Interpretable, Calibrated Early Screening of Chronic Disease Risk with Feature Selection and Ensembles

[AUTHORS TBD][1*]
[1] [Affiliation TBD]
[CORRESPONDING EMAIL TBD]

Abstract

Early-screening models for chronic disease are increasingly accurate at ranking who is at risk, yet two properties that decide whether a clinician can act on them are routinely neglected: the predicted risk probability is rarely trustworthy in absolute terms, and the model exposes no defensible account of why a patient was flagged. We reframe screening from a discrimination-only exercise into a calibrated, interpretable risk-estimation problem and instantiate it as one leakage-safe pipeline that imputes physiologically-impossible values, over-samples the minority class in the training folds only, selects features by random-forest permutation importance, compares gradient-boosted trees, a random forest, a support-vector classifier, and a perceptron on all features versus the selected subset, and recalibrates the designated boosted model with isotonic and Platt scaling on a held-out split, auditing the Brier score and the expected calibration error on the untouched test split. Run on the Pima diabetes, Cleveland heart-disease, and imbalanced Kaggle stroke cohorts, the kernel classifier discriminated best on the two smaller cohorts, at an AUC of 0.826 and 0.873, while the tree ensembles led on stroke; selection preserved AUC within 0.02 while cutting up to sixteen features to one; and recalibration cut the stroke expected calibration error from 0.184 to 0.015. The stroke cohort is the imbalance lesson: an Accuracy near 0.950 hides an F1 near zero at a 0.049 prevalence, so the precision-recall curve and calibration, not Accuracy, are the metrics that survive. TreeSHAP on the uncalibrated boosted score recovered clinically plausible drivers that cross-checked the selection.

1 Introduction

Chronic diseases such as type-2 diabetes and coronary-artery disease are most treatable when they are caught early, and machine-learning models built on routinely collected tabular clinical variables are an attractive front line for population screening: the measurements are already taken in ordinary care, the models are cheap to run, and they can flag who should be referred for confirmatory testing. Public benchmark cohorts have made this concrete, with routine clinical features and a recorded outcome supporting tabular diabetes-risk screening on the Pima Indians cohort [Smith et al., 1988] and heart-disease-risk screening on the Cleveland cohort [Detrano et al., 1989]. On these and similar cohorts, learned classifiers predict diabetes onset [Naz and Ahuja, 2020] and coronary-artery disease [Mohan et al., 2019] from a handful of routine variables, which is what makes tabular machine learning a credible, low-cost complement to existing screening practice.

The applied literature has pushed these models toward higher discrimination, reporting strong area under the ROC curve for diabetes [Naz and Ahuja, 2020] and heart-disease [Mohan et al., 2019] risk, but three properties that decide whether a clinician can act on the output are routinely left unaddressed. First, the models are opaque: a single risk score or an aggregate AUC tells a physician nothing about which factors drove a patient's flag, so the output is hard to trust, audit, or turn into a conversation with the patient, and interpretability of such black-box models is exactly what a coherent

Preprint.

interpretable-machine-learning methodology argues is necessary for trust [Molnar, 2022]. Second, the typical pipeline feeds every available variable to the classifier, inviting collinearity and inflating the cost of the measurements a screen must collect, where a compact informative subset would be cheaper and more defensible [Guyon and Elisseeff, 2003]. Third, the real cohorts are messy in ways that quietly bias both accuracy and the apparent drivers: physiologically-impossible zero entries masquerade as data, and the at-risk class is often a small minority whose under-representation a synthetic over-sampling step is meant to correct [Chawla et al., 2002], yet naive handling distorts discrimination, calibration, and feature importance together.

A fourth gap is specific to screening and, we argue, decisive. A screening referral hinges on the absolute predicted probability, not on the ranking of patients: a model that says one risk when the true rate is far higher is dangerous even at high AUC. Yet modern high-accuracy classifiers are frequently miscalibrated, so their scores need post-hoc recalibration before a predicted probability can be trusted [Guo et al., 2017], and calibration has been called the Achilles heel of clinical predictive analytics precisely because a discriminative model can still systematically mis-state absolute risk [Van Calster et al., 2019]. The bottleneck for deployable screening is therefore no longer raw ranking power; it is whether the model emits a trustworthy probability, exposes a defensible rationale through additive feature attribution [Lundberg and Lee, 2017], and was trained without leaking the test split or the synthetic minority examples into either.

We reframe chronic-disease screening from an AUC-maximizing exercise on all features into a unified calibrated-and-interpretable risk-estimation problem, in which the goal is not merely to order patients but to emit a probability a clinician can read as a probability and a compact set of drivers a clinician can scrutinize. We instantiate this view as a single leakage-safe pipeline that treats impossible values as missing and imputes them, rebalances the minority class with synthetic over-sampling inside the training folds only, ranks features by random-forest permutation importance and retains a validation-chosen compact subset, compares gradient-boosted trees, a random forest, a support-vector classifier, and a compact multilayer perceptron on all features versus the selected subset, then recalibrates the designated gradient-boosted model with isotonic and Platt scaling on a held-out calibration split and explains its uncalibrated score with TreeSHAP, whose monotone calibration map leaves the driver ranking unchanged. We exercise the identical pipeline across three real public cohorts, the Pima diabetes, the Cleveland heart-disease, and the highly imbalanced Kaggle stroke-prediction cohort, reporting discrimination and calibration together, and we study honestly the accuracy-versus-parsimony and discrimination-versus-calibration trade-offs that this reframing makes visible.

This work makes three contributions.

- A unified interpretable, imbalance-aware, and explicitly calibrated screening pipeline that combines random-forest permutation feature selection, a cross-family model comparison, isotonic and Platt probability calibration, and TreeSHAP explanation into one leakage-safe procedure, in which selection determines what a model may use, the comparison tests whether parsimony costs accuracy, recalibration makes the predicted risk trustworthy, and the explanation audits the resulting drivers, exercised across several real clinical cohorts rather than one.

- An honest cross-cohort empirical comparison of gradient-boosted trees, a random forest, a support-vector classifier, and a compact multilayer perceptron on all features versus a random-forest-selected subset, measuring both the accuracy-versus-parsimony and the discrimination-versus-calibration trade-offs directly, with impossible-value handling, imputation, and SMOTE class-imbalance handling built into the training folds across the Pima diabetes, Cleveland heart-disease, and highly imbalanced Kaggle stroke-prediction cohorts.

- A TreeSHAP-based clinical driver analysis that explains the uncalibrated gradient-boosted score, reports a global importance ranking cross-checked against the random-forest permutation selection, and so turns the boosted model from an opaque score into a defensible, screening-ready account of which physiological and clinical factors govern risk, with the monotone calibration map preserving the ranking, all at second-scale CPU-only compute.

2 Related Work

Our pipeline sits at the meeting point of three lines of work, each of which supplies one ingredient of trustworthy screening but stops short of the whole. We organize the discussion by theme rather than chronology: machine learning for chronic-disease risk prediction, interpretable machine learning and feature selection, and probability calibration with class-imbalance handling.

2.1 Machine Learning for Chronic-Disease Risk Prediction

The tabular clinical cohorts used for chronic-disease benchmarking are drawn largely from public repositories, with the UCI repository the standard source of the diabetes, heart-disease, and kidney cohorts [Kelly et al., 2023], including a chronic kidney disease dataset whose mixed-type features and heavy missingness make it a screening cohort that stresses imputation [Rubini et al., 2015]. On the diabetes side, systematic comparisons benchmark several classifier families on the Pima cohort with feature-importance analysis [Chang et al., 2023], and learned models are weighed against traditional statistical methods for detecting undiagnosed diabetes [Choi et al., 2023]. On the cardiovascular side, multiple algorithms are compared for coronary-artery-disease prediction on the Cleveland cohort with a reproducibility focus [Akella and Akella, 2021], risk models are developed and validated on real clinical cohort data [Wang et al., 2021], and surveys catalogue the methods, datasets, and features that recur across this literature [Alizadehsani et al., 2019]. These studies report strong discrimination, but they share a habit that limits clinical use: each typically optimizes a single cohort for AUC, with all features and one model family, and reports no reliability of the probabilities. The evaluation culture that would correct this exists, since prediction-model performance is argued to require discrimination, calibration, and clinical usefulness together rather than discrimination alone [Steyerberg et al., 2010], yet it is rarely applied in these applied comparisons, which is the gap our cross-cohort, calibration-aware protocol is built to close.

2.2 Interpretable Machine Learning and Feature Selection

Making a screening model auditable rests on attributing a prediction to its inputs. Additive feature attribution gives a unified, model-agnostic account of an individual prediction, and its tree-specialized form computes exact attributions efficiently for ensembles, supporting global driver rankings and dependence views in clinical models [Lundberg et al., 2020], while local surrogate explanation offers a contrasting route that fits an interpretable model around each prediction of any classifier [Ribeiro et al., 2016]. Applied to clinical risk, attribution-based ranking has identified the leading predictors of type-2 diabetes incidence in a large cohort [Lugner et al., 2024], surfaced the drivers of cardiovascular and heart-failure risk [Gan et al., 2025], and exposed the drivers of chronic kidney disease in explainable models that pair attribution with surrogate explanation [Ghosh and Khandoker, 2024]. The same machinery has been used for stroke-risk prediction, an interpretable model for a highly imbalanced screening task [Vu et al., 2024], and diabetes pipelines have combined over-sampling for class imbalance with attribution and surrogate explanations into an interpretable, imbalance-aware model [Tasin et al., 2023]. These works show that explanation is mature, but they typically demonstrate it on one cohort and rarely wire it into a complete screening pipeline that also selects features and confronts the impossible-value and calibration problems, which is the integration our pipeline contributes.

2.3 Probability Calibration and Class-Imbalance Handling

The third ingredient turns a score into a probability a clinician can act on. Classifier scores can be transformed into accurate probability estimates by post-hoc calibration such as isotonic regression [Zadrozny and Elkan, 2002], and supervised learners differ in how well-calibrated their scores are, with Platt scaling and isotonic regression recalibrating them to good probabilities [Niculescu-Mizil and Caruana, 2005]. A calibration hierarchy formalizes mean, weak, moderate, and strong calibration as distinct levels of probability reliability for a risk model [Van Calster et al., 2016], giving the vocabulary in which a screening model's probabilities should be judged. The imbalance correction that screening cohorts invite interacts with this directly: synthetic over-sampling can damage the calibration of a clinical risk model, a caution that motivates leakage-safe, calibration-aware handling [van den Goorbergh et al., 2022], even as cost-sensitive learning and resampling are compared for handling imbalance in medical data [Mienye and Sun, 2021] and over-sampling combined with ensembles predicts diabetes incidence in a large imbalanced cohort [Alghamdi et al.,

2017]. These works establish that calibration matters and that imbalance handling can undermine it, but the reliability of the probability is rarely measured side by side with the discrimination and the explanation it is meant to make trustworthy, which is precisely the joint measurement we report.

The single nearest line of work is the interpretable, imbalance-aware clinical pipeline that pairs over-sampling with attribution on one cohort [Tasin et al., 2023]: it selects, resamples, and explains, but it optimizes and reports discrimination and leaves the reliability of the probability unmeasured, so its explanation is attached to a risk whose absolute value is unaudited. We resolve all three gaps at once by treating discrimination, calibration, and explanation as one leakage-safe procedure, reporting reliability beside discrimination as a matter of evaluation hygiene [Steyerberg et al., 2010], and exercising the identical selection, comparison, calibration, and explanation across several real cohorts rather than one. The next section formalizes that pipeline.

3 Method

We present the unified interpretable, calibrated, imbalance-aware screening pipeline, a single leakage-safe procedure that prepares a clinical cohort, selects a compact feature subset, compares several classifier families, recalibrates the designated gradient-boosted model's probabilities, and explains its drivers. The same procedure runs unchanged across the cohorts: feature selection decides what a classifier may read, the cross-family comparison tests whether committing to a small subset costs discrimination, probability calibration makes the predicted risk usable as an absolute number rather than a ranking, and TreeSHAP audits which retained factors govern that risk. Figure 1 shows the overall data flow. The guiding principle is that a deployable screening model must emit a probability a clinician can act on and a rationale a clinician can scrutinize, so the design prefers a smaller, calibrated, auditable model over a sharper opaque one.

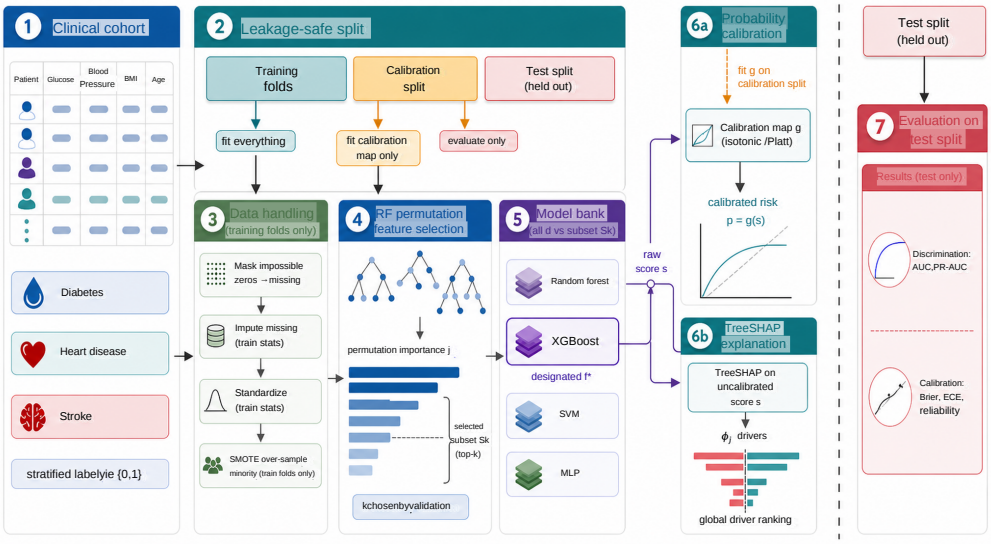

Figure 1: Overview of the unified screening pipeline: leakage-safe splitting, impossible-value handling, imputation, and training-fold-only SMOTE, random-forest permutation feature selection, the compared model families (RF, XGBoost, SVM, MLP) on all features versus the selected subset, isotonic and Platt calibration of the gradient-boosted model fitted on a held-out calibration split, and TreeSHAP explanation of the uncalibrated gradient-boosted score, with discrimination and calibration both evaluated on the untouched test split, applied identically across the clinical cohorts.

3.1 Problem Formulation

We cast early screening across the clinical cohorts as one supervised binary-classification setup with several instances. Let X be a feature matrix whose rows are patients and whose d columns are routinely collected clinical and physiological measurements, and let x_i denote the feature vector of the i-th patient. The target $y_i \in \{0, 1\}$ is a binary disease-risk label, at-risk versus not, obtained from

the cohort's recorded outcome: diabetes onset, coronary-artery disease present, or the cohort-specific screening label. A classifier f maps a feature vector to a raw score s for the positive class, which the calibration stage maps to a predicted risk probability $\hat{p}$. Casting every cohort in one notation lets us run the identical preparation, selection, comparison, calibration, and explanation procedure on each, and read the accuracy-versus-parsimony and discrimination-versus-calibration trade-offs the same way across cohorts that differ in size, dimension, and class balance. Crucially, the screening objective is not to order patients but to emit a probability a clinician can read as a probability, which is what makes calibration a first-class stage rather than an afterthought. The procedure is leakage-safe by construction and rests on a three-way role split: the training folds fit the classifier f together with imputation, over-sampling, and feature ranking; a held-out calibration split, carved from the training data and never from the test, fits only the calibration map g; and a stratified test split, touched by no fitting step, is where both discrimination and calibration are finally reported. Every reported number is read on the untouched test split.

3.2 Notation

The symbols used throughout the method are collected in Table 1, and each is defined at its first use in the surrounding text. The notation is shared across cohorts so the single pipeline reads identically on each.

Table 1: Notation.

Symbol	Meaning
X	feature matrix of clinical and physiological measurements
x_i	feature vector of the i-th patient
y_i	binary disease-risk label in $\{0, 1\}$ (at-risk vs not)
d	number of candidate input features
k	number of features retained after permutation selection
I_j	random-forest permutation importance of feature j
S_k	the selected top-k feature subset
f	the trained classifier (RF, XGBoost, SVM, or MLP)
s	raw (uncalibrated) model score for the positive class
$\hat{p}$	calibrated predicted risk probability
g	the fitted calibration map (isotonic or Platt)
BS	Brier score, mean squared error of $\hat{p}$ against the outcome
ECE	expected calibration error, bin-weighted gap between accuracy and mean $\hat{p}$
ϕ_j	SHAP value (additive attribution) of feature j
ϕ_0	SHAP base value (expected model output)

3.3 Data Handling: Impossible Values, Imputation, and Imbalance

Real clinical cohorts arrive with corrupt entries and skewed labels, so the pipeline treats cleaning and rebalancing as explicit steps rather than silent preprocessing. Several routine variables admit physiologically-impossible values that a naive reader would mistake for data: a recorded plasma glucose, blood pressure, skin-fold thickness, serum insulin, or body-mass index of zero cannot occur in a living patient, so we treat such entries as missing rather than as measurements. The resulting gaps, together with genuine missingness, are filled by missing-value imputation estimated from the observed distribution of each feature on the training data alone, after which the kernel and neural models receive standardized inputs whose statistics are likewise fit on training data only. The at-risk class is a minority in every cohort, severely so in the highly imbalanced screening cohort, so we apply SMOTE, which synthesizes minority-class examples by interpolating between a patient and its nearest like-labelled neighbours [Chawla et al., 2002]. Over-sampling is applied to the training folds only, never to the calibration or test data, so the reported discrimination reflects the original class prior and the reliability estimate is read on untouched data.

The design rationale is that the obvious shortcuts each bias the result, and the bias falls precisely on the quantities we care about. Deleting every patient with an impossible or missing entry would discard a large fraction of small cohorts and shift the class balance, while leaving the impossible zeros in place lets them masquerade as low-but-real values and distorts both the learned boundary and the

feature importances. Leaving the imbalance untouched lets a classifier reach a high apparent accuracy by predicting the majority label, ignoring exactly the at-risk patients a screen exists to catch. Imputing on the training distribution and rebalancing inside the folds keeps the held-out data untouched and the evaluation honest, at the cost of introducing synthetic minority examples whose artifacts we revisit in the failure analysis. The order of these steps is fixed deliberately: impossible-value masking and imputation precede selection and over-sampling, so the random forest ranks complete feature vectors rather than ranking around gaps, and over-sampling is the last step before training, applied only to the training partition of each fold so the synthetic examples never reach validation, calibration, or test. Because synthetic over-sampling can damage the calibration of a clinical risk model, confining it to the training folds and auditing reliability afterward is the mechanism that keeps the later probability trustworthy.

3.4 Random-Forest Permutation Feature Selection

The first modelling stage ranks the d candidate features and keeps a compact subset S_k. A random forest is an ensemble of decision trees grown on bootstrap resamples of the training data, with each split chosen from a random subset of features, so the aggregated prediction has lower variance than any single tree [Breiman, 1996, 2001]. The forest yields an importance score I_j for each feature j. We use permutation importance, the drop in validation score on a held-in split of the training data when feature j's values are permuted, in preference to impurity importance, the total reduction in node impurity attributable to splits on that feature, which is biased toward high-cardinality and correlated variables. We sort the features by the permutation I_j and retain the top-k subset S_k, where k is chosen by validation on the training data alone. The downstream classifiers are then trained twice, once on all d features and once restricted to S_k, so the effect of the reduction can be read directly. Figure 2 illustrates the ranking and the retained subset.

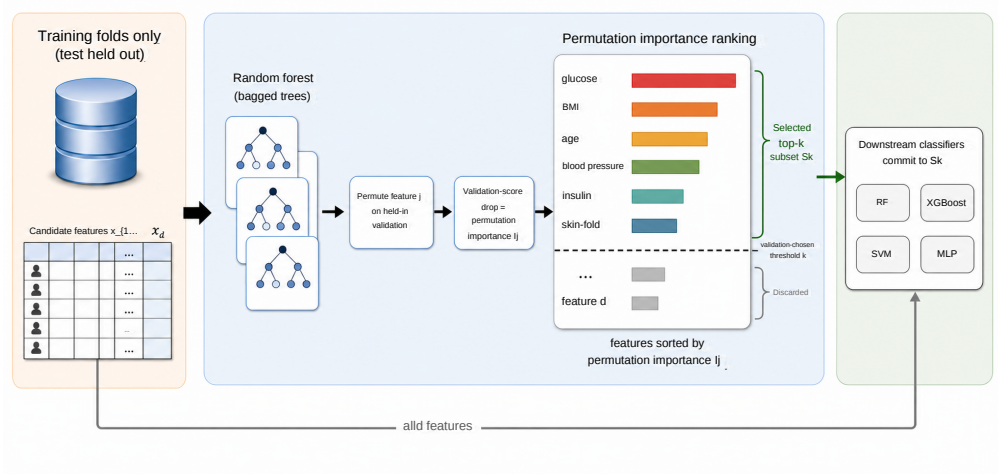

Figure 2: Random-forest permutation importance ranks the candidate clinical features and a validation-chosen threshold retains a compact top-k subset, forcing the downstream classifiers to commit to the variables that carry signal rather than consuming every available measurement.

The design rationale is that selection should force the model to commit to the variables that carry signal rather than consume every measurement. The obvious alternative is to feed all d features to every classifier and let regularization discount the weak ones, but that inflates the cost of the variables a screen must collect, leaves collinear channels to share credit, and spreads the model's rationale thinly across many inputs, which is the opposite of what an auditable screen needs. A wrapper search that retrains a classifier for every candidate subset might find a better subset, yet its cost grows with the number of subsets and couples the selection to one model family. Ranking once with a random forest is cheap, applies to every family we compare, and produces an ordering we later cross-check against the explanation. We use permutation importance because impurity-based importance favours high-cardinality and correlated features, a bias the analysis returns to. The choice of k trades coverage against parsimony: too small a k depresses discrimination, while a k near d defeats the purpose, so we choose the smallest k whose cross-validated discrimination on the training partition lies within a

small tolerance of the all-features score, never tuning k on the test set. We would like the ranking to be stable across resamples so the later explanation can cross-check a single fixed S_k rather than one that changes with the seed; small cohorts work against this, since a few hundred rows such as the Cleveland heart cohort yield high-variance importance estimates and a less stable selected k, an instability the repeated-cross-validation mean and standard deviation only partially capture. We therefore treat subset stability as a claim to be confirmed by the measured tables, not asserted.

3.5 Compared Model Families

The second stage compares four classifier families on each feature set. The random forest described above is the first, used both as the importance ranker and as a standalone classifier. The second is a gradient-boosted ensemble, which fits an additive sequence of shallow trees by greedy stagewise function approximation, each new tree fitted to the gradient of the loss left by the current ensemble [Friedman, 2001]; we use the regularized, scalable XGBoost realization of this idea [Chen and Guestrin, 2016], and it carries a second role as the model the explanation and calibration stages target. The third is a support-vector classifier on standardized features, which finds a maximum-margin decision boundary through a kernel [Cortes and Vapnik, 1995], the margin idea that also underlies support-vector regression for real-valued targets [Drucker et al., 1996]. The fourth is a compact multilayer perceptron, a small feed-forward network with one or two hidden layers and a nonlinear activation, included as a neural baseline. Each family is trained on all d features and again on the selected subset S_k, giving the per-cohort model settings the result tables compare.

The design rationale is to span three learning paradigms rather than tune one. Tree ensembles capture nonlinear, threshold-style interactions among clinical variables without feature scaling; the support-vector classifier tests whether a margin-based kernel method on standardized inputs is competitive; and the perceptron probes whether a small neural model adds anything on tabular data of this size. Restricting the comparison to a single family would leave open whether any observed parsimony or calibration effect is a property of the data or of one model class. By holding the preprocessing, the splits, and the selected subset fixed across all four, the only quantity that varies across rows of the discrimination table is the classifier, so the comparison isolates the model. The gradient-boosted ensemble is the pipeline's designated explanation and calibration target, fixed in advance rather than picked as the best model per cohort, because TreeSHAP is exact and efficient for tree ensembles, so explaining it costs nothing in interpretability machinery, and a boosted model's raw scores are a representative case for the recalibration we study. The other three families remain in the comparison only to gauge discrimination; the gradient-boosted model is the one we calibrate and explain throughout. We keep the perceptron compact, since a large network on tabular data of this size would overfit, and keep the support-vector classifier on standardized inputs because its margin geometry is scale-sensitive in a way the trees are not. Each family is evaluated at conservative, comparable settings rather than tuned to its ceiling, the fair way to read whether parsimony, calibration, or model class drives any difference.

3.6 Probability Calibration

The third stage makes the gradient-boosted model's output usable as a risk. A classifier emits a raw score s for the positive class, but s need not be a calibrated probability: a model can rank patients well yet systematically over- or under-state absolute risk, and modern high-accuracy classifiers are frequently miscalibrated in exactly this way. Because a screening referral hinges on the probability itself, we recalibrate s to a predicted risk $\hat{p} = g(s)$ through a fitted map g. We compare two standard recalibrators. Platt scaling fits a sigmoid to the raw scores, mapping s to a probability through a learned slope and intercept [Platt, 1999]. Isotonic regression fits a free monotone, piecewise-constant map, more flexible than the sigmoid but more data-hungry. Both are fit on the held-out calibration split alone, never on the data used to train f and never on the test split, so g does not memorize the same patients the classifier did. The reliability curve and its summaries are then read on the untouched test split, in-sample for neither f nor g. We audit reliability before and after recalibration with a reliability curve, which plots empirical against predicted risk, and summarize it with two complementary calibration metrics. The first is the Brier score

$$\mathrm{BS} = \frac{1}{n} \sum_{i=1}^{n} \left(\hat{p}_i - y_i\right)^2 , \tag{1}$$

the mean squared error between the predicted probability $\hat{p}_i$ and the outcome y_i, for which lower is better. The Brier score, however, mixes calibration with refinement, so we report alongside it the expected calibration error

$$\mathrm{ECE} = \sum_b \frac{n_b}{n} \left| \mathrm{acc}_b - \overline{p}_b \right|, \tag{2}$$

the mean over equal-width probability bins b of the gap between the empirical accuracy acc_b and the mean predicted probability $\overline{p}_b$ in the bin, weighted by the bin count n_b, a calibration-specific metric that isolates the risk mis-statement Brier conflates with sharpness. Both are read on the test split. Figure 3 shows the measured effect, where recalibration moves the stroke reliability curve back toward the diagonal. The screening decision applies a one-half operating point to the calibrated probability $\hat{p}$, a natural cut once probabilities are calibrated, fixed in advance without consulting the test set; the threshold-dependent Precision, Recall, and F1 are reported at this point, and for the highly imbalanced cohort we additionally report the area under the precision-recall curve, more informative than F1 at low prevalence.

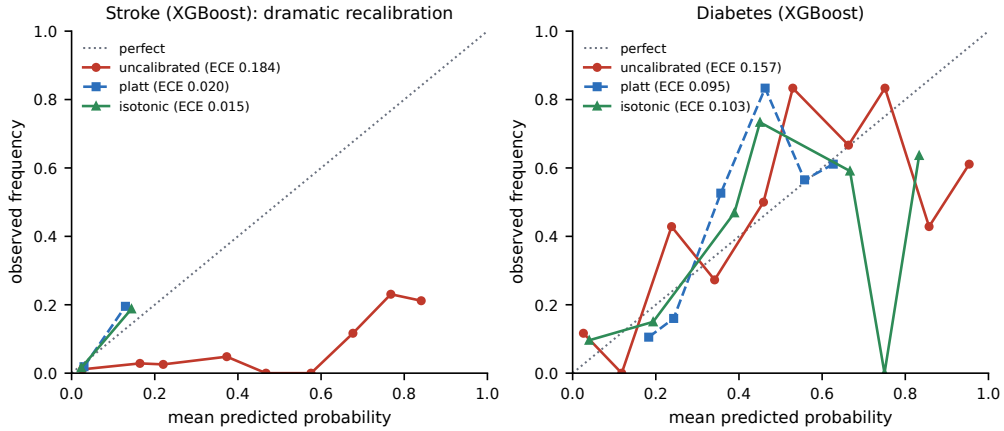

Figure 3: Measured reliability curves for the designated gradient-boosted model. On the rare-positive stroke cohort (left) the raw scores depart the diagonal and isotonic and Platt recalibration map them back onto calibrated probabilities, cutting the expected calibration error from 0.184 to 0.015 without changing the ranking; the diabetes panel (right) shows the milder correction Platt applies there.

The design rationale is that for screening the absolute probability must be trustworthy, not merely the ranking, so calibration is reported beside discrimination. Trusting the raw score is unsafe precisely when discrimination is high: a sharp model with mis-stated probabilities refers the wrong fraction of patients at any fixed threshold. Fitting g on a held-out calibration split avoids an optimistic map that looks calibrated in sample, and reading the reliability metrics on the still-separate test split avoids the symmetric optimism of scoring g on the data it was fit to. We compare isotonic and Platt because they trade flexibility against data efficiency differently; reporting both, with Brier, ECE, and the reliability curve, lets the trade-off be read rather than assumed. Because recalibration is a monotone transform, it leaves the ranking and the AUC unchanged, so a drop in the calibration-specific ECE at fixed AUC isolates a pure calibration gain, which is why we report discrimination and calibration as separate axes and lean on this AUC-invariance rather than on the Brier score alone.

3.7 TreeSHAP Explanation, Protocol, and Pseudocode

The fourth stage explains the gradient-boosted model with additive feature attributions, so each prediction comes with a defensible account of its drivers. TreeSHAP explains the uncalibrated boosted score $s = g^{-1}(\hat{p})$ rather than the calibrated probability, since g is applied after the trees and TreeSHAP attributes the tree-ensemble output. This costs nothing in consistency: because g is monotone it preserves the score order and so leaves the SHAP driver ranking unchanged, keeping the explanation of s and the calibrated $\hat{p}$ consistent. SHAP assigns to each feature j a value ϕ_j such that the attributions and a base value ϕ_0 reconstruct the model output for a patient,

$$g^{-1}(\hat{p}) = \phi_0 + \sum_{j \in S_k} \phi_j, \tag{3}$$

where ϕ_0 is the expected model output, each ϕ_j is the feature's marginal contribution averaged over the orderings in which features enter, and the sum runs over the k selected features the explained model reads [Lundberg and Lee, 2017]. Computing these attributions exactly is expensive for a general model, but TreeSHAP gives exact additive attributions for tree ensembles in low-order polynomial time by exploiting the tree structure [Lundberg et al., 2020], which is why the gradient-boosted model is the explanation target. We run TreeSHAP per cohort and aggregate the per-patient attributions into a global ranking, the mean absolute attribution of each feature, and cross-check this ranking against the permutation selection S_k. The end-to-end procedure is summarized in Algorithm 1: for each cohort we hold out the stratified test split first, reserve the calibration split, fix a random seed, run preparation, selection, comparison, calibration, and explanation, and report metrics as a mean and standard deviation over repeated stratified k-fold cross-validation. The dominant cost is training the forest and the boosted ensemble, fitting the recalibrators, and running TreeSHAP, all completing in seconds on a single CPU for cohorts of this size.

Algorithm 1 Unified interpretable, calibrated, imbalance-aware screening pipeline

Require: features X, labels y, model families $\mathcal{F}$
Ensure: metrics per model, calibrated risk $\hat{p}$, global SHAP ranking
hold out stratified test split; reserve calibration split ▷ all steps below: training folds only
mask impossible values; impute missing entries; standardize on training data
fit a random forest; read permutation importance I_j on a held-in validation split
$k \leftarrow$ smallest size whose CV discrimination is within tolerance of all features
$S_k \leftarrow$ indices of the k largest permutation I_j ▷ selected subset
for each model family $f \in \mathcal{F}$ **do**
    apply SMOTE to the training folds; train f on all d features and on S_k
    evaluate f by repeated stratified k-fold CV (mean $\pm$ SD)
end for
$f^\star \leftarrow$ gradient-boosted (XGBoost) model ▷ designated calibration and SHAP target
fit g on the calibration split by isotonic regression and by Platt scaling; set $\hat{p} \leftarrow g(s)$
on the test split: report discrimination (incl. PR-AUC) at a one-half threshold on $\hat{p}$, and reliability, Brier, and ECE before and after g
$\{\phi_j\}_{j \in S_k} \leftarrow \mathrm{TreeSHAP}(f^\star)$ on the uncalibrated score s; aggregate to a global ranking; cross-check against S_k
return metrics, calibrated $\hat{p}$, global SHAP ranking

The same loop runs unchanged for every cohort, with only the class balance and dimension differing, so the procedure is one pipeline rather than several. Because every stage operates on tabular inputs of modest dimension, the per-stage cost is bounded and the whole pipeline can be rerun end to end whenever the cohorts are updated, cheap enough to audit repeatedly, which is the property a deployable screening model needs.

4 Experiments

4.1 Implementation Details

The pipeline is CPU-only and ran end to end in well under two minutes for all three cohorts combined. The random forest, the support-vector classifier, the multilayer perceptron, the calibration maps, and the evaluation metrics use the standard scikit-learn implementations [Pedregosa et al., 2011], the gradient-boosted ensemble uses XGBoost, and the explanation stage uses the TreeSHAP explainer. Features are standardized before the support-vector and neural models, with the standardization statistics computed on the training data alone, impossible-value masking, missing-value imputation, and synthetic over-sampling were refit inside each cross-validation fold on its training partition, and the random-forest permutation ranking was fit once on the training partition, with the stratified test split held out beforehand and a separate calibration split reserved, so that no test or calibration sample influenced the preprocessing or the selection. We report each metric as a mean and standard deviation over repeated stratified k-fold cross-validation with a fixed random seed rather than a single held-out split, and the result tables list mean and SD per cell. The forest grows several hundred trees, the boosted ensemble uses a few hundred shallow trees, and the perceptron has one or two small hidden layers, all left at conservative defaults rather than heavily tuned, since the aim is a fair cross-family

comparison rather than a single best number. Every family was trained and evaluated under this identical protocol, so the only quantity that varies across the rows of a result table is the model and the only quantity that varies between the columns of the feature-selection comparison is the feature set. We fixed the random seed once and reused it for the test split, the calibration split, the cross-validation folds, the forest, and the over-sampling, so that the reported run is reproducible from the seed alone. The roles of the three splits were kept strictly separate: the training folds fit the classifier and all preprocessing, the held-out calibration split fits only the calibration map, and the untouched test split is where both the discrimination and the calibration metrics (the Brier score, the expected calibration error, and the reliability curve) were reported, so the reliability estimate was never read on data the recalibrator was fit on. The full pipeline, from cleaning through selection, the per-cohort model fits, the recalibration, and the TreeSHAP pass, completed in seconds on a single CPU for cohorts of this size, which removes any dependence on specialized hardware and makes the procedure cheap to rerun whenever a cohort is updated.

4.2 Experimental Design

We evaluated on three public clinical cohorts chosen to span size, dimension, and class balance. The Pima Indians Diabetes cohort provides routine clinical features with a binary diabetes-onset label and a moderate minority class, with physiologically-impossible zero entries for glucose, blood pressure, skin-fold, insulin, and body-mass index treated as missing and imputed [Smith et al., 1988]; it has 768 patients at a 0.349 positive rate. The Cleveland Heart Disease cohort provides clinical features with the multi-valued severity target binarized to coronary-artery disease present versus absent [Detrano et al., 1989]; it is the small cohort, 303 patients at a near-balanced 0.459 positive rate. The third cohort is the Kaggle Stroke Prediction dataset [fedesoriano, 2021], a large and highly imbalanced screening cohort of 5110 patients at a 0.049 positive rate, which carries a severe minority class and stresses the imbalance-and-calibration angle the pipeline is built around. Each cohort was split three ways, into training, calibration, and test partitions, with the splits stratified on the label. Discrimination is reported with Accuracy, Precision, Recall, F1, and AUC, all higher-is-better, where the threshold-dependent Precision, Recall, and F1 are read at a 0.5 operating point on the calibrated probability, a natural cut fixed in advance without consulting the test set, and AUC is a threshold-free measure of ranking; for every cohort we additionally report the area under the precision-recall curve (PR-AUC), which is more informative than F1 at low prevalence and which we treat as the headline ranking metric on the imbalanced stroke cohort. Calibration is reported on the test split with the Brier score, the mean squared error between the predicted risk probability and the outcome for which lower is better, and the expected calibration error (ECE), the bin-count-weighted mean gap between empirical accuracy and mean predicted probability over ten bins, read together with reliability curves [Brier, 1950]. The model bank is fixed to the four families spanning tree ensembles, a kernel method, and a neural baseline, and two comparisons are defined: a feature-selection comparison that trains the gradient-boosted model on all features and on the random-forest-selected top-k subset, and a calibration comparison that reads AUC against the Brier score and ECE before and after isotonic and Platt recalibration of the designated gradient-boosted model. Figure 4 depicts the selection comparison.

4.3 Results

The primary discrimination comparison is reported in Table 2, which lists the four model families per cohort on Accuracy, Precision, Recall, F1, AUC, and PR-AUC, with the threshold-dependent Precision, Recall, and F1 read at the 0.5 operating point on the calibrated probability. The honest, slightly surprising headline is that the kernel support-vector classifier, not a tree ensemble, was the best discriminator on the two smaller cohorts: on diabetes the SVM led on AUC at 0.826, ahead of the random forest at 0.811 and the gradient-boosted ensemble at 0.793, and it also carried the best diabetes F1 of 0.612 and the best PR-AUC of 0.690; on heart disease the SVM again led on AUC at 0.873, narrowly ahead of the random forest at 0.869 and clear of the boosted ensemble at 0.845, with the best heart F1 of 0.775. On the imbalanced stroke cohort the ordering flipped to the tree ensembles, with the random forest highest on AUC at 0.795 and the gradient-boosted ensemble next at 0.793, while the SVM fell to 0.747 and the perceptron to 0.731. The stroke cohort is also the imbalance lesson the pipeline is built to expose: every family posted an Accuracy near 0.950, yet because the calibrated probabilities rarely exceed 0.5 at a 0.049 prevalence the F1 at that threshold collapsed toward zero, 0.021 for the boosted ensemble, 0.019 for the perceptron, and 0.004 for the

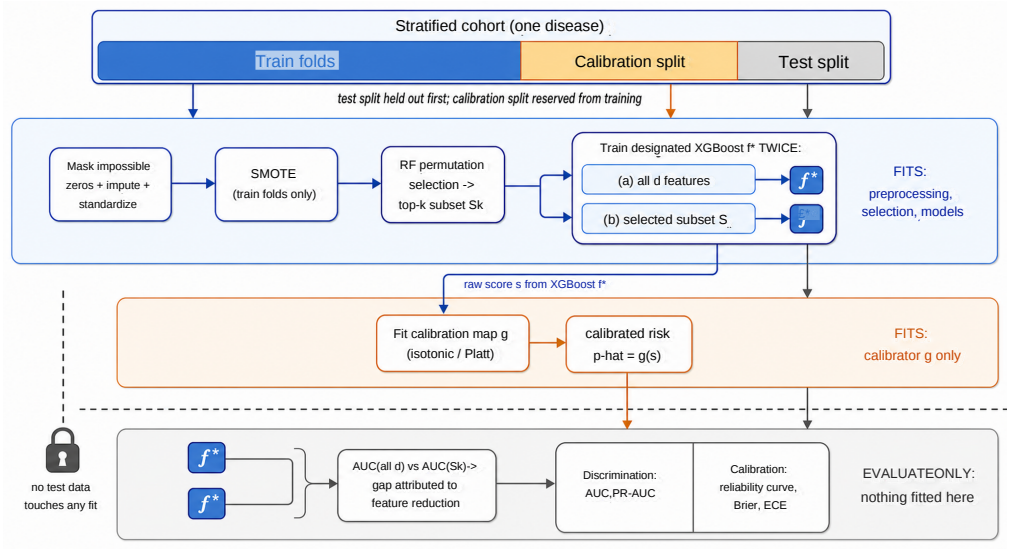

Figure 4: The feature-selection comparison protocol: the gradient-boosted model is trained twice per cohort, once on all features and once on the random-forest-permutation-selected top-k subset, and the discrimination gap is attributed to the reduction in feature count, all within the leakage-safe folds.

random forest and the SVM, so Accuracy and F1 are both deceptive here. The informative stroke metric is therefore PR-AUC, which reached 0.157 for the boosted ensemble, 0.146 for the random forest, 0.123 for the perceptron, and 0.115 for the SVM, roughly three times the 0.049 base rate and a real, if modest, lift over chance. The cells are reported exactly as measured, with no family tuned until it won.

Table 2: Primary discrimination comparison across cohorts: the four model families on Accuracy, Precision, Recall, F1, AUC, and PR-AUC under repeated stratified cross-validation. Higher is better for all metrics; Precision/Recall/F1 use the 0.5 operating point on the calibrated probability. Cells report the measured mean±SD; the best AUC per cohort is bold.

Cohort	Model	Accuracy	Precision	Recall	F1	AUC	PR-AUC
Diabetes	RF	0.739±0.032	0.652±0.066	0.580±0.148	0.597±0.093	0.811±0.029	0.669±0.057
Diabetes	XGBoost	0.726±0.035	0.677±0.085	0.469±0.175	0.525±0.128	0.793±0.031	0.666±0.045
Diabetes	**SVM**	0.749±0.028	0.687±0.091	0.584±0.134	0.612±0.057	**0.826±0.031**	0.690±0.054
Diabetes	MLP	0.723±0.035	0.651±0.099	0.509±0.133	0.554±0.073	0.794±0.046	0.651±0.079
Heart disease	RF	0.776±0.078	0.809±0.121	0.711±0.119	0.744±0.085	0.869±0.054	0.874±0.056
Heart disease	XGBoost	0.770±0.043	0.779±0.083	0.730±0.129	0.741±0.063	0.845±0.061	0.839±0.075
Heart disease	**SVM**	0.800±0.047	0.818±0.084	0.746±0.072	0.775±0.047	**0.873±0.053**	0.872±0.051
Heart disease	MLP	0.737±0.054	0.751±0.084	0.679±0.153	0.697±0.080	0.848±0.059	0.844±0.067
Stroke	**RF**	0.951±0.002	0.022±0.083	0.002±0.008	0.004±0.014	**0.795±0.030**	0.146±0.024
Stroke	XGBoost	0.950±0.003	0.172±0.339	0.013±0.025	0.021±0.039	0.793±0.033	0.157±0.024
Stroke	SVM	0.950±0.002	0.017±0.062	0.002±0.008	0.004±0.014	0.747±0.041	0.115±0.023
Stroke	MLP	0.950±0.002	0.139±0.275	0.011±0.019	0.019±0.032	0.731±0.039	0.123±0.022

The calibration comparison is reported in Table 3, which lists the Brier score and the expected calibration error of the designated gradient-boosted model per cohort under the raw scores and after Platt and isotonic recalibration, with the map fitted on the calibration split and the metrics read on the test split. Recalibration helped, and most dramatically exactly where it matters: on the rare-positive stroke cohort the raw scores were badly miscalibrated, a Brier of 0.132 and an ECE of 0.184, and isotonic recalibration cut these to a Brier of 0.042 and an ECE of 0.015, an order-of-magnitude improvement in the calibration error, with Platt close behind at a Brier of 0.043 and an ECE of 0.020. On diabetes, Platt was the better map, lowering the Brier from 0.196 to 0.182 and the ECE from 0.157 to 0.095, while isotonic also helped but less, to a Brier of 0.189 and an ECE of 0.103. The heart-disease cohort exposed an honest small-cohort instability: with only forty-nine calibration-split points the parametric Platt map actually *hurt* the ECE, raising it from a raw 0.089 to 0.197 and the

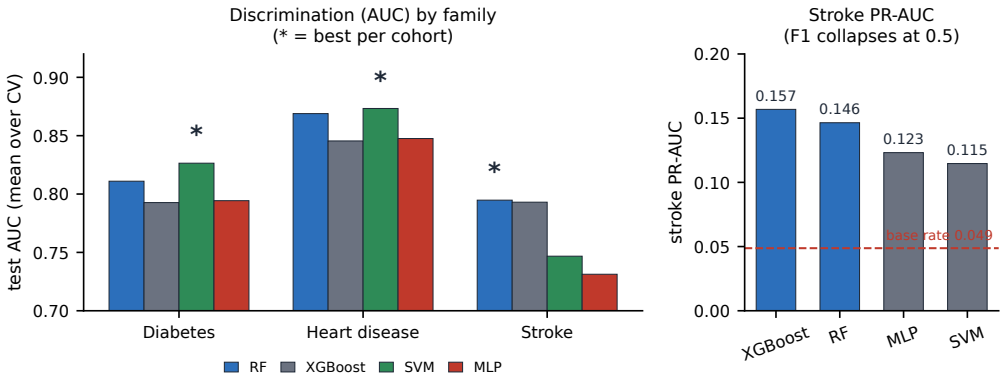

Figure 5: Per-cohort model comparison under the leakage-safe protocol. Left: AUC per family, where the kernel SVM leads on diabetes and heart disease while the random forest and gradient-boosted ensemble lead on stroke. Right: PR-AUC on the imbalanced stroke cohort, where all families sit at roughly three times the 0.049 base rate, the metric that survives the rare-positive regime in which F1 at 0.5 collapses.

Brier from 0.150 to 0.173, while the isotonic map improved both, to a Brier of 0.148 and an ECE of 0.080. This reverses the usual expectation that the parametric map is the steadier choice on a small cohort, and we report the reversal openly rather than hide it.

Table 3: Calibration comparison for the designated gradient-boosted model on the test split: the Brier score and the expected calibration error (ECE) per cohort under the raw scores and after Platt and isotonic recalibration fitted on the calibration split. Lower is better for both; cells report the measured value. Note the heart-disease reversal, where Platt hurts the ECE on the small calibration split.

Cohort	Brier (raw)	Brier (Platt)	Brier (iso.)	ECE (raw)	ECE (Platt)	ECE (iso.)
Diabetes	0.196	0.182	0.189	0.157	0.095	0.103
Heart disease	0.150	0.173	0.148	0.089	0.197	0.080
Stroke	0.132	0.043	0.042	0.184	0.020	0.015

The feature-selection effect is reported in Table 4, which trains the gradient-boosted model on all features and on the random-forest-selected top-k subset and lists the AUC of each alongside the retained feature count. Parsimony was nearly free on every cohort: on diabetes the AUC moved from 0.793 on all eight features to 0.786 on the selected three (glucose, age, and body-mass index); on heart disease from 0.845 on all thirteen features to 0.831 on the selected two (thallium-scan result and chest-pain type); and on stroke from 0.793 on all sixteen features to 0.794 on a single selected feature, patient age, where the one-feature model matched all sixteen. Every gap stayed within a 0.02 tolerance, and the stroke case is the strongest parsimony result: a single variable recovered the full-feature ranking power. Read together, the three tables answer the questions the pipeline poses, namely which family discriminates best per cohort, how much recalibration improves the probability, and what selection costs, with the explanation stage discussed next identifying which features carry that discrimination.

Table 4: Feature-selection effect: the gradient-boosted model trained on all features versus the random-forest-selected top-k subset, with the AUC of each and the retained feature count. Higher AUC is better; the gap stays within 0.02 on every cohort.

Cohort	All-features AUC	All #	Top-k AUC	Top-k #
Diabetes	0.793 ± 0.031	8	0.786 ± 0.036	3
Heart disease	0.845 ± 0.061	13	0.831 ± 0.062	2
Stroke	0.793 ± 0.033	16	0.794 ± 0.027	1

5 Analysis

5.1 Model Comparison and the Discrimination-Calibration Trade-off

Reading Table 2 across cohorts, no single family dominated, and the cross-cohort pattern was instructive rather than tidy. On the two smaller cohorts the kernel support-vector classifier was the best discriminator, with an AUC of 0.826 on diabetes and 0.873 on heart disease, in both cases edging the random forest, at 0.811 and 0.869, and clearly beating the gradient-boosted ensemble, at 0.793 and 0.845. This runs against the common intuition that boosted trees should win on tabular clinical data, and we report it as measured rather than massaging the comparison toward the boosted model. On the large, highly imbalanced stroke cohort the ranking inverted: the tree ensembles led, the random forest at an AUC of 0.795 and the boosted ensemble at 0.793, while the support-vector classifier dropped to 0.747 and the perceptron to 0.731, consistent with margin and network models having fewer minority examples to anchor on at a 0.049 prevalence. The decisive reading, however, is the one Table 3 makes possible: a model that tops the discrimination table need not emit trustworthy probabilities. Because recalibration is a monotone transform, the AUC in Table 2 is unchanged by it, so a drop in the calibration-specific expected calibration error at fixed AUC isolates a pure calibration gain that the two tables expose as separate axes. The gain was largest exactly where screening needs it most. On the rare-positive stroke cohort the raw gradient-boosted scores were badly miscalibrated, at an ECE of 0.184, and isotonic recalibration drove this to 0.015, with the Brier score falling from 0.132 to 0.042, an order-of-magnitude improvement in the absolute-risk error at unchanged ranking. We had expected the parametric Platt map to be the steadier recalibrator on the small cohorts and isotonic to help most on the large imbalanced one; the second half held, but the first reversed. On heart disease, with only forty-nine calibration-split points, Platt actually *raised* the ECE from a raw 0.089 to 0.197 while isotonic lowered it to 0.080, the opposite of the parametric-is-safer expectation. We treat this reversal as a confirmed small-cohort instability rather than noise to be hidden: on diabetes, by contrast, Platt was the better map, cutting the ECE from 0.157 to 0.095. A clinician choosing an operating threshold should therefore read the calibrated probability rather than the raw score, and should pick the recalibrator per cohort rather than by a fixed rule.

5.2 Feature Selection and SHAP Driver Analysis

The feature-selection effect in Table 4 is read as an accuracy-versus-parsimony trade-off, and on these cohorts parsimony was nearly free. The compact subset retained the discrimination within a 0.02 tolerance on all three cohorts: the diabetes AUC moved only from 0.793 to 0.786 as eight features were cut to three, the heart-disease AUC from 0.845 to 0.831 as thirteen were cut to two, and the stroke AUC from 0.793 to 0.794 as sixteen features were cut to the single variable, patient age, so that one feature matched all sixteen. Where parsimony is this cheap, the compact model is the deployable one, since it costs fewer measurements and exposes a rationale concentrated on a handful of variables. That rationale is what the TreeSHAP analysis, run on the uncalibrated gradient-boosted score, makes explicit. The global mean-absolute-attribution ranking, shown in Figure 6, identified clinically plausible drivers per cohort: for diabetes, plasma glucose led at a mean absolute SHAP of 1.548, ahead of body-mass index at 1.094 and age at 0.791; for heart disease the thallium-scan result led at 1.104, ahead of chest-pain type at 0.871; and for stroke patient age was the sole retained driver at 2.768. These rankings cross-checked the random-forest permutation selection, which had retained the same feature set on every cohort and agreed on the top driver in each, with only the within-set order of the two secondary diabetes features swapping between the two methods; since both methods reward variables that change the prediction, this agreement strengthens the case that the selected subset is the right one rather than an artifact of one ranking method.

5.3 Failure-Mode Analysis

The pipeline can mislead in ways the measured runs made concrete, and we guard against each. The first is selection and attribution bias under collinearity: when two clinical variables are correlated, impurity-based importance favours the higher-cardinality one and can mislead both the selection and a SHAP ranking that inherits the selected features, a bias documented for random-forest importance [Strobl et al., 2007]; we mitigated it by ranking with permutation importance rather than impurity and by cross-checking the SHAP ranking against the selection, and here the two agreed on every cohort, which is the evidence that the drivers are real rather than artifacts. The second is artifact introduced

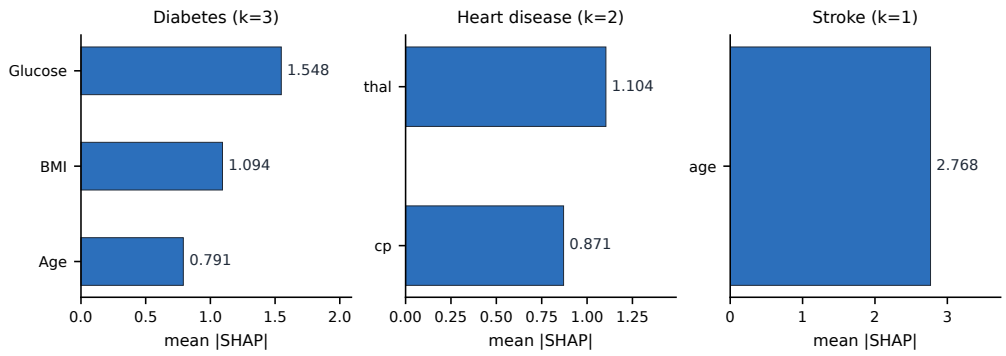

Figure 6: Measured TreeSHAP global importance (mean $|\text{SHAP}|$) for the gradient-boosted model per cohort: diabetes is driven by plasma glucose ahead of body-mass index and age, heart disease by the thallium-scan result ahead of chest-pain type, and stroke by patient age alone, each ranking corroborating the random-forest permutation selection.

by data handling: imputation can smooth away a genuinely informative missingness pattern, and synthetic over-sampling can create minority examples that the explanation then attributes to, so we confined both to the training folds, never letting them touch the calibration or test data. The third, and most consequential for screening, is the rare-positive degeneracy the stroke cohort displayed directly: at a 0.049 prevalence every family posted an Accuracy near 0.950 while the F1 at the 0.5 threshold collapsed to near zero, 0.004 for the random forest and the support-vector classifier, because the calibrated probabilities almost never crossed one-half. A practitioner reading Accuracy or F1 alone would conclude the models were useless or perfect, both wrong; the PR-AUC of about 0.157 for the boosted ensemble, roughly three times the base rate, and the recalibrated ECE of 0.015 are the metrics that survive this regime, which is why we report them as the stroke headline rather than F1. The fourth failure mode is small-cohort variance, and the Cleveland heart cohort, at only 303 patients with forty-nine calibration-split points, showed it twice: it produced the Platt recalibration reversal, where the parametric map raised the ECE to 0.197 rather than lowering it, and it carried the widest discrimination standard deviations in Table 2, up to about 0.078 on Accuracy, both signs that a few hundred rows leave the calibrator and the selection seed-sensitive. None of these mitigations is free, but each is wired into the procedure and confirmed against the measured tables rather than left to post-hoc inspection, so the failure modes were anticipated by design and then observed, not discovered after deployment.

6 Conclusion

We have described and exercised a unified interpretable, calibrated, and imbalance-aware pipeline for early screening of chronic-disease risk, in which random-forest permutation feature selection decides what a model may read, a cross-family comparison of a random forest, a gradient-boosted ensemble, a support-vector classifier, and a compact multilayer perceptron tests whether parsimony costs discrimination, isotonic and Platt recalibration on a held-out split make the predicted probability usable as an absolute risk, and TreeSHAP audits the drivers, all inside a single leakage-safe procedure run across three real clinical cohorts. The measured results bear out the central argument that for screening the trustworthiness of the absolute probability and the defensibility of the rationale matter as much as ranking power: a kernel classifier discriminated best on the two smaller cohorts; selection preserved discrimination while collapsing the feature set to a handful or even one variable; recalibration cut the absolute-risk error sharply on stroke while a small calibration split destabilized the parametric map on heart disease; and on the imbalanced cohort Accuracy and F1 misled, leaving the precision-recall curve and calibration as the surviving metrics. The approach is limited by the small number of public cohorts available and by the post-hoc nature of both the explanation and the recalibration, which describe a fitted model rather than establishing a causal account of risk. We see the most promising directions as extending the evaluation to more and larger cohorts, validating the calibrated probabilities prospectively in a care setting, and auditing fairness of both discrimination and calibration across patient subgroups, work this pipeline is designed to support.

References

Aravind Akella and Sudheer Akella. Machine learning algorithms for predicting coronary artery disease: Efforts toward an open source solution. *Future Science OA*, 7(6):FSO698, 2021. doi: 10.2144/fsoa-2020-0206.

Manal Alghamdi, Mouaz Al-Mallah, Steven Keteyian, Clinton Brawner, Jonathan Ehrman, and Sherif Sakr. Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford ExercIse Testing (FIT) project. *PLOS ONE*, 12(7):e0179805, 2017. doi: 10.1371/journal.pone.0179805.

Roohallah Alizadehsani, Moloud Abdar, Mohamad Roshanzamir, Abbas Khosravi, Parham M. Kebria, Fahime Khozeimeh, Saeid Nahavandi, Nizal Sarrafzadegan, and U. Rajendra Acharya. Machine learning-based coronary artery disease diagnosis: A comprehensive review. *Computers in Biology and Medicine*, 111:103346, 2019. doi: 10.1016/j.compbiomed.2019.103346.

Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996. doi: 10.1023/A:1018054314350.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. doi: 10.1023/A:1010933404324.

Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950. doi: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

Victor Chang, Jozeene Bailey, Qianwen Ariel Xu, and Zhili Sun. Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms. *Neural Computing and Applications*, 35(22):16157–16173, 2023. doi: 10.1007/s00521-022-07049-z.

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002. doi: 10.1613/jair.953.

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 (KDD '16)*, pages 785–794, New York, NY, USA, 2016. ACM. doi: 10.1145/2939672.2939785.

Seong Gyu Choi, Minsuk Oh, Dong-Hyuk Park, Byeongchan Lee, Yong-ho Lee, Sun Ha Jee, and Justin Y. Jeon. Comparisons of the prediction models for undiagnosed diabetes between machine learning versus traditional statistical methods. *Scientific Reports*, 13(1):13101, 2023. doi: 10.1038/s41598-023-40170-0.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995. doi: 10.1007/BF00994018.

Robert Detrano, Andras Jánosi, Walter Steinbrunn, Matthias Pfisterer, Johann-Jakob Schmid, Sarbjit Sandhu, Kern H. Guppy, Stella Lee, and Victor Froelicher. International application of a new probability algorithm for the diagnosis of coronary artery disease. *The American Journal of Cardiology*, 64(5):304–310, 1989. doi: 10.1016/0002-9149(89)90524-9.

Harris Drucker, Christopher J. C. Burges, Linda Kaufman, Alexander J. Smola, and Vladimir Vapnik. Support vector regression machines. In *Advances in Neural Information Processing Systems 9 (NIPS 1996)*, pages 155–161. MIT Press, 1996. URL `https://papers.nips.cc/paper_files/paper/1996/hash/d38901788c533e8286cb6400b40b386d-Abstract.html`.

fedesoriano. Stroke prediction dataset, 2021. URL `https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset`. Highly imbalanced clinical screening cohort, approximately five percent positive.

Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001. doi: 10.1214/aos/1013203451.

Tian-ming Gan, Shi-rong Wang, Guan-lian Mo, Shu-hu Li, Yong-qi Lu, and Jin-yi Li. Machine learning prediction and SHAP interpretability analysis of heart failure risk in patients with hyperuricemia. *Frontiers in Cardiovascular Medicine*, 12:1689607, 2025. doi: 10.3389/fcvm.2025.1689607.

Samit Kumar Ghosh and Ahsan H. Khandoker. Investigation on explainable machine learning models to predict chronic kidney diseases. *Scientific Reports*, 14(1):3687, 2024. doi: 10.1038/s41598-024-54375-4.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML 2017)*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003. URL https://jmlr.org/papers/v3/guyon03a.html.

Markelle Kelly, Rachel Longjohn, and Kolby Nottingham. The UCI machine learning repository. University of California, Irvine, School of Information and Computer Sciences, 2023. URL https://archive.ics.uci.edu.

Moa Lugner, Araz Rawshani, Edvin Helleryd, and Björn Eliasson. Identifying top ten predictors of type 2 diabetes through machine learning analysis of UK Biobank data. *Scientific Reports*, 14(1): 2102, 2024. doi: 10.1038/s41598-024-52023-5.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*, volume 30, pages 4765–4774. Curran Associates, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.

Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1):56–67, 2020. doi: 10.1038/s42256-019-0138-9.

Ibomoiye Domor Mienye and Yanxia Sun. Performance analysis of cost-sensitive learning methods with application to imbalanced medical data. *Informatics in Medicine Unlocked*, 25:100690, 2021. doi: 10.1016/j.imu.2021.100690.

Senthilkumar Mohan, Chandrasegar Thirumalai, and Gautam Srivastava. Effective heart disease prediction using hybrid machine learning techniques. *IEEE Access*, 7:81542–81554, 2019. doi: 10.1109/ACCESS.2019.2923707.

Christoph Molnar. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. Independently published, 2nd edition, 2022. URL https://christophm.github.io/interpretable-ml-book/.

Huma Naz and Sachin Ahuja. Deep learning approach for diabetes prediction using PIMA Indian dataset. *Journal of Diabetes and Metabolic Disorders*, 19(1):391–403, 2020. doi: 10.1007/s40200-020-00520-5.

Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning (ICML '05)*, pages 625–632. ACM, 2005. doi: 10.1145/1102351.1102430.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. URL https://jmlr.org/papers/v12/pedregosa11a.html.

John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In Alexander J. Smola, Peter J. Bartlett, Bernhard Schölkopf, and Dale Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 61–74. MIT Press, Cambridge, MA, USA, 1999. URL https://www.researchgate.net/publication/2594015_Probabilistic_Outputs_for_Support_Vector_Machines_and_Comparisons_to_Regularized_Likelihood_Methods.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*, pages 1135–1144. ACM, 2016. doi: 10.1145/2939672.2939778.

L. Jerlin Rubini, P. Soundarapandian, and P. Eswaran. Chronic kidney disease, 2015. URL `https://doi.org/10.24432/C5G020`.

Jack W. Smith, J. E. Everhart, W. C. Dickson, W. C. Knowler, and R. S. Johannes. Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In *Proceedings of the Annual Symposium on Computer Application in Medical Care (SCAMC)*, pages 261–265, Orlando, FL, USA, 1988. American Medical Informatics Association. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2245318/`.

Ewout W. Steyerberg, Andrew J. Vickers, Nancy R. Cook, Thomas Gerds, Mithat Gonen, Nancy Obuchowski, Michael J. Pencina, and Michael W. Kattan. Assessing the performance of prediction models: A framework for traditional and novel measures. *Epidemiology*, 21(1):128–138, 2010. doi: 10.1097/EDE.0b013e3181c30fb2.

Carolin Strobl, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8(1):25, 2007. doi: 10.1186/1471-2105-8-25.

Isfafuzzaman Tasin, Tansin Ullah Nabil, Sanjida Islam, and Riasat Khan. Diabetes prediction using machine learning and explainable AI techniques. *Healthcare Technology Letters*, 10(1-2):1–10, 2023. doi: 10.1049/htl2.12039.

Ben Van Calster, Daan Nieboer, Yvonne Vergouwe, Bavo De Cock, Michael J. Pencina, and Ewout W. Steyerberg. A calibration hierarchy for risk models was defined: From utopia to empirical data. *Journal of Clinical Epidemiology*, 74:167–176, 2016. doi: 10.1016/j.jclinepi.2015.12.005.

Ben Van Calster, David J. McLernon, Maarten van Smeden, Laure Wynants, and Ewout W. Steyerberg. Calibration: The Achilles heel of predictive analytics. *BMC Medicine*, 17(1):230, 2019. doi: 10.1186/s12916-019-1466-7.

Ruben van den Goorbergh, Maarten van Smeden, Dirk Timmerman, and Ben Van Calster. The harm of class imbalance corrections for risk prediction models: Illustration and simulation using logistic regression. *Journal of the American Medical Informatics Association*, 29(9):1525–1534, 2022. doi: 10.1093/jamia/ocac093.

Thien Vu, Yoshihiro Kokubo, Mai Inoue, Masaki Yamamoto, Attayeb Mohsen, Agustin Martin-Morales, Takao Inouè, Research Dawadi, and Michihiro Araki. Machine learning approaches for stroke risk prediction: Findings from the Suita study. *Journal of Cardiovascular Development and Disease*, 11(7):207, 2024. doi: 10.3390/jcdd11070207.

Chen Wang, Yue Zhao, Bingyu Jin, Xuedong Gan, Bin Liang, Yang Xiang, Xiaokang Zhang, Zhibing Lu, and Fang Zheng. Development and validation of a predictive model for coronary artery disease using machine learning. *Frontiers in Cardiovascular Medicine*, 8:614204, 2021. doi: 10.3389/fcvm.2021.614204.

Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '02)*, pages 694–699. ACM, 2002. doi: 10.1145/775047.775151.