Advances in Precision Medicine, 2026, Volume 11, Issue 7
https://bbwpublisher.com/journal/APM
Online ISSN: 2424-9106
Print ISSN: 2424-8592

Two-Stage Constrained Model Selection and Cost-Sensitive Threshold Control for Explainable Diabetes Readmission Prediction

Aminu Usman Jibril[1]*, A. Senthil Kumar[2], Imran Dauda Aliyu[3]

[1]Department of Computer Science, Skyline University Nigeria, Kano, Kano State, Nigeria
[2]School of Science and Information Technology, Skyline University Nigeria, Kano, Kano State, Nigeria
[3]Department of Computer Science, Faculty of Computing, Bayero University Kano, Kano, Kano State, Nigeria

Corresponding author: Aminu Usman Jibril, 2788@sun.edu.ng

Co-author email addresses: senthil.kumar@sun.edu.ng; idaliyu89@gmail.com

Copyright: © 2026 Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), permitting distribution and reproduction in any medium, provided the original work is cited.

Abstract: Model architecture and decision threshold jointly determine the behavior of a clinical prediction system. This study presents two complementary analyses of diabetes readmission prediction. Stage 1 applies Pareto enumeration, an area under the receiver operating characteristic curve (AUROC) tolerance, normalized utility, a worst-objective regret diagnostic, and 100,000 preference-weight simulations to five candidates with complete reported metrics. Stage 2 examines archived results for seven classifiers and explores calibration and cost-sensitive threshold selection for a separately refitted stack. In Stage 1, XGBoost and stacking were non-dominated; an absolute AUROC tolerance of 0.0005 selected XGBoost, which maximized utility in 99.88% of five-objective weight draws. In Stage 2, the stack achieved AUROC 0.70349 and average precision 0.66796. Platt scaling marginally reduced the refitted stack's Brier score from 0.217859 to 0.217845. With illustrative false-negative and false-positive costs of 5,000 and 1,000 units, the test-grid minimum occurred at 0.20. Relative to 0.50, recall increased from 0.56475 to 0.99519 and total cost fell by 56.28%; however, 96.66% of encounters were flagged, and cost was only 3.74% lower than flagging all encounters. Shapley additive explanations (SHAP) described the individual LightGBM model. Preprocessing before splitting, encounter-level partitioning, test-informed calibration and threshold selection, and unresolved cohort provenance limit inference. The framework makes selection preferences and operating trade-offs explicit, but requires a locked, patient-separated evaluation before clinical use.

Keywords: calibration; cost-sensitive classification; decision-curve analysis; diabetes readmission; epsilon-constraint; multi-objective optimization; threshold control; explainable artificial intelligence

Online publication: November 9, 2026

1. Introduction

Diabetes mellitus remains a major source of avoidable illness, premature death, and health-system expenditure worldwide [1, 2]. Readmission is one measurable part of this burden, although it is shaped by much

more than disease severity. Multimorbidity, medication access, discharge coordination, social circumstances and continuity of care can all influence whether a patient returns to hospital. A readmission score therefore cannot establish why an individual returned or assign responsibility to a patient or clinician. Its appropriate role is narrower: to help a clinical team identify encounters that may warrant additional review.

The public Diabetes 130-US Hospitals dataset has become a widely used benchmark for this problem [3]. It contains inpatient encounters collected from 130 United States hospitals between 1999 and 2008 and includes demographic, diagnostic, utilization, laboratory, and medication variables. The dataset is valuable for methodological comparison, but it is also difficult. It contains missing values, repeated encounters, high-cardinality clinical codes, heterogeneous care settings, and an outcome whose operational meaning depends on whether only early readmission or any recorded readmission is considered. These properties make a high headline score insufficient evidence of clinical usefulness.

Published diabetes-readmission studies have examined regression, tree ensembles, boosting and other machine-learning approaches, but their reported performance varies with cohort construction, outcome definition and validation design [4–8]. Ensemble methods are attractive because electronic health records contain nonlinear relationships and mixtures of numerical and categorical information. Random forests reduce variance through decorrelated trees [9], gradient boosting builds an additive learner [10], XGBoost combines boosting with regularization and systems optimizations [11], and LightGBM uses histogram-based learning and leaf-wise growth [12]. Stacking adds a further layer by learning how to combine base-model outputs [13]. These methods do not necessarily rank in the same order across AUROC, recall, F1-score, average precision and implementation complexity.

Clinical prediction requires several distinct performance assessments. AUROC summarizes ranking across thresholds but does not describe workload or error rates at a particular operating point [14]. Precision–recall analysis complements it when positive-case retrieval and false alerts are central concerns [15]. Calibration concerns the agreement between predicted probabilities and observed outcomes [16], while decision-curve analysis summarizes net benefit across threshold probabilities [17]. The TRIPOD+AI reporting statement and PROBAST+AI assessment tool emphasize transparent specification of participants, outcomes, predictors, analysis, validation, performance, and potential bias [18,19].

Deployment introduces further objectives. Edge inference can reduce dependence on continuous connectivity and may limit the transfer of raw records, but it does not automatically provide privacy, security, or clinical safety. Converting a fitted model may change numerical outputs; hardware affects latency, energy, and memory; and explanation generation adds computational cost. Explainability methods such as SHAP can expose influential features [20,21], but explanations require stability checks and domain review. Privacy claims require a specified threat model and empirical or formal evidence; local execution alone is not proof against inversion, extraction, or membership inference [22–25].

Model selection is therefore naturally multi-objective. Pareto analysis identifies candidates for which no objective can improve without degrading another [26–28]. Scalar utility, ε-constraint methods, and sensitivity analysis then make decision preferences explicit [29–32]. This is preferable to choosing a model after informally inspecting a table because it produces a traceable decision rule and reveals how robust that decision is to alternative priorities.

This study examines model choice and threshold choice through two complementary analyses. Stage 1 reconstructs a finite-candidate selection problem from historical model-level metrics, combining four

predictive objectives with a structural simplicity proxy. Stage 2 examines a separate archived experiment with calibration, threshold, decision-curve, and LightGBM explanation outputs. The second analysis uses a stack rather than the XGBoost candidate preferred in Stage 1, so the study does not evaluate a single sequentially optimized pipeline. Its contribution is a transparent account of how stated preferences alter model selection and how an illustrative error-cost ratio alters the operating point, together with explicit limits on what the available evidence establishes.

2. Related work

2.1. Diabetes readmission prediction

Strack et al. introduced the source dataset in an analysis of the relationship between HbA1c measurement and readmission [3]. Subsequent work framed the data as a binary or multiclass prediction task and examined traditional statistical models, tree ensembles, boosting, resampling and feature selection [4–8]. Results differ because investigators use different cohorts, missing-data rules, encounter exclusions, outcome definitions, and validation designs. A performance number is therefore interpretable only together with its analytical protocol.

The distinction between encounters and patients is particularly important. Randomly assigning encounters to training and test sets can place records from the same person in both sets, allowing information leakage through stable patient characteristics or care patterns even when identifiers are excluded from the predictors. Patient-grouped splitting is appropriate when evaluating performance on previously unseen patients. Temporal evaluation addresses performance in future practice, where coding, treatments, pathways, and prevalence may change. External evaluation assesses transportability across settings [18,19,33].

Readmission studies also face a choice of target. Thirty-day readmission is commonly used for quality measurement, whereas any recorded readmission combines early and later events. Combining “less than 30 days” and “more than 30 days” increases the positive class but changes the clinical question: the model predicts whether another admission appears in the dataset rather than imminent readmission alone. This study preserves the outcome used in the source analysis and labels that outcome precisely.

2.2. Explainability, calibration, and clinical utility

Interpretability is not a single property. Intrinsic interpretability concerns whether a model’s mechanism can be understood directly; post-hoc explanation concerns an approximation or attribution produced after fitting. SHAP connects feature attribution to cooperative game theory and supports local and global summaries [20,21]. For tree ensembles, TreeSHAP provides efficient exact or consistent computations under specified dependence assumptions. However, feature attribution is not a causal effect, and correlated predictors can redistribute attribution. A clinically useful explanation should be checked for direction, stability, plausible ranges, and consistency with the data-generating and care processes.

Probability quality also matters. Calibration plots, calibration intercept and slope, and the Brier score should accompany discrimination when predicted risks inform decisions [16,33]. Decision curves compare model-guided action with acting on all or no encounters at stated threshold probabilities [17]. These analyses require individual predictions and outcomes, which were unavailable for the Stage 1 reconstruction. Stage 2 supplies archived calibration and decision-curve outputs, although the underlying prediction vectors were not retained in the materials examined.

2.3. Edge artificial intelligence and privacy

Edge artificial intelligence places computation near the data source, potentially improving responsiveness and operational resilience [34,35]. In a hospital workflow, an edge implementation could transform locally available variables, generate a risk estimate, and present a concise explanation without transmitting raw inputs to a third-party inference service. Nevertheless, deployability is an empirical property of a specified model, conversion pipeline, device, runtime, and workload. File size, peak memory, warm and cold latency, throughput, power, numerical drift, and failure behavior must be measured on target hardware.

Privacy must be evaluated separately. Model inversion can reconstruct sensitive attributes under certain assumptions [22], model extraction can reproduce a service through queries [23], and membership inference can estimate whether a record contributed to training [24]. Differential privacy offers a formal framework when its mechanisms and privacy budget are correctly implemented [25]. Edge placement may reduce one exposure channel—external transmission of raw inputs—but it does not remove risks from compromised endpoints, logs, outputs, insiders, or the fitted model. Any future local implementation should therefore be described as exposure-minimizing, not privacy-preserving, unless an explicit privacy mechanism and threat-model evaluation are supplied.

2.4. Multi-objective model selection

Multi-objective optimization seeks solutions that balance competing objectives. A candidate dominates another if it is no worse on every objective and strictly better on at least one; non-dominated candidates form the Pareto set [26]. Evolutionary methods such as NSGA-II and particle-swarm approaches are useful when the search space is large or continuous [27,28]. Here, the candidate set and reported metrics are finite, so exhaustive enumeration is exact and more transparent than an unnecessary stochastic search.

After identifying the Pareto set, a decision maker still needs a compromise. Weighted sums are simple but sensitive to scale and preferences and may not recover every point in a non-convex continuous frontier [29,30]. The ε-constraint method optimizes secondary criteria while requiring the primary objective to remain within a stated margin [31]. Multicriteria decision analysis emphasizes explicit value judgments and sensitivity analysis [32]. In the present finite problem, all candidates can be evaluated directly under each rule and Monte Carlo weights reveal how often a model is optimal across a broad preference space.

3. Materials and methods

3.1. Study design and evidence boundary

This secondary methodological study uses two evidence layers. Stage 1 applies post hoc decision rules to historical metrics reported in Jibril's doctoral thesis manuscript [36], principally Chapter 4, Tables 4.8–4.11. Stage 2 examines the code, saved outputs, four results CSV files, and figures in the archived notebook diabetes_readmission_experiment.ipynb. The archive provides an additional illustrative experiment, not independent reproduction of the historical cohort or performance estimates. The original input extract, split identifiers, fitted estimators, and individual prediction and SHAP arrays were absent. The present study audits the available records and recomputes summary-level decision analyses; it does not retrain the predictive models.

The distinction matters for interpretation. Stage 1 begins with a finite set of already evaluated candidates, denoted by $\mathcal{M}$. Its normalization, dominance, utility, regret, and weight-sensitivity calculations

are reproducible from the reported metrics and decision rules. Stage 2 documents what the archived implementation and outputs show. Neither layer establishes external validity, an independently assessed clinical operating threshold, device performance, or privacy protection.

3.2. Data source, cohort, and outcome

The stated source is the Diabetes 130-US Hospitals for Years 1999–2008 dataset in the University of California Irvine (UCI) Machine Learning Repository [3,37]. The repository lists 101,766 encounters and 47 predictor features, in addition to identifier and outcome fields. The archived notebook loads a local file named diabetes_data.csv and displays 40,000 rows and 26 input fields. The extract itself and the rule used to derive it from the UCI resource were not supplied. Matching cohort size and class totals do not establish identical encounter membership or partitions across Stages 1 and 2. Table 1 distinguishes the historical report from the notebook's documented dimensions.

The binary outcome combined both recorded readmission categories:

$$y_i = I\left(c_i \in \{\text{“<30”}, \text{“>30”}\}\right), \tag{1}$$

Here c_i is the recorded category for encounter i, I is the indicator function, and $y_i = 1$ denotes either recorded readmission category. The notebook implements the target as 0 when the label equals "No" and 1 for every other nonmissing value. Without the input extract, exact correspondence between that implementation and the intended categories in Equation (1) cannot be independently checked. This is an any-recorded-readmission endpoint, not a fixed 30-day outcome.

The saved target counts comprise 18,727 positive and 21,273 negative encounters. The historical report gives 32,000 training encounters (14,982 positive and 17,018 negative) and 8,000 test encounters (3,745 positive and 4,255 negative). The Stage 2 code specifies an 80:20 outcome-stratified encounter split with random_state = 42, without patient grouping. Removing patient_nbr from predictors does not prevent the same patient from contributing encounters to both partitions. The exact split membership and extent of patient overlap remain unverified.

Table 1. Cohort characteristics and evidence available for the two analyses

Item	Value and evidence
Original resource	101,766 encounters; 47 predictor features (UCI)
Analytical input	40,000 encounters; 26 input fields (notebook output)
Positive encounters	18,727 (46.82%)
Engineered variables	21 (archived code; Appendix A)
Final predictors	43 in Stage 2; historical count of 44 is unreconciled
Training partition	32,000; 17,018 negative and 14,982 positive (historical report)
Test partition	8,000; 4,255 negative and 3,745 positive
Calibration refit	25,600 fitting encounters; 6,400 calibration encounters
Partition rule	Outcome-stratified encounters; seed 42 in Stage 2; no patient grouping
Outcome	Any recorded readmission versus none; see target-mapping limitation

3.3. Preprocessing and candidate models

The Stage 2 code replaces "?", "Unknown/Invalid", and empty strings with missing values, then removes columns with more than 80% missingness. The saved output identifies max_glu_serum (94.7150%) and

A1Cresult (83.1125%) for removal. It adds 21 engineered variables, replaces three diagnosis fields with broad categories, integer-encodes categorical predictors, removes the outcome and identifier fields, median-fills numerical columns, and standardizes the resulting matrix. The saved matrix has 43 predictors, correcting the historical count of 44. Missing or unparseable diagnoses become an "unknown" category; the diagnosis mode-imputation lines are commented out. Appendix A specifies the engineered variables, coding rules, and final predictor inventory. Crucially, feature construction, missingness decisions, encoding, median imputation, and scaling precede the split. Standardization therefore uses full-cohort statistics:

$$z_{ij} = \frac{x_{ij} - \mu_{j,cohort}}{\sigma_{j,cohort}}. \quad (2)$$

In Equation (2), x_{ij} is the value of predictor j for encounter i, and $\mu_{j,cohort}$ and $\sigma_{j,cohort}$ are its full-cohort mean and standard deviation. This preprocessing order exposes the evaluation partitions to information used to construct the predictors and limits generalization claims. A corrected evaluation must fit data-dependent transformations within training folds. Integer encoding can impose artificial order on nominal categories, and scaling is not necessary for tree splits. The seven candidate configurations are summarized in Table 2. The stack combines LightGBM, XGBoost, and ExtraTrees through logistic regression and includes the input predictors through passthrough=True:

$$p_{stack}(x) = \sigma\left(\beta_0 + \sum_{i=1}^{3} \beta_i p_i(x) + \sum_{j=1}^{d} \gamma_j x_j\right) \quad (3)$$

Here $p_i(x)$ is the probability from base learner i, x_j is an input predictor, and σ is the logistic function. Stage 2 specifies five-fold StackingClassifier fitting (cv = 5), which generates out-of-fold base-model predictions for the meta-learner. This learner-level cross-validation does not undo preprocessing performed on the full cohort. Equivalent fold-level evidence was unavailable for the historical Stage 1 stack. Average precision was missing for Stage 1 logistic regression and random forest; they remain in the descriptive table but are excluded from optimization requiring all five objectives.

Table 2. Candidate configurations reported in the thesis and specified in the archived notebook

Model	Reported specification
Logistic regression	C=1.0; max_iter=2000; class_weight=balanced
Random forest	400 trees; max_depth=15; min_samples_leaf=5; class_weight=balanced
ExtraTrees	400 trees; max_depth=15; min_samples_leaf=3; class_weight=balanced
XGBoost	500 trees; max_depth=5; learning_rate=0.03; subsample=0.8; colsample_bytree=0.8; scale_pos_weight≈1.14
LightGBM	500 trees; learning_rate=0.03; num_leaves=31; class_weight=balanced
Tuned / inherited LightGBM	305 trees; learning_rate=0.0458; num_leaves=77; max_depth=5; subsample=0.996; colsample_bytree=0.691
Stacked ensemble	LightGBM, XGBoost, ExtraTrees; logistic-regression meta-learner; passthrough=True; cv=5 in Stage 2

All Stage 2 stochastic estimators use random_state = 42. The stack uses cv = 5 and a logistic-regression meta-learner with max_iter = 2000 and predictor passthrough. LightGBM with inherited settings also uses class_weight = balanced. Historical fold construction remains unverified.

The thesis reports 30 Optuna trials with AUROC evaluated on the test set [36]. The historical tuned LightGBM candidate is consequently test-informed and exploratory. It is retained as a labeled candidate

in this post hoc decision analysis, not as an unbiased estimate of the benefit of tuning. In Stage 2, RUN_OPTUNA_TUNING is False and the notebook reuses the inherited configuration; it does not execute a new search. The CSV label "Tuned LightGBM" denotes those inherited settings. Stage 2 uses random_state = 42 for its stochastic estimators and splits. The notebook metadata records Python 3.10.19, but a pinned package environment was not supplied.

No missing average-precision value was imputed from AUROC or prevalence. The five complete Stage 1 candidates are ExtraTrees, LightGBM, tuned LightGBM, XGBoost, and stacking. Because the historical tuned candidate is dominated by XGBoost, its inclusion does not change the Pareto set. All selection results remain conditional on the available, rounded historical metrics and the stated candidate set.

3.4. Performance measures

Stage 1 accuracy, precision, recall, F1-score, AUROC, and average precision were taken from the historical test-set tables. Stage 2 values were checked against the archived CSV files. Precision, recall, and F1-score are

$$Precision = \frac{TP}{TP + FP}, Recall = \frac{TP}{TP + FN}, F_1 = \frac{2PR}{P + R}. \tag{4}$$

TP, FP, and FN denote true-positive, false-positive, and false-negative counts; P and R in Equation (4) denote precision and recall. AUROC estimates the probability that a positive encounter receives a higher score than a negative encounter, with half credit for ties. For N_+ positive and N_- negative encounters, prediction score s, and indicator function I,

$$AUROC = \frac{1}{N_+ N_-} \sum_{i:y_i=1} \sum_{k:y_k=0} [I(s_i > s_k) + \frac{1}{2} I(s_i = s_k)] \tag{5}$$

Average precision (AP) sums precision at each increase in recall, with K denoting the number of increments:

$$AP = \sum_{r=1}^{K} (Recall_r - Recall_{r-1}) Precision_r. \tag{6}$$

Recall describes positive-case retrieval, whereas F1-score combines precision and recall at the evaluated operating point. AUROC and AP summarize ranking [14,15]. Accuracy was reported descriptively but excluded from optimization. The AP objective P_m below is distinct from precision P in Equation (4). Threshold-dependent differences cannot establish that one architecture is superior at every operating point.

3.5. Multi-objective formulation

For candidate m, the objective vector was

$$\max_{m \in M} \mathbf{g}(m) = (A_m, R_m, F_m, P_m, S_m), \tag{7}$$

A, R, F, P, and S denote AUROC, recall, F1-score, AP, and structural simplicity, respectively; every objective is maximized. Complexity C counts fitted predictive components required for inference: C = 1 for an individual estimator and C = 4 for the three base learners plus meta-learner in the stack. This is an architectural proxy, not a measurement of tree count, latency, memory, energy, or model size. Simplicity is

$$S_m = 1 - \frac{C_m - C_{min}}{C_{max} - C_{min}}. \tag{8}$$

Thus, S = 1 for the individual estimators and S = 0 for the stack. This binary proxy represents the observed architectural distinction. Measured computational costs would be preferable for deployment decisions.

Candidate m_a dominated m_b when

$$m_a \succ m_b \Leftrightarrow g_j(m_a) \geq g_j(m_b)\ \forall j \wedge \exists j : g_j(m_a) > g_j(m_b). \tag{9}$$

The non-dominated candidates constituted the exact Pareto set. Since only five complete candidates were available, all pairwise dominance relations were enumerated.

Equivalently, the Pareto set was written as

$$P = \{m \in M : \nexists k \in M \text{such that} k \succ m\}. \tag{10}$$

3.6. ε-constraint and compromise selection

AUROC was the constrained discrimination objective. For a permitted absolute AUROC tolerance ε, the feasible set was

$$F_\varepsilon = \{m \in M : A_m \geq A_{max} - \varepsilon\}. \tag{11}$$

Within that set, candidates were compared lexicographically by recall, F1-score, average precision, simplicity, and AUROC:

$$m_\varepsilon^* = arg\ \max_{m \in F_\varepsilon} (R_m, F_m, P_m, S_m, A_m)_{lex}. \tag{12}$$

The rule first retains discrimination within ε, then favors recall, F1-score, AP, simplicity, and residual AUROC in that order. Tolerances of 0, 0.00025, 0.0005, 0.001, 0.003, and 0.01 were evaluated. These are descriptive decision tolerances, not statistically established non-inferiority or equivalence margins.

For a complementary compromise analysis, each objective was min–max normalized across complete candidates:

$$\tilde{g}_{jm} = \frac{g_{jm} - \min_{k \in M} g_{jk}}{\max_{k \in M} g_{jk} - \min_{k \in M} g_{jk}}. \tag{13}$$

The weighted utility was

$$U_m(\mathbf{w}) = \sum_{j=1}^{5} w_j \tilde{g}_{jm}, w_j \geq 0, \sum_{j=1}^{5} w_j = 1. \tag{14}$$

Equal weights provided a simple reference score, followed by sensitivity analysis. They remain a value judgment, particularly because predictive objectives can be correlated. Min–max normalization also depends on the candidate set.

To make the compromise interpretation more transparent, normalized objective regret was also defined as the distance from the best observed value,

$$r_{jm} = 1 - \tilde{g}_{jm}, \tag{15}$$

and the worst-objective regret of a candidate was

$$R_m^{max} = \max_{j \in \{1,\ldots,5\}} r_{jm}. \tag{16}$$

This worst-objective regret diagnostic complements the stated utility and ε-constraint rules. It identifies

candidates that accept a large shortfall on at least one normalized objective.

3.7. Weight-sensitivity analysis

For preference sensitivity, 100,000 weight vectors were sampled from Dirichlet(1,1,1,1,1) using NumPy default_rng with seed 2788. Candidate order was ExtraTrees, LightGBM, tuned LightGBM, XGBoost, and stacking; argmax selects the first candidate if utilities tie. The selection frequency was

$$\hat{\pi}_m = \frac{1}{B}\sum_{b=1}^{B} I\left[m = arg\max_{k\in M} U_k\left(\mathbf{w}^{(b)}\right)\right]. \tag{17}$$

A second analysis omitted simplicity and sampled four-dimensional Dirichlet weights after resetting the generator to seed 2788. The reported frequencies were reproduced during the numerical audit using NumPy 2.3.5. They describe the specified mathematical weight distribution, not the distribution of clinical preferences. The analysis is reproducible from the tabulated objectives, candidate order, tie convention, seed, and numerical environment.

3.8. Explainability and edge-validation protocol

No fitted Stage 1 model or encounter-level explanation array was available. The following protocol therefore concerns future validation of the selected XGBoost candidate. The separate Stage 2 LightGBM analysis is described in Sections 3.9 and 4.11. For a locked model, a SHAP attribution can be expressed as

$$\phi_j = \sum_{S\subseteq F\setminus\{j\}} \frac{|S|!\left(p-|S|-1\right)!}{p!}\left[f\left(S\cup\{j\}\right)-f\left(S\right)\right], \tag{18}$$

F is the feature set, p its size, S a subset excluding feature j, and f(S) the coalition value under a specified background-data and feature-dependence convention. The additive representation is $f(x) = \phi_0 + \Sigma_j\phi_j$ [20,21]. A future study should report the explanation scale, background population, dependence assumptions, global and local results, and stability under resampling or perturbation. Correlated engineered predictors require particular care.

The selected model should then be converted only after the preprocessing and model artifacts are frozen. Numerical fidelity should be measured by

$$\Delta p_{max} = \max_i \left| p_{i,converted} - p_{i,source} \right|. \tag{19}$$

On each target device, cold start, warm-up, and repeated inference should be separated. Latency percentiles are

$$L_q = percentile_q\left(\{t_{i,end} - t_{i,start}\}_{i=1}^{N}\right), q \in \{50,95,99\}. \tag{20}$$

Peak resident memory, serialized size, throughput, energy per inference, failure rate, and explanation latency should be recorded. To quantify only the communication exposure reduced by local inference, one may report

$$R_{exposure} = 1 - \frac{B_{edge\to external}}{B_{raw\to external}}, \tag{21}$$

B denotes transmitted bytes under a defined workflow. The denominator must be positive for this ratio to be meaningful. Reduced transmission is not a general privacy guarantee; threat-specific security and privacy evaluation remains necessary.

3.9. Archived calibration and threshold analysis

Stage 2 examines the hybrid stack because the archived calibration and threshold outputs were generated for that architecture. It does not apply threshold control to the Stage 1-selected XGBoost model. For **Table 7**, each candidate is fitted on the 32,000-encounter training partition. For calibration, the notebook splits that partition again, using outcome stratification, test_size = 0.20, and random_state = 42. It fits a new five-fold stack on 25,600 encounters and then fits Platt and isotonic mappings on the remaining 6,400. These calibration encounters are excluded from fitting this stack, although they already contributed to full-cohort preprocessing. The raw calibration comparator is the same 25,600-encounter refit, not the **Table 7** stack.

For predicted probability p_i and binary outcome y_i, probability quality was summarized by the Brier score

$$BS = \frac{1}{n}\sum_{i=1}^{n}\left(p_i - y_i\right)^2. \tag{22}$$

The archived expected calibration error (ECE) calculation uses ten equal-width bins B_b over [0,1),

$$ECE = \sum_{b=1}^{10}\frac{\left|B_b\right|}{n}\left|\overline{y}_b - \overline{p}_b\right|, \tag{23}$$

where n is the number of evaluated encounters, $|B_b|$ is the bin count, $\hat{y}_b$ is its observed positive proportion, and $\hat{p}_b$ is its mean predicted probability. Empty bins are skipped. The implementation excludes scores exactly equal to 1 from the bin sum while retaining n as the denominator; whether such scores occurred cannot be checked without the predictions. **Table 8** therefore labels these values as archived ECE. The reliability diagram instead uses ten quantile bins, a distinct binning scheme.

Both calibration mappings are fitted on the 6,400-encounter calibration partition. The notebook then compares raw, Platt, and isotonic probabilities using Brier scores on the 8,000-encounter test partition and selects the lowest-scoring variant for subsequent analyses. Thus, although the mappings are not fitted to test outcomes, their selection is test-informed. The resulting calibration and threshold comparisons are exploratory.

Let c_{FN} and c_{FP} be illustrative costs assigned to a missed positive encounter and a false-positive alert. The notebook selects the threshold with the smallest total empirical test-set cost over the discrete grid T = {0.01, 0.02, …, 0.99}:

$$t^* = \arg\min_{t\in T}\left[c_{FN}FN\left(t\right) + c_{FP}FP\left(t\right)\right]. \tag{24}$$

The costs are c_{FN} = 5,000 and c_{FP} = 1,000 units. These scenario weights do not represent verified monetary costs or intervention effects. The threshold grid is evaluated on the same test outcomes used to select the calibrator, so its minimum is a descriptive sample result. The implementation predicts positive when $p \geq t$ and chooses the first grid threshold if costs tie. Decision-curve net benefit is calculated as

$$NB\left(t\right) = \frac{TP\left(t\right)}{n} - \frac{FP\left(t\right)}{n}\frac{t}{1-t}. \tag{25}$$

Decision curves compare model-guided action, acting on all encounters, and acting on none for threshold probabilities from 0.01 to 0.50 in increments of 0.01. Net benefit uses the threshold-specific weighting t/(1−t), which differs from holding the illustrative 5:1 error-cost ratio fixed across the grid. The SHAP analysis calls TreeExplainer on the individual LightGBM baseline fitted on 32,000 encounters and samples 2,000 test

encounters with random_state = 42. No explicit background dataset, feature_perturbation setting, model_output setting, or SHAP version is recorded. These attributions describe that individual estimator on its default output scale; they are not a validated explanation of the stack or calibrated probabilities.

3.10. Reproducibility and integrity checks

The archive was checked at code, saved-output, CSV, and figure levels. The seven-model metrics, calibration summaries, and threshold comparisons were reconciled with the four CSV files. Confusion counts were checked arithmetically, and Stage 1 dominance, normalization, utility, regret, and weight frequencies were recalculated. Stage 1 figures were regenerated from its tables. Stage 2 plot data were retained from the archived figures; titles and labels were corrected without changing plotted results. Dense local SHAP force plots were omitted from the main manuscript. These checks establish internal consistency of the retained records, not independent reproduction of model training or data provenance.

4. Results

4.1. Predictive comparison

Table 3 reproduces the historical Stage 1 point estimates. The stack has the largest reported accuracy (0.6461), precision (0.6377), AUROC (0.7042), and AP (0.6678). XGBoost and LightGBM share the largest recall (0.6323), while XGBoost has the largest F1-score (0.6246). Relative to XGBoost, stacking gains 0.0005 AUROC and 0.0014 AP but loses 0.0673 recall and 0.0254 F1-score at the reported operating points. Individual paired predictions were unavailable, preventing uncertainty estimation for these differences.

The thesis also reports the stack's threshold-specific confusion counts: 3,053 true negatives, 1,202 false positives, 1,629 false negatives, and 2,116 true positives. These values reproduce its recall of 0.5650 and imply specificity of 0.7175. The stack therefore behaved conservatively at the evaluated threshold: it rejected more negative encounters correctly but failed to identify 43.50% of the positive encounters. This operating pattern reinforces the need to evaluate model choice jointly rather than infer clinical suitability from AUROC alone.

Table 3. Historical Stage 1 test-set performance

Model	Accuracy	Precision	Recall	F1	AUROC	AP
Logistic regression	0.6268	0.6130	0.5498	0.5797	0.6702	NR
Random forest	0.6415	0.6296	0.5688	0.5976	0.6963	NR
ExtraTrees	0.6324	0.6144	0.5765	0.5948	0.6862	0.6538
LightGBM	0.6411	0.6132	0.6323	0.6226	0.7009	0.6622
Tuned LightGBM	0.6425	0.6160	0.6275	0.6217	0.7015	0.6631
XGBoost	0.6441	0.6170	0.6323	0.6246	0.7037	0.6664
Stacked ensemble	0.6461	0.6377	0.5650	0.5992	0.7042	0.6678

AP, average precision; AUROC, area under the receiver operating characteristic curve; NR, not reported. Tuned LightGBM is an exploratory, test-informed historical configuration. All values are reported point estimates.

4.2. Pareto structure

The Stage 1 workflow is shown in **Figure 1**, and **Figure 2** projects the complete candidates onto AUROC and recall. XGBoost dominates ExtraTrees, LightGBM, and tuned LightGBM across the five objectives. Neither XGBoost nor stacking dominates the other: stacking retains small AUROC and AP advantages, while XGBoost has greater recall, F1-score, and structural simplicity. The exact Pareto set therefore contains XGBoost and the stack.

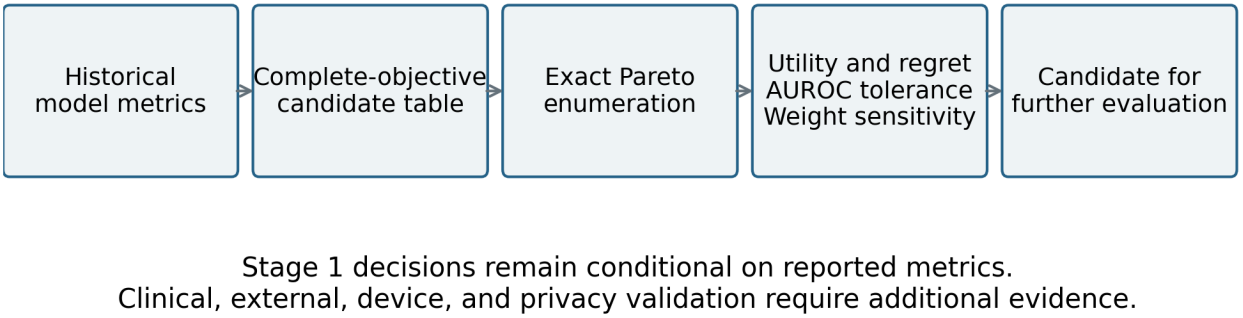


Figure 1. Stage 1 finite-candidate workflow. Decisions are conditional on the historical summary metrics. Calibration, external validation, device performance, and clinical readiness cannot be inferred from this analysis.

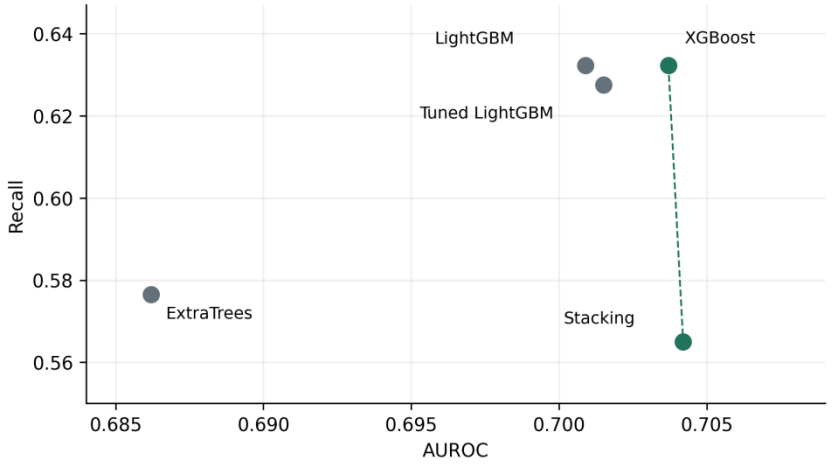


Figure 2. AUROC–recall projection of the five complete candidates. The dashed segment connects the two non-dominated candidates in the full five-objective analysis. A two-dimensional projection cannot by itself establish dominance; the complete objective table was used.

4.3. Normalized compromise utility

Table 4 gives the normalized objectives and equal-weight utility; **Figure 3** shows the corresponding profiles. XGBoost achieves the largest utility (0.9744), combining near-maximum AUROC and AP with maximum recall, F1-score, and simplicity. The stack's maximum AUROC and AP do not offset its lower recall, F1-score, and simplicity under equal weights.

Maximum normalized regret is 1.0000 for ExtraTrees, 0.4000 for LightGBM, 0.3357 for tuned LightGBM, 0.1000 for XGBoost, and 1.0000 for stacking. XGBoost therefore also minimizes the worst observed objective shortfall. This supplementary calculation agrees with the utility comparison without changing the stated decision rule.

Table 4. Min–max normalized objectives and equal-weight utility

Model	AUROC	Recall	F1	AP	Simplicity	Utility
ExtraTrees	0.0000	0.1709	0.0000	0.0000	1.0000	0.2342
LightGBM	0.8167	1.0000	0.9329	0.6000	1.0000	0.8699
Tuned LightGBM	0.8500	0.9287	0.9027	0.6643	1.0000	0.8691
XGBoost	0.9722	1.0000	1.0000	0.9000	1.0000	0.9744
Stacked ensemble	1.0000	0.0000	0.1477	1.0000	0.0000	0.4295

Min–max normalization is sample-dependent: the values describe these candidates, not universal clinical scales. The utility is therefore a decision aid rather than an estimate of patient benefit.

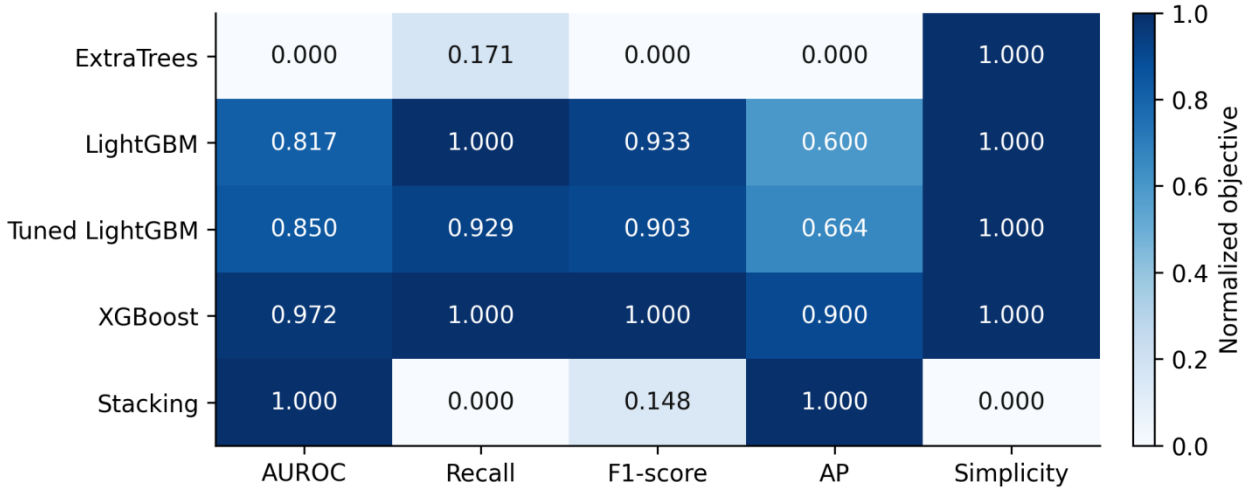


Figure 3. Normalized objective profiles used in the compromise calculation. XGBoost remains close to the observed best value on every objective; the stack reaches the maximum AUROC and average precision but has larger deficits in recall, F1-score, and structural simplicity.

4.4. ε-constraint results

At ε = 0 and 0.00025, only stacking satisfies the AUROC constraint. At ε = 0.0005, XGBoost becomes feasible and is selected by the stated lexicographic rule because it has greater recall. XGBoost remains selected at all wider evaluated tolerances (**Table 5**; **Figure 4**). The switching point is conditional on the four-decimal historical inputs: allowing a 0.0005 absolute AUROC shortfall changes the preferred model at the reported operating points. It does not establish statistical non-inferiority or clinical equivalence.

Table 5. ***ε-constraint selection results***

ε	Minimum AUROC	Feasible candidates	Selected
0	0.70420	Stacked ensemble	Stacked ensemble
0.00025	0.70395	Stacked ensemble	Stacked ensemble
0.00050	0.70370	XGBoost; Stacked ensemble	XGBoost
0.00100	0.70320	XGBoost; Stacked ensemble	XGBoost
0.00300	0.70120	Tuned LightGBM; XGBoost; Stacked ensemble	XGBoost
0.01000	0.69420	LightGBM; Tuned LightGBM; XGBoost; Stacked ensemble	XGBoost

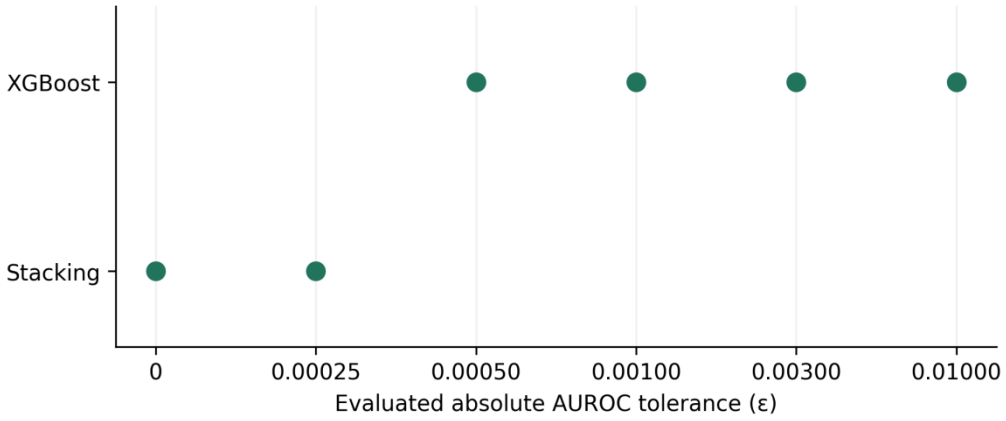


Figure 4. Model selected at each evaluated absolute AUROC tolerance. The switch is descriptive, based on the rounded historical metrics; it is not a statistical non-inferiority result.

4.5. Robustness to preference weights

XGBoost maximizes utility in 99,880 of 100,000 five-objective draws (99.88%), and stacking in 120 (0.12%). With simplicity removed, the corresponding counts are 98,693 (98.69%) and 1,307 (1.31%). **Figure 5** summarizes both analyses. No other candidate is selected. The preference for XGBoost therefore does not depend solely on component count; stacking is favored only when its AUROC and AP advantages receive sufficient weight relative to its recall and F1-score deficits.

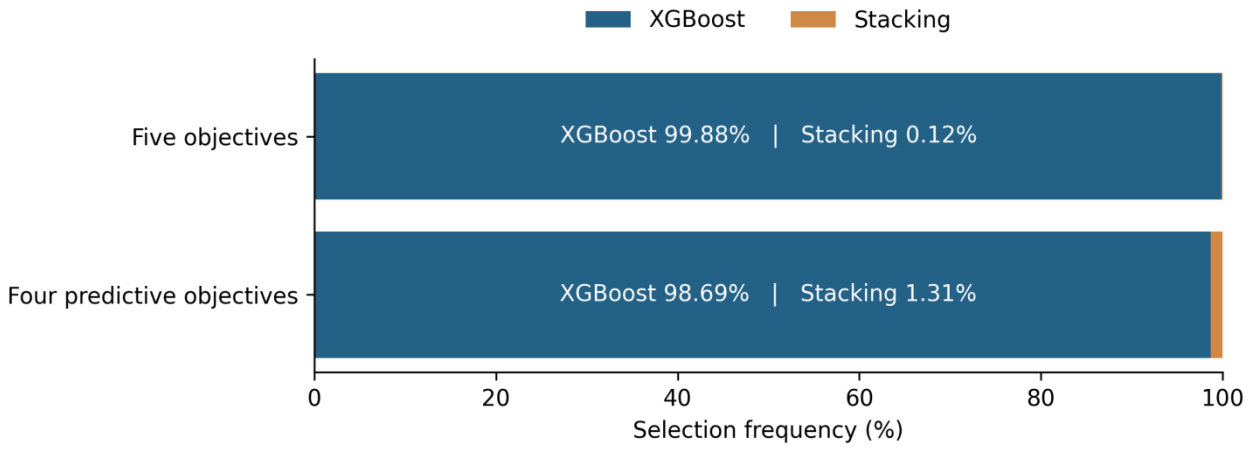


Figure 5. Selection frequencies for 100,000 uniformly distributed Dirichlet weight vectors with seed 2788. "Five objectives" includes structural simplicity; "predictive objectives only" excludes it.

4.6. Dominance interpretation

Figure 6 displays all five pairwise dominance relations: XGBoost dominates ExtraTrees, LightGBM, and tuned LightGBM; LightGBM and tuned LightGBM each also dominate ExtraTrees. Neither XGBoost nor stacking dominates the other. Non-dominance identifies a trade-off rather than equal preference: in the historical table, stacking exchanges recall, F1-score, and simplicity for small AUROC and AP gains.

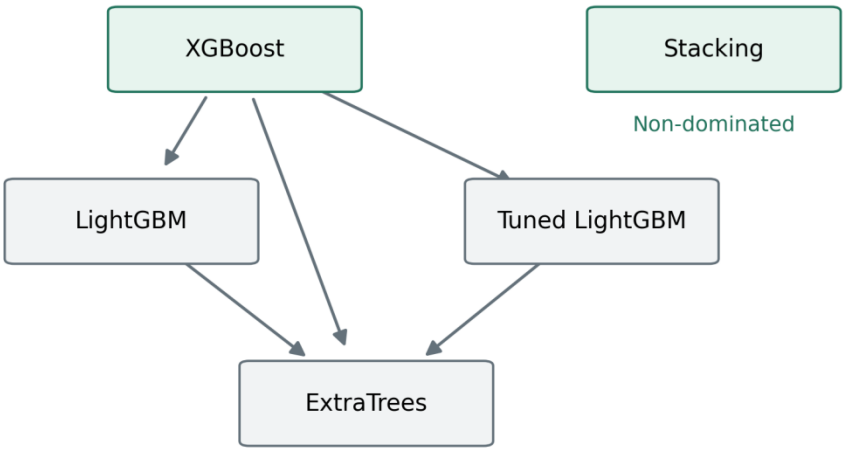


Figure 6. Complete Stage 1 dominance relation. Each arrow points from a dominating candidate to a dominated candidate. XGBoost and stacking form the non-dominated set; all five dominance edges are shown.

4.7. Edge-validation requirements

The analysis selects XGBoost as the more stable candidate for subsequent implementation studies; it does not establish edge readiness. **Table 6** specifies the minimum next-stage evidence. Acceptance thresholds should be set before testing according to the intended hardware, workflow, and clinical use. Mean latency or model file size alone would be insufficient.

Table 6. Proposed validation protocol for the selected candidate

Domain	Required measurement	Evidence needed
Conversion fidelity	Maximum/mean probability drift; label agreement	Same locked cases in source and converted runtimes
Performance	AUROC, AP, recall, F1 after conversion	No material degradation with paired uncertainty
Calibration	Plot, intercept, slope, Brier score	Locked and external cohorts
Latency	Cold start; p50, p95, p99; explanation latency	Named devices, runtime, threads, batch size
Resources	File size, peak memory, throughput, energy	Repeated device measurements
Reliability	Failures, missing inputs, offline behavior	Stress and interruption tests
Explainability	Global/local SHAP and stability	Clinical review on locked cases
Privacy/security	Data-flow map and threat-model tests	Inversion, extraction, membership, access/log review
Clinical utility	Decision curves and workflow simulation	Prespecified thresholds and interventions
Generalization	Temporal and external validation	Patient-grouped independent cohorts

4.8. Additional archived experiment

Table 7 reports Stage 2 CSV values separately from the historical Stage 1 inputs. The 32,000-encounter stack has the largest AUROC (0.703494) and AP (0.667963). LightGBM with inherited settings has the largest accuracy, recall, and F1-score. Its AUROC differs from the stack by 0.000042, computed from the full-precision CSV values. Such small point differences do not establish superiority without paired uncertainty analysis. **Figure 7** presents the archived discrimination curves. Similar rankings do not establish identical

cohorts or independent reproduction of the historical experiment.

Table 7. Stage 2 performance recorded in the archived experiment

Model	Accuracy	Precision	Recall	F1	AUROC	AP
LR	0.62625	0.61192	0.55113	0.57994	0.66829	0.63577
RF	0.64200	0.62929	0.57250	0.59955	0.68963	0.65379
ET	0.63238	0.61350	0.58024	0.59640	0.68234	0.64987
LGBM	0.64838	0.62231	0.63311	0.62766	0.70249	0.66658
XGB	0.64638	0.61958	0.63364	0.62653	0.70313	0.66737
Inherited LGBM	0.64925	0.62268	0.63632	0.62942	0.70345	0.66687
Stack	0.64763	0.64064	0.56315	0.59940	0.70349	0.66796

LR, logistic regression; RF, random forest; ET, ExtraTrees; LGBM, LightGBM; XGB, XGBoost. "Inherited LGBM" uses the historical tuned settings; no new Optuna search was run. Threshold-dependent metrics use the notebook's default model predictions. These values concern models fitted on 32,000 encounters, unlike the calibration refit in **Table 8**.

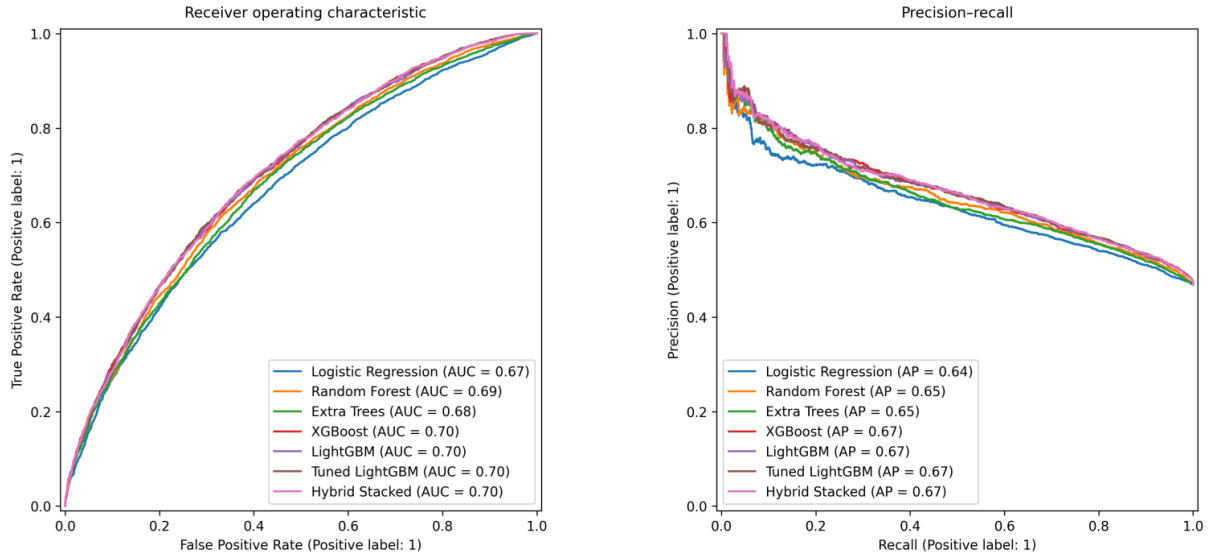


Figure 7. Archived Stage 2 receiver operating characteristic and precision–recall curves. "Tuned LightGBM" is the archive's label for inherited settings, not a new tuning experiment. The plotted curves are preserved from the archive.

For the uncalibrated stack fitted on 32,000 encounters, **Figure 8** contains 3,072 true negatives, 1,183 false positives, 1,636 false negatives, and 2,109 true positives. These counts reproduce recall 2,109/3,745 = 0.56315 and precision 2,109/3,292 = 0.64064. They do not describe the separately refitted and Platt-calibrated model used in **Table 9**.

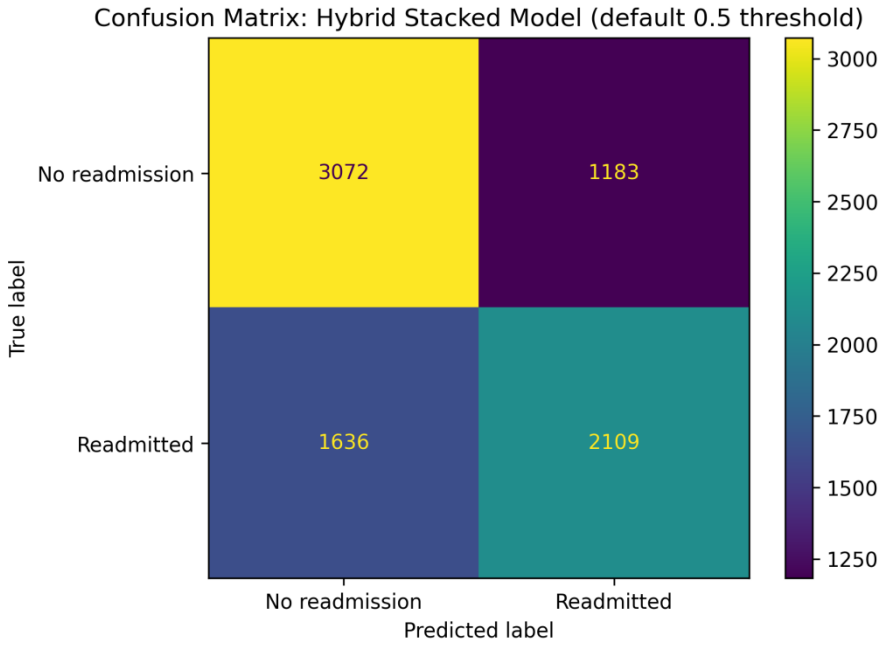


Figure 8. Archived confusion matrix for the uncalibrated 32,000-encounter stack at threshold 0.50.

4.9. Calibration results

For the stack refitted on 25,600 encounters, Platt scaling gives the smallest test-set Brier score and archived ECE (**Table 8**; **Figure 9**). Its Brier improvement over that same refit's raw probabilities is approximately 0.000014. Isotonic regression has larger values on both measures. The differences are descriptive: the test scores informed method selection, and no repeated-split or uncertainty analysis was available. The ECE implementation caveat in Section 3.9 also applies.

Table 8. Calibration results for the separately refitted Stage 2 stack

Method	Brier score	Archived ECE
Raw refitted stack	0.217859	0.011906
Platt scaling	0.217845	0.010220
Isotonic regression	0.218445	0.016716

The raw comparator is the stack refitted on 25,600 encounters; both mappings are fitted on 6,400 separate calibration encounters. All summary scores are evaluated on 8,000 test encounters. Archived ECE uses the implementation described in Section 3.9.

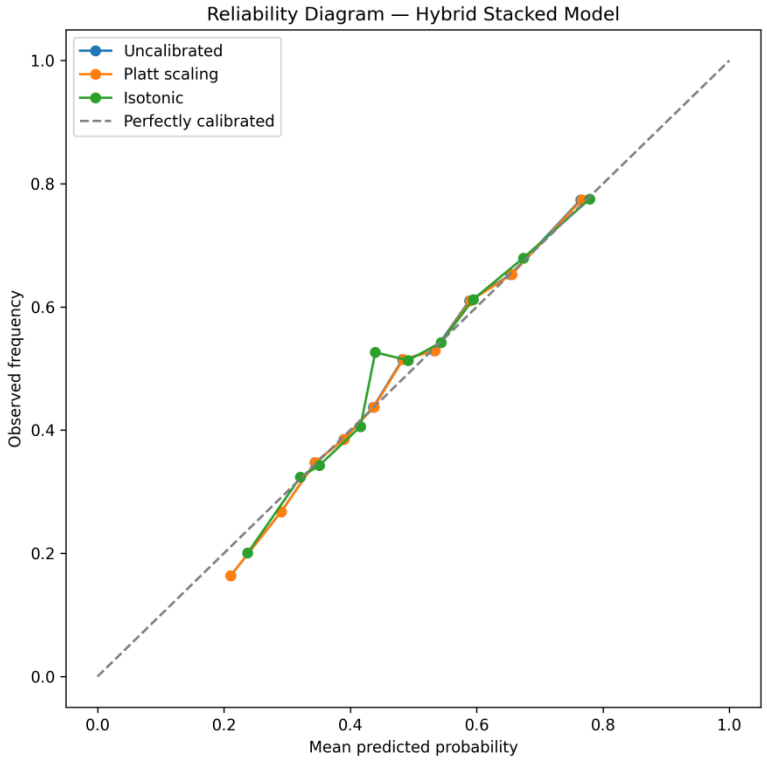


Figure 9. Archived reliability curves for the 25,600-encounter stack before and after calibration. Curves use ten quantile bins; the ECE values in Table 8 use equal-width bins.

4.10. Cost-sensitive threshold control

With the illustrative 5:1 false-negative-to-false-positive cost ratio, the lowest test-grid cost occurs at t = 0.20 (Figure 10). Relative to 0.50, recall rises from 0.56475 to 0.99519 and F1-score from 0.59754 to 0.64942, while precision falls from 0.63437 to 0.48196 (Table 9). Total empirical test-set cost decreases by 5,273,000 units (56.28%). This result is conditional on the cost ratio, sample, preprocessing, and test-informed selection procedure.

The CSV metrics and known class totals imply TP = 2,115, FP = 1,219, FN = 1,630, and TN = 3,036 at 0.50; the corresponding counts at 0.20 are 3,727, 4,006, 18, and 249. Thus, lowering the threshold increases flagged encounters from 3,334 (41.68%) to 7,733 (96.66%) and reduces specificity from 71.35% to 5.85%. Acting on all 8,000 encounters would cost 4,255,000 units under the same weights. The 0.20 setting saves 159,000 units (3.74%) relative to that comparator. **Figure 11** adds an exploratory decision-curve view, but does not provide an independent validation of the selected operating policy.

Table 9. Threshold metrics and workload for the Platt-calibrated refitted stack

Measure	Default 0.50	Test-grid minimum 0.20	Act on all
Precision	0.63437	0.48196	0.46813
Recall	0.56475	0.99519	1.00000
F1-score	0.59754	0.64942	0.63772
True positives	2,115	3,727	3,745
False positives	1,219	4,006	4,255

Measure	Default 0.50	Test-grid minimum 0.20	Act on all
False negatives	1,630	18	0
True negatives	3,036	249	0
Flagged encounters	3,334 (41.68%)	7,733 (96.66%)	8,000 (100%)
Specificity	71.35%	5.85%	0.00%
Total cost (units)	9,369,000	4,096,000	4,255,000

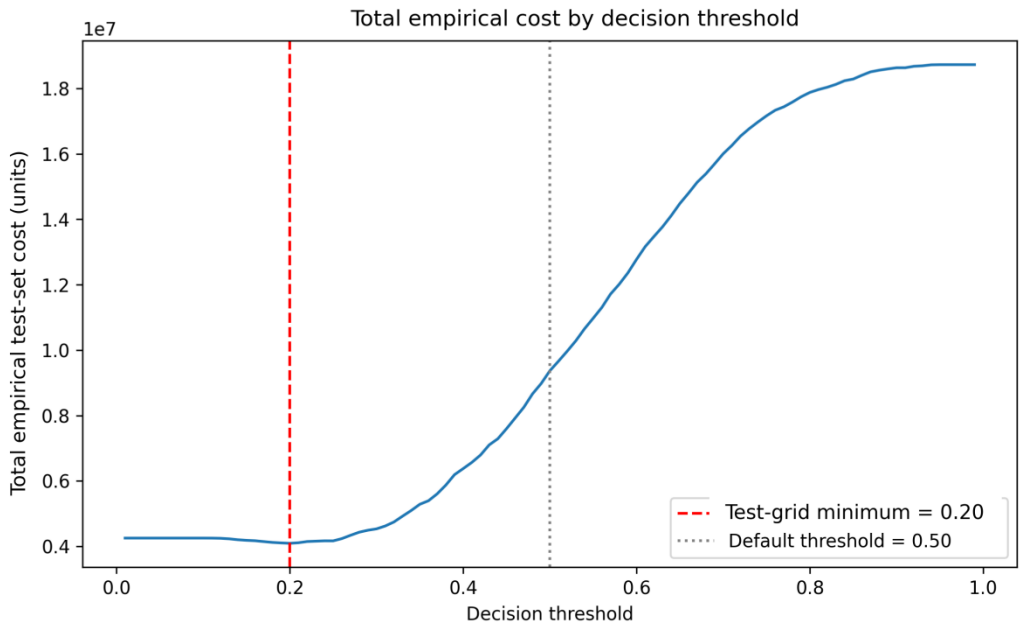


Figure 10. Archived total empirical test-set cost over thresholds 0.01–0.99, with false-negative and false-positive costs of 5,000 and 1,000 units. The marked 0.20 minimum was selected on these same test outcomes.

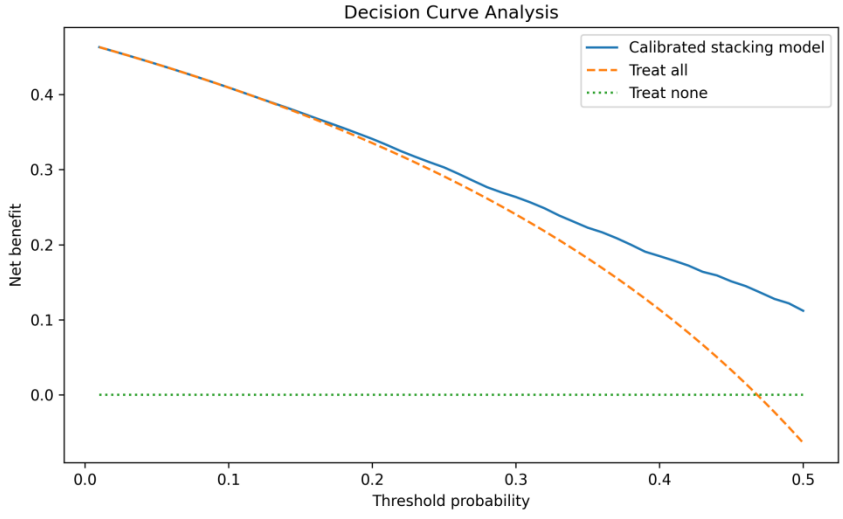


Figure 11. Archived exploratory decision curves for the Platt-calibrated refitted stack and the act-on-all and act-on-none reference strategies. Selection and evaluation share the test outcomes, so these curves do not establish independently validated clinical benefit.

4.11. Explainability evidence

The individual LightGBM analysis in **Figure 12** ranks clinical_risk_score and total_visits most prominently. Higher values generally shift that model's output upward. Discharge disposition, admission source, diagnosis count, medication variables, and diagnosis categories also contribute. These are predictive associations, not causal effects. The composite risk feature contains utilization components by construction (Appendix A), and correlated predictors can redistribute attribution. The figure does not explain the complete stack, the calibration mapping, or the thresholded decision. Full-pipeline explanation, feature-ablation analysis, stability assessment, and clinical review remain necessary.

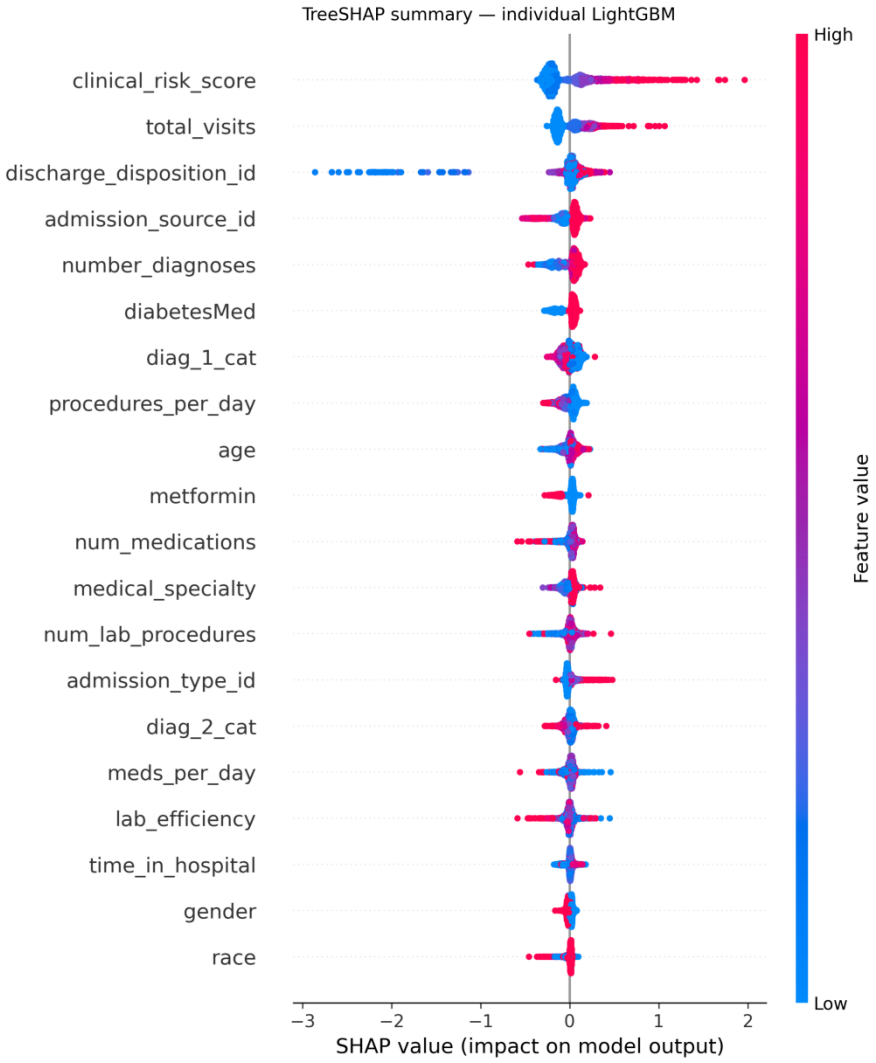


Figure 12. Archived TreeSHAP summary for the individual LightGBM baseline on 2,000 sampled test encounters. Color indicates the transformed feature value; horizontal position indicates attribution on the model's default output scale. This is not a whole-stack explanation or a calibrated probability contribution.

5. Discussion

5.1. Principal finding

In Stage 1, the stack's AUROC and AP advantages do not settle the selection question. XGBoost has greater

recall and F1-score at the reported operating points and uses one predictive component rather than four. It is preferred under equal-weight utility, the worst-objective regret diagnostic, and most sampled preference weights. These are conditional decision-analysis results; they are not independent evidence of clinical superiority.

The ε-constraint makes the decisive preference explicit. Requiring the largest reported AUROC retains stacking, whereas allowing an absolute AUROC tolerance of 0.0005 permits XGBoost and changes the lexicographic choice. Whether that shortfall is acceptable requires clinical judgment and uncertainty analysis. Because the inputs are rounded and lack paired predictions, the switching tolerance should not be treated as an exact population property.

5.2. Methodological contribution

Many model-comparison papers rank candidates by one metric or invoke "best performance" after considering several metrics informally. The present formulation converts that process into an auditable optimization problem. Exact enumeration is appropriate because the candidate set is small; a stochastic evolutionary algorithm would add variability without improving the search. NSGA-II, multi-objective particle swarm optimization, Bayesian optimization, or random search would become appropriate if model families, hyperparameters, thresholds, feature subsets, and compression settings were optimized jointly [27,28,38].

The framework makes compensatory trade-offs explicit. Pareto analysis identifies candidates dominated under the chosen objectives, utility analysis summarizes stated preferences, and the ε-constraint limits the permitted AUROC shortfall. Weight simulation shows how choices vary across a mathematical preference space. These tools improve transparency, but they cannot correct biased performance estimates, establish clinical values for the objectives, or prevent poor decisions when the inputs are unsuitable.

The simplicity objective is intentionally modest. Counting predictive components does not equal latency or energy, and XGBoost can contain many trees. However, it accurately records an architectural difference supported by the manuscript: a single estimator versus three base estimators plus a meta-estimator. Because the predictive-only sensitivity analysis still selected XGBoost in 98.69% of draws, the main conclusion is not an artifact of this proxy.

5.3. Clinical interpretation

At the Stage 1 operating points, XGBoost has precision 0.6170 and recall 0.6323, compared with 0.6377 and 0.5650 for stacking. The preferred balance depends on the proposed intervention and available capacity. Stage 2 illustrates the same issue more sharply: the threshold with high recall also flags nearly the entire test cohort. Alert burden, specificity, the act-on-all comparator, and a justified intervention model are therefore necessary alongside recall and cost summaries.

An operational threshold should be selected on development or validation data using explicit capacity, harm, and benefit assumptions, then locked before independent assessment. Probability calibration should be evaluated when risks inform action. Decision curves require a clinically defensible action and threshold range [16,17]. The present data and analyses do not establish such an intervention policy; prospective users would also need a specified prediction time, outcome horizon, eligible population, and failure-handling process.

The outcome here is any recorded readmission, not specifically 30-day readmission. A patient readmitted after 30 days is grouped with a patient readmitted within 30 days. This broader endpoint may support general

follow-up prioritization but is less aligned with interventions targeted specifically at the early transition after discharge. Future work should compare early, late, and competing outcomes under clinically motivated horizons.

5.4. Explainability, edge deployment, and privacy

XGBoost is compatible with exact tree-attribution methods, but compatibility is not completed explainability. The fitted model, preprocessing map, feature definitions, and representative cases are required. Global mean absolute SHAP values can summarize influence; dependence plots can reveal nonlinear patterns; and local waterfall views can support case review. Explanations should be tested across resamples, sites, and demographic groups. They must not be described as causal, especially where utilization variables reflect both patient health and care processes.

An edge workflow is plausible because tree inference can be implemented in compact runtimes, but the word "edge" describes location, not measured performance. Conversion may alter preprocessing, missing-value semantics, categorical handling, floating-point arithmetic, or tree traversal. A faithful study must compare source and converted probabilities on locked cases before benchmarking. It should report percentile latency rather than only an average and should include explanation generation, because an interface that promises explanations incurs that cost.

A future local implementation could reduce external transmission of raw inputs and dependence on connectivity. It would also require device authentication, secure storage and updates, rollback, restricted logging, and monitoring. Model and output attacks would remain possible [22–25]. Exposure reduction and privacy protection must therefore be measured separately. No implemented edge architecture or privacy guarantee is demonstrated in this study.

5.5. Comparison with prior literature

The observed AUROC values are modest, which is credible for a heterogeneous retrospective benchmark and cautions against deployment based on internal performance. Prior diabetes-readmission studies report variable results because they use different cohorts, outcomes, resampling methods, and leakage controls [4–8]. Direct numerical comparison is therefore less informative than comparison of study design. The present work adds value by exposing the decision structure and evidence gaps rather than presenting a larger metric as sufficient validation.

Recent optimization-oriented clinical machine-learning research increasingly combines feature selection, hyperparameter search and classification. Such studies demonstrate the relevance of metaheuristics and multicriteria methods but also risk optimistic bias if search and assessment share data. Random or Bayesian search should be nested inside training data, with the test set untouched until the pipeline is locked [38,39]. Sample-size planning for prediction models should consider outcome prevalence, predictor parameters, anticipated performance, and overfitting rather than apply a universal events-per-variable rule [40].

Transparent documentation must continue through the clinical lifecycle. FUTURE-AI organizes trustworthy deployment around fairness, universality, traceability, usability, robustness, and explainability [41], while DECIDE-AI specifies what should be reported during early-stage clinical evaluation of AI-supported decision systems [42]. Applying these principles would make the cohort-selection rule, patient grouping, missing-data strategy, human–AI interaction, failure handling, and deployment assumptions open to review.

5.6. Coupling optimization with threshold control

Model selection determines a scoring function; calibration changes the score-to-risk mapping; threshold selection determines which encounters trigger action. Stage 1 and Stage 2 illustrate these choices through different models and evidence records. Their results cannot be combined into a validated XGBoost-plus-threshold pipeline. In the archive, both calibrator choice and threshold choice use test outcomes, so the reported operating results are exploratory. A prospective evaluation should fit transformations and models within training data, reserve appropriate data for calibration and threshold selection, and assess the locked system once on patient-separated test or external data.

The illustrative cost ratio is not a health-economic estimate. A subsequent study should vary the false-negative-to-false-positive cost ratio and impose an explicit capacity limit, for example $\Sigma_i I(p_i \geq t) \leq K$ when K encounters can be reviewed. It should compare the resulting policies with acting on all or no encounters under the same costs. At the reported 0.20 threshold, the modest cost advantage over acting on all encounters must be weighed against flagging 96.66% of the cohort.

6. Limitations

The exact 40,000-encounter extract, inclusion and exclusion rules, and correspondence between the historical and archived cohorts remain unresolved. The Stage 2 input file was not supplied, so target-label mapping, patient overlap, subgroup structure, and availability of predictors at the intended prediction time cannot be independently audited. The encounter split does not enforce patient separation. Full-cohort preprocessing precedes splitting and exposes test and calibration encounters to data-dependent transformations; its effect on the performance estimates cannot be quantified from summary outputs.

Historical test-informed LightGBM tuning, Stage 2 selection of calibration by test-set Brier score, and threshold minimization on those same test outcomes prevent confirmatory interpretation. The 0.20 result concerns a separately refitted stack, not the Stage 1-selected XGBoost model or the Table 7 stack. Stage 2 specifies five-fold stacking, but the corresponding historical fold-level evidence is absent. Stage 1 lacks AP for logistic regression and random forest, narrowing its complete-objective candidate set. Rounded metrics and absent prediction vectors prevent paired inference, while unpinned dependencies and the ECE endpoint implementation further limit exact reconstruction.

The structural simplicity measure is not a device benchmark, and the Dirichlet frequencies are not estimates of clinical preferences. No temporal, external, or prospective evaluation is available; the source data reflect United States care from 1999–2008. Stage 2 SHAP outputs explain only an individual LightGBM model, with no retained attribution arrays or stability study. Stage 1 fitted artifacts, full-pipeline explanation, feature-ablation evidence, fairness assessment, device measurements, and security or privacy testing are absent. The study therefore supports an exploratory decision framework and a validation agenda, not clinical or edge-deployment readiness.

7. Conclusion

Explicit preferences are needed to choose both a prediction model and its operating threshold. In the historical Stage 1 comparison, XGBoost and stacking are non-dominated, and XGBoost is preferred under the stated compromise rules once an absolute AUROC tolerance of 0.0005 is allowed. The separate archived

experiment shows similarly small discrimination differences among boosting and stacking configurations. For its refitted stack, Platt scaling offers a marginal descriptive Brier improvement, and a test-selected threshold of 0.20 reduces illustrative cost relative to 0.50 while flagging 96.66% of encounters. Its cost advantage over acting on all encounters is 3.74%. The contribution is a transparent separation of model preference, probability calibration, and threshold choice. Cohort recovery, training-only preprocessing, patient-separated validation, independent assessment of the selected pipeline, uncertainty analysis, and clinically justified costs and capacity are required before stronger claims or clinical use.

Disclosure statement

The authors declare no conflict of interest.

Data and code availability

The original UCI dataset is publicly available under DOI:10.24432/C5230J [37]. The secondary analysis examined the archived experiment notebook, four summary CSV files, and associated figures. The exact 40,000-encounter input extract, cohort-selection script, split identifiers, fitted estimators, and individual prediction and attribution arrays were not included in the available archive. Full computational reproduction therefore requires recovery of those materials; no public repository for the study-specific archive is identified in the supplied record.

Appendix A. Predictor construction in the archived experiment

The following definitions transcribe the archived code; they are not a new feature-ablation experiment or a validated clinical score. I(condition) equals 1 when the condition holds and 0 otherwise. Visit counts refer to the year preceding the encounter in the source dataset; procedure, medication, laboratory, and length-of-stay totals refer to the index encounter. Availability at a specified clinical prediction time still requires verification.

Table A1. The 21 engineered variables and their implemented definitions

Variable	Implemented definition
total_visits	number_outpatient + number_emergency + number_inpatient
has_prior_visits	I(total_visits > 0)
inpatient_ratio	number_inpatient / (total_visits + 1)
outpatient_ratio	number_outpatient / (total_visits + 1)
emergency_ratio	number_emergency / (total_visits + 1)
high_medication	I(num_medications > 15)
meds_per_day	num_medications / (time_in_hospital + 1)
procedures_per_day	num_procedures / (time_in_hospital + 1)
high_procedures	I(num_procedures > 3)
high_lab_burden	I(num_lab_procedures > 50)
lab_efficiency	num_lab_procedures / (time_in_hospital + 1)
high_diag_burden	I(number_diagnoses > 10)

Variable	Implemented definition
clinical_risk_score	0.4 × number_inpatient + 0.3 × number_emergency + 0.2 × high_medication + 0.1 × high_diag_burden
high_risk_flag	I(clinical_risk_score > full-cohort median)
change_made	I(change = “Ch”)
insulin_up	I(insulin = “Up”)
insulin_steady	I(insulin = “Steady”)
metformin_yes	I(metformin = “Up”); a dose-increase indicator, not any metformin use
age_code	Decade category: [0,10) → 0, …, [90,100) → 9
elderly	I(age_code ≥ 7)
young	I(age_code ≤ 3)

Names such as lab_efficiency, clinical_risk_score, elderly, and young are code labels, not validated clinical constructs. The +1 denominators are retained exactly; these variables are not literal per-day rates. The composite coefficients and indicator cutoffs are fixed heuristic choices. The high_risk_flag cutoff is the full-cohort median, another data-dependent step performed before splitting.

Diagnosis coding and predictor inventory

The three diagnosis fields are replaced by diag_1_cat, diag_2_cat, and diag_3_cat. Codes beginning with V map to supplemental and those beginning with E to external. Numeric codes map to infectious (1–139), neoplasms (140–239), endocrine (240–279), blood (280–289), mental (290–319), nervous (320–359), circulatory (390–459), respiratory (460–519), digestive (520–579), genitourinary (580–629), skin (680–709), musculoskeletal (710–739), and injury (800–999); other numeric codes map to other. Missing or unparseable codes map to unknown. This is the implemented coarse grouping, not a complete clinical taxonomy. The saved output contains 118 unknown secondary diagnoses and 535 unknown tertiary diagnoses.

Fourteen fields are integer-encoded with LabelEncoder: race, gender, age, admission_type_id, discharge_disposition_id, admission_source_id, medical_specialty, metformin, insulin, change, diabetesMed, and the three diagnosis categories. The fitted category mappings are not included in the archive. Other unhandled missing numerical values are median-filled before all predictors are standardized.

The final 43 predictors comprise 22 retained or recoded source predictors plus the 21 variables in Table A1. The 22 are race, gender, age, admission_type_id, discharge_disposition_id, admission_source_id, time_in_hospital, medical_specialty, num_lab_procedures, num_procedures, num_medications, number_outpatient, number_emergency, number_inpatient, number_diagnoses, metformin, insulin, change, diabetesMed, diag_1_cat, diag_2_cat, and diag_3_cat. The target, readmitted, and patient_nbr are excluded from the predictor matrix; encounter_id is also dropped if present. The notebook’s saved input has no encounter_id column.

References

[1] International Diabetes Federation, 2025, IDF Diabetes Atlas (11th ed.). Brussels, Belgium: International Diabetes Federation.

[2] World Health Organization, 2016, Global Report on Diabetes. Geneva, Switzerland: World Health Organization.

[3] Strack B, DeShazo JP, Gennings C, et al., 2014, Impact of HbA1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records. BioMed Res Int, 2014: 781670. DOI:10.1155/2014/781670.

[4] Bhuvan MS, Kumar A, Zafar A, Kishore V, 2016, Identifying diabetic patients with high risk of readmission. arXiv:1602.04257. DOI:10.48550/arXiv.1602.04257.

[5] Rubin DJ, 2015, Hospital readmission of patients with diabetes. Curr Diab Rep, 15(4): 17. DOI:10.1007/s11892-015-0584-7.

[6] Song Y, Park J, Kim D, Kim C, 2024, Prediction of hospital readmission for type 2 diabetes patients using machine learning models. J Korean Data Inf Sci Soc, 35(1): 99-110. DOI:10.7465/jkdi.2024.35.1.99.

[7] Mishra V, Tanniru MR, Sreedharan J, 2025, Prediction of 30-day readmission in diabetes management using machine learning. Comput Biol Med, 195: 110616. DOI:10.1016/j.compbiomed.2025.110616.

[8] Artetxe A, Beristain A, Graña M, 2018, Predictive models for hospital readmission risk: a systematic review of methods. Comput Methods Programs Biomed, 164: 49-64. DOI:10.1016/j.cmpb.2018.06.006.

[9] Breiman L, 2001, Random forests. Mach Learn, 45: 5-32. DOI:10.1023/A:1010933404324.

[10] Friedman JH, 2001, Greedy function approximation: a gradient boosting machine. Ann Stat, 29(5): 1189-1232. DOI:10.1214/aos/1013203451.

[11] 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. ACM. DOI:10.1145/2939672.2939785.

[12] Ke G, Meng Q, Finley T, et al., 2017, LightGBM: a highly efficient gradient boosting decision tree. In Advances in Neural Information Processing Systems 30, pp. 3146-3154. Curran Associates.

[13] Wolpert DH, 1992, Stacked generalization. Neural Netw, 5(2): 241-259. DOI:10.1016/S0893-6080(05)80023-1.

[14] Fawcett T, 2006, An introduction to ROC analysis. Pattern Recognit Lett, 27(8): 861-874. DOI:10.1016/j.patrec.2005.10.010.

[15] Saito T, Rehmsmeier M, 2015, The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One, 10(3): e0118432. DOI:10.1371/journal.pone.0118432.

[16] Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW, 2019, Calibration: the Achilles heel of predictive analytics. BMC Med, 17: 230. DOI:10.1186/s12916-019-1466-7.

[17] Vickers AJ, Elkin EB, 2006, Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making, 26(6): 565-574. DOI:10.1177/0272989X06295361.

[18] Collins GS, Moons KGM, Dhiman P, et al., 2024, TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ, 385: e078378. DOI:10.1136/bmj-2023-078378.

[19] Moons KGM, Damen JAA, Kaul T, et al., 2025, PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. BMJ, 388: e082505. DOI:10.1136/bmj-2024-082505.

[20] Lundberg SM, Lee SI, 2017, A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems 30, pp. 4765-4774. Curran Associates.

[21] Lundberg SM, Erion G, Chen H, et al., 2020, From local explanations to global understanding with explainable AI for trees. Nat Mach Intell, 2: 56-67. DOI:10.1038/s42256-019-0138-9.

[22] Fredrikson M, Jha S, Ristenpart T, 2015, Model inversion attacks that exploit confidence information and basic countermeasures. In Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, pp. 1322-1333. ACM. DOI:10.1145/2810103.2813677.

[23] Tramèr F, Zhang F, Juels A, Reiter MK, Ristenpart T, 2016, Stealing machine learning models via prediction APIs.

In 25th USENIX Security Symposium, pp. 601-618. USENIX Association.

[24] Shokri R, Stronati M, Song C, Shmatikov V, 2017, Membership inference attacks against machine learning models. In 2017 IEEE Symposium on Security and Privacy, pp. 3-18. IEEE. DOI:10.1109/SP.2017.41.

[25] Dwork C, Roth A, 2014, The algorithmic foundations of differential privacy. Found Trends Theor Comput Sci, 9(3-4): 211-407. DOI:10.1561/0400000042.

[26] Miettinen K, 1999, Nonlinear Multiobjective Optimization. Kluwer Academic Publishers.

[27] Deb K, Pratap A, Agarwal S, Meyarivan T, 2002, A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE Trans Evol Comput, 6(2): 182-197. DOI:10.1109/4235.996017.

[28] Coello Coello CA, Pulido GT, Lechuga MS, 2004, Handling multiple objectives with particle swarm optimization. IEEE Trans Evol Comput, 8(3): 256-279. DOI:10.1109/TEVC.2004.826067.

[29] Ehrgott M, 2005, Multicriteria Optimization (2nd ed.). Springer.

[30] Marler RT, Arora JS, 2004, Survey of multi-objective optimization methods for engineering. Struct Multidiscip Optim, 26: 369-395. DOI:10.1007/s00158-003-0368-6.

[31] Chankong V, Haimes YY, 1983, Multiobjective Decision Making: Theory and Methodology. North-Holland.

[32] Belton V, Stewart TJ, 2002, Multiple Criteria Decision Analysis: An Integrated Approach. Kluwer Academic Publishers.

[33] Steyerberg EW, 2019, Clinical Prediction Models (2nd ed.). Springer. DOI:10.1007/978-3-030-16399-0.

[34] Zhou Z, Chen X, Li E, Zeng L, Luo K, Zhang J, 2019, Edge intelligence: paving the last mile of artificial intelligence with edge computing. Proc IEEE, 107(8): 1738-1762. DOI:10.1109/JPROC.2019.2918951.

[35] Shi W, Cao J, Zhang Q, Li Y, Xu L, 2016, Edge computing: vision and challenges. IEEE Internet Things J, 3(5): 637-646. DOI:10.1109/JIOT.2016.2579198.

[36] Jibril AU, 2026, An explainable hybrid AI framework for diabetes hospital readmission risk prediction: clinically informed feature engineering, ensemble learning and edge-readiness design. Doctoral thesis manuscript submitted to Skyline University Nigeria, September 2026. Chapter 4; Tables 4.8–4.11. Unpublished manuscript.

[37] Clore J, Cios K, DeShazo J, Strack B, 2014, Diabetes 130-US Hospitals for Years 1999–2008 [Dataset]. UCI Machine Learning Repository. DOI:10.24432/C5230J.

[38] Bergstra J, Bengio Y, 2012, Random search for hyper-parameter optimization. J Mach Learn Res, 13: 281-305.

[39] Snoek J, Larochelle H, Adams RP, 2012, Practical Bayesian optimization of machine learning algorithms. In Advances in Neural Information Processing Systems 25, pp. 2951-2959. Curran Associates.

[40] Riley RD, Snell KIE, Ensor J, et al., 2019, Minimum sample size for developing a multivariable prediction model: part II—binary and time-to-event outcomes. Stat Med, 38(7): 1276-1296. DOI:10.1002/sim.7992.

[41] Lekadir K, Frangi AF, Porras AR, et al., 2025, FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. BMJ, 388: e081554. DOI:10.1136/bmj-2024-081554.

[42] Vasey B, Nagendran M, Campbell B, et al., 2022, Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nat Med, 28(5): 924-933. DOI:10.1038/s41591-022-01772-9.

Publisher's note

Bio-Byword Scientific Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.