold, the logistic regression (LR) meta-learner is trained on the out-of-fold (OOF) predicted probabilities from the calibrated base models. Each column in the meta-feature matrix corresponds to a base model's probability output. This ensures that the meta-learner does not see the outer-test data during training, preserving strict leakage-free evaluation.
  - *Weighted Stacking*: The OOF probabilities of each base model are multiplied by performance-derived weights (from ROC-AUC, Brier Score, and ECE) before constructing the meta-feature matrix. This weighted matrix is then used to train the LR meta-learner. By weight-

ing base models according to their discrimination and calibration, higher-quality models have a stronger influence on the ensemble while maintaining leakage-free evaluation.
- *Soft Voting*: It averages predicted probabilities from selected ML and DL models.
- *Weighted Voting*: It assigns weights to models and averages the sum of weighted predictions as shown in Equation 1.

$$\hat{y}_{ensemble} = \sum_{i=1}^{N} w_i \hat{y}_i, \quad \sum_{i=1}^{N} w_i = 1 \tag{1}$$

where $\hat{y}_i$ is the predicted probability from model i and w_i is its normalized weight derived from validation performance.

This framework allows a flexible integration of diverse ML and DL models while maintaining rigorous evaluation protocols and producing well-calibrated, interpretable risk estimates for cardiovascular disease prediction.

4. EXPERIMENTS AND RESULTS

4.1 Model Performance

TABLE 5, presents the different metrics used in this study to evaluate the performance of the proposed model.

Table 5: Performance Metrics for Model Evaluation.

Metric	Formula	Purpose
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$	Overall correctness
Precision	$\frac{TP}{TP+FP}$	Correct positive predictions
Recall	$\frac{TP}{TP+FN}$	Sensitivity of the model
Specificity	$\frac{TN}{TN+FP}$	Detection of negative cases
F1 Score	$\frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}}$	Harmonic mean of precision and recall
ROC-AUC	AUC(ROC)	Class separation measure
PR-AUC	AUC(Precision-Recall)	Suitable for imbalanced datasets
Brier Score	$\frac{1}{N}\sum_{i=1}^{N}(p_i - y_i)^2$	Probability accuracy
Log Loss	$-\frac{1}{N}\sum_{i=1}^{N}[y_i \log(p_i) + (1-y_i)\log(1-p_i)]$	Penalizes incorrect probability predictions
ECE	$\sum_{m=1}^{M}\frac{\lvert B_m\rvert}{N}\lvert\text{acc}(B_m) - \text{conf}(B_m)\rvert$	Measures calibration quality; lower ECE indicates better calibration

4.2 Comparative Analysis

Table 6: Comparative Analysis of Performance Metrics and Probability Calibration for standalone ML and DL Models.

Model	Calibration	ROC_AUC	PR_AUC	Sensitivity	F1	Accuracy	Brier_Score	Log_Loss	Calib_Slope	Calib_Intercept	ECE
LGBMLGBMLGBMpt< -LGBMpt>	none	80.07	78.20	69.04	72.01	73.45	0.1807	0.5416	0.9887	-0.11	0.84
	isotonic	**80.10**	**78.23**	**69.33**	**72.12**	**73.48**	**0.1806**	**0.5413**	1.0050	-0.00182	**0.0083**
ETETETpt< -ETpt>	none	79.50	77.52	68.05	71.38	73.00	0.1837	0.5497	0.9343	-0.0043	0.0145
	sigmoid	**79.65**	**77.66**	**68.83**	**71.79**	**73.24**	**0.1830**	**0.5470**	1.03	-0.00081	0.0114
RFRFRFpt< -RFpt>	none	79.30	77.46	68.57	71.50	72.96	0.1846	0.5518	0.9126	-0.0021	0.0162
	sigmoid	**79.55**	**77.72**	**69.00**	**71.71**	**73.07**	**0.1830**	**0.5480**	1.04	0.003	**0.0127**
LRLRLRpt< -LRpt>	none	79.52	**77.37**	**68.50**	**71.52**	73.01	0.1838	0.5507	0.9984	-0.0003	0.0191
	isotonic	**79.53**	77.19	66.30	70.86	**73.04**	**0.1830**	**0.5480**	0.9900	-0.00093	**0.0099**
KNNKNNKNNpt< -KNNpt>	none	78.44	75.63	69.30	**71.34**	72.46	0.1891	0.6460	0.6114	-0.0182	0.0287
	isotonic	**78.98**	**76.90**	**67**	70.82	**72.72**	**0.1855**	**0.5540**	1.0500	0.00099	**0.0114**
XGB	none	**80.12**	**78.23**	**68.72**	**71.96**	**73.50**	**0.1805**	**0.5411**	1.0145	0.00014	**0.0091**
FT-Transformer	none	80.10	78.22	68.71	71.95	73.50	0.1807	0.5417	0.9950	-0.01397	0.0123
MLP	none	80.00	77.94	68.21	71.73	73.40	0.1813	0.5432	0.97	0.02690	0.0157
TabNet	none	79.73	77.78	67.60	71.32	73.10	0.1823	0.5455	0.99	0.00720	0.0128

TABLE 6, summarizes the comparative analysis of the standalone ML and DL models and their raw and calibrated variants. Each model was evaluated with no calibration, Platt scaling (sigmoid), or isotonic regression applied separately, whichever resulted in better Brier Score, Log Loss, and Expected Calibration Error (ECE). XGB is intrinsically well calibrated, and little gain from calibration can be seen. However, models like LR, KNN, ET, and LGBM can gain significantly from isotonic or sigmoid calibration in terms of ECE, Brier, and Log Loss metrics. The deep learning models (MLP, TabNet, FT-Transformer) also show moderate improvements in calibration, while their discrimination (ROC-AUC and PR-AUC) is largely unchanged. In general, this table demonstrates that selective pre-ensemble calibration significantly improves probability reliability for models that start miscalibrated, whereas well-calibrated models continue to perform consistently without the need for additional adjustments.

4.3 Calibration Analysis of Ensembles

To further investigate probability reliability, FIGURES 2 and 3, show the impact of pre-ensemble calibration on predicted probabilities. Isotonic regression or Platt scaling as shown in TABLE 6, were used to calibrate the base models within the inner folds of the cross-validation before building the ensemble. This method consistently decreased the Brier Score and Expected Calibration Error (ECE) for all ensembles, thus increasing the reliability of the probabilities. The most improvement was observed in non-XGB ensembles (e.g., LR, KNN, ET stacks/votes), which were also the worst calibrated to begin with. On the contrary, ensembles with XGB were naturally well-calibrated and changed little after calibration. This shows that calibration works best for poorly calibrated base models, and that high-quality base learners like XGB already produce reliable probability estimates without any further adjustment.

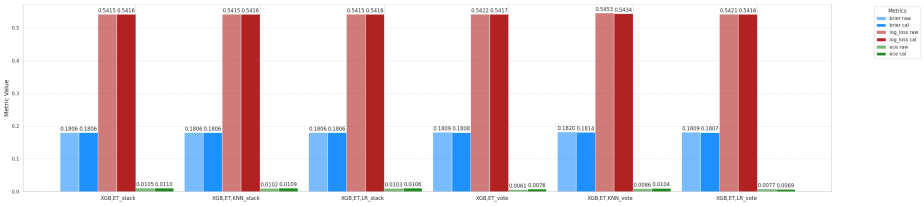


Figure 2: Calibration plots of raw and calibrated ensembles without XGB.

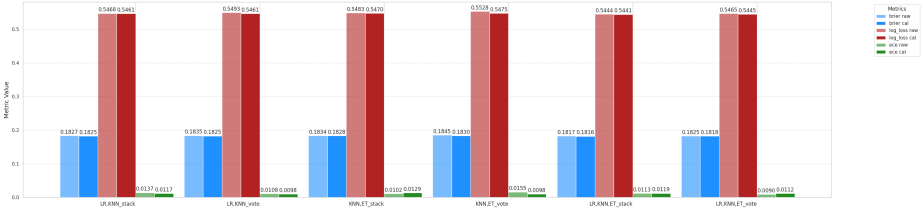


Figure 3: Calibration plots of raw and calibrated ensembles including XGB.

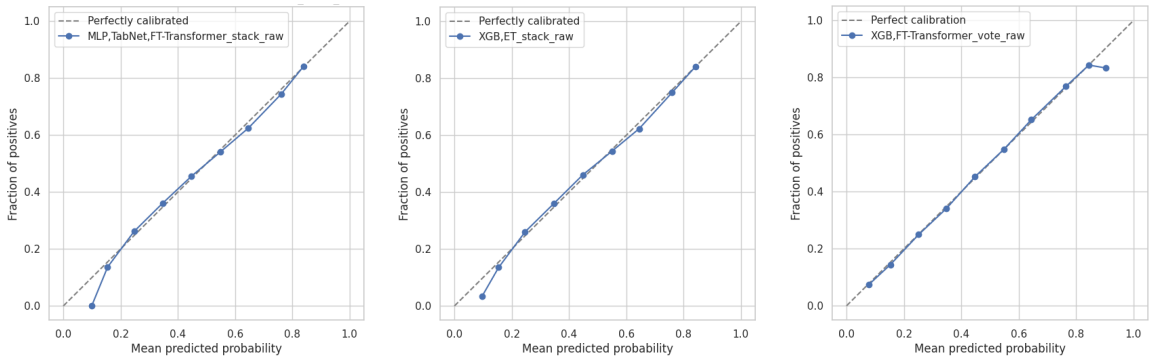
Table 7: Performance Metrics of Best ML, DL, and Hybrid Ensemble Metrics

Ensemble	ROC_AUC	PR_AUC	Sensitivity	F1	Accuracy	Brier Score	Log Loss	Calib Slope	Calib Intercept	ECE
Soft Voting XGB, FT-Transformer	**80.16**	78.22	68.72	71.99	73.55	**0.1803**	**0.5405**	1.01302	-0.00661	**0.004**
Raw Stacking XGB, ET	80.14	**78.23**	**69.03**	**72.1**	73.52	0.1806	0.5415	0.9999	0.0003	0.0105
Stacking MLP, TabNet, FT-Transformer	80.12	78.13	68.94	72.06	**73.56**	0.18064	0.5417	1.0004	-0.00007	0.0111

The performance of the best ML, DL and hybrid ensembles on multiple metrics such as discrimination (ROC-AUC, PR-AUC) and calibration (Brier Score, Log Loss, Calibration Slope/Intercept and ECE) is summarized in TABLE 7. The hybrid ensemble with soft voting of XGB and FT-Transformer got the highest ROC-AUC (80.16%) and the best calibration (Brier = 0.1803, Log Loss = 0.505, ECE = 0.004), demonstrating a balance between predictive accuracy and probability reliability. The raw stacking ML ensemble (XGB+ET) slightly outperforms the best DL stacking ensemble (MLP+TabNet+FT-Transformer) in terms of sensitivity and F1-score, but with a higher ECE, showing that calibration further enhances the reliability of probabilities. Calibration curves for these ensembles are shown in FIGURES 4 (a–c). All ensembles track the diagonal well, suggesting good calibration, with the largest gain for non-XGB models following pre-ensemble calibration. The XGB-based ensembles were already well-calibrated, which is consistent with their low ECE values. These results point out that selective pre-ensemble calibration can be beneficial to improve the probability estimates, especially for the base models that are poorly calibrated initially, while the high-performing base learners such as XGB can preserve the reliable probability predictions without further adjustment.

CONCLUSION AND FUTURE WORK

A calibration-aware hybrid machine learning and deep learning ensemble for the prediction of cardiovascular disease is presented in this paper. It combines four machine learning models, in-



(a) Stacking of MLP, TabNet, FT-Transformer (DL ensemble)

(b) Raw stacking of XGB and ET (ML ensemble)

(c) Soft voting of XGB and FT-Transformer (Hybrid ensemble)

Figure 4: Calibration plots for top-performing ML, DL, and hybrid ensembles.

cluding LR, ET, XGB, and KNN, and three deep learning models, namely MLP, TabNet, and FT-Transformer. These models are integrated into the ensemble using stacking and soft voting, including both weighted and unweighted variants. A comparative analysis is conducted to identify the best machine learning ensemble, the best deep learning ensemble, and subsequently to construct the best hybrid ensemble. The findings of this work show that soft voting of XGB and FT-Transformer achieved the highest discrimination ROC-AUC = 80.16% and best calibration metrics with Brier Score = 0.1803, Log Loss = 0.505, and ECE = 0.004. These results are based on a single public benchmark dataset. Future work will include model validation in independent cohorts, longitudinal and temporal integration of patient data, advanced feature engineering, and statistical significance analysis for better predictive performance and explainability to support clinical adoption.

Author Contribution

All authors have equally contributed to this work in conceptualization, data analysis, writing, reviewing, and editing.

References

[1] https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).

[2] Alsabhan W, Alfadhly A. Effectiveness of Machine Learning Models in Diagnosis of Heart Disease: A Comparative Study. Sci Rep. 2025;15:24568.

[3] Kumar R, Garg S, Kaur R, Johar MG, Singh S, et al. A Comprehensive Review of Machine Learning for Heart Disease Prediction: Challenges Trends Ethical Considerations and Future Directions. Front Artif Intell. 2025;8:1583459.

[4] Janosi A, Steinbrunn W, Pfisterer M, Detrano R. Heart Disease [dataset]. UCI machine learning repository. 1989.

[5] Korkmaz S. Bioleak: Leakage-Aware Modeling and Diagnostics for Machine Learning in R. 2026. Arxiv preprint: https://arxiv.org/pdf/2604.10965.

[6] https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset.

[7] https://www.kaggle.com/datasets/aasheesh200/framingham-heart-study-dataset.

[8] Yasmeen R, Khan L, Choi A. Heart Disease Prediction Using Hybrid Tabnet Architecture With Stacked Ensemble Learning. Front Physiol. 2025;16:1665128.

[9] Ganie SM, Pramanik PK, Zhao Z. Ensemble Learning With Explainable AI for Improved Heart Disease Prediction Based on Multiple Datasets. Sci Rep. 2025;15:13912.

[10] Zaidi SA, Ghafoor A, Kim J, Abbas Z, Lee SW. Heartensemblenet: An Innovative Hybrid Ensemble Learning Approach for Cardiovascular Risk Prediction. Healthcare. 2025;13:507.

[11] Sadr H, Salari A, Ashoobi MT, Nazari M. Cardiovascular Disease Diagnosis: A Holistic Approach Using the Integration of Machine Learning and Deep Learning Models. Eur J Med Res. 2024;29:455.

[12] Oyewola DO, Dada EG, Misra S. Diagnosis of Cardiovascular Diseases by Ensemble Optimization Deep Learning Techniques. Int J Healthc Inf Syst Inform. 2024;19:1-21.

[13] Setiawan DW, Lapatta NT, Amriana A, Nugraha DW, Lamasitudju CA. Performance Com Parison of Multilayer Perceptron (MLP) and Random Forest for Early Detection of Cardiovascular Disease. J Appl Inform Comput. 2025;9:3012-3024.

[14] Asadi F, Homayounfar R, Mehrali Y, Masci C, Talebi S, et al. Detection of Cardiovascular Disease Cases Using Advanced Tree-Based Machine Learning Algorithms. Sci Rep. 2024;14:22230.

[15] Habib M, Al Ali O, Merhi Z. Earthquake Tweets Analysis System Using Machine Learn Ing Techniques. In 2026 International Conference on Artificial Intelligence, Computer, Data Sciences and Applications (ACDSA). IEEE. 2026:1-6.

[16] Talukder SH, Mondal SK, Aljaidi M, Sulaiman RB, Islam T. Heart Disease Risk Assessment and Prediction: A Robust Ensemble Approach With Extra Tree Classifier. In 2023 2nd International Engineering Conference on Electrical, Energy, and Artificial Intelligence (EICEEAI). IEEE. 2023:1-6.

[17] Habib M, Ayoubi MA, Kanaan M, Merhi Z. Joint Forecasting of Residential Energy Consumption and Solar Generation Using Advanced AI Architectures. Adv Artif Intell Mach Learn. 2026;6:5160-5175.

[18] Sawlikar A, Behera P, Dingankar SU, Yadav RK, Kumari N, et al. Heart disease prediction using light GBM and real-time patient data analysis. In AIP Conference Proceedings.AIP Publishing LLC. 2026;3386:020139.

[19] Komma SK, Vullam N, Sankar GS, Bhaskar PV, Lakshmanarao A, et al. Novel AP Proaches to Heart Disease Detection With Tabnet and Ft Transformer Models. In 2025 International Conference on Visual Analytics and Data Visualization (ICVADV). IEEE. 2025:768-772.