CAUCHY – Jurnal Matematika Murni dan Aplikasi
Volume 11 (2) (2026), Pages 2653–2677
p-ISSN: 2086-0382; e-ISSN: 2477-3344

Comparing Cost-Sensitive LightGBM and TabPFN for Imbalanced Coronary Heart Disease Risk Prediction

Arfian Jauhar Himawan[1] and A'yunin Sofro*[2]

[1] *Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Negeri Surabaya, Indonesia*
[2] *Department of Actuarial Science, Faculty of Mathematics and Natural Sciences, Universitas Negeri Surabaya, Indonesia*

Abstract

Ten-year coronary heart disease (CHD) risk prediction remains challenging due to class-imbalanced cohorts and evaluations lacking statistical uncertainty estimates. This study compares a fully tuned, cost-sensitive LightGBM pipeline with a zero-tuning tabular foundation model (TabPFN) on the imbalanced Framingham cohort of 4,240 records with a 15.19% event rate, under a leakage-free protocol that includes an eight-scenario ablation, a DeLong test, bootstrap confidence intervals, a five-seed robustness check, calibration, and decision-curve analysis. TabPFN achieved higher discrimination, with an area under the receiver operating characteristic curve of 0.713 compared with 0.677 for the tuned ensemble (DeLong $p = 0.0037$), and better calibration, with a Brier score of 0.120 compared with 0.199. It also attained the strongest imbalance-aware detection, with an F1-score of 0.380 and a Matthews correlation coefficient of 0.245, and remained superior under all five seeds, averaging 0.737 compared with 0.706. Threshold optimization raised minority-class recall from 0.031 to at least 0.47 without resampling, and ten-predictor curation did not harm the foundation model, leaving its AUC unchanged and slightly increasing its threshold-dependent F1-score and MCC. Normalized SHAP-based relative feature-importance weights (0–100) identified age, smoking, and blood pressure as dominant drivers, consistent with established CHD risk factors. All reported performance reflects internal validation on the Framingham cohort; the paired-bootstrap 95% CI of the Brier-score difference ([0.061, 0.098]) confirms the calibration advantage over the tuned cost-sensitive ensemble, constituting preliminary evidence toward SDG 3 (Target 3.4) screening applications that warrants external validation before any clinical use.

Keywords: catastrophic disease; class imbalance; coronary heart disease; sustainable development goals; TabPFN

Copyright © 2026 by Authors, Published by CAUCHY Group. This is an open access article under the CC BY-SA License (https://creativecommons.org/licenses/by-sa/4.0)

1. Introduction

Cardiovascular diseases (CVDs) continue to top the global mortality ranking: the World Health Organization estimates that 19.8 million lives were lost to CVDs in 2022 alone, roughly one in three deaths worldwide [1]. Beyond its clinical dimension, this burden behaves as a non-natural disaster in the sense of disaster-risk research [2], and its reduction is explicitly targeted

*Corresponding author. E-mail: ayuninsofro@unesa.ac.id

Received: September 1, 2026 Revised: October 28, 2026 Accepted: November 6, 2026
DOI: https://doi.org/10.18860/cauchy.v11i2.45897

by Sustainable Development Goal (SDG) 3, Target 3.4, which commits to cutting premature mortality from non-communicable diseases by one third by 2030 [3]. Meeting this target requires screening instruments that run on routinely collected data in resource-limited settings and whose outputs are statistically validated, calibrated, and interpretable.

Formally, let $x \in \mathbb{R}^d$ denote the predictor vector of a subject and $y \in \{0, 1\}$ the ten-year CHD indicator; the screening task is to estimate $p(x) = P(y = 1 \mid x)$ and to act through a decision rule $\hat{y}(x) = \mathbf{1}\{p(x) \geq \tau\}$ with cutoff $\tau \in (0, 1)$. Gradient-boosted tree ensembles such as XGBoost [4] and LightGBM [5] estimate $p(x)$ with an additive function $f(x) = \sum_{t=1}^{T} f_t(x)$ fitted by regularized empirical-risk minimization [6], generalizing the linear logit specification of classical cohort models [7]. When the positive prior $\pi_+ = P(y = 1)$ is small, however, the empirical risk is dominated by the majority class, so the ranking f may remain informative while the default operating point $\tau = 1/2$ is clinically useless [8]. Two standard remedies act on this decoupling: cost-sensitive class weighting, which penalizes false negatives w_+ times more heavily inside the training objective—on a fixed ranking its weighted decision rule would correspond to a lowered cutoff $\tau_c = w_0/(w_0 + w_+) < 1/2$, but because the weights enter the loss the trees are refitted and the ranking itself can change [8, 9], and direct threshold optimization, $\tau^* = \arg\max_\tau F_1(\tau)$, which moves only the operating point on a fixed ranking [10]. A complementary route resamples the training measure: SMOTE interpolates between minority neighbors inside their convex hull [11]. On the same cohort studied here, Amiroch et al. coupled SMOTE with XGBoost and reported an AUC of 0.6502 with a 95% bootstrap interval of [0.6017, 0.6981] under a single pre-specified 80:20 split, leaving the cutoff at its default and the comparison without paired inference [12].

The most recent development recasts tabular prediction as amortized Bayesian inference. The Tabular Prior-data Fitted Network (TabPFN) is a transformer trained once to minimize the cross-entropy over more than 10^8 synthetic datasets drawn from a prior defined by structural causal models [13]; by construction, the minimizer approximates the posterior predictive distribution $p(y_{\text{test}} \mid X_{\text{test}}, D_{\text{train}})$ under that prior [14, 15], so at deployment the model maps an entire dataset to predictions in a single forward pass, with neither parameter updates nor hyperparameter search. We use the term *zero-tuning* in this strict sense: the foundation model itself is never fitted or tuned on the study data, although feature curation and threshold optimization remain part of the surrounding evaluation protocol. Hollmann et al. showed that this predictor dominates tree ensembles tuned for up to four hours on datasets of at most 10,000 samples [13]; whether the same dominance survives a small, imbalanced clinical cohort, and whether the posterior approximation benefits from the feature curation that tree ensembles require, remain open empirical questions.

Three mathematical objects—the weighted risk functional, the cutoff as an optimization variable, and the amortized posterior predictive distribution—have thus been studied in isolation, yet none has been brought together on CHD data in a head-to-head comparison in which paired significance, repeated-partition robustness, calibration, and decision-analytic utility are quantified rather than assumed [10, 12, 13]. The Framingham Heart Study, the cohort behind the classical Framingham risk score [7], serves as a suitable test-bed rather than a source of novelty: its public extract exhibits exactly the properties that stress-test the competing paradigms—a strongly imbalanced target, missing entries, and heavy upper-tail dispersion [12]—and mirrors the moderate epidemiological imbalance typical of real screening programs rather than the extreme ratios of fraud-style benchmarks [10].

This study therefore compares cost-sensitive LightGBM and TabPFN under a common leakage-free protocol and states its claims explicitly. The objective is to determine whether a zero-tuning foundation model can match or exceed a fully tuned, imbalance-aware tree ensemble on the imbalanced Framingham cohort, both statistically and decision-analytically. The novelty comprises (i) an ablation hierarchy of eight scenarios that separates threshold optimization, cost-sensitive weighting, hyperparameter tuning, and feature curation, including a supplementary all-feature ablation of TabPFN; (ii) an inference framework that treats the paired DeLong test

as the primary comparison, supported by bootstrap confidence intervals, a five-seed robustness check, calibration analysis, and decision-curve analysis [16–19]; and (iii) normalized SHAP-based relative feature-importance weights (0–100) that summarize the global attribution structure of the final tree-based pipeline—presented as model-explanatory importance, not as a validated patient-level clinical score. The scientific contribution is a quantified, reproducible answer to whether foundation models surpass tuned ensembles on imbalanced clinical tabular data—a question aligned with SDG 3 (Target 3.4)—extending prior CHD prediction work [12] from point estimates to statistical and clinical-utility inference. The remainder of the paper details the methodology, reports and discusses the results, and closes with concluding remarks.

2. Methods

This section outlines the methodological framework employed to compare two modeling paradigms, namely cost-sensitive LightGBM and the tabular foundation model TabPFN, for predicting ten-year coronary heart disease (CHD) risk from an epidemiological cohort. The research procedure encompasses seven sequential stages: data acquisition and preparation, leakage-free preprocessing, model construction, threshold optimization, ablation analysis, performance evaluation with statistical validation, and model explainability. Each stage is designed to ensure a fair and reproducible comparison between the two paradigms while maintaining consistency in preprocessing and evaluation.

2.1. Research Procedure and Dataset

The empirical material is a de-identified public extract of the Framingham Heart Study, a longitudinal cardiovascular cohort that has served for decades as a reference population for risk-factor research.[1] The released file comprises 4,240 records described by fifteen candidate predictors and a single binary outcome; all fifteen predictors were retained for modeling, with operational definitions listed in Table 1.

The outcome of interest, `TenYearCHD`, records whether a coronary event occurred during the ten-year observation window. For subject i, it is formalized as the binary indicator in (1),

$$y_i = \begin{cases} 1, & \text{if a CHD event occurs within 10 years of baseline,} \\ 0, & \text{otherwise,} \end{cases} \tag{1}$$

so that $y_i = 1$ marks a positive case and $y_i = 0$ an event-free subject. The retained covariates fall naturally into two families—nine continuous clinical or sociodemographic measurements and six binary indicators—as listed in Table 1.

The data were partitioned using stratified sampling with a fixed random seed of 42 into training (70%), validation (15%), and test (15%) subsets, corresponding to 2,968, 636, and 636 records, respectively, such that the class proportion is preserved in each subset. The validation subset was used exclusively for decision-threshold optimization, whereas the test subset was held out and consulted only once for the final evaluation. To assess the sensitivity of the conclusions to this single partition, the split–fit–evaluate procedure was additionally repeated under four further random seeds; in these repetitions the top-10 feature set and the LightGBM hyperparameters were held fixed at the values selected under the primary seed, so that the exercise isolates the effect of the data partition rather than repeating the full model-development procedure. Accordingly, this multi-seed exercise should be read as a sensitivity analysis to data partitioning rather than as nested or fully repeated validation; a fully nested redesign of feature selection and hyperparameter search within each split would provide stronger evidence and is acknowledged as a limitation of the present study. All experiments were executed under fixed seeds and deterministic training settings, so that re-running the released code reproduces the

[1]The extract is publicly available in the Kaggle repository: https://www.kaggle.com/datasets/noeyislearning/framingham-heart-study.

reported numbers exactly; this guarantees computational reproducibility but does not, by itself, establish external generalizability.

Table 1: Data description used in the analysis ($n = 4{,}240$).

No.	Attribute	Description	Explanation
1	male	Biological sex	0: female, 1: male
2	age	Age at the baseline examination	Years
3	education	Self-reported education level	Score 1–4
4	currentSmoker	Active tobacco use at baseline	0: no, 1: yes
5	cigsPerDay	Average cigarettes smoked per day	Cig./day
6	BPMeds	Ongoing antihypertensive therapy	0: no, 1: yes
7	prevalentStroke	History of cerebrovascular accident	0: no, 1: yes
8	prevalentHyp	Established hypertension	0: no, 1: yes
9	diabetes	Prior diabetes diagnosis	0: no, 1: yes
10	totChol	Serum total cholesterol	mg/dL
11	sysBP	Systolic blood pressure	mmHg
12	diaBP	Diastolic blood pressure	mmHg
13	BMI	Body mass index	kg/m^2
14	heartRate	Resting pulse frequency	Beats/min
15	glucose	Blood glucose level	mg/dL
16	TenYearCHD	Coronary event within the 10-year window	0: no event, 1: event

2.2. Leakage-Free Data Preprocessing

Before model construction, all research variables were prepared through a preprocessing pipeline whose governing principle is the prevention of information leakage between data partitions. Concretely, every statistic used to transform the data—imputation medians, winsorization cutoffs, and standardization parameters—was estimated exclusively on the training subset and subsequently applied, without re-fitting, to the validation and test subsets. This train-only fitting protocol ensures that the empirical distribution of the evaluation sets remains untouched and that no evaluation information can influence the training process, in line with the leakage-aware practice advocated in recent CHD-prediction studies [12].

Missing entries were first imputed using the feature-wise median computed on the training subset, a robust choice that is insensitive to the extreme values present in several clinical variables. Continuous features were then winsorized using quartile-based cutoffs derived from the training subset, formulated in Eq. (2),

$$\tilde{x}_j = \min\big(\max(x_j,\ Q1_j - 1.5\,\mathrm{IQR}_j),\ Q3_j + 1.5\,\mathrm{IQR}_j\big), \tag{2}$$

where x_j denotes the value of feature j, $Q1_j$ and $Q3_j$ are the first and third quartiles computed on the training subset, and $\mathrm{IQR}_j = Q3_j - Q1_j$ is the interquartile range. Binary covariates were deliberately excluded from this transformation and left unmodified, so that their original 0/1 signal is preserved.

Following imputation and winsorization, all fifteen predictors were standardized using a z-score transform, with the mean and standard deviation estimated on the training subset only, consistent with the leakage-free protocol above; the resulting standardized matrix served as the input for both modeling paradigms. Standardization is not strictly required for the tree-based LightGBM model, whose splits are invariant to monotonic rescaling under exact split search, and TabPFN performs its own internal normalization of the input space [13]; external standardization was nevertheless retained so that both models share a single, uniform preprocessing pipeline, consistent with common practice in the CHD-prediction literature [12]. A direct consequence of this choice, addressed explicitly in the explainability subsection of this section, is that the SHAP-based explanations reported in the results are computed and displayed in standardized units rather than the original clinical units.

2.3. Cost-Sensitive LightGBM

The primary model was constructed using LightGBM, a gradient boosting decision tree framework that achieves high efficiency through histogram-based learning and leaf-wise tree growth [5]. Following the additive formulation of gradient boosting machines [6], the ensemble is written as in Eq. (3):

$$f(x) = \sum_{t=1}^{T} f_t(x), \tag{3}$$

where each f_t denotes a regression tree and T the number of boosting iterations; for a convex loss ℓ, each tree is fitted sequentially to the first- and second-order statistics of ℓ evaluated at the current ensemble, so that the fitting procedure is a second-order analog of the regularized empirical risk introduced in the Introduction.

The empirical class distribution is strongly skewed toward event-free subjects, as quantified in the exploratory analysis. Rather than balancing the data through synthetic oversampling, we handled this imbalance through cost-sensitive learning [8, 9], in which misclassifications of the minority class are assigned a larger penalty during training. Specifically, the positive class was assigned a weight equal to the ratio of negative to positive samples in the training subset, as defined in Eq. (4),

$$w_+ = \frac{N_{\text{neg}}}{N_{\text{pos}}}, \tag{4}$$

and the weight was incorporated into the regularized boosting objective in Eq. (5),

$$L(\theta) = -\frac{1}{N} \sum_{i=1}^{N} \left[w_+ \, y_i \ln(p_i) + w_- \, (1 - y_i) \ln(1 - p_i) \right] + \Omega(\theta), \tag{5}$$

where $p_i = \sigma(f(x_i))$ is the predicted probability for sample i, $y_i \in \{0, 1\}$ is the observed label, $w_- = 1$, and $\Omega(\theta)$ denotes the regularization term on the tree ensemble [4, 6]. This formulation ensures that false-negative predictions, which are clinically far more dangerous in a screening context, are penalized approximately w_+ times more heavily than false-positive predictions, shifting the fitted leaves toward higher positive-class probabilities without any modification of the data.

2.4. Tabular Foundation Model (TabPFN)

The second paradigm is the Tabular Prior-data Fitted Network (TabPFN), the zero-tuning tabular foundation model compared in this study. This transformer performs prediction through in-context learning, so it operates without conventional training or hyperparameter tuning on the target dataset [13]. Following Hollmann et al. [13], we use the term *zero-tuning* in this strict, model-internal sense, where no gradient update of the network parameters and no hyperparameter search are performed on the study data at any stage. The term does not imply that the surrounding pipeline is tuning-free—the leakage-free preprocessing, the split-based feature curation, and the validation-based threshold optimization are applied identically to both paradigms.

Bayesian prediction problem. Let $D_{\text{train}} = (X_{\text{train}}, y_{\text{train}})$ denote the training subset and suppose the data are generated by a probabilistic model with parameter θ. Given a prior distribution $p(\theta)$, Bayes' rule yields the posterior $p(\theta \mid D_{\text{train}})$, and the prediction for a query point x_{test} is the posterior predictive distribution

$$p(y_{\text{test}} \mid x_{\text{test}}, D_{\text{train}}) = \int p(y_{\text{test}} \mid x_{\text{test}}, \theta) \, p(\theta \mid D_{\text{train}}) \, \mathrm{d}\theta, \tag{6}$$

which averages the predictions of every model θ consistent with the training data, weighted by its posterior plausibility [13, 14]. Eq. (6) is the standard Bayesian answer to tabular prediction,

but the integral is intractable except in conjugate models, and numerical approximations such as Markov chain Monte Carlo or deep ensembles are computationally expensive when repeated across many datasets. This intractability motivates replacing the per-dataset computation with a learned, amortized mapping.

Amortization via prior-data fitted networks. TabPFN approximates the map $D_{\text{train}} \mapsto p(\cdot \mid x_{\text{test}}, D_{\text{train}})$ with a transformer q_φ trained once over synthetic datasets. Following [14, 15], a prior over datasets $p(D)$ is defined by a generative process based on structural causal models, and a synthetic dataset $D = D_{\text{train}} \cup D_{\text{test}}$ is drawn from $p(D)$ with the targets of D_{test} masked. The network parameter is then obtained as

$$\varphi^\star = \arg\min_{\varphi} \; \mathbb{E}_{D \sim p(D)}\big[-\log q_\varphi(y_{\text{test}} \mid X_{\text{test}}, D_{\text{train}})\big], \tag{7}$$

where the expectation runs over a synthetic corpus of roughly 10^8 datasets [13]. Because the negative log-likelihood is a strictly proper scoring rule, the minimizer of Eq. (7) satisfies, for every dataset in the support of the prior,

$$q_{\varphi^\star}(y_{\text{test}} \mid X_{\text{test}}, D_{\text{train}}) \approx p(y_{\text{test}} \mid X_{\text{test}}, D_{\text{train}}), \tag{8}$$

where the right-hand side is the posterior predictive distribution under the prior $p(D)$ [14]. Eq. (8) is the theoretical link that ties TabPFN to Eq. (6): the posterior predictive is not imposed on the model but *obtained* as the optimal solution of the meta-training objective, so the trained network amortizes Bayesian inference in a single forward pass.

Architecture and deployment. The transformer adapts the standard encoder to tabular data through two attention sublayers—attention over the features within each sample and attention over the samples within each feature—which renders the mapping invariant to the order of samples and features [13]. At deployment—the only stage performed in this study—the pretrained, frozen network is queried directly for every test-set subject i,

$$\hat{p}_i = q_{\varphi^\star}(y_i = 1 \mid x_i, D_{\text{train}}), \tag{9}$$

obtained in a single forward pass with no further optimization, hyperparameter search, or gradient update. In practice, $q_{\varphi^\star}$ was implemented through the `priorlabs tabpfn` package (v3 classifier checkpoint) with its default configuration and a fixed random seed, executed on the available Google Colab compute accelerator: a GPU when one was allocated to the runtime, and CPU otherwise. Because TabPFN internally performs normalization and handles missing values, applying external standardization beforehand does not interfere with its own normalization step.

2.5. Hyperparameter Tuning

Building on the cost-sensitive objective introduced above, the boosting hyperparameters of LightGBM were tuned in two independent rounds, each using a randomized search with forty candidates and five-fold cross-validation, following the screening protocol of He et al. [20]. In the first round (*Round-1*), the search was run on the training subset using the full fifteen-feature standardized space. The resulting model's feature importances were then used to rank and retain the ten most informative attributes (age, diastolic and systolic blood pressure, cigarettes per day, total cholesterol, glucose, body mass index, sex, heart rate, and diabetes status). In the second round (*Round-2*), an independent randomized search using the same protocol and hyperparameter grid was run restricted to this ten-feature subset, yielding a second, potentially different, set of hyperparameters.

Both hyperparameter configurations were retained rather than discarding the first: as detailed in the ablation design, the top-10 feature set is evaluated under the Round-1 hyperparameters in

Scenario 5 and under the Round-2 hyperparameters in Scenario 6, which directly isolates whether re-tuning specifically for the reduced feature space yields any improvement beyond the effect of feature curation alone. Consistent with the leakage-free protocol, the importance ranking and both rounds of hyperparameter search were computed exclusively from the training subset, so that the validation and test subsets played no role whatsoever in model selection. This two-round design also foregrounds a deliberate methodological contrast with the second paradigm evaluated in this study: whereas LightGBM requires this extensive, two-stage tuning effort, the tabular foundation model introduced next requires none.

Each round therefore selects the hyperparameter configuration

$$\hat{h} = \arg\max_{h \in \Lambda} \mathrm{AUC}_{\mathrm{CV}}(h), \tag{10}$$

where Λ denotes the forty-candidate randomized grid and $\mathrm{AUC}_{\mathrm{CV}}$ the five-fold cross-validated AUC computed on the training subset only.

2.6. Threshold Optimization

Because the default cutoff of 0.5 is generally inappropriate for imbalanced screening tasks, the probabilistic outputs of every model variant were converted into binary decisions using a decision threshold optimized on the validation subset. The optimal threshold τ^* was defined in Eq. (11),

$$\tau^* = \arg\max_{\tau} \frac{2\,P(\tau)\,R(\tau)}{P(\tau) + R(\tau)}, \tag{11}$$

where $P(\tau)$ and $R(\tau)$ denote the precision and recall obtained at threshold τ. This criterion balances precision and recall and prevents the model from operating at an overly conservative point that would miss true CHD cases. The identical optimization protocol was applied to every model variant to preserve the fairness of the comparison, and the resulting thresholds were then fixed and applied only to the held-out test subset.

2.7. Ablation Design

To investigate the influence of each methodological component within the proposed framework, this study incorporates an ablation-based experimental design consisting of eight classification scenarios evaluated on the same held-out test subset in Table 2. The ladder opens with a deliberately plain reference configuration (S1), and the ablation is interpreted through designated pairwise contrasts, each of which isolates exactly one component: (S1, S2) isolates threshold optimization under the unweighted objective; (S1, S3) isolates cost-sensitive weighting at the default threshold; (S3, S4) isolates threshold optimization on top of the weighted objective; (S4, S5) isolates feature curation under the weighted, threshold-optimized regime; (S5, S6) isolates hyperparameter re-tuning on the curated feature space; and (S7, S8) isolates feature curation for the foundation model. Transitions that alter more than one component at once (for example S2 to S3, which swaps threshold optimization for class weighting) are deliberately not interpreted as single-component effects, while the two finished pipelines, S6 and S8, constitute the primary head-to-head comparison of the study. Throughout the remainder of the manuscript, these two pipelines are referred to as the final LightGBM model (S6) and the final TabPFN model (S8), respectively.

Two protocol choices govern how these scenarios may legitimately be compared. First, every decision threshold was optimized on the validation subset alone and frozen before the test subset was evaluated, so no scenario can leak test-set information into its operating point. Second, the two paradigms were treated according to their nature: LightGBM received the positive-class weight of Eq. (5) together with the two-round hyperparameter tuning, whereas TabPFN was used exactly as a foundation model is meant to be used—without tuning, re-weighting, or resampling—so that any difference between S6 and S8 reflects the modeling paradigm itself rather than an imbalance of experimental effort.

Table 2: Eight-scenario ablation design. Opt. = decision threshold optimized on the validation subset via Eq. (11); CS = cost-sensitive weighting via Eq. (5). All LightGBM scenarios retain the Round-1 hyperparameters except S6, which uses the Round-2 configuration re-tuned on the reduced feature set. The primary head-to-head comparison is between the final LightGBM model and the final TabPFN model.

ID	Model	Features	CS	Threshold	Component introduced
S1	LightGBM	15 Features	No	0.5	Reference configuration
S2	LightGBM	15 Features	No	Opt.	Threshold optimization
S3	LightGBM	15 Features	Yes	0.5	Cost-sensitive weighting
S4	LightGBM	15 Features	Yes	Opt.	Threshold optimization under CS
S5	LightGBM	Top-10 Features	Yes	Opt.	Feature curation
S6	LightGBM	Top-10 Features	Yes	Opt.	Re-tuning on the reduced set (final LightGBM)
S7	TabPFN	15 Features	-	Opt.	Paradigm swap (foundation model)
S8	TabPFN	Top-10 Features	-	Opt.	Feature curation under TabPFN

2.8. Performance Evaluation and Statistical Validation

The models were evaluated on the test subset using discrimination, detection, calibration, and clinical-utility measures. Discrimination was summarized by the area under the ROC curve (AUC) and the area under the precision–recall curve (PR-AUC), whereas detection performance was summarized by precision, recall, the F1-score, and accuracy, defined in Eq. (12), Eq. (13), and Eq. (14),

$$\text{Precision} = \frac{TP}{TP + FP}, \qquad \text{Recall} = \frac{TP}{TP + FN}, \tag{12}$$

$$F_1 = \frac{2TP}{2TP + FP + FN}, \tag{13}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \tag{14}$$

where TP, TN, FP, and FN denote the four entries of the confusion matrix. In the imbalanced screening context, recall receives particular emphasis because false negatives correspond to missed CHD cases, whereas the F1-score aggregates precision and recall into a single harmonic measure. The Matthews correlation coefficient (MCC) was additionally computed as in Eq. (15),

$$\text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}, \tag{15}$$

as a balanced single-value summary that remains informative under skewed class priors [21].

Calibration was assessed using the Brier score in Eq. (16),

$$BS = \frac{1}{n} \sum_{i=1}^{n} (p_i - y_i)^2, \tag{16}$$

which measures the mean squared deviation between predicted probabilities and observed outcomes [18], together with a graphical reliability diagram. Because the cohort prevalence is approximately 15%, an equal-width binning of the predicted-probability axis leaves very few test-set observations in the high-probability bins, producing unstable, visually misleading calibration curves. The reliability diagrams were therefore constructed using *quantile-based* binning into five bins of approximately equal sample size, rather than a larger number of equal-width bins, to obtain a stable estimate of the observed event rate in every bin.

Calibration was further summarized by logistic calibration, in which the binary outcome is regressed on the logit of the predicted probabilities; the ideal values are a calibration intercept of 0 and a calibration slope of 1, with a negative intercept indicating systematic over-prediction of risk and a slope below one indicating over-confident probabilities.

Clinical utility was assessed using decision curve analysis through the net benefit in Eq. (17),

$$NB(\tau) = \frac{TP(\tau)}{n} - \frac{FP(\tau)}{n} \cdot \frac{\tau}{1-\tau}, \tag{17}$$

evaluated over threshold probabilities between 1% and 40% [19].

To quantify the uncertainty of the point estimates, 95% confidence intervals were constructed using the nonparametric bootstrap percentile method with 1,000 resamples of the test subset [17], as shown in Eq. (18),

$$CI_{95\%} = [\hat{\theta}_{(0.025)}, \hat{\theta}_{(0.975)}]. \tag{18}$$

The primary test of the AUC difference between the final LightGBM model (Scenario 6) and the final TabPFN model (Scenario 8) was the DeLong test [16], which compares two correlated ROC curves estimated on the same test subset by deriving the covariance of the paired AUC estimates analytically from the underlying rank statistics, rather than through resampling. Writing $\widehat{\mathrm{AUC}}_A$ and $\widehat{\mathrm{AUC}}_B$ for the two paired AUC estimates and $\widehat{\mathrm{Var}}(\widehat{\mathrm{AUC}}_A - \widehat{\mathrm{AUC}}_B)$ for their DeLong-estimated covariance-based variance, the test statistic is given in Eq. (19),

$$z = \frac{\widehat{\mathrm{AUC}}_A - \widehat{\mathrm{AUC}}_B}{\sqrt{\widehat{\mathrm{Var}}(\widehat{\mathrm{AUC}}_A - \widehat{\mathrm{AUC}}_B)}}, \tag{19}$$

which is asymptotically standard normal under the null hypothesis of equal AUC. The DeLong test was adopted as the primary inferential test because it operates directly on the single held-out test subset and correctly accounts for the correlation between the two models' predictions (both models are scored on the same patients), rather than relying on resampling.

To quantify the uncertainty of the between-model comparison without relying on an independence assumption, a paired bootstrap confidence interval for the AUC difference was constructed by scoring both final pipelines on each of the 1,000 bootstrap resamples of the test subset and recording the paired difference; the 95% percentile interval of these differences is reported alongside the DeLong test. The same paired resamples were used to obtain 95% percentile intervals for the Brier-score difference and for the calibration intercept and slope of each model.

Finally, to assess whether the head-to-head comparison is stable across data partitions rather than an artifact of the single primary split, the entire split–preprocess–fit–evaluate procedure described above was repeated under four additional random seeds, holding the Round-2 (top-10) LightGBM hyperparameters and the top-10 feature set fixed at the values selected under the primary seed, so that only the effect of the data partition is varied. For each of the five seeds, the final LightGBM model (Scenario 6) and the final TabPFN model (Scenario 8) were refit on the corresponding training subset and evaluated on the corresponding held-out test subset; the mean and standard deviation of the AUC and PR-AUC across the five replicates are reported alongside the primary single-split results.

2.9. Model Explainability and SHAP-Based Feature Importance

The internal decision mechanism of the final LightGBM model (Scenario 6) was explained using SHapley Additive exPlanations (SHAP), which assigns each feature a contribution value derived from cooperative game theory [22]. For a feature j, the Shapley value is computed as in Eq. (20),

$$\varphi_j = \sum_{S \subseteq N \setminus \{j\}} \frac{|S|!\,(n - |S| - 1)!}{n!} \Big[f(S \cup \{j\}) - f(S) \Big], \tag{20}$$

where N is the set of all features, S is a subset of features, $n = |N|$, and $f(S)$ is the model prediction using only the features in S [23]. SHAP values were computed with a TreeExplainer applied directly to the standardized feature matrix that the final model was trained on, so that the explainer operates on exactly the same input representation as the model itself. The TreeExplainer was configured with a probability-scale output, so every Shapley value, summary plot, and waterfall bar is expressed in probability units. A direct consequence is that the beeswarm and waterfall plots reported in the results display feature values in standardized units rather than original clinical units; this trade-off was accepted in favor of a single, uniformly preprocessed pipeline shared between LightGBM's training procedure and its explanation, and the resulting plots should be read with this caveat in mind. Importantly, the normalized relative feature-importance weights derived below are unaffected by this choice: they are computed from the absolute magnitude of each feature's Shapley value alone and do not depend on the units in which that feature is displayed.

The mean absolute Shapley values over the test subset were then normalized into 0–100 relative feature-importance weights as in Eq. (21),

$$W_j = \frac{|\varphi_j|}{\sum_k |\varphi_k|} \times 100, \tag{21}$$

These weights summarize the global attribution structure of the final tree-based pipeline in a single transparent profile; they are intended as model-explanatory importance, not as a validated patient-level clinical score, and no patient-level scoring rule is derived or validated from them.

All analyses were implemented in Python 3.13.15 on a Google Colab runtime using scikit-learn 1.6.1, LightGBM 4.6.0, SHAP 0.52.0, and the priorlabs tabpfn package 9.0.0 with the TabPFN-3 classifier checkpoint explicitly pinned via model_path='tabpfn-v3-classifier-v3_default.ckpt'. TabPFN inference used an NVIDIA Tesla T4 GPU when allocated and CPU otherwise; the actual device was logged for each run. All experiments used a fixed random seed of 42 with deterministic LightGBM training settings, so LightGBM results are exactly reproducible; TabPFN outputs may show minor hardware-dependent floating-point variation, disclosed here for transparency.

3. Results and Discussion

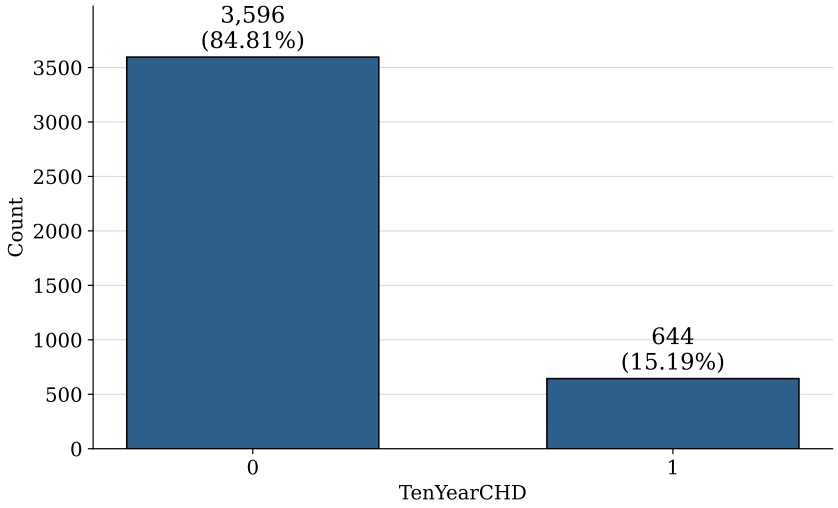
The results are organized into eight components: an exploratory characterization of the cohort; the convergence of the two-round hyperparameter search and the feature curation it produces; an eight-scenario ablation under class imbalance; the head-to-head comparison between the final LightGBM model and the final TabPFN model, supported by paired significance and multi-seed robustness evidence; calibration and decision-analytic utility; SHAP-based explainability together with the derived 0–100 relative feature-importance weights; a synthesis with prior literature; and the limitations of the study. Each component is presented as statistical and clinical-utility inference rather than routine algorithm benchmarking, aligning with the study's comparative objective. Unless stated otherwise, all test-set quantities refer to the held-out subset consulted once under the leakage-free protocol of the Methods, and the two finished pipelines are reported under the scenario labels of Table 2.

3.1. Exploratory Characteristics of the Dataset

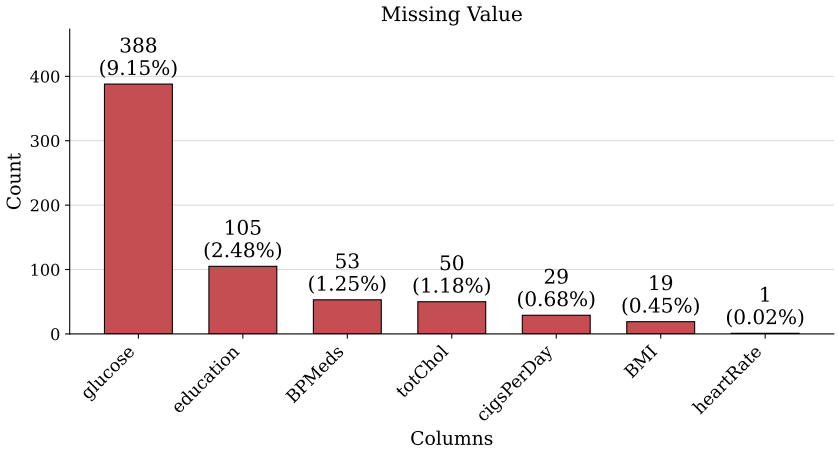
The initial stage of the analysis examines the empirical structure of the cohort, covering class distribution, missingness, and outlier behavior. These diagnostics establish the methodological motivation for the imbalance-handling and robust-preprocessing steps adopted in this study. Together, they confirm that the cohort is suitable for modeling but requires careful treatment of class imbalance, missing entries, and extreme physiological readings before any classifier is fitted.

Fig. 1(a) shows that the target variable `TenYearCHD` is strongly imbalanced: the majority class (no CHD event) accounts for 84.81% of the 4,240 observations, while the minority class

(CHD event within ten years) accounts for only 15.19% ($n = 644$), yielding a negative-to-positive ratio of approximately 5.6 : 1. This skew is methodologically important because empirical learners optimized for accuracy tend to collapse toward the majority class, thereby missing the high-risk patients that a screening instrument is supposed to catch. It directly motivates the cost-sensitive weighting and threshold-calibration strategies evaluated later in this study.



(a) Class distribution of `TenYearCHD`.



(b) Missing-value count per column.

Fig. 1: Exploratory characteristics of the cohort: (a) target imbalance and (b) missingness pattern.

Fig. 1(b) summarizes the missing-value pattern before imputation. The largest gaps occur in glucose (9.15%), education (2.48%), BP medication (1.25%), and total cholesterol (1.18%), with the remaining features complete or nearly so. From a statistical perspective, this profile argues against case deletion, which would reduce the effective sample size and potentially distort the empirical distribution, and in favor of the median-imputation strategy adopted in the preprocessing pipeline.

Fig. 2 displays the marginal distributions of the nine numerical predictors. Several variables, notably glucose, systolic blood pressure, diastolic blood pressure, total cholesterol, and BMI, exhibit pronounced upper-tail dispersion consistent with the clinical reality that extreme physiological readings, while rare, do occur in the cohort. These patterns justify the IQR-based winsorization step applied before model fitting, which caps extreme values at training-derived bounds so that a handful of outliers cannot dominate the learned decision boundaries.

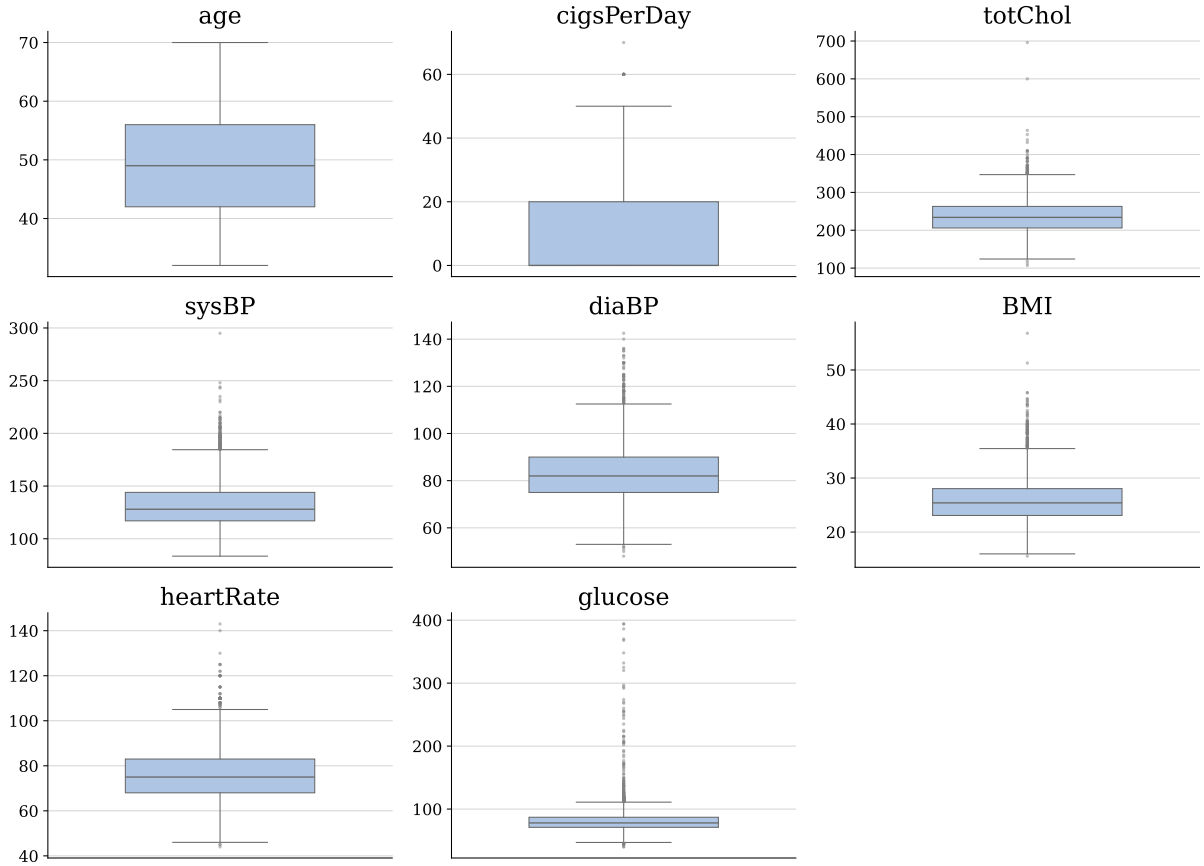


Fig. 2: Boxplots of eight of the nine numerical predictors (education omitted). Glucose, systolic and diastolic blood pressure, total cholesterol, and BMI exhibit pronounced upper-tail dispersion, motivating IQR-based winsorization.

3.2. Hyperparameter Tuning Convergence and Feature Curation

The Round-1 randomized search over the full fifteen-feature space solved Eq. (10), drawing each of its forty candidates uniformly at random from the hyperparameter space summarized in Table 3, and identified the configuration reported in Table 4, attaining a five-fold cross-validated AUC of 0.7086 on the training subset. The associated split-based importances in Fig. 3 concentrate on a compact set of clinically familiar variables: age and the two arterial pressure readings dominate, followed by tobacco exposure (cigarettes per day), total cholesterol, glucose, body mass index, sex, heart rate, and diabetes status. Retaining these ten predictors discards only five weak attributes—education, current-smoker status, blood-pressure medication, prevalent stroke, and prevalent hypertension, the last of whose signal is largely carried by the pressure readings themselves.

The Round-2 search, run independently on the reduced ten-feature space, returned a configuration identical to Round-1 in every entry of Table 4, with a cross-validated AUC of 0.700. Two readings of this convergence are worth stating. First, the exact agreement of the two forty-candidate searches over different feature spaces is reassuring evidence that, within this cohort and this candidate grid, the selected configuration is not an artifact of a particular grid draw; however, because both searches share the same training subset, this agreement does not by itself establish that the optimization landscape is stable in general. Second, re-tuning adds no improvement beyond curation itself; this is precisely the contrast that Scenarios 5 and 6 exploit to isolate the marginal effect of hyperparameter re-optimization, and the near-identical cross-validated scores of the two rounds indicate that the discarded five attributes contribute little reusable signal once the leading ten are in the model.

Table 3: LightGBM hyperparameter search space used in both randomized-search rounds. Each round drew forty candidates uniformly at random from these candidate sets and scored them with five-fold cross-validated ROC-AUC on the training subset (Eq. (10)); the winning configuration of each round is reported in Table 4.

Hyperparameter	Candidate values
n_estimators	{100, 200, 300, 500}
learning_rate	{0.01, 0.03, 0.05, 0.1}
max_depth	{−1, 4, 6, 8, 10}
num_leaves	{15, 31, 63, 127}
min_child_samples	{10, 20, 30, 50}
subsample	{0.7, 0.8, 0.9, 1.0}
colsample_bytree	{0.7, 0.8, 0.9, 1.0}
reg_lambda	{0, 0.1, 0.5, 1.0}
reg_alpha	{0, 0.1, 0.5, 1.0}

Table 4: LightGBM hyperparameter configuration selected by the randomized search. The Round-1 (fifteen-feature) and Round-2 (top-10) searches returned this configuration in every entry, so a single table serves both rounds.

Hyperparameter	Value
n_estimators	100
learning_rate	0.03
max_depth	4
num_leaves	63
min_child_samples	50
subsample	0.9
colsample_bytree	1.0
reg_lambda	0
reg_alpha	0

Because the top-10 set was derived from the split-based importances of a single Round-1 fit, it may in principle be sensitive to the particular training split and to the known tendency of split-count importance to favor high-range predictors; the five-seed robustness check holds the feature set fixed and therefore assesses partition sensitivity of the final comparison rather than selection stability, which we leave to future work.

The ranking is also coherent with classical Framingham epidemiology and, as shown later, with the SHAP attribution of the final model: age, arterial pressure, and smoking behavior sit at the top of both the tree-based importance and the game-theoretic attribution, providing an early face-validity check on the curation step before any test-set evidence is consulted.

3.3. Ablation Study under Class Imbalance

Table 5 reports the test-set performance of the eight scenarios defined in Table 2. The most striking pattern is the decoupling between ranking and detection: the six LightGBM configurations occupy a narrow AUC band (0.677–0.686), yet their recall spans more than an order of magnitude (0.031–0.670). The imbalance remedies therefore act almost entirely on the operating point rather than on the quality of the underlying risk ranking, which is precisely the behavior threshold moving and cost-sensitive weighting are designed to produce.

The reference configuration (S1) displays the classic majority-collapse signature: at the default cutoff it attains 0.847 accuracy and 0.500 precision but a recall of only 0.031, i.e. it silently passes over roughly 97 of every 100 true CHD cases. Lowering the threshold well below 0.5 (S2, Table 5) rescues the recall without retraining the model, confirming that the ranking learned by S1 was informative even though its default operating point was clinically useless.

Cost-sensitive weighting alone (S3) produces a similar rescue through a different mechanism: with the positive class penalized $w_+ = 5.596$ times more heavily (Eq. (5)), the predicted

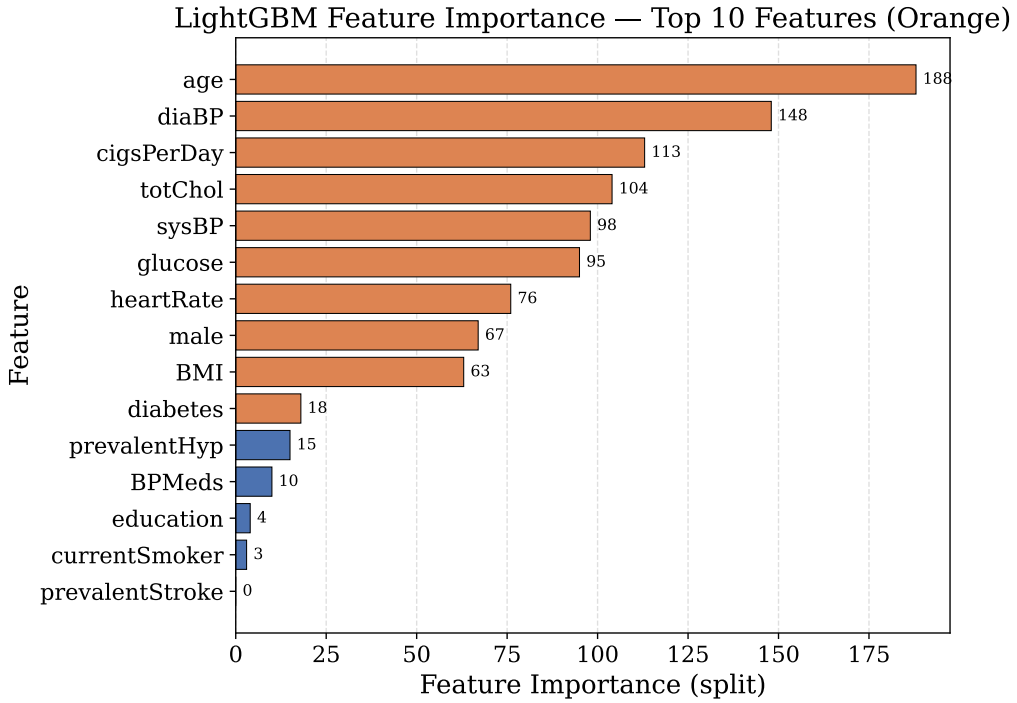


Fig. 3: LightGBM split-based feature importance from the Round-1 model; orange bars denote the ten curated predictors retained for Scenarios 5–8.

probability mass shifts upward so that even the unadjusted 0.5 cutoff now yields a recall of 0.577. Optimizing the threshold on top of the weighted objective (S4) moves the cutoff above 0.5 (Table 5), consistent with the upward shift of the probability mass—and trades a little recall for precision. Two caveats merit attention. First, the combination does not raise MCC: the coefficient falls from 0.211 in S3 to 0.195 in S4, an over-correction signature in which the stacked remedies push slightly past the best precision–recall compromise. Second, the AUC of S4 reaches only 0.681, compared with 0.686 for S1, reminding us that re-weighting optimizes a cost-aware objective, not the ROC.

Feature curation (S5–S6) delivers the strongest detection performance of the LightGBM family: recall 0.670 and MCC 0.214 at its optimized threshold, with the two hyperparameter rounds producing identical predictions under Scenarios 5 and 6, so re-tuning on the reduced space adds nothing beyond curation itself. The accompanying AUC dip, from 0.681 in S4 to 0.677, is well inside the bootstrap intervals and is the expected price of discarding five weak predictors: a slightly noisier ranking exchanged for a simpler, more recall-oriented instrument.

TabPFN (S7–S8) occupies a clearly higher discrimination band, with AUC 0.713 and PR-AUC 0.314 against 0.677–0.686 for the best LightGBM configurations, and the final TabPFN model (S8) achieves the best imbalance-aware scores of the entire study (F1 0.380, MCC 0.245). Feature curation did not harm TabPFN: its AUC and PR-AUC are unchanged at the reported precision (0.713 and 0.314 for both S7 and S8), while the threshold-dependent F1-score rises from 0.365 to 0.380 and the MCC from 0.223 to 0.245 (recall 0.577 to 0.619, precision 0.267 to 0.274). Because these detection metrics depend on the validation-optimized threshold (0.153 for S7, 0.142 for S8), were evaluated on a single test subset, and were not subjected to a significance test, we read the change as a small and possibly threshold-driven improvement rather than as evidence that the foundation model benefits from curation; its ranking quality is unaffected.

The confusion matrices in Fig. 4 translate these trade-offs into clinically readable terms. The final LightGBM model (S6) behaves as a recall-leaning screener, capturing 65 of the 97 true events at the cost of 203 false alarms, whereas the final TabPFN model (S8) is more conservative, giving up five detected cases in exchange for 44 fewer false alarms and hence a higher precision

Table 5: Test-set performance of the eight ablation scenarios of Table 2, reported at each scenario's operating threshold τ^*. AUC and PR-AUC are threshold-independent; the detection metrics are evaluated at τ^*.

ID	τ^*	AUC	PR-AUC	F1	Recall	Precision	Accuracy	MCC
S1	0.500	0.686	0.279	0.058	0.031	0.500	0.847	0.094
S2	0.183	0.686	0.279	0.335	0.474	0.258	0.712	0.184
S3	0.500	0.681	0.268	0.357	0.577	0.258	0.682	0.211
S4	0.523	0.681	0.268	0.345	0.526	0.256	0.695	0.195
S5	0.448	0.677	0.273	0.356	0.670	0.243	0.631	0.214
S6	0.448	0.677	0.273	0.356	0.670	0.243	0.631	0.214
S7	0.153	0.713	0.314	0.365	0.577	0.267	0.693	0.223
S8	0.142	0.713	0.314	0.380	0.619	0.274	0.692	0.245

(0.274 against 0.243). The modest precision of both pipelines is the unavoidable consequence of a 15.19% event rate rather than a modeling failure—at this prevalence, any screener that raises recall must accept a large false-positive burden—and it motivates the decision-curve analysis that follows, where the practical value of acting on these probabilities is assessed directly.

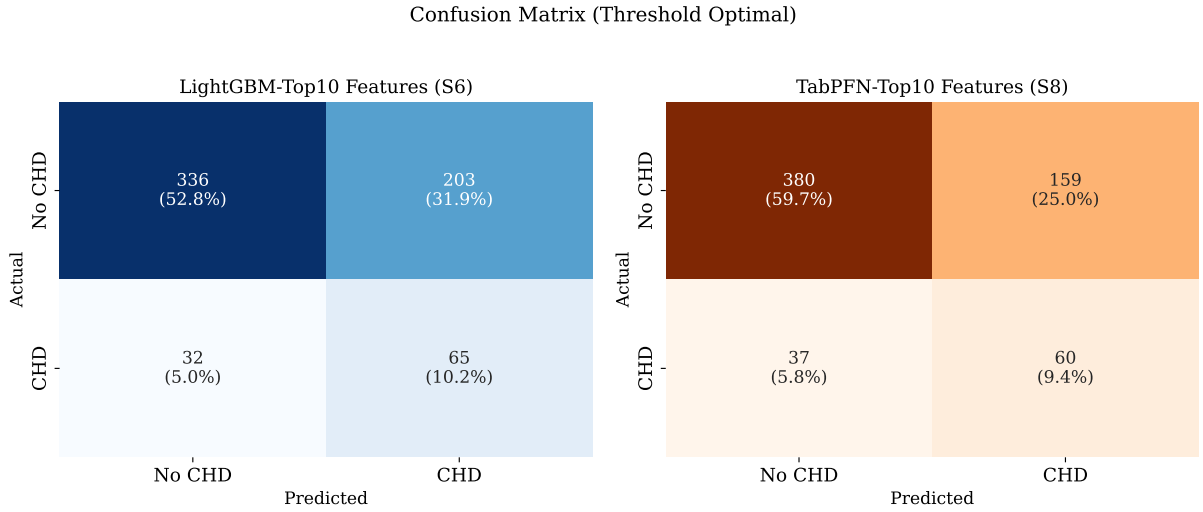


Fig. 4: Confusion matrices at the optimized thresholds on the held-out test subset: (left) the final LightGBM model (Scenario 6); (right) the final TabPFN model (Scenario 8).

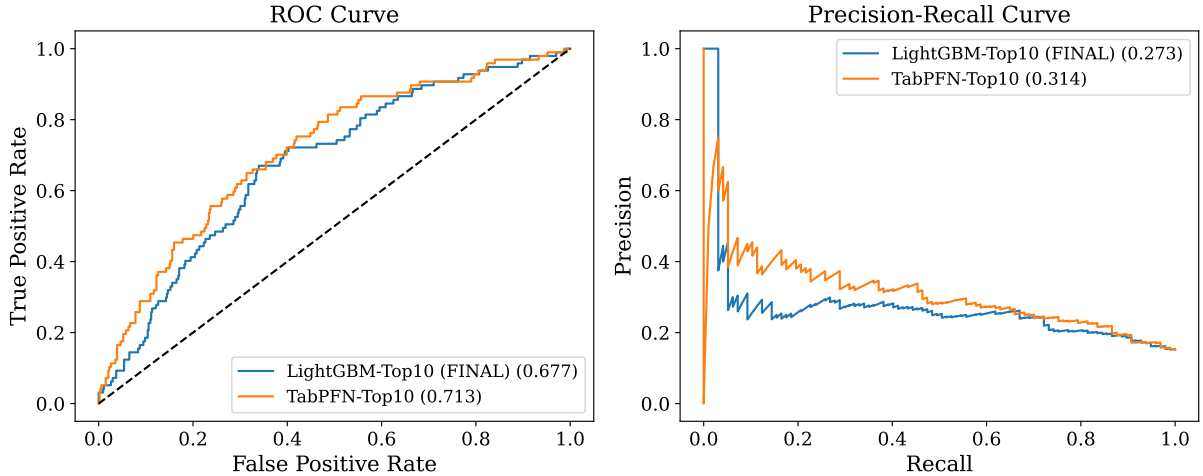


Fig. 5: Receiver operating characteristic (left) and precision–recall (right) curves of the two final pipelines on the held-out test subset; values in parentheses are the AUC and PR-AUC.

3.4. Head-to-Head Comparison: Statistical Significance and Robustness

The receiver operating characteristic and precision–recall curves of the two final pipelines on the held-out test subset are plotted in Fig. 5, where the TabPFN curve lies above the LightGBM curve over most of the false-positive rate range, though the two curves draw closer near the extremes. The descriptive gap between the two final pipelines is confirmed as statistically significant by the designated primary test: the DeLong comparison of the correlated ROC curves yields $z = -2.899$ and $p = 0.0037$, so the null hypothesis of equal AUC is rejected at any conventional level. It is worth emphasizing what this test does and does not say: because DeLong's procedure accounts for the covariance between two AUC estimates computed on the same test cases, the significance reflects a consistent, case-by-case re-ranking advantage of TabPFN rather than an artifact of comparing two independent noise draws.

The bootstrap intervals in Table 6 overlap between the two models, as expected for a test set of 636 observations; this is precisely why interval overlap was never used as the decision rule and the paired DeLong test was designated the primary criterion. The resampling nonetheless places TabPFN above LightGBM on every reported metric—AUC, PR-AUC, and F1—so the advantage extends from pure ranking to the precision–recall behavior that governs screening utility.

As a complement to the DeLong test, the paired bootstrap replicates yield a confidence interval for the between-model AUC difference. Because each resample scores both pipelines on the same subjects, the paired differences respect the correlation structure of the comparison. The mean difference is 0.036 in favor of TabPFN, with a 95% percentile interval of [0.012, 0.061]; the interval excludes zero, corroborating the DeLong result, while its width confirms that the practical magnitude of the advantage is modest rather than large.

The multi-seed robustness check extends the conclusion from one partition to the population of plausible partitions in Table 7. Across the five seeds, TabPFN exceeds LightGBM on AUC and on PR-AUC in *all* replications. Two aspects deserve attention. First, the mean between-model gap is of the same order as the between-seed standard deviation of either model, so the advantage is real but modest in magnitude—a finding we state deliberately, since overstating it would contradict the very intervals we report. Second, the ordering of the two models never reverses under any of the five seeds. Because the top-10 feature set and the LightGBM hyperparameters were held fixed at their primary-seed selections, this exercise isolates partition sensitivity of refitting and evaluation rather than repeating the full development pipeline; it therefore supports the direction of the comparison without establishing that a fully nested redesign would reproduce every magnitude, which we acknowledge as a limitation of the present study.

Table 6: Bootstrap percentile 95% confidence intervals for the two final pipelines on the held-out test subset (1,000 resamples).

Model	Metric	Point estimate	Lower 95% CI	Upper 95% CI
LightGBM (S6)	AUC	0.677	0.618	0.732
	PR-AUC	0.273	0.204	0.353
	F1	0.356	0.297	0.417
TabPFN (S8)	AUC	0.713	0.655	0.766
	PR-AUC	0.314	0.241	0.414
	F1	0.380	0.310	0.448

Taken together, the three layers of inference—paired significance on the held-out test, resampling intervals, and repeated partitions—answer the second research gap: the comparison between a tuned, imbalance-aware tree ensemble and a zero-tuning tabular foundation model is quantified with statistical and robustness evidence rather than a single point estimate.

Table 7: Multi-seed robustness check: AUC and PR-AUC of the two final pipelines under five stratified partitions, with hyperparameters and the feature set fixed at the seed-42 selections.

Seed	LGBM AUC	LGBM PR-AUC	TabPFN AUC	TabPFN PR-AUC
42	0.677	0.273	0.713	0.314
1	0.736	0.339	0.768	0.350
7	0.733	0.353	0.767	0.403
123	0.700	0.379	0.732	0.382
2024	0.684	0.273	0.706	0.315
Mean±SD	0.706±0.024	0.323±0.043	0.737±0.026	0.353±0.036

3.5. Calibration and Decision-Analytic Utility

Beyond discrimination and detection, a screening instrument is only actionable if its predicted probabilities Eq. (9) track observed event rates. The Brier score summarizes this property in a single number, and the contrast between the two final pipelines is sharp: 0.199 for the final LightGBM model (S6), compared with 0.120 for the final TabPFN model (S8). The reliability diagrams in Fig. 6 make the source of the gap visible. The final TabPFN model follows the identity line closely over its entire operating range, with predicted probabilities that never exceed roughly 0.30 and a top bin whose observed event rate of about one third matches its prediction; such a ceiling is sensible for a cohort whose prevalence is 15.19%. The curve of the final LightGBM model, by contrast, departs steadily from the diagonal: subjects assigned probabilities of 0.5–0.7 exhibit observed event rates of only 0.24–0.28, the classic signature of an overconfident scorer whose probability scale is stretched far beyond the range the data support.

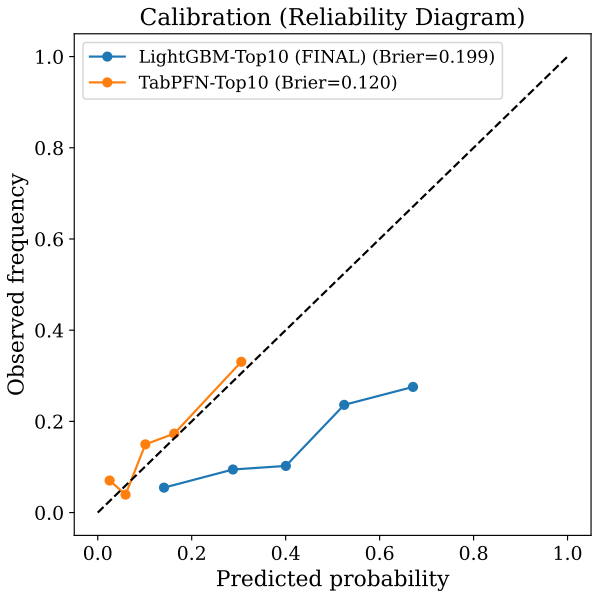


Fig. 6: Reliability diagrams (five quantile bins) of the two final pipelines on the held-out test subset; the dashed line denotes perfect calibration. The final TabPFN model (S8) tracks the identity line (Brier 0.120), whereas the final LightGBM model (S6) overestimates risk in the high-probability region (Brier 0.199).

The logistic calibration coefficients quantify this contrast against their ideal values of 0 (intercept) and 1 (slope). The final LightGBM model yields an intercept of −1.5164 and a slope of 0.7243, whereas the final TabPFN model yields an intercept of −0.1249 and a slope of 0.8073. To quantify the statistical robustness of these calibration differences, paired bootstrap confidence intervals were constructed from 1,000 resamples of the held-out test subset. The mean Brier score difference (LightGBM minus TabPFN) is 0.079 with a 95% percentile interval of [0.061,

0.098], excluding zero and confirming that TabPFN's calibration advantage is not attributable to sampling variability of this test set. The calibration intercept for LightGBM is −1.516 (95% CI: [−1.784, −1.314]), strictly negative and therefore indicating significant systematic over-prediction of absolute risk, whereas for TabPFN the intercept is −0.125 (95% CI: [−0.645, 0.399]), comfortably containing zero and consistent with unbiased prediction in the large. Calibration slopes are 0.724 (95% CI: [0.446, 1.042]) for LightGBM and 0.807 (95% CI: [0.562, 1.090]) for TabPFN; the paired bootstrap interval for the difference in slope deviation from one, $|1 - \text{slope}|$, is $[-0.074, 0.192]$ and includes zero, so no slope difference is claimed, whereas the paired interval for the difference in |intercept| is $[0.908, 1.596]$, excluding zero and supporting the intercept-based claim. The calibration advantage therefore rests on the intercept and the Brier score. The strongly negative intercept of the final LightGBM model confirms a systematic over-statement of absolute risk (miscalibration-in-the-large); the final TabPFN model, with an intercept close to zero, is essentially unbiased in the large. Slope point estimates lie below one for both pipelines, but because the paired interval for the slope-deviation difference includes zero, no slope-based claim is made. Two readings of the LightGBM miscalibration are worth stating. First, it is the price of the cost-sensitive objective: penalizing false negatives w_+ times more heavily (Eq. (5)) pushes the boosting procedure toward higher predicted probabilities for suspected cases, which improves recall but distorts the probability scale. Second, the distortion is systematic rather than erratic—the curve of the final LightGBM model lies below the identity line across the entire high-risk region—so it could in principle be removed by a post-hoc recalibration step, at the cost of the detection behavior deliberately selected in the ablation. The two pipelines are therefore best read as complementary operating points: a recall-oriented screener whose probabilities are deliberately aggressive (the final LightGBM model), and a near-unbiased probabilistic scorer (the final TabPFN model).

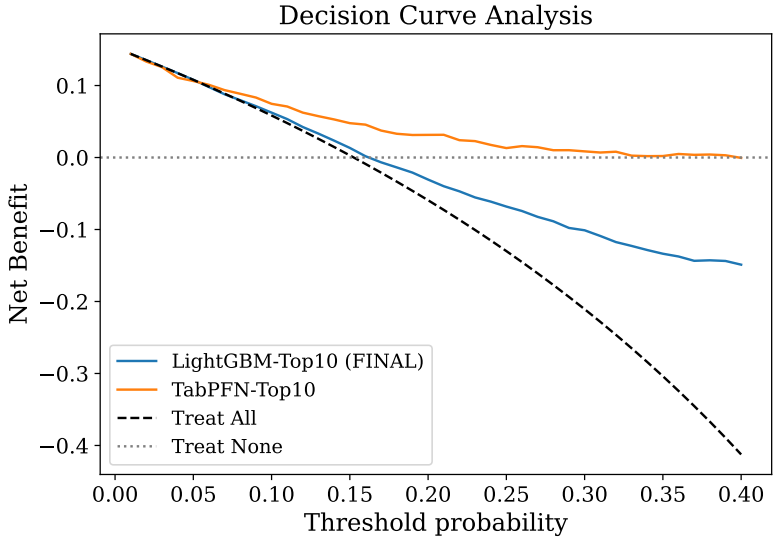
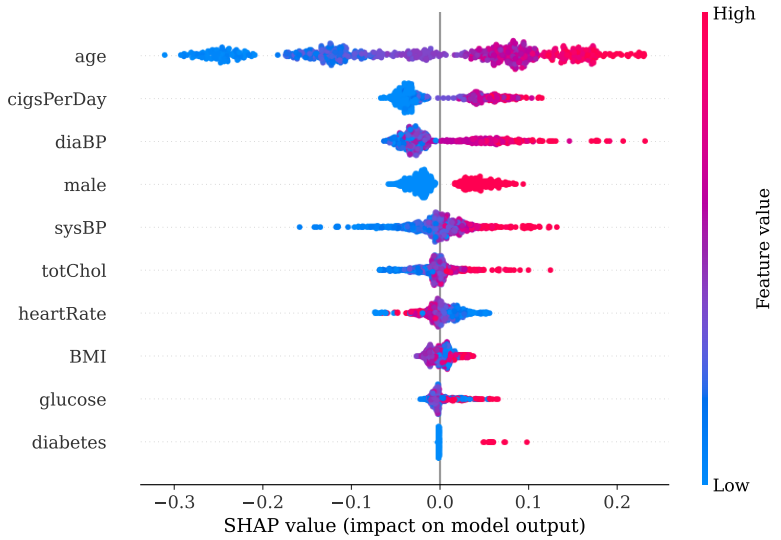


Fig. 7: Decision curve analysis over threshold probabilities 1–40%, comparing the net benefit of the two final pipelines against the treat-all and treat-none strategies.

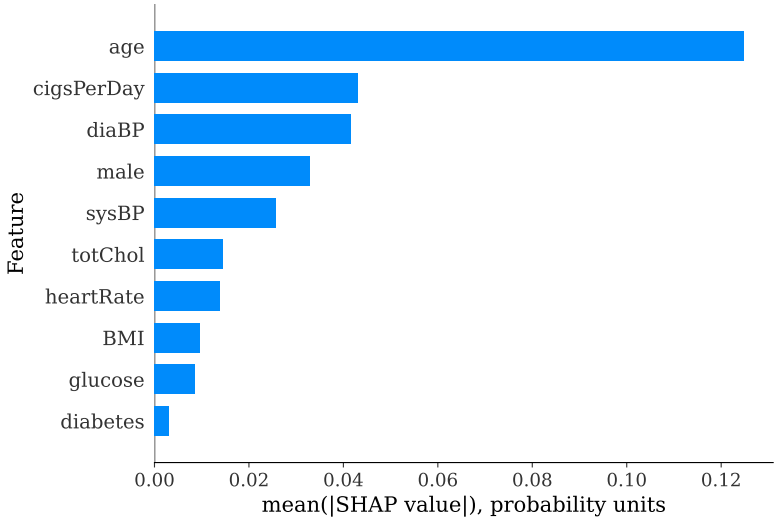
The decision curve analysis in Fig. 7 translates these differences into clinical utility. The emphasized threshold window, 10–20% predicted ten-year risk, was fixed before evaluation because it directly brackets the observed cohort prevalence of 15.19%, the region in which net benefit is most decision-relevant for this specific cohort. At threshold probabilities below roughly 5% all strategies coincide, but from about 6% upward the final TabPFN model separates favorably, keeping its net benefit above both the treat-all and treat-none strategies across approximately the 5–30% window and only approaching the treat-none line near 35–40%. The final LightGBM model is competitive only at the lower end of this range: beyond a threshold of about 16–17% its net benefit turns negative and falls below treat-none, because its overconfident probabilities

would trigger intervention on too many false-positive subjects. Within the 10–20% window itself, the final TabPFN model therefore delivers a positive net benefit, whereas the final LightGBM model is best reserved for low-threshold triage or rule-out use. Because the decision curves rest on a single held-out test subset and on threshold-derived harm–benefit weights that were not elicited from clinicians, we treat them as exploratory decision-analytic support rather than as definitive evidence of clinical utility. Combined with the DeLong result, these exploratory decision-curve results are consistent with the calibration findings but do not establish clinical utility.

3.6. Explainability and SHAP-Based Feature Importance



(a) Beeswarm plot in standardized units; color indicates the feature value and the horizontal axis the Shapley contribution.

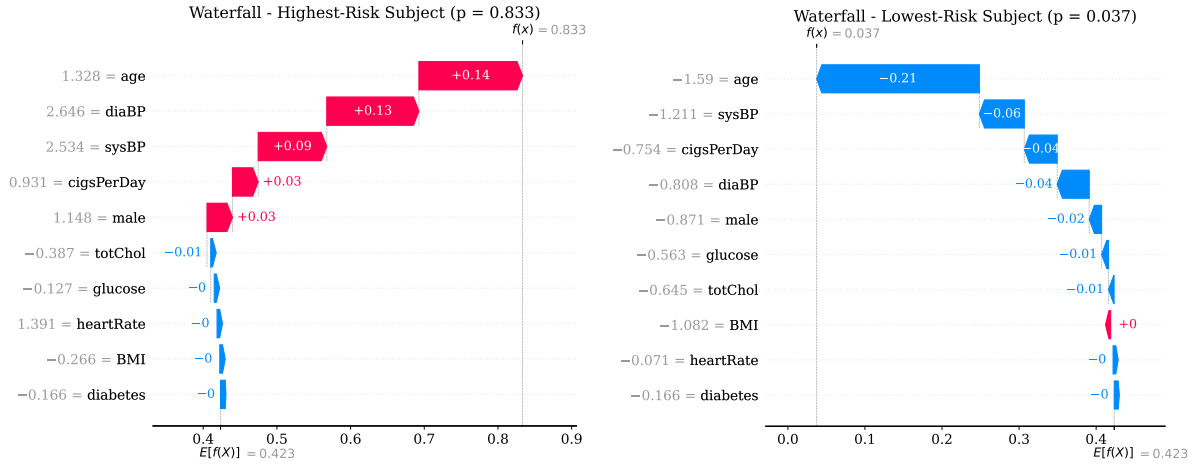


(b) Global importance as mean absolute Shapley value.

Fig. 8: Global SHAP interpretation of the final LightGBM model (Scenario 6).

The TreeExplainer attribution of the final LightGBM model (Scenario 6), configured to return contributions in probability units, concentrates on a small number of clinically familiar predictors in Fig. 8. Age alone accounts for more of the attribution mass than the next three predictors combined, and the remaining weight is shared among tobacco exposure, arterial pressure, sex, and the metabolic markers—precisely the variables that dominate classical Framingham risk functions.

The beeswarm panel additionally shows the direction of each effect: higher age, more cigarettes per day, and elevated arterial pressure consistently push the predicted probability upward, while low values of the same features pull it toward the event-free side. This directional coherence provides a strong associational face-validity check: the model has recovered the epidemiological structure of the cohort rather than spurious artifacts of the preprocessing pipeline.



(a) Highest-risk test observation: age, tobacco, and pressure contributions lift the prediction above baseline. **(b)** Lowest-risk test observation: favorable values of the same predictors drive the prediction downward.

Fig. 9: Local SHAP decompositions in probability space for the two extreme test observations.

Two comparisons reinforce the credibility of this attribution structure. First, the SHAP ranking is broadly consistent with the split-based importance used for curation: the same ten predictors emerge at the top under both criteria, so the feature set retained for Scenarios 5–8 is not an artifact of one particular importance measure. Second, the ordering closely matches the probability-space SHAP analysis reported by Amiroch et al. on the same cohort with XGBoost and SMOTE, in which age, cigsPerDay, male, heartRate, and sysBP were the dominant contributors. That two different learners, trained under different imbalance-handling regimes—synthetic resampling there, cost-sensitive weighting here—converge on essentially the same attribution skeleton suggests that the signal in this cohort is stable rather than model-specific.

The local explanations behave equally sensibly in Fig. 9, where every contribution is expressed in probability units. Each waterfall reconstructs the subject's predicted probability as an additive chain of per-predictor contributions: positive bars raise the predicted probability, negative bars lower it. For the highest-risk test observation, older age, tobacco consumption, and elevated arterial pressure contribute a combined +0.410 in probability units, and the model's predicted probability reaches 0.833; for the lowest-risk observation, the protective values of those same predictors contribute −0.386, and the predicted probability is 0.037. Read in this way, each waterfall is directly auditable in the units a clinician reasons in—absolute changes in predicted risk—rather than in an untransformed model score.

Normalizing the mean absolute Shapley values through Eq. (21) yields the 0–100 relative feature-importance weights presented in Table 8. Unlike a validated patient-level clinical score, these weights summarize the global attribution structure of the tree-based pipeline and serve as a transparent, associational summary of model behavior. They confirm that age, smoking, and blood pressure dominate the model's decision-making process, aligning with established CHD epidemiology, but they do not constitute a deployable diagnostic instrument; any translation into a patient-level screening tool would require prospective clinical validation, explicit scoring rules, and external calibration.

Table 8: Normalized SHAP-based relative feature-importance weights (0–100) of the final LightGBM model (Scenario 6), obtained by normalizing mean absolute Shapley values computed in probability space on the test subset via Eq. (21). The sign of each attribution is displayed in the beeswarm panel of Fig. 8; all ten predictors act in the clinically expected direction.

Predictor	Weight
age	39.24
cigsPerDay	13.57
diaBP	13.10
male	10.36
sysBP	8.09
totChol	4.59
heartRate	4.34
BMI	3.05
glucose	2.72
diabetes	0.94

3.7. Synthesis with Prior Literature

The most direct point of comparison is the corrected analysis of the same cohort by Amiroch et al. [12], who reported an XGBoost model with SMOTE attaining 79.36% accuracy, 22.48% recall, 24.89% F1-score, and an ROC–AUC of 0.6502 with a 95% bootstrap interval of [0.6017, 0.6981] under a pre-specified 80:20 split. The final TabPFN model (0.713) lies above that interval, while the final LightGBM model (0.677) falls within it; because the cohorts, splits, and imbalance treatments differ, no formal cross-study significance claim is made, and their recalls (0.670 and 0.619, Table 5) are roughly three times the value obtained under SMOTE with a default threshold, indicating that algorithmic imbalance handling combined with threshold optimization extracts substantially more minority-class signal from this cohort than resampling alone. The convergence extends to the explanation layer: the SHAP ranking of Amiroch et al. (age, cigsPerDay, male, heartRate, sysBP) overlaps largely with the one obtained here (age, cigsPerDay, diaBP, male, sysBP) despite different learners, different imbalance treatments, and different feature sets, which reinforces the view that the attribution structure of this cohort is stable rather than model-specific. What the present study adds beyond that prior work is precisely the inferential layer that their corrected protocol identified as missing: a paired significance test, repeated-partition robustness, calibration, and decision-analytic evaluation, assembled together with an ablation hierarchy that attributes the gains to specific methodological components.

The threshold behavior observed here also provides a clinical-cohort replication of the systematic findings of Leevy et al. [10] on imbalanced credit-card data. They showed that the default threshold performs poorly under imbalance, that optimal thresholds rise as the positive-class prior probability increases, and that the best threshold-selection results are obtained without resampling. All three patterns appear in Table 5: the default cutoff misses nearly all cases in S1 (recall 0.031), whereas S3 still misses a substantial share (recall 0.577), the optimized thresholds under cost-sensitive weighting (0.448–0.523) lie well above those obtained without it (0.142–0.183), mirroring the prior-probability effect induced by their random undersampling, and the strongest imbalance-aware scores are achieved without any synthetic resampling. The present results therefore extend their evidence from an extreme imbalance ratio (1:578) to the moderate epidemiological imbalance (approximately 1:5.6) typical of clinical screening cohorts.

Finally, the head-to-head result situates the tabular foundation-model claim of Hollmann et al. [13] within a clinically relevant, imbalanced setting. In their benchmarks, default TabPFN outperformed tree ensembles tuned for up to four hours on datasets of up to 10,000 samples; here, a zero-tuning TabPFN outperforms a two-round tuned, cost-sensitive LightGBM with statistical significance (DeLong $p = 0.0037$). The ranking advantage is modest in absolute magnitude, as quantified in the head-to-head comparison, but its direction is preserved under every seed of the

robustness check, and TabPFN additionally produces better-calibrated probabilities (Brier 0.120 versus 0.199) and a positive net benefit over a wider decision-analytic window. Feature curation, in turn, did not harm the foundation model: the comparison of Scenarios 7 and 8 shows an unchanged AUC and slightly higher threshold-dependent F1-score and MCC, so removing the five weakest predictors is at least not detrimental; this single-split evidence is, however, too limited to qualify the raw-input default recommended for TabPFN. Read alongside recent medical-screening comparisons that rely on accuracy alone [20, 24], the present protocol illustrates how paired tests, resampling intervals, and decision curves can convert a method comparison into statistical and clinical-utility inference.

3.8. Limitations

Several limitations should be acknowledged. First, the analysis rests on a single, historically collected cohort whose demographic composition does not represent contemporary or non-Western populations, and no external validation was performed; the reported performance should therefore be interpreted as internal-validity evidence rather than as transportable accuracy. Second, the discrimination achieved, although numerically higher than prior analyses of the same cohort, remains moderate in absolute terms, and the modest positive predictive values reflect the low event prevalence; the models are best viewed as transparent screening aids and careful baselines rather than as standalone diagnostic instruments.

On the methodological side, the SHAP explanations display feature values in standardized units while the contributions themselves are expressed in probability units, and the normalized SHAP-based relative feature-importance weights inherit the associational nature of the underlying cohort, so they should not be read as causal effects nor as a validated patient-level score. The top-10 feature set was derived from the split-based importances of a single Round-1 fit; split-count importance is known to favor high-range predictors and may be sensitive to the particular training split, and selection stability across resamples was not assessed. Relatedly, the five-seed robustness check held the feature set and the hyperparameters fixed at their primary-seed selections, so it assesses partition sensitivity of refitting and evaluation rather than a fully nested redevelopment of the pipeline. The decision-threshold optimization was performed on a single validation subset, and the calibration and decision-curve analyses were conducted on the same test subset used for the primary comparison, albeit with bootstrap confidence intervals; because the decision-curve harm–benefit weights were not elicited from clinicians, the net-benefit findings remain exploratory. Independent confirmation on external cohorts therefore remains necessary. Finally, TabPFN inference is computationally heavier and more memory-demanding than LightGBM and may exhibit minor hardware-dependent floating-point variation, so the tree-boosting pipeline retains practical advantages in resource-constrained settings.

These limitations point to concrete directions for future work: external and temporal validation on independent cohorts, formal recalibration of both pipelines, ensembles that combine the recall-oriented tree booster with the calibrated foundation model, constrained threshold optimization, feature-selection stability analysis across resamples, fully nested cross-validation, clinician-elicited decision-analytic weights, and application of the same protocol to other imbalanced clinical datasets.

4. Conclusion

This study compared a cost-sensitive LightGBM ensemble with the tabular foundation model TabPFN for ten-year CHD risk prediction on the imbalanced Framingham cohort, under a common leakage-free protocol with validation-based threshold optimization, an eight-scenario ablation hierarchy, and normalized SHAP-based relative feature-importance weights. The comparison was positioned as statistical and clinical-utility inference rather than routine algorithm benchmarking.

Under the pre-specified 70/15/15 partition, the final TabPFN model achieved an AUC of 0.713 against 0.677 for the final LightGBM model, a difference confirmed by the paired DeLong

test ($z = -2.899$, $p = 0.0037$) and accompanied by a paired bootstrap difference of 0.036 (95% percentile interval [0.012, 0.061]); the direction of the advantage was preserved under all five seeds of a partition-sensitivity check (mean AUCs 0.737 versus 0.706), although the absolute magnitude of the gap remains modest. The final TabPFN model was also better calibrated (Brier 0.120 versus 0.199) and retained a positive net benefit over a wider decision-analytic window in an exploratory decision-curve comparison, while the ablation showed that threshold optimization and cost-sensitive weighting raised the recall of the final LightGBM model from 0.031 to 0.670 without resampling, and that feature curation did not harm the final TabPFN model (unchanged AUC, slightly higher F1-score and MCC).

The SHAP analysis identified age, cigsPerDay, diaBP, male, and sysBP as the dominant contributors, consistent with established CHD risk structure, and the normalized relative feature-importance weights (0–100) summarize this attribution structure as a transparent, associational profile of model behavior rather than as a patient-level clinical score. Overall, the findings indicate that a zero-tuning foundation model can surpass a fully tuned, imbalance-aware tree ensemble on small, imbalanced clinical tabular data; because all evidence is internal to a single historical cohort and absolute discrimination remains moderate, both pipelines are best viewed as careful screening baselines and hypothesis-generating instruments rather than clinically ready decision tools. Read against Sustainable Development Goal 3 (Target 3.4), the two pipelines offer complementary candidates for screening programs—the calibrated foundation model where tuning expertise is unavailable and the lightweight cost-sensitive ensemble where compute is constrained—pending external, temporal, and prospective validation. External validation, formal recalibration, feature-selection stability analysis, and ensemble combinations are the natural next steps.

CRediT Authorship Contribution Statement

Arfian Jauhar Himawan: Conceptualization, Methodology, Software, Formal analysis, Writing–Original Draft. **A'yunin Sofro:** Supervision, Validation, Writing–Review & Editing.

Declaration of Generative AI and AI-assisted technologies

Generative AI was used solely for language editing and proofreading. The authors reviewed and verified the revised text and remain fully responsible for all content presented.

Declaration of Competing Interest

The authors declare no competing interests.

Ethics Approval and Consent to Participate

Ethics approval was not required for this study.

Funding and Acknowledgments

This research received no external funding.

Data and Code Availability

The dataset analyzed during the current study is publicly available in the Kaggle repository[2]. The code used to preprocess the data under the leakage-free protocol, train the cost-sensitive LightGBM and TabPFN pipelines, optimize the decision thresholds, estimate the bootstrap

[2] https://www.kaggle.com/datasets/noeyislearning/framingham-heart-study

confidence intervals and calibration coefficients, and generate the SHAP analyses is available from the corresponding author upon reasonable request, together with the random seeds and environment specifications required to reproduce the reported runs.

References

[1] World Health Organization. *Cardiovascular Diseases (CVDs): Fact Sheet*. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds). Accessed: 2026-01-29. 2023.

[2] Republik Indonesia. *Undang-Undang Republik Indonesia Nomor 24 Tahun 2007 tentang Penanggulangan Bencana*. Lembaran Negara RI Tahun 2007 Nomor 66. 2007.

[3] United Nations General Assembly. *Transforming Our World: The 2030 Agenda for Sustainable Development*. https://sdgs.un.org/2030agenda. Resolution A/RES/70/1. 2015.

[4] Tianqi Chen and Carlos Guestrin. "XGBoost: A Scalable Tree Boosting System". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016, pp. 785–794. DOI: 10.1145/2939672.2939785.

[5] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. "LightGBM: A Highly Efficient Gradient Boosting Decision Tree". In: *Advances in Neural Information Processing Systems 30 (NeurIPS)*. 2017, pp. 3146–3154.

[6] Jerome H. Friedman. "Greedy Function Approximation: A Gradient Boosting Machine". In: *Annals of Statistics* 29.5 (2001), pp. 1189–1232. DOI: 10.1214/aos/1013203451.

[7] Lisa M. Sullivan, Joseph M. Massaro, and Ralph B. D'Agostino. "Presentation of Multivariate Data for Clinical Use: The Framingham Study Risk Score Functions". In: *Statistics in Medicine* 23.10 (2004), pp. 1631–1660. DOI: 10.1002/sim.1742.

[8] Haibo He and Edwardo A. Garcia. "Learning from Imbalanced Data". In: *IEEE Transactions on Knowledge and Data Engineering* 21.9 (2009), pp. 1263–1284. DOI: 10.1109/TKDE.2008.239.

[9] Charles Elkan. "The Foundations of Cost-Sensitive Learning". In: *Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI)*. 2001, pp. 973–978.

[10] Joffrey L. Leevy, Justin M. Johnson, John Hancock, and Taghi M. Khoshgoftaar. "Threshold optimization and random undersampling for imbalanced credit card data". In: *Journal of Big Data* 10.1 (2023), p. 58. DOI: 10.1186/s40537-023-00738-z.

[11] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. "SMOTE: Synthetic Minority Over-sampling Technique". In: *Journal of Artificial Intelligence Research* 16 (2002), pp. 321–357. DOI: 10.1613/jair.953.

[12] Siti Amiroch, Fitri Nur Laili, Awawin Mustana Rohmah, and Dicka Yale Kardono. "Coronary Heart Disease Risk Prediction under Class Imbalance Using XGBoost with SHAP-Based Interpretation". In: *CAUCHY: Jurnal Matematika Murni dan Aplikasi* 11.1 (2026), pp. 937–955. DOI: 10.18860/cauchy.v11i1.37938.

[13] Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. "Accurate Predictions on Small Data with a Tabular Foundation Model". In: *Nature* 637 (2025), pp. 319–326. DOI: 10.1038/s41586-024-08328-6.

[14] Samuel Müller, Noah Hollmann, Sebastian Pineda Arango, Josif Grabocka, and Frank Hutter. "Transformers can do Bayesian inference". In: *Proceedings of the Tenth International Conference on Learning Representations (ICLR)*. 2022. DOI: 10.48550/arXiv.2112.10510.

[15] Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. "TabPFN: A transformer that solves small tabular classification problems in a second". In: *Proceedings of the Eleventh International Conference on Learning Representations (ICLR)*. 2023. DOI: 10.48550/arXiv.2207.01848.

[16] Elizabeth R. DeLong, David M. DeLong, and Daniel L. Clarke-Pearson. "Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach". In: *Biometrics* 44.3 (1988), pp. 837–845. DOI: 10.2307/2531595.

[17] Bradley Efron. "Bootstrap Methods: Another Look at the Jackknife". In: *The Annals of Statistics* 7.1 (1979), pp. 1–26. DOI: 10.1214/aos/1176344552.

[18] Glenn W. Brier. "Verification of Forecasts Expressed in Terms of Probability". In: *Monthly Weather Review* 78.1 (1950), pp. 1–3. DOI: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

[19] Andrew J. Vickers and Elena B. Elkin. "Decision Curve Analysis: A Novel Method for Evaluating Prediction Models". In: *Medical Decision Making* 26.6 (2006), pp. 565–574. DOI: 10.1177/0272989X06295361.

[20] Bing He et al. "Development of Machine Learning-Based Differential Diagnosis Model and Risk Prediction Model of Organ Damage for Severe *Mycoplasma pneumoniae* Pneumonia in Children". In: *Scientific Reports* 15 (2025), p. 9431. DOI: 10.1038/s41598-025-92089-3.

[21] Brian W. Matthews. "Comparison of the Predicted and Observed Secondary Structure of T4 Phage Lysozyme". In: *Biochimica et Biophysica Acta (BBA) — Protein Structure* 405.2 (1975), pp. 442–451. DOI: 10.1016/0005-2795(75)90109-9.

[22] Scott M. Lundberg and Su-In Lee. "A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems 30 (NeurIPS)*. 2017, pp. 4768–4777.

[23] Lloyd S. Shapley. "A Value for n-Person Games". In: *Contributions to the Theory of Games, Volume II*. Ed. by Harold W. Kuhn and Albert W. Tucker. Princeton University Press, 1953, pp. 307–317. DOI: 10.1515/9781400881970-018.

[24] Ardi Kurniawan, Nur Azizah, and Sheila Sevira Asteriska Naura. "Comparing Multivariate Adaptive Regression Splines and Machine Learning Methods for Classifying Pneumonia in Indonesian Toddlers". In: *CAUCHY: Jurnal Matematika Murni dan Aplikasi* 10.1 (2025), pp. 467–479. DOI: 10.18860/cauchy.v10i1.32418.