ISSN: 2181-3337 | SCIENTISTS.UZ VOLUME 5 ISSUE 10 | OCTOBER 2026

A LEAKAGE-SAFE STACKING ENSEMBLE OF GRADIENT BOOSTING MODELS FOR HEART DISEASE PREDICTION: ABLATION, CLINICAL METRICS AND SUBGROUP ANALYSIS

Tolibova Nafisa[1], Kholdorov Shokhrukhmirzo[2], Eshmurodova Dildorakhon[3], Sulaymonova Fozila[4]

Trainee Lecturer, Department of Modern Information Technologies and Artificial Intelligence, Uzbekistan State World Languages University[1]
Acting Associate Professor (PhD), Tashkent University of Information Technologies named after Muhammad al-Khwarizmi[2]
Master's Student, Faculty of Computer Engineering, Tashkent University of Information Technologies named after Muhammad al-Khwarizmi[3]
PhD Student, Tashkent University of Information Technologies named after Muhammad al-Khwarizmi[3]

DOI: 10.67620/sai.510d0004

ABSTRACT

Heart disease remains the leading cause of death worldwide, and much of the information needed to flag patients at risk is already collected during routine cardiac work-ups. This study evaluates a stacking ensemble that combines XGBoost, LightGBM and CatBoost through a logistic regression meta-learner. The experiments use the 630,000-record synthetic dataset of the Kaggle Playground Series (Season 6, Episode 2), which was generated by a deep learning model trained on an original clinical dataset with the Statlog (Heart) schema. To rule out data leakage, all preprocessing was placed inside the pipeline of each base learner, and three additional features were computed from fixed formulas that require no fitting. Six models were compared on identical five-fold stratified splits. The ensemble reached a ROC-AUC of 0.9553; at a threshold of 0.5 its sensitivity was 0.867, specificity 0.906, PPV 0.883 and NPV 0.894, and the Youden threshold (0.40) raised sensitivity to 0.887. The ensemble exceeded the best single model, CatBoost, by only 0.00004 in AUC and logistic regression by 0.0022. An ablation study showed that the engineered features did not improve performance ($\Delta AUC = -0.00003$, $p = .058$). Permutation importance and SHAP values identified the thallium test, chest pain type, Health_Score and maximum heart rate as the leading predictors. In women the AUC equalled that of men (0.947), but sensitivity fell to 0.70, which reflects threshold choice in a low-prevalence group. The findings apply to synthetic data and require external validation on real patient records before clinical use.

ACCEPTED

07.11.26

Keywords: *heart disease, machine learning, stacking ensemble, gradient boosting, data leakage, SHAP, subgroup analysis.*



Introduction

Cardiovascular disease accounted for an estimated 18.6 million deaths in 2019, and its burden continues to rise in low- and middle-income countries as populations age (Roth et al., 2020). Much of this

burden is due to ischaemic heart disease, which often progresses without symptoms until a first acute event. Identifying patients at risk earlier is one of the most direct ways to reduce mortality.

A cardiologist rarely relies on a single test. The decision draws on age and sex, the type of chest pain, resting blood pressure and cholesterol, the resting electrocardiogram and the response to an exercise test. These variables are inexpensive to record and are present in most clinical records, but their interactions are non-linear and differ between patient groups. For such structured data, ensembles of gradient-boosted decision trees still tend to outperform deep neural networks of comparable cost (Grinsztajn et al., 2022; Shwartz-Ziv & Armon, 2022). Stacked generalisation goes a step further by training a second-level model on the predictions of several first-level models (Wolpert, 1992).

Stacking is attractive, but it is also easy to evaluate incorrectly. A gain from stacking is meaningful only when it is measured against the individual base learners on the same splits, when every fitted transformation is learned from training data alone, and when the results are reported with metrics that matter clinically rather than accuracy alone. This paper studies a stack of XGBoost, LightGBM and CatBoost under these conditions. Its contributions are:

1. A leakage-safe design in which preprocessing is fitted inside each base-learner pipeline and the engineered features use fixed constants only.
2. A direct comparison of the stack with each base learner and with logistic regression on identical folds, with paired significance tests, and an ablation of the three engineered features.
3. A clinically oriented evaluation: sensitivity, specificity, predictive values, calibration, threshold analysis, model-level explanations and subgroup performance by sex and age.

Literature Review

The attribute set used in most heart disease studies goes back to Detrano et al. (1989), who combined clinical and exercise-test variables into a probability model for coronary artery disease and tested it on patients from four centres. The same thirteen attributes, with the thallium scan coded as normal, fixed or reversible defect, form the UCI *Statlog (Heart)* dataset (*Statlog (Heart)*, n.d.), a widely used benchmark.

Recent work has moved from single classifiers to ensembles and larger cohorts. Bhatt et al. (2023) compared random forest, decision tree, XGBoost and a multilayer perceptron on about 70,000 records and reported accuracies close to 87% with AUC values of 0.94–0.95. Wu et al. (2024) proposed a stacking model with KNN, decision tree and SVM base learners and an attention-based BiLSTM meta-learner, reporting that dimensionality reduction raised AUC on the Cleveland data from 0.76 to 0.92. Qiu et al. (2024) tuned a CatBoost classifier with opposition-based cuckoo search and reached an AUC of 0.80 on a cardiovascular risk task. These studies show that boosting and stacking are competitive, but they seldom report how much a stack adds over its strongest member, and several rely on a single train–test split.

The three base learners used here grow trees differently. XGBoost uses second-order gradient information and explicit regularisation of leaf weights (Chen & Guestrin, 2016). LightGBM grows trees leaf-wise on histogram-binned features, which makes it fast on large data (Ke et al., 2017). CatBoost applies ordered boosting with target statistics computed on permutations, reducing prediction shift on categorical inputs (Prokhorenkova et al., 2018). Because their errors are only partly correlated, they are natural stacking candidates.

Methodological reviews warn that leakage, the use of information during training that would not be available at prediction time, is one of the most common sources of over-optimistic results (Kaufman et al., 2012). The TRIPOD+AI statement accordingly asks that prediction model studies describe the data source, report discrimination together with calibration, and examine performance across relevant subgroups (Collins et al., 2024). The present study follows these recommendations.

ISSN: 2181-3337 | SCIENTISTS.UZ

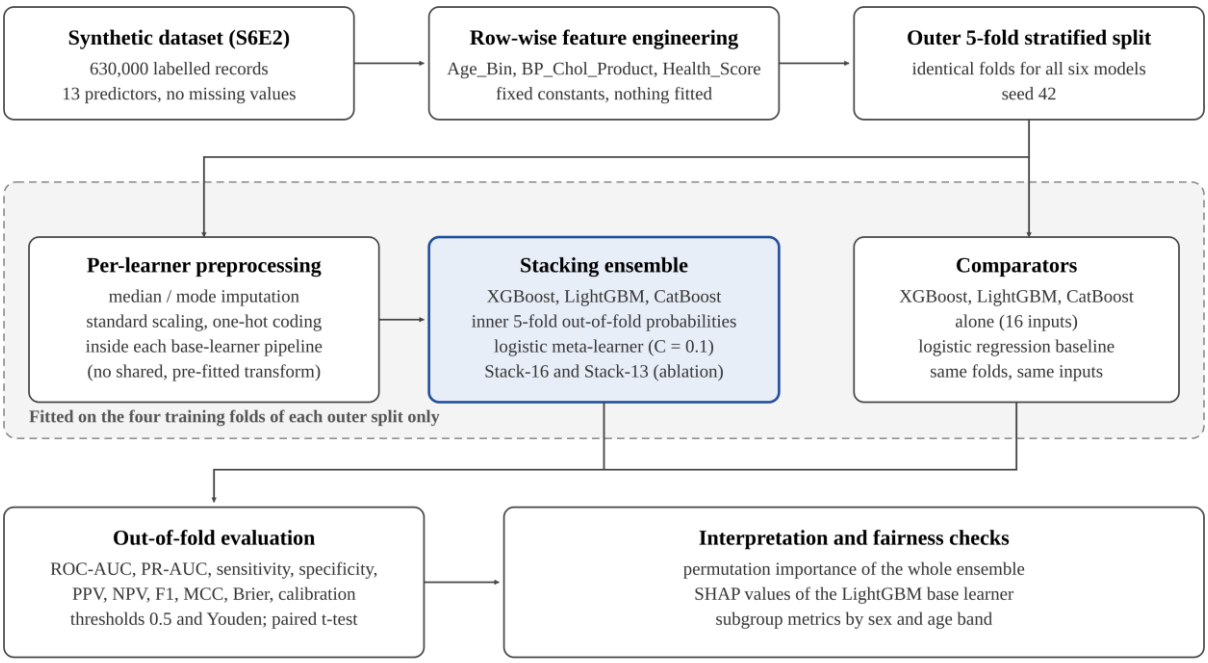
Research Methodology

Study Design

The study is a retrospective analysis of tabular data, and Figure 1 summarises the workflow. Three features are first derived row by row with fixed constants. The data are then divided into five stratified outer folds that are shared by all models. Within each outer split, every fitted component, including imputation, scaling, encoding, the base learners and the meta-learner, is trained on the four training folds only and evaluated on the fifth. All experiments were run in Python 3.12 with scikit-learn 1.9, XGBoost 3.4, LightGBM 4.7 and CatBoost 1.2 on a single NVIDIA RTX 6000 Ada GPU; the complete experiment, covering six models over five folds, took 15.9 minutes.

Figure 1

Study Workflow



Note. Study workflow. All fitted steps are confined to the training folds of each outer split; the six models share identical folds.

Dataset And Its Provenance

The data come from the Kaggle Playground Series, Season 6, Episode 2, “Predicting Heart Disease” (Kaggle, 2026). According to the competition description, the records were generated by a deep learning model trained on an original heart disease dataset, so feature distributions are close to, but not identical with, the original clinical data. The data are distributed under the CC BY 4.0 licence. Only the labelled training file was used: 630,000 records, 13 predictors and a binary target with the values Presence and Absence. The predictor coding follows the UCI *Statlog (Heart)* schema (*Statlog (Heart)*, n.d.) (Table 1). No values were missing. The classes were moderately balanced, with 347,546 Absence (55.2%) and 282,454 Presence (44.8%) records, so no resampling was applied.

Table 1

Predictors Used in the Study

Variable	Description	Type / coding
Age	Age in years	Numerical
Sex	1 = male, 0 = female	Categorical
Chest pain type	Typical angina, atypical angina, non-anginal pain, asymptomatic	Categorical (1–4)
BP	Resting blood pressure, mm Hg	Numerical
Cholesterol	Serum cholesterol, mg/dl	Numerical
FBS over 120	Fasting blood sugar > 120 mg/dl	Binary
EKG results	Resting electrocardiogram	Categorical (0–2)
Max HR	Maximum heart rate achieved	Numerical
Exercise angina	Exercise-induced angina	Binary (0/1)
ST depression	ST depression induced by exercise relative to rest	Numerical
Slope of ST	Slope of the peak exercise ST segment	Categorical (1–3)
Number of vessels fluro	Major vessels coloured by fluoroscopy	Categorical (0–3)
Thallium	Thallium stress test: 3 normal, 6 fixed, 7 reversible defect	Categorical

Because the records are synthetic, they contain no personal data, but they also do not represent any specific patient population. All results in this paper therefore describe performance on the synthetic distribution; their transfer to real clinical cohorts is discussed in the "Subgroup Analysis and Discussion" section.

Feature Engineering Without Fitted Statistics

Three features were added on clinical grounds. Each one is computed from a single record using constants fixed before the analysis, so no information can pass between folds.

1. **Age_Bin** groups age into four bands: up to 40 (young), 41–55 (middle-aged), 56–65 (senior) and over 65 (elderly).
2. **BP_Chol_Product** expresses the joint burden of hypertension and hypercholesterolaemia on a log scale:

$$\text{BP_Chol_Product} = \ln(1 + \text{BP} \cdot \text{Chol}) \qquad (1)$$

3. **Health_Score** summarises the exercise-test and fluoroscopy findings. Each component is divided by the upper bound of its clinical range in the Statlog schema (6.2 mm for ST depression, 3 for the number of vessels) rather than by a minimum and maximum estimated from the data:

$$HS = 0.4\,\frac{\min(\max(x_{ST}, 0),\ 6.2)}{6.2} + 0.4\,\frac{\min(\max(x_V, 0),\ 3)}{3} + 0.2\,x_A \qquad (2)$$

Here x_{ST} is ST depression, x_V the number of vessels and x_A exercise-induced angina coded 0/1. An earlier version of this pipeline mapped angina from text labels although the data store it as 0/1, which set the angina term to zero for every record. After correction, the mean score is 0.078 for patients without and 0.381 for patients with exercise angina, confirming that all three components contribute. With these features each record has 16 inputs: 7 numerical and 9 categorical.

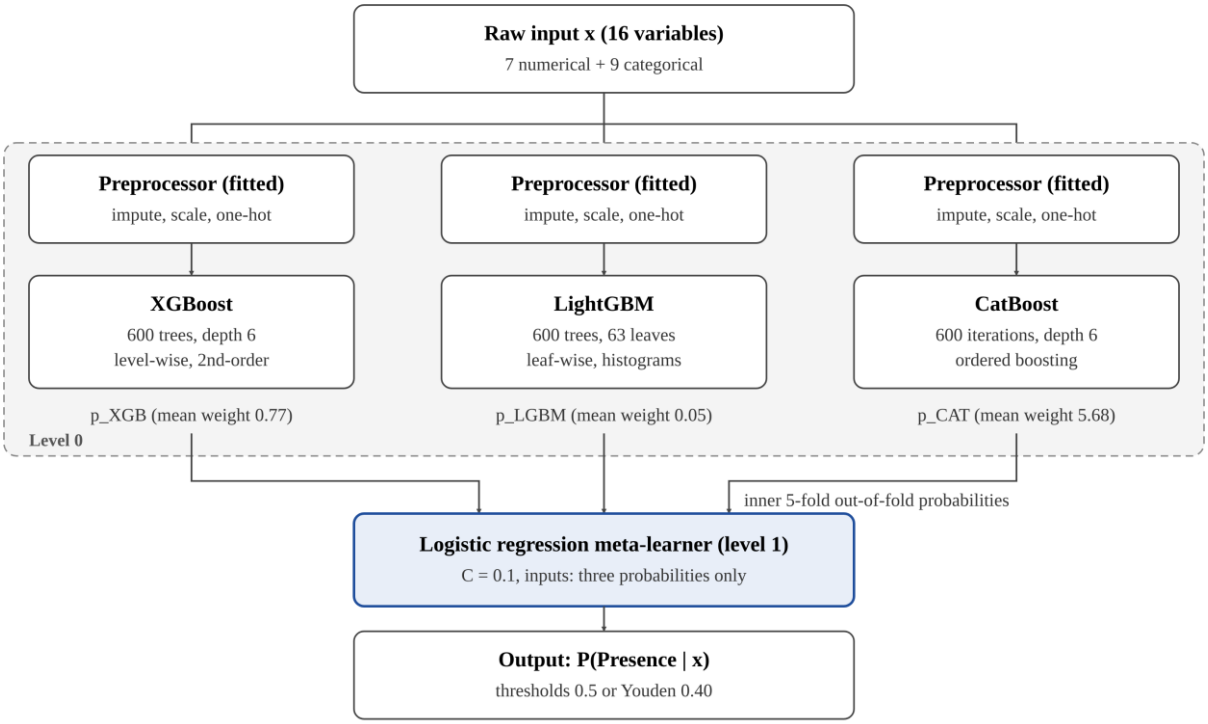
Preprocessing And Model Configuration

Numerical inputs were imputed with the median and standardised; categorical inputs were imputed with the most frequent value and one-hot encoded, with unseen categories ignored. The preprocessor is not shared: each base learner is a separate scikit-learn pipeline (Pedregosa et al., 2011) that contains its own copy of the preprocessor (Figure 2). When the stacking classifier creates its internal folds, it clones these pipelines, so the preprocessing statistics used to produce each out-of-fold probability are learned from the inner training part only.

ISSN: 2181-3337 | SCIENTISTS.UZ

Figure 2

Architecture of the Stacking Ensemble



Note. Architecture of the stacking ensemble. Each base learner carries its own preprocessor, fitted only on the data it is trained on; the meta-learner receives the three out-of-fold probabilities. Mean meta-learner weights over the five outer folds are shown.

The level-1 model is an L2-regularised logistic regression that receives only the three base probabilities (no passthrough of the original inputs). Hyperparameters were fixed in advance (Table 2) and the random seed was 42 throughout.

Table 2

Settings of the Base Learners, the Meta-Learner and the Baseline

Model	Role	Main settings
XGBoost	Level 0	600 trees, learning rate 0.05, max depth 6, subsample 0.8, column sample 0.8, $\gamma = 0.1$, $\alpha = 0.05$, $\lambda = 1.0$, histogram method
LightGBM	Level 0	600 trees, learning rate 0.05, 63 leaves, subsample 0.8, column sample 0.8, $\alpha = 0.05$, $\lambda = 1.0$, min. 30 samples per leaf
CatBoost	Level 0	600 iterations, learning rate 0.05, depth 6, L2 leaf regularisation 3, Bernoulli bootstrap, subsample 0.8
Logistic regression	Level 1	C = 0.1, L2 penalty, inputs: three out-of-fold probabilities
Logistic regression	Baseline	C = 1.0, L2 penalty, same preprocessing and 16 inputs

The meta-learner estimates the probability of disease as

$$\hat{p}(y = 1 \mid \mathbf{x}) = \sigma(w_0 + w_1 p_{XGB} + w_2 p_{LGBM} + w_3 p_{CAT}), \qquad \sigma(z) = \frac{1}{1 + e^{-z}} \qquad (3)$$

Validation Procedure And Compared Models

Algorithm 1 gives the procedure. Six models were evaluated on the same outer folds: the stack with 16 inputs (Stack-16), the stack with the 13 original inputs (Stack-13, ablation), XGBoost, LightGBM and CatBoost alone with 16 inputs, and logistic regression with 16 inputs. For the single boosters, the out-of-fold predictions are those of the base learners inside Stack-16, which were refitted on the full training part

of each outer fold; this ensures that the stack and its members are compared on exactly the same fitted models.

Algorithm 1. Leakage-safe evaluation of the stacking ensemble and comparators

Input : D = {(x_i, y_i)}, i = 1..N; base learners B = {XGB, LGBM, CAT}
Output: out-of-fold probabilities p_m for every model m; metrics

```
D <- AddFeatures(D)        // Age_Bin, BP_Chol_Product, Health_Score;
                           // fixed constants, no fitting
split D into 5 stratified folds F_1..F_5 (shuffle, seed = 42)
for k = 1 to 5 do
    D_tr <- D \ F_k ;  D_va <- F_k
    for each b in B do
        pipe_b <- Pipeline(Preprocessor, b)       // own preprocessor
        z_b <- inner 5-fold out-of-fold probabilities of pipe_b on D_tr
        refit pipe_b on all of D_tr
    end for
    fit meta-learner M on [z_XGB, z_LGBM, z_CAT]
    p_Stack16[F_k] <- M(pipe_XGB(D_va), pipe_LGBM(D_va), pipe_CAT(D_va))
    p_b[F_k] <- pipe_b(D_va) for each b in B
    repeat lines 5-11 with 13 inputs -> p_Stack13[F_k]
    fit Pipeline(Preprocessor, LogReg) on D_tr -> p_LogReg[F_k]
end for
compute metrics on all out-of-fold probabilities and per fold
paired t-test on fold-level ROC-AUC (Stack-16 vs each model)
```

Evaluation Metrics

Discrimination was measured with the area under the ROC curve (ROC-AUC), which equals the probability that a randomly chosen patient with disease receives a higher score than one without (Fawcett, 2006), and with the area under the precision–recall curve (PR-AUC). Threshold-dependent metrics were computed from the confusion matrix:

$$\text{Sens} = \frac{TP}{TP + FN}, \quad \text{Spec} = \frac{TN}{TN + FP}, \quad \text{PPV} = \frac{TP}{TP + FP}, \quad \text{NPV} = \frac{TN}{TN + FN} \qquad (4)$$

The F1-score of the Presence class is the harmonic mean of PPV and sensitivity, and the weighted F1 averages the per-class F1 values with weights equal to class support:

$$F1 = \frac{2\,\text{PPV} \cdot \text{Sens}}{\text{PPV} + \text{Sens}}, \qquad F1_w = \sum_c \frac{n_c}{N} F1_c \qquad (5)$$

Neither F1 value is the same as accuracy, $(TP + TN)/N$, and the two are reported separately. Metrics were evaluated at the conventional threshold of 0.5 and at the threshold that maximises Youden's index (Youden, 1950):

$$J = \max_t [\text{Sens}(t) + \text{Spec}(t) - 1] \qquad (6)$$

The Matthews correlation coefficient (MCC) was added as a balanced summary of the full confusion matrix (Chicco & Jurman, 2020). Calibration was assessed with the Brier score (Brier, 1950) and with reliability curves over ten quantile bins (Van Calster et al., 2019):

$$\text{Brier} = \frac{1}{N}\sum_{i=1}^{N}(\hat{p}_i - y_i)^2 \qquad (7)$$

ISSN: 2181-3337 | SCIENTISTS.UZ

Metrics are reported as the mean ± standard deviation over the five outer folds. Differences in ROC-AUC between Stack-16 and each comparator were tested with a paired *t*-test on the five fold-level values. Because training sets in k-fold cross-validation overlap, the naive test understates the variance; the corrected resampled *t*-test of Nadeau and Bengio (2003) was therefore also applied.

Explainability And Subgroup Analysis

Two complementary views of feature influence were used. Permutation importance (Breiman, 2001) was computed on the whole Stack-16 ensemble from the last outer fold: each input was shuffled in a random sample of 20,000 validation records, five times, and the mean drop in ROC-AUC was recorded. Because the inputs are shuffled before preprocessing, a categorical variable is permuted as a whole rather than column by column. SHAP values (Lundberg & Lee, 2017) were computed with TreeSHAP for the LightGBM base learner on 3,000 validation records, and the absolute values of one-hot columns were summed per original variable. Finally, out-of-fold metrics of Stack-16 were computed separately for women and men and for the four age bands.

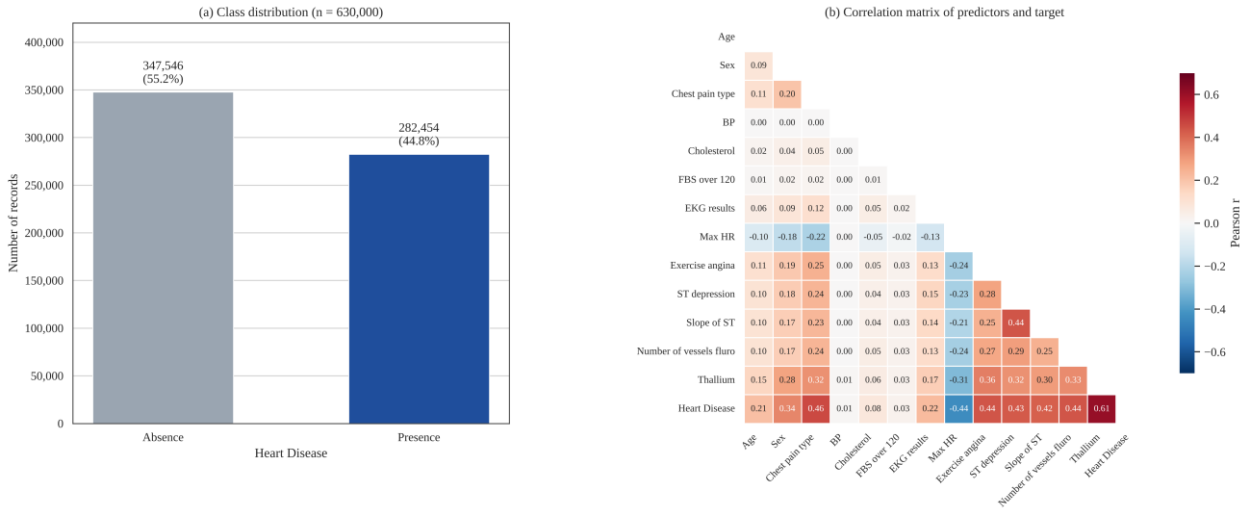
Analysis And Results

Exploratory Analysis

The thallium stress-test result showed the strongest linear association with the target ($r = 0.61$), followed by chest pain type, the exercise-test variables and the number of vessels, all between 0.42 and 0.46 in absolute value (Figure 3b). Maximum heart rate was the only predictor with a clear negative correlation ($r = -0.44$). Resting blood pressure, cholesterol and fasting blood sugar were almost uncorrelated with the outcome on their own ($|r| \leq 0.08$).

Figure 3

Class Distribution and Predictor Correlations



Note. Exploratory analysis: (a) class distribution; (b) Pearson correlation matrix of the predictors and the target. Categorical variables enter with their integer codes, so the values indicate direction and approximate strength only.

Comparison With Base Learners And Baseline

Table 3 reports the performance of all six models at the 0.5 threshold. All three boosters and both stacks lie within 0.0002 of each other in ROC-AUC, between 0.9551 and 0.9553. Logistic regression is about 0.002 lower (0.9531) and has the lowest sensitivity (0.860). Accuracy (0.8887) and the F1-score of the Presence class (0.8748) differ by 0.014 for the stack, which illustrates that F1 should not be read as accuracy.

Table 3

Out-of-Fold Performance at Threshold 0.5

Model	ROC-AUC	Acc.	Sens.	Spec.	PPV	NPV	F1	MCC	Brier
Stack-16	0.9553	0.8887	0.8673	0.9062	0.8825	0.8936	0.8748	0.7748	0.0831
Stack-13	0.9553	0.8887	0.8671	0.9062	0.8825	0.8935	0.8748	0.7747	0.0831
XGBoost	0.9551	0.8885	0.8677	0.9055	0.8818	0.8939	0.8747	0.7744	0.0814
LightGBM	0.9551	0.8884	0.8674	0.9054	0.8817	0.8937	0.8745	0.7741	0.0815
CatBoost	0.9552	0.8886	0.8669	0.9062	0.8825	0.8934	0.8747	0.7745	0.0813
LogReg	0.9531	0.8858	0.8601	0.9067	0.8822	0.8886	0.8710	0.7687	0.0833

Note. Values are means over five folds (n = 630,000). SD of ROC-AUC was 0.0004 for every model; F1 refers to the Presence class.

The standard deviations of the threshold metrics were between 0.0004 and 0.0022 for all models. The fold-by-fold comparison (Table 4) shows that the ranking is stable: Stack-16 was ahead of every single booster and of logistic regression in all five folds. The differences were statistically significant under both the naive and the corrected t-test, but they are very small in absolute terms. The advantage over CatBoost, 0.00004 in AUC, corresponds to fewer than one correctly re-ordered pair in twenty thousand and has no practical importance. The meta-learner weights explain this result: averaged over folds, CatBoost received a coefficient of 5.68, XGBoost 0.77 and LightGBM 0.05, so the stack behaves almost as a recalibrated CatBoost.

Table 4

Paired Comparison of Fold-Level ROC-AUC

Comparison	Mean ΔAUC	Folds Stack-16 better	p (paired t)	p (corrected)
vs Stack-13 (ablation)	−0.000031	1 / 5	.058	.153
vs XGBoost	+0.000137	5 / 5	< .001	< .001
vs LightGBM	+0.000212	5 / 5	< .001	< .001
vs CatBoost	+0.000043	5 / 5	< .001	< .001
vs logistic regression	+0.002235	5 / 5	< .001	< .001

Note. Differences are Stack-16 minus the comparator. The corrected test follows Nadeau and Bengio (2003).

Ablation Of The Engineered Features

Removing Age_Bin, BP_Chol_Product and Health_Score did not reduce performance. Stack-13 had the same mean ROC-AUC as Stack-16 (0.9553), was marginally better in four of five folds, and the difference was not significant (p = .058; corrected p = .153). The engineered features therefore re-express information that the boosters already extract from the original variables; they are retained in the final model only because they make the explanations in the “Explainability” section easier to interpret, not because they improve accuracy.

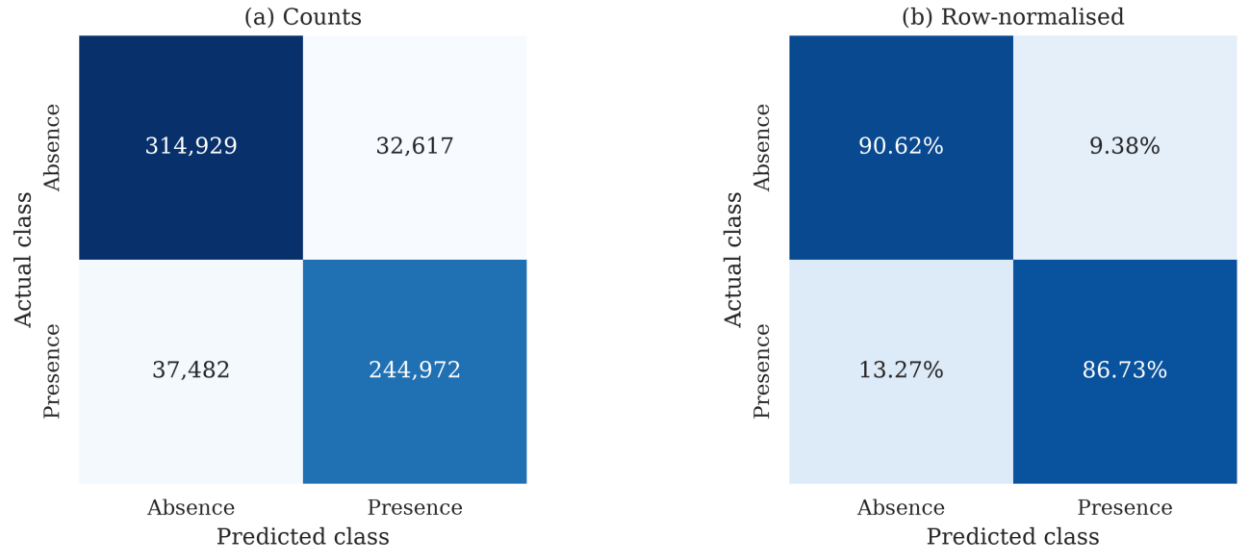
Clinical Operating Characteristics And Calibration

At the 0.5 threshold, Stack-16 correctly identified 244,972 of 282,454 patients with disease (sensitivity 0.867) and 314,929 of 347,546 without (specificity 0.906); 37,482 cases were missed and 32,617 were false alarms (Figure 4). The ROC curves of all six models practically coincide (Figure 5a).

ISSN: 2181-3337 | SCIENTISTS.UZ

Figure 4

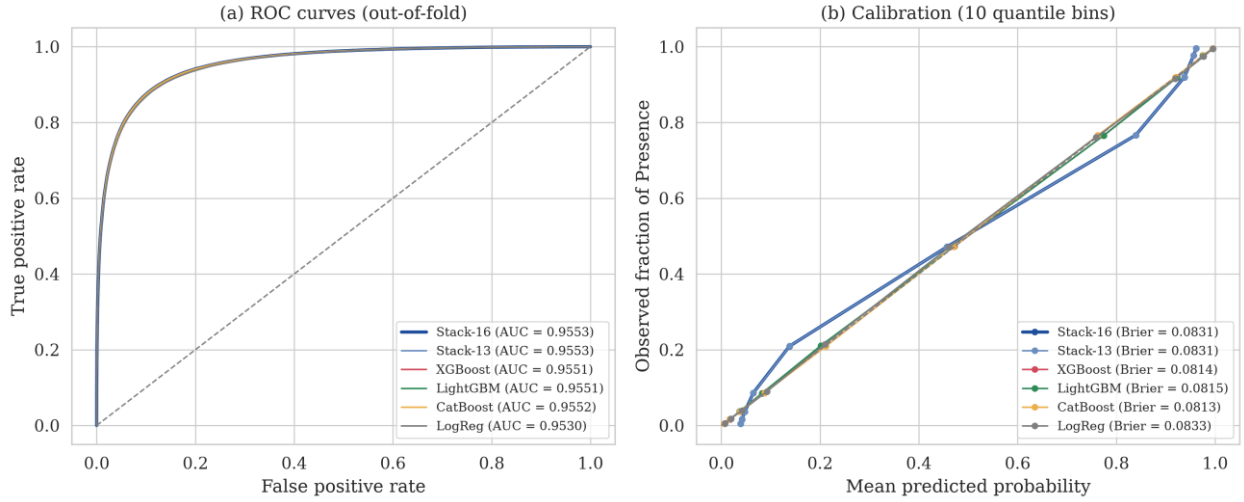
Confusion Matrix of Stack-16 at Threshold 0.5



Note. Confusion matrix of Stack-16 on all out-of-fold predictions at threshold 0.5: (a) counts; (b) row-normalised rates.

Figure 5

ROC Curves and Calibration



Note. (a) Out-of-fold ROC curves of the six models; (b) reliability curves over ten quantile bins with Brier scores.

The base learners were well calibrated, with Brier scores of 0.0813–0.0815. The stack had a slightly higher Brier score (0.0831): its reliability curve over-predicts risk in the 0.1–0.2 range and under-predicts it near 0.8 (Figure 5b). This is a side effect of the strongly regularised meta-learner, which compresses the base probabilities onto a single logistic curve. When calibrated probabilities are needed, as in risk communication, CatBoost alone or a stack with a recalibration step is preferable.

For screening, missing a patient with disease is usually costlier than a false alarm. The Youden threshold of 0.402 raised sensitivity from 0.867 to 0.887 and NPV from 0.894 to 0.906, while specificity fell from 0.906 to 0.889 (Table 5, Figure 6). A threshold of 0.3 raised sensitivity to 0.907 with a specificity of 0.864. MCC changed by less than 0.01 across thresholds of 0.3–0.6, so the choice of threshold is a clinical trade-off rather than a question of overall model quality.

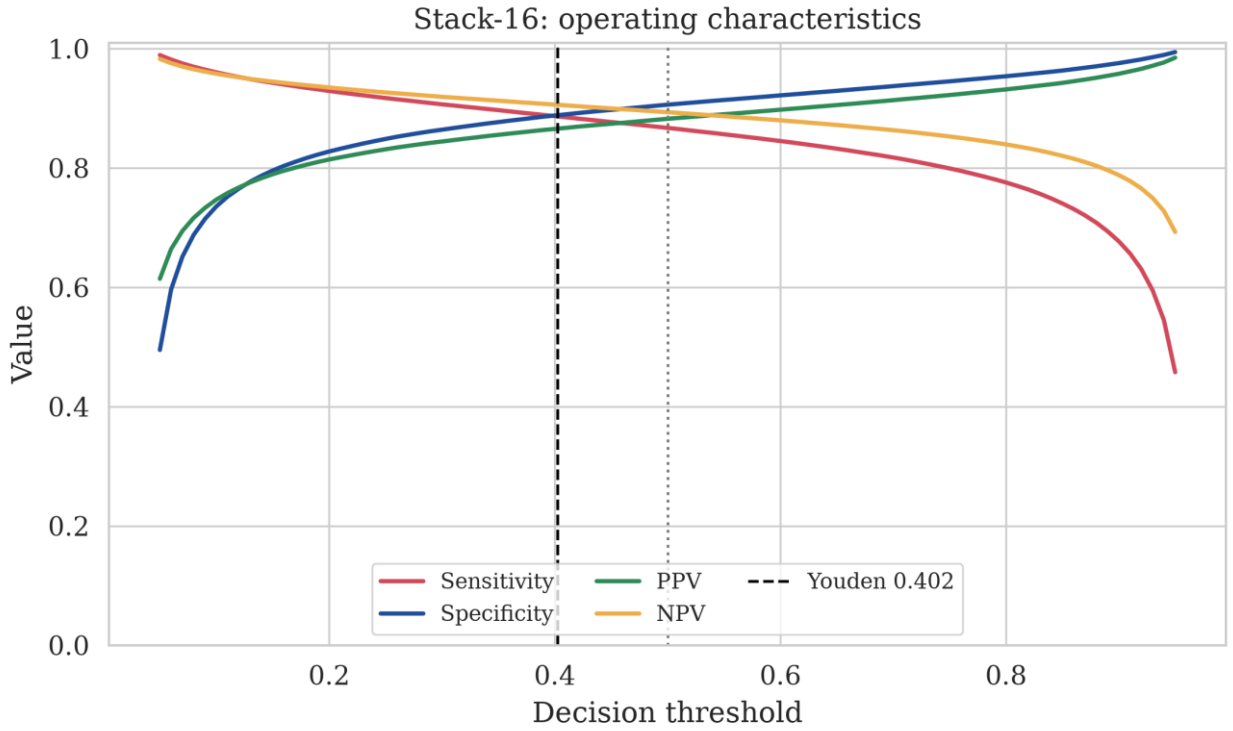
Table 5

Operating Characteristics of Stack-16 at Different Thresholds

Threshold	Sensitivity	Specificity	PPV	NPV	Accuracy	F1 (Presence)	MCC
0.30	0.9068	0.8642	0.8444	0.9194	0.8833	0.8745	0.7674
0.40	0.8872	0.8880	0.8656	0.9064	0.8876	0.8762	0.7735
0.402 (Youden)	0.8868	0.8884	0.8659	0.9061	0.8877	0.8762	0.7736
0.50	0.8673	0.9062	0.8825	0.8936	0.8887	0.8748	0.7748
0.60	0.8453	0.9218	0.8978	0.8799	0.8875	0.8707	0.7724
0.70	0.8176	0.9373	0.9137	0.8635	0.8836	0.8630	0.7660

Figure 6

Operating Characteristics Across Decision Thresholds



Note. Sensitivity, specificity, PPV and NPV of Stack-16 as functions of the decision threshold. Dashed line: Youden threshold (0.402); dotted line: 0.5.

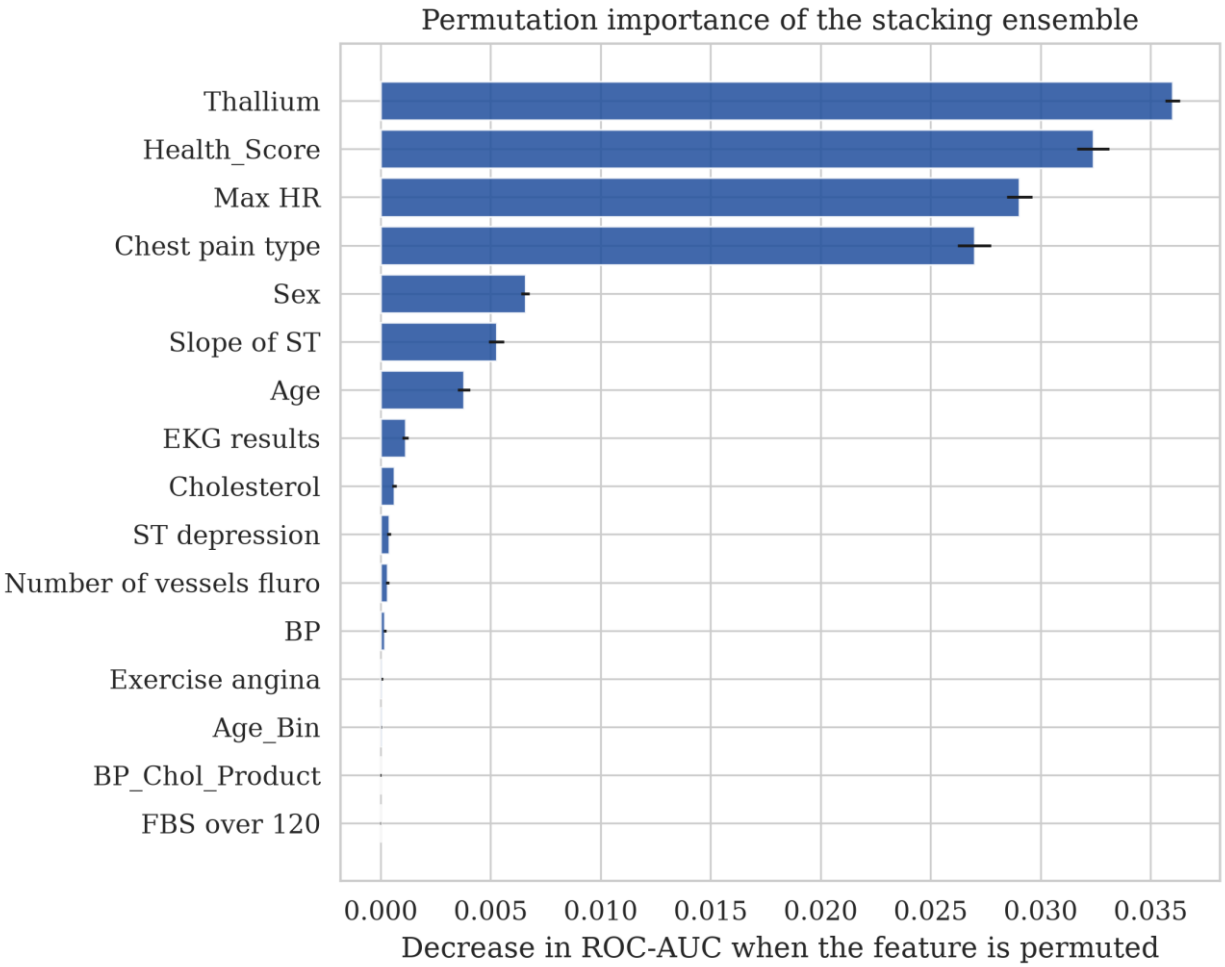
Explainability

Permutation importance of the whole ensemble (Figure 7) ranked four inputs far above the rest: thallium (ΔAUC = 0.036), Health_Score (0.032), maximum heart rate (0.029) and chest pain type (0.027). Sex, slope of ST and age followed with drops of 0.004–0.007, and the remaining inputs each caused a drop below 0.002. SHAP values of the LightGBM base learner gave the same leading group (Figure 8): thallium (mean |SHAP| 1.25), chest pain type (0.98), Health_Score (0.95) and maximum heart rate (0.71). The SHAP plot also shows the clinical direction of the effects: a normal thallium result (code 3) and a high maximum heart rate lower the predicted risk, whereas asymptomatic chest pain (code 4) and a high Health_Score raise it.

ISSN: 2181-3337 | SCIENTISTS.UZ

Figure 7

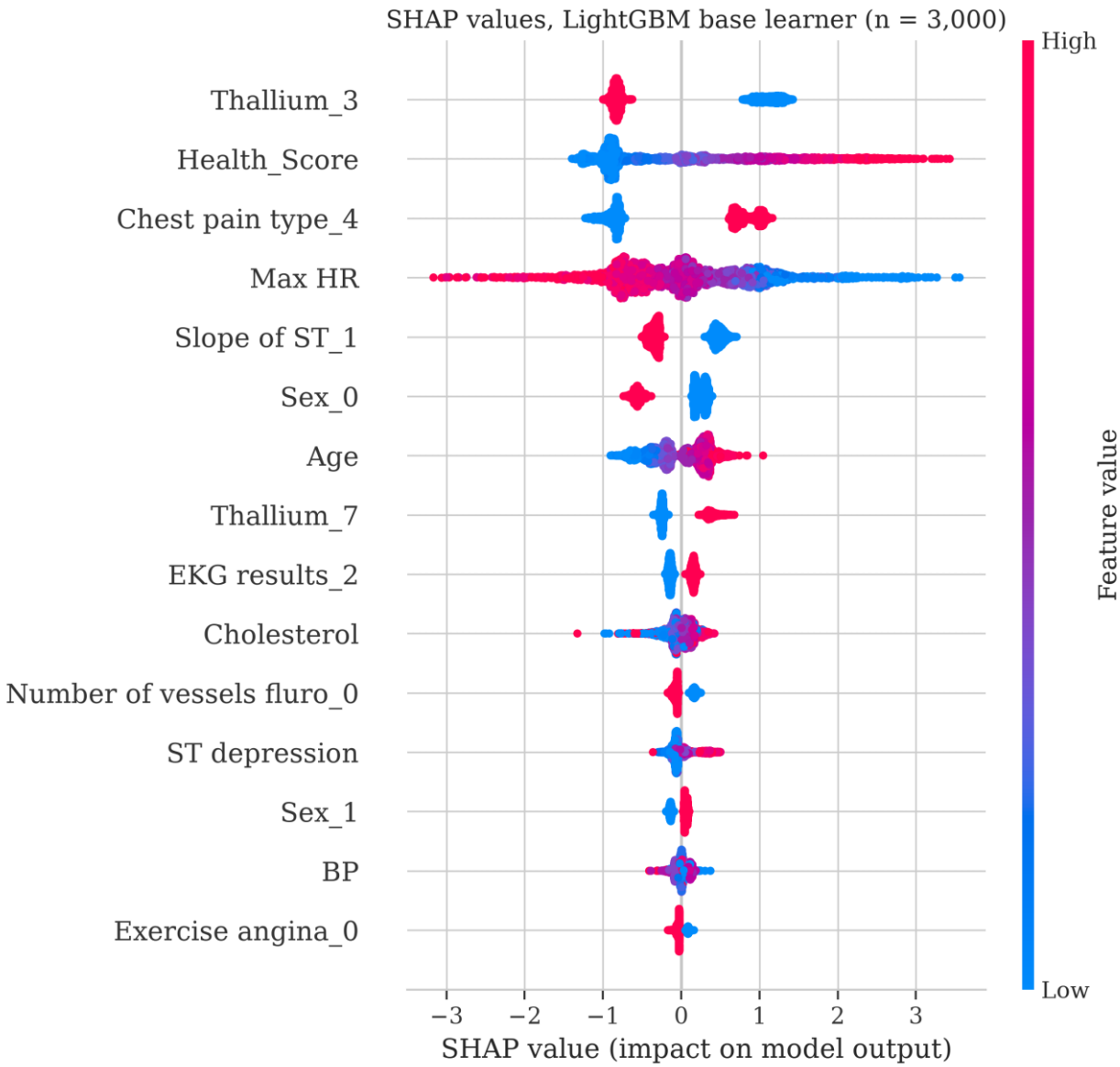
Permutation Importance of the Stack-16 Ensemble



Note. Permutation importance of the Stack-16 ensemble (mean drop in ROC-AUC over five repeats on 20,000 validation records; bars show ± SD).

Figure 8

SHAP Summary Plot for the LightGBM Base Learner



Note. SHAP summary plot for the LightGBM base learner (3,000 validation records). Each point is one patient; colour shows the feature value.

ST depression, the number of vessels and exercise angina have low importance in both analyses even though each is correlated with the outcome ($|r| = 0.42–0.46$). These three variables are the components of Health_Score. When correlated inputs carry the same information, permutation of one of them leaves the others intact, and the importance is attributed to the composite. The low ranks of these variables therefore reflect shared information rather than irrelevance. This also agrees with the ablation result: Health_Score concentrates existing signal in a single input without adding new information.

Subgroup Analysis And Discussion

Table 6 reports Stack-16 performance by sex and age band. Discrimination was nearly identical across groups: ROC-AUC was 0.947 for women and men and 0.951–0.955 across the age bands. Threshold-dependent metrics, however, differed considerably. At 0.5, sensitivity was 0.70 in women and 0.89 in men, and 0.80 in patients up to 40 years of age compared with 0.90 in those over 65.

ISSN: 2181-3337 | SCIENTISTS.UZ

Table 6

Subgroup Performance of Stack-16 at Threshold 0.5

Group	n	Prevalence	ROC-AUC	Sensitivity	Specificity	PPV	NPV
Women (Sex = 0)	179,717	0.179	0.947	0.699	0.969	0.830	0.937
Men (Sex = 1)	450,283	0.556	0.947	0.889	0.860	0.888	0.861
Age ≤ 40	21,168	0.292	0.954	0.795	0.947	0.860	0.918
Age 41–55	311,656	0.357	0.955	0.831	0.930	0.869	0.908
Age 56–65	246,900	0.553	0.951	0.894	0.866	0.892	0.868
Age > 65	50,276	0.567	0.951	0.898	0.859	0.893	0.865

Since the AUC is the same in all groups, the model ranks patients within each group equally well. The lower sensitivity in women and younger patients follows from their lower disease prevalence (17.9% in women against 55.6% in men). Sex is itself a predictor (Figure 7), so the predicted probabilities of women are shifted downward and fewer of them exceed a single global threshold. In practice, this argues for group-specific or sensitivity-targeted thresholds rather than a different model: for example, choosing the threshold so that sensitivity in women reaches the overall level. Such a choice must be made with clinicians, because it trades fewer missed cases for more false alarms.

The overall ROC-AUC of about 0.955 is in the range reported for large heart disease datasets (Bhatt et al., 2023), but direct comparison is limited because cohorts, attribute sets and validation schemes differ. The main practical finding is negative: on this dataset, stacking three strong gradient boosting models adds almost nothing to the best of them, while the simpler single model is better calibrated and cheaper to train. Logistic regression, by contrast, was clearly weaker, which confirms that non-linear interactions contribute measurable signal.

The study has the following limitations. First, the data are synthetic. A generative model reproduces the joint distribution of its training data only approximately and may smooth out rare combinations, so performance on real patients may be lower. Second, external validation was not possible because the competition test labels are not public; all results are out-of-fold estimates on one data source. Third, hyperparameters were fixed rather than tuned, which may favour some learners. Fourth, SHAP values were computed for one base learner rather than for the full stack, although the permutation analysis of the full ensemble gave the same leading features.

Conclusion And Recommendations

A stacking ensemble of XGBoost, LightGBM and CatBoost with a logistic meta-learner was evaluated for heart disease prediction on 630,000 synthetic records under a leakage-safe design. The ensemble reached a ROC-AUC of 0.9553, sensitivity of 0.867 and specificity of 0.906 at the 0.5 threshold, and sensitivity of 0.887 at the Youden threshold. Its advantage over CatBoost alone was statistically significant but negligible in size (0.00004 AUC), the three engineered features gave no gain in an ablation study, and the base learners were slightly better calibrated than the stack. Thallium, chest pain type, Health_Score and maximum heart rate drove the predictions. Discrimination was equal across sex and age groups, but sensitivity at a single threshold was markedly lower in women.

Based on these findings, the following recommendations are made:

1. Report the gain of any stacking ensemble against its individual base learners on identical folds, together with a paired test, before attributing improvements to stacking.
2. Keep every fitted transformation inside the cross-validation loop, and use fixed clinical ranges rather than data-derived minima and maxima in composite scores.
3. Report sensitivity, specificity, predictive values and calibration in addition to ROC-AUC and F1, and select the decision threshold according to the clinical cost of missed cases.

ISSN: 2181-3337 | SCIENTISTS.UZ

4. Check threshold-dependent metrics in subgroups with different prevalence, especially by sex, and consider group-specific thresholds where sensitivity differs.
5. Validate the model on anonymised patient records from clinics in Uzbekistan before any use in practice.

REFERENCES

1. Bhatt, C. M., Patel, P., Ghetia, T., & Mazzeo, P. L. (2023). Effective heart disease prediction using machine learning techniques. *Algorithms, 16*(2), Article 88. https://doi.org/10.3390/a16020088
2. Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324
3. Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review, 78*(1), 1–3. https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2
4. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939785
5. Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics, 21*, Article 6. https://doi.org/10.1186/s12864-019-6413-7
6. Collins, G. S., Moons, K. G. M., Dhiman, P., Riley, R. D., Beam, A. L., Van Calster, B., Ghassemi, M., Liu, X., Reitsma, J. B., van Smeden, M., Boulesteix, A.-L., Camaradou, J. C., Celi, L. A., Denaxas, S., Denniston, A. K., Glocker, B., Golub, R. M., Harvey, H., Heinze, G., . . . Logullo, P. (2024). TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ, 385*, Article e078378. https://doi.org/10.1136/bmj-2023-078378
7. Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J. J., Sandhu, S., Guppy, K. H., Lee, S., & Froelicher, V. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. *The American Journal of Cardiology, 64*(5), 304–310. https://doi.org/10.1016/0002-9149(89)90524-9
8. Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010
9. Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data? *Advances in Neural Information Processing Systems, 35*. https://doi.org/10.48550/arXiv.2207.08815
10. Kaggle. (2026). *Predicting heart disease: Playground series, season 6, episode 2* [Data set]. https://www.kaggle.com/competitions/playground-series-s6e2
11. Kaufman, S., Rosset, S., Perlich, C., & Stitelman, O. (2012). Leakage in data mining: Formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data, 6*(4), Article 15. https://doi.org/10.1145/2382577.2382579
12. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems, 30*, 3146–3154. https://papers.nips.cc/paper/6907-lightgbm-a-highly-efficient-gradient-boosting-decision-tree
13. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems, 30*, 4765–4774.
14. Nadeau, C., & Bengio, Y. (2003). Inference for the generalization error. *Machine Learning, 52*(3), 239–281. https://doi.org/10.1023/A:1024068626366
15. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M.,

ISSN: 2181-3337 | SCIENTISTS.UZ

Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830. https://jmlr.org/papers/v12/pedregosa11a.html

16. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. *Advances in Neural Information Processing Systems, 31*, 6638–6648. https://doi.org/10.48550/arXiv.1706.09516
17. Qiu, Z., Qiao, Y., Shi, W., & Liu, X. (2024). A robust framework for enhancing cardiovascular disease risk prediction using an optimized category boosting model. *Mathematical Biosciences and Engineering, 21*(2), 2943–2969. https://doi.org/10.3934/mbe.2024131
18. Roth, G. A., Mensah, G. A., Johnson, C. O., Addolorato, G., Ammirati, E., Baddour, L. M., Barengo, N. C., Beaton, A. Z., Benjamin, E. J., Benziger, C. P., Bonny, A., Brauer, M., Brodmann, M., Cahill, T. J., Carapetis, J., Catapano, A. L., Chugh, S. S., Cooper, L. T., Coresh, J., . . . Fuster, V. (2020). Global burden of cardiovascular diseases and risk factors, 1990–2019: Update from the GBD 2019 Study. *Journal of the American College of Cardiology, 76*(25), 2982–3021. https://doi.org/10.1016/j.jacc.2020.11.010
19. Shwartz-Ziv, R., & Armon, A. (2022). Tabular data: Deep learning is not all you need. *Information Fusion, 81*, 84–90. https://doi.org/10.1016/j.inffus.2021.11.011
20. *Statlog (Heart)* [Data set]. (n.d.). UCI Machine Learning Repository. https://doi.org/10.24432/C57303
21. Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L., & Steyerberg, E. W. (2019). Calibration: The Achilles heel of predictive analytics. *BMC Medicine, 17*, Article 230. https://doi.org/10.1186/s12916-019-1466-7
22. Wolpert, D. H. (1992). Stacked generalization. *Neural Networks, 5*(2), 241–259. https://doi.org/10.1016/S0893-6080(05)80023-1
23. Wu, Y., Xia, Z., Feng, Z., Huang, M., Liu, H., & Zhang, Y. (2024). Forecasting heart disease risk with a stacking-based ensemble machine learning method. *Electronics, 13*(20), Article 3996. https://doi.org/10.3390/electronics13203996
24. Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer, 3*(1), 32–35. https://doi.org/10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3