Jurnal Riset Rumpun Seni, Desain dan Media
Volume 2, Nomor 1, April 2023
e-ISSN: 2829-0186; p-ISSN: 2829-0283, Hal. 192-214
DOI: https://doi.org/10.55606/jurrsendem.v2i1.9197
Tersedia: https://prin.or.id/index.php/JURRSENDEM

Explainable Clinical Risk Prediction from EHR Tabular Data using Monotonic Constraints and Calibrated Probabilities

Danang[1*], Toni Wijanarko Adi Putra[2]
[1-2]Universitas Sains dan Teknologi Komputer, Indonesia
Email: danang150787@gmail.com [1], toni.wijanarko@stekom.ac.id [2]
**Corresponding Author: danang150787@gmail.com*

Abstract. Tabular-based clinical risk prediction models are extensively applied in medical decision support systems; however, two major challenges often reduce their reliability: predictions that contradict basic clinical logic and poorly calibrated probability outputs that weaken threshold-based decision making. This study investigates explainable binary risk prediction using the processed Cleveland subset of the UCI Heart Disease dataset as a public clinical benchmark. A lightweight and CPU-efficient pipeline is proposed by employing an XGBoost classifier integrated with monotonic constraints on clinically relevant features, followed by probability calibration through post-hoc methods, including Platt scaling, temperature scaling, and isotonic regression on a separate validation set. Model performance is assessed in terms of discrimination capability using AUROC, AUPRC, F1-score, sensitivity, and specificity, while probability reliability is evaluated using ECE and Brier score metrics. A monotonicity audit is also conducted through counterfactual feature sweeps to measure violation rates. In addition, the model is applied for risk stratification into low-, medium-, and high-risk categories with corresponding event-rate reporting. The findings demonstrate that isotonic regression improves probability reliability without degrading discrimination performance. Furthermore, the monotonicity audit reveals no observed violations for constrained features. Overall, the integration of monotonic constraints and probability calibration produces more decision-ready risk estimates for threshold-based clinical decision support while maintaining transparency through SHAP-based analysis.

Keywords: Clinical Risk Prediction; Monotonic Constraints; Probability Calibration; SHAP; XGBoost.

1. INTRODUCTION

Clinical prediction models estimate an individual patient's risk of an adverse outcome and are frequently used to support screening, triage, and preventive interventions. In operational settings, decision support depends on at least two distinct properties: discrimination (the ability to rank higher-risk patients above lower-risk patients) and calibration (the agreement between predicted probabilities and observed event frequencies) (Van Calster et al., 2019). While discrimination is often summarized with AUROC, calibration is the property that determines whether a predicted probability can be used as a trustworthy decision input. A model can achieve high AUROC yet still output systematically overconfident or underconfident probabilities, which in practice can translate into over-referral, missed high-risk cases, and unstable actions when clinical thresholds are applied.

The practical consequence is straightforward: if a system is deployed to trigger follow-up tests or specialist referral above a probability threshold, the threshold assumes that the reported probabilities correspond to real event rates. When that assumption fails, downstream actions become hard to justify and may harm both patient outcomes and resource efficiency. This is why calibration has been described as a recurring weak point for predictive analytics in

Naskah Masuk: 18 Februari 2023; Revisi: 16 Maret 2023; Diterima: 25 April 2023; Terbit: 30 April 2023

medicine (Van Calster et al., 2019). In addition to reliability diagrams and summary calibration errors (Nixon et al., 2019), proper scoring rules such as the Brier score provide a principled way to evaluate probabilistic accuracy (Brier, 1950), and recent work discusses modified variants under practical evaluation constraints (Yang et al., 2022). Post-hoc calibration techniques such as temperature scaling and related approaches are widely used to improve probability reliability when raw model scores are miscalibrated (Guo et al., 2017; Silva Filho et al., 2023).

Beyond calibration, clinicians also expect basic directional behavior: for certain variables, increasing values should not reduce predicted risk. Examples include age or markers of disease burden and severity. Standard gradient-boosted trees are attractive for tabular clinical prediction because they offer strong performance and are efficient on CPU hardware (Chen & Guestrin, 2016). However, they optimize predictive loss without structural constraints, so they can fit non-intuitive local patterns, especially in small or noisy clinical datasets. These local reversals can reduce face validity and complicate deployment even when headline discrimination metrics look strong. Monotonicity constraints are a pragmatic way to encode minimal domain structure by forcing the prediction function to be non-decreasing (or non-increasing) with respect to selected features (Pei et al., 2016; Wang et al., 2022). Importantly, monotonic constraints do not automatically guarantee clinical correctness, but they can prevent a class of behaviors that are difficult to defend in real decision support.

This paper studies a lightweight, CPU-friendly approach to more trustworthy tabular risk prediction using an open benchmark. We use the processed Cleveland subset of the UCI Heart Disease repository as a public clinical tabular dataset (Dua & Graff, 2019; Niculescu-Mizil & Caruana, 2006). We treat the task as binary risk prediction and develop a pipeline that combines two mechanisms aimed at deployment-relevant trustworthiness: (i) explicit monotonic constraints reflecting clinically plausible ordering for a subset of variables, and (ii) post-hoc probability calibration selected on a held-out validation set to improve reliability without changing the underlying ranking. For calibration, we consider standard options including Platt scaling, temperature scaling, and isotonic regression (Guo et al., 2017; Platt, 1999). The end-to-end workflow is illustrated in Figure 1.

Research questions. This work is organized around three practical questions that matter for real-world clinical deployment. First, to what extent can monotonic constraints enforce clinically plausible directional behavior in a boosted-tree risk model without materially degrading discrimination on a standard tabular benchmark? Second, when a strong discriminative model is trained, how much can post-hoc calibration improve probabilistic

reliability as measured by proper scoring rules and calibration summaries, and which calibration option is preferable under validation-based selection (Brier, 1950; Nixon et al., 2019; Silva Filho et al., 2023)? Third, can the resulting model be made auditable in a way that is useful for practitioners by combining counterfactual monotonicity checks with well-established explanation methods for tree ensembles without over-claiming what explanations can guarantee in healthcare settings (Ghassemi et al., 2021; Lundberg &Lee, 2017; Lundberg et al., 2020)?

Contributions. We make several contributions intended to be practical, reproducible, and aligned with how clinical models are assessed. We provide a fully reproducible pipeline for monotone-constrained gradient boosting using XGBoost, including an explicit construction of the monotonic constraint vector that remains consistent with the final feature representation after preprocessing. We integrate multiple calibration techniques and perform model selection based strictly on held-out validation criteria to avoid optimistic bias, then report discrimination and reliability on the test set using complementary metrics and visual diagnostics (Chen & Guestrin, 2016; Guo et al., 2017; Nixon et al., 2019; Platt, 1999). To verify that monotonic constraints translate into the expected behavior, we implement an empirical monotonicity audit based on counterfactual feature sweeps and summarize results via a violation-rate measure, providing an actionable check rather than relying on qualitative claims (Pei et al., 2016; Wang et al., 2022). For transparency, we add integrated explainability using SHAP for both global importance and patient-level explanations, while explicitly treating explanations as an auditing aid rather than a proof of causality (Arrieta et al., 2020; Lundberg & Lee, 2017; Lundberg et al., 2020). Finally, we connect probabilistic outputs to decision support by presenting risk stratification into actionable bands and reporting empirical event rates per band, a step that is closer to operational decision-making than discrimination-only reporting (Vickers & Elkin, 2006; Vickers et al., 2019). These choices also support clearer reporting and bias-aware evaluation in line with clinical prediction modeling guidance (Moons et al., 2019). This emphasis on an auditable and adaptive pipeline is also consistent with broader framework-oriented system design that prioritizes modular integration for dependable deployment.

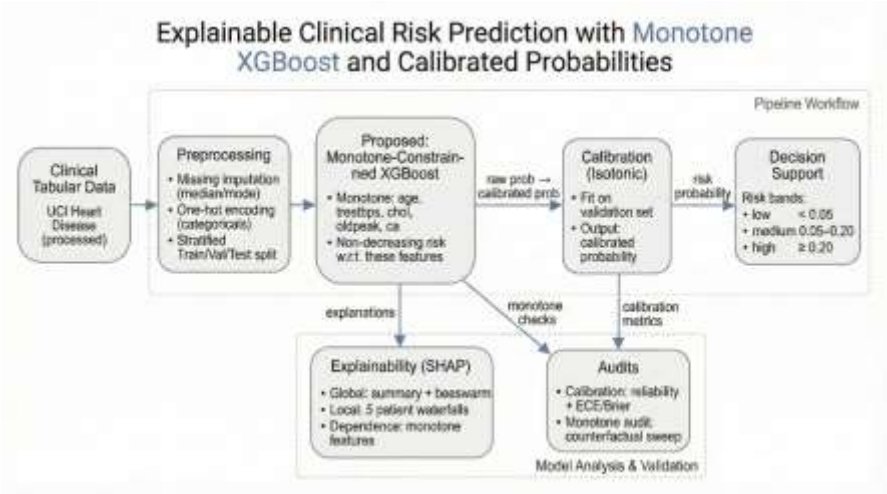


Figure 1. End-to-end workflow: preprocessing, monotone-constrained XGBoost, validation-based probability calibration, evaluation of discrimination and reliability, explainability, and monotonicity audits for decision support.

2. LITERATURE REVIEW

Clinical Prediction Modeling and Reporting

Clinical prediction models aim to estimate an individual's probability of experiencing a future clinical outcome over a defined horizon, and they are frequently used to support screening, triage, and preventive interventions. In EHR-based settings, such models must operate under distribution shift, missingness, proxy variables, and complex care pathways that can induce spurious associations. Although deep learning has demonstrated impressive results at scale for certain EHR applications (Rajkomar et al., 2018), practical tabular prediction in many institutions still relies heavily on strong classical baselines because they are easier to train, deploy, monitor, and govern. Tree-based gradient boosting remains a particularly strong option for structured clinical prediction due to its competitive accuracy and implementation simplicity (Chen & Guestrin, 2016; Ke et al., 2017; Prokhorenkova et al., 2018).

However, high predictive performance alone is insufficient in clinical contexts, where models can cause harm if their claims are overstated or if their evaluations are biased. For this reason, clinical prediction research places strong emphasis on methodological rigor, including appropriate validation, avoidance of leakage, and explicit assessment of bias and applicability. PROBAST provides a structured framework to assess risk of bias and applicability across domains such as participants, predictors, outcomes, and analysis (Moons et al., 2019). More recently, TRIPOD+AI updates reporting guidance specifically for AI-based clinical prediction models, highlighting the need to clearly document data handling, preprocessing, model development choices, and validation strategy. PROBAST+AI further updates risk-of-bias assessment for AI prediction models, reflecting the additional failure modes introduced by

complex pipelines. Evidence from EHR prediction model reviews continues to show common reporting weaknesses and bias risks, reinforcing the need for careful methodology even when using seemingly standard datasets and models (Stevenson et al., 2021).

From a decision-support perspective, evaluations must separate two properties that are often conflated: discrimination and calibration. Discrimination indicates whether higher-risk patients tend to receive higher scores than lower-risk patients, while calibration assesses whether predicted probabilities reflect observed outcome frequencies. Calibration has been described as a recurring weak point of predictive analytics, yet it is central when models are used to trigger actions at thresholds (Van Calster et al., 2019). This is especially relevant in clinical settings where the same score may be used for referral decisions, follow-up scheduling, or risk stratification; if the probability scale is unreliable, threshold-based policies can become brittle. Therefore, contemporary clinical prediction pipelines increasingly treat probability reliability as a first-class objective, rather than an optional afterthought.

Uncertainty Quantification and Conformal Prediction

In manyclinical risk estimation scenarios, domain knowledge implies directional relationships between certain variables and risk. For instance, higher age, higher blood pressure, or more severe physiological measurements are often expected to correspond to non-decreasing risk, at least locally and within clinically plausible ranges. Monotonic constraints provide a simple but powerful mechanism to encode such expectations: they enforce that the prediction function is non-decreasing (or non-increasing) with respect to selected features while holding other features fixed. The formal study of monotonicity-constrained decision trees and related multivariate structures has a long history, with work showing how monotonic constraints can be incorporated into tree induction and how they affect expressivity and interpretability (Pei et al., 2016). In a medical context, the appeal is pragmatic: monotonic constraints can prevent counterintuitive local reversals that are difficult to justify to clinicians, reduce the governance burden associated with “surprising” model behavior, and create models that better align with basic face-validity expectations.

Modern gradient boosting systems implement monotonic constraints directly in training. XGBoost supports specifying monotone constraints per feature, influencing split selection and tree construction such that the ensemble respects the desired ordering (Chen & Guestrin, 2016). Related boosting systems (e.g., LightGBMandCatBoost) have also been used extensively in tabular modeling (Ke et al., 2017; Prokhorenkova et al., 2018). Recent work on monotonic gradient boosting for risk prediction further demonstrates that monotone constraints can be incorporated into high-performing boosted models to impose domain structure, typically

with modest trade-offs in flexibility (Wang et al., 2022). Importantly, monotonic constraints are not a guarantee of causal correctness. They encode a limited form of structural prior, and they can be inappropriate when the true relationship is non-monotone (e.g., U-shaped effects) or when the selected feature is a proxy for multiple mechanisms. This motivates two practical considerations in constrained modeling: (1) constraints should be applied only where domain knowledge is strong and the feature definition supports a monotone relationship, and (2) constrained behavior should be empirically audited rather than assumed.

In operational clinical pipelines, monotonic constraints also interact with preprocessing and feature representation. When categorical predictors are one-hot encoded, applying monotonicity to the raw categorical variable is not straightforward; constraints must be mapped carefully to the expanded feature space (often assigning zero constraints to one-hot components). This mapping step is often under-described in applied papers, yet it is essential for reproducibility and correct implementation. Our work emphasizes an explicit construction of the constraint vector aligned to the final feature representation, and evaluates monotonic behavior using counterfactual sweeps to verify that constraints translate into the expected directional behavior.

Probability Calibration

Probability calibration addresses the mismatch between a model's output scores and true event probabilities. Even when a classifier ranks patients well, its probability scale may be systematically distorted, producing overconfident or underconfident predictions. This issue is widely documented in modern predictive systems (Guo et al., 2017; Silva Filho et al., 2023), and has been highlighted as particularly problematic in clinical decision support where thresholds correspond to interventions (Van Calster et al., 2019). Post-hoc calibration techniques attempt to correct the probability scale by learning a mapping from raw scores to calibrated probabilities, typically fitted on a held-out validation set to avoid optimistic bias.

Several calibration approaches are commonly used. Platt scaling fits a logistic regression mapping from modelscorestoprobabilities andwasoriginallyintroducedfor SVMoutputs(Platt, 1999). Temperaturescaling, popularized in the neural network calibration literature, is a simple form of scaling that can substantially reduce miscalibration while minimally affecting ranking (Guo et al., 2017). Isotonic regression provides a non-parametric monotone mapping that can capture more complex calibration curves, though it may overfit in small validation sets. For multi-class settings, methods such as Dirichlet calibration generalize calibration to richer transformations (Kull et al., 2019). A broad review summarizes calibration

methods and evaluation practices across classification models, discussing when calibration is likely to be effective and how it should be assessed (Silva Filho et al., 2023).

Calibration quality is usually evaluated using both visualization and scalar metrics. Reliability diagrams provide an intuitive view of the relationship between predicted probability bins and observed frequencies. Scalar summaries include expected calibration error (ECE), which measures deviations between predicted and empirical frequencies across bins, and proper scoring rules such as the Brier score, which penalizes miscalibration and lack of sharpness in a single measure (Brier, 1950; Nixon et al., 2019). Variants of the Brier score have been discussed for practical evaluation contexts and outcome settings (Yang et al., 2022). Because different calibration methods can behave differently depending on model family and dataset size, a robust practice is to select the calibration approach using validation criteria (e.g., negative log-likelihood or Brier score) and report final metrics on a held-out test set. This principle is consistent with clinical prediction modeling guidance that stresses separation of model selection from final evaluation (Moons et al., 2019).

Explainability in Healthcare

Explainability is frequently cited as a requirement for clinical adoption, but its role must be framed carefully. In high-stakes healthcare settings, explanations should primarily support auditing, debugging, and communication, not serve as evidence of causal mechanisms. Broader surveys define explainable AI, its taxonomies, and its potential benefits and limitations, including the tension between transparency and performance (Arrieta et al., 2020). Critiques specifically in healthcare argue that post-hoc explanation methods can create a false sense of security if the model is biased, poorly validated, or relies on spurious correlations (Ghassemi et al., 2021). At the same time, some argue that, when feasible, inherently interpretable models should be preferred in high-stakes decisions rather than explaining black-box models after the fact (Rudin, 2019). These perspectives motivate a balanced approach: use explanation tools as part of a rigorous evaluation and governance process, not as a substitute for it.

For tree ensembles, SHAP has become a standard approach to feature attribution, providing a unified framework for interpreting model predictions (Lundberg & Lee, 2017). Subsequent work develops tree specific SHAP methods that enable efficient computation and connects local explanations to global model understanding (Lundberg et al., 2020). In clinical prediction settings, SHAP can support multiple practical needs: identifying dominant predictors, verifying that known risk factors behave plausibly, and providing patient-level attributions that help clinicians understand why a patient received a high predicted risk. However, SHAPexplanations still reflect the model's internal logic rather than causal effects;

therefore, they should be interpreted alongside calibration diagnostics, monotonicity audits, and bias-aware evaluation.

Finally, explainability is most useful when coupled to decision-making. Risk stratification into actionable probability bands can be combined with empirical event rates to communicate performance in operational terms. To evaluate whether a model is likely to improve decisions across thresholds, decision curve analysis provides a principled framework that links predicted probabilities to net benefit, complementing discrimination and calibration metrics (Vickers & Elkin, 2006; Vickers et al., 2019). In summary, modern clinical prediction pipelines increasingly emphasize a triad of properties: accurate ranking, reliable probabilities, and auditable behavior supported by careful reporting and bias assessment (Moons et al., 2019; Van Calster et al., 2019).

3. PROPOSED METHOD

Task Definition

Let $\mathbf{x} \in R^d$ denote a vector of tabular patient features and $y \in \{0,1\}$ denote a binary outcome indicating clinical risk (a proxy for an adverse event). We learn a probabilistic classifier $f(x) \in [0,1]$ that estimates $\hat{p}(\, y = 1 \mid x\,)$ a proxy for an adverse event). We learn a probabilistic classifier f (x) ∈ [0,1] that estimates ˆ $p(y$ = 1 | x). The predicted probability is used in two ways. First, it induces a ranking for prioritization (patients with larger $\hat{}p$ are deemed higher risk). Second, it supports threshold-based decision policies and risk stratification, where actions depend on whether $\hat{}p$ crosses predefined clinical thresholds. Because threshold-based decisions implicitly assume probability reliability, we treat calibration as a primary objective alongside discrimination (Silva Filho et al., 2023; Van Calster et al., 2019).

Dataset and Preprocessing

Weuse the processed Cleveland heart disease dataset from the UCI Machine Learning Repository (Dua & Graff, 2019; Niculescu-Mizil & Caruana, 2006). The dataset contains a mixture of numeric and categorical predictors. Since clinical tabular datasets often exhibit missing values, we apply a simple and reproducible preprocessing pipeline designed to minimize leakage and remain compatible with lightweight deployment. Missing values are imputed using statistics computed only from the training split to avoid information leakage. Specifically, numeric features are imputed by the training-set median and categorical features by the training-set mode. After imputation, categorical variables are one-hot encoded to produce a final feature representation ˜ x ∈ R ˜d. For Logistic Regression, we standardize

numeric columns (after encoding) to improve optimization and comparability across features; for boosted trees, scaling is not required because split-based learners are invariant to monotone transformations of individual features (Chen & Guestrin, 2016). All preprocessing steps are applied consistently across train/validation/test using parameters fitted on the training split only.

Data Splitting

To ensure a fair evaluation and a clean separation between model fitting, calibration, and final reporting, we create stratified train/validation/test splits that preserve class proportions. Stratification is important because the Cleveland subset is relatively small and class imbalance can otherwise introduce instability in both discrimination and calibration estimates.

The validation set serves two roles. First, it is used for early stopping in gradient boosting to prevent overfitting. Second, it is reserved for fitting calibration mappings and selecting the best calibration method according to a validation criterion. The test set is used only once for final evaluation after all model and calibrator choices are fixed, consistent with recommended practice in clinical prediction research to avoid optimistic bias (Moons et al., 2019). This split design is particularly important for calibration, because choosing a calibrator based on test performance would indirectly tune the pipeline to the test distribution.

Baselines

We compare the proposed constrained-and-calibrated pipeline against two baselines that reflect common practice in tabular clinical modeling. Logistic Regression (LR). Logistic Regression is a classical baseline with a linear decision boundary and direct probabilistic outputs. It is widely used in clinical risk modeling due to transparency, simplicity, and ease of calibration analysis. We train LR using the preprocessing pipeline described above (imputation, one-hot encoding, and standardization). This baseline provides a reference point for both discrimination and probability reliability.

XGBoost without constraints. Gradient-boosted trees are strong baselines for structured data and often outperform linear models in tabular settings (Chen & Guestrin, 2016). We train an unconstrained XGBoost classifier with early stopping on the validation set. This baseline isolates the effect of monotonic constraints and post-hoc calibration: any improvements in reliability or monotone behavior can be attributed to the additional components rather than the underlying learner.

Monotone-constrained XGBoost (proposed)

We incorporate domain knowledge by enforcing monotonic constraints on a subset of features where a consistent directional relationship with risk is clinically plausible (e.g., age, resting blood pressure, cholesterol, ST depression, and number of major vessels). Monotonic constraints require that, holding other features fixed, increasing a constrained feature cannot decrease the predicted risk. This aims to improve face validity and reduce counterintuitive local reversals that can occur in unconstrained boosted trees.

Reduce counterintuitive local reversals that can occur in unconstrained boosted trees. In practice, constraints must align with the final feature representation after preprocessing. Because categorical variables are one-hot encoded, a constraint specified on a raw categorical variable does not directly translate into a monotone constraint in the expanded space. Therefore, we construct a monotone constraint vector over the final features ˜ x by assigning non-zero monotonicity only to selected numeric features and assigning zero to one-hot features and any variables for which monotonicity is not justified. The resulting constraint vector is passed to XGBoost so that monotonicity is respected during tree construction (Chen &Guestrin, 2016). This procedure follows the practical spirit of monotonicity-constrained tree induction: encode minimal and defensible domain structure, preserve predictive strength, and then empirically verify that the learned model behaves as expected (Pei et al., 2016; Wang et al., 2022).

Calibration (Proposed)

Even with monotonic constraints, boosted-tree probabilities may remain miscalibrated. We therefore apply post-hoc calibration to improve probability reliability for decision support. Given a trained monotone XGBoost model, we obtain uncalibrated scores on the validation set and fit calibration mappings using three standard methods: (1) Platt scaling, which fits a sigmoid mapping from raw scores to probabilities (Platt, 1999); (2) temperature scaling, which rescales logits (or score proxies) by a single parameter to reduce overconfidence (Guo et al., 2017); and (3) isotonic regression, which learns a non-parametric monotone mapping and can correct more complex miscalibration patterns.

Because calibration methods can behave differently depending on dataset size and score distribution, we select the final calibrator based on validation performance. Specifically, we select the method that minimizes validation negative log-likelihood (NLL), then apply this fixed calibrator to test predictions for final reporting. This validation-based selection avoids test-set tuning and aligns with broader recommendations on transparent model development and evaluation (Moons et al., 2019; Silva Filho et al., 2023). For completeness, we also report

calibration diagnostics (reliability diagrams and summary metrics) to ensure that improvements are not artifacts of a single measure.

Evaluation Metrics

We evaluate both ranking performance and probability reliability, since both are required for safe threshold based decision support (Van Calster et al., 2019). Discrimination is reported using AUROC and AUPRC, and we also report F1-score, sensitivity (recall), and specificity at a chosen threshold. AUROC summarizes ranking quality across thresholds, while AUPRC can be more informative under class imbalance.

For calibration, we report the Brier score and expected calibration error (ECE). The Brier score is a proper scoring rule for binary probabilistic predictions (Brier, 1950):

$$\text{Brier} = \frac{1}{n}\sum_{i=1}^{n}(\hat{p}_i - y_i)^2. \quad (1)$$

Lower Brier values indicate better probabilistic accuracy. We also report ECE, which bins predictions into intervals and summarizes the average discrepancy between empirical accuracy and mean predicted confidence across bins (Guo et al., 2017; Nixon et al., 2019). While ECE is sensitive to binning choices, it provides an interpretable summary that complements the Brier score. Finally, reliability diagrams visualize calibration by plotting predicted risk against observed event rates per bin, enabling qualitative inspection of under- or over-confidence patterns**.**

Monotonicity Audit

To verify that monotonic constraints translate into the intended directional behavior in practice, we perform an empirical monotonicity audit based on counterfactual sweeps. For each constrained feature, we sample instances (typically from the test set) and generate counterfactual variants by sweeping the feature across a predefined grid while holding all other features fixed. For each sweep, we evaluate the model's predicted probabilities along the grid.

Amonotonicity violation is recorded when the predicted probability decreases for an increase in a feature that is constrained to be non-decreasing. We summarize audit outcomes using a violation rate, defined as the fraction of sweeps that contain at least one decrease. This audit provides an actionable verification step that goes beyond assuming constraint correctness, and it is consistent with the goal of producing models that are not only accurate but also behaviorally auditable (Pei et al., 2016; Wang et al., 2022).

Decision Support Evaluation Via Risk Stratification

To connect probabilistic predictions to operational decision support, we report a simple risk stratification analysis. Using calibrated probabilities, we define three risk bands: low ($p < 0.05$), medium ($0.05 \leq p < 0.20$), and high ($p \geq 0.20$). These bands represent a pragmatic triage view: low-risk cases may require minimal follow-up, medium-risk cases may warrant monitoring, and high-risk cases may trigger additional assessment or intervention.

For each band, we report the empirical event rate and the mean predicted probability. This communicates whether the probability scale meaningfully separates groups with different observed outcome frequencies and whether predicted risk aligns with observed risk in each operational regime. Such stratified reporting complements aggregate discrimination and calibration metrics by providing an end-user-friendly summary. In addition, decision curve analysis is a standard framework to evaluate net benefit across thresholds (Vickers & Elkin, 2006; Vickers et al., 2019), while our primary focus is risk stratification, the same calibrated probabilities can be evaluated under decision-analytic criteria in future work.

4. RESULT AND DISCUSSION

Experimental setup

All experiments were conducted on CPU to reflect a lightweight deployment setting in which the full pipeline can be executed without specialized hardware. To support reproducibility, we export random seeds, configuration files, and package versions alongside model artifacts. This includes the preprocessing schema (feature order after one-hot encoding), the monotonic constraint vector aligned to that schema, and all plots and tables used in reporting. The dataset source and preprocessing steps follow Section 3.2, and the evaluation protocol follows the stratified train/validation/test splitting described in Section 3.3. Importantly, the validation set is used for early stopping and calibration fitting/selection, while the test set is reserved strictly for final reporting, consistent with recommended practice to reduce optimistic bias in clinical prediction studies (Moons et al., 2019).

Because the benchmark is relatively small, we emphasize transparent reporting of both discrimination and calibration rather than relying on a single metric. In particular, calibration estimates can exhibit higher variance under limited sample sizes; therefore, we interpret results conservatively and highlight the need for external validation before any clinical use (Stevenson et al., 2021; Van Calster et al., 2019).

Overall Performance and Ablation

Wefirst summarize discrimination and calibration on the held-out test set. Table 1 reports overall performance for Logistic Regression, unconstrained XGBoost, and the proposed monotone-and-calibrated pipeline. Table 2 then isolates the contribution of two additions beyond the unconstrained tree baseline: (1) monotonic constraints and (2) post-hoc probability calibration.

Across models, boosted trees achieve strong ranking performance typical for structured tabular data (Chen &Guestrin, 2016). Thekeyobservation is that improving trustworthiness is not only about increasing AUROC; rather, it requires ensuring that predicted probabilities are meaningful for downstream threshold-based actions (Van Calster et al., 2019). The ablation study supports this point: adding monotonic constraints primarily targets behavioral coherence (preventing counterintuitive local reversals), whereas calibration primarily targets probability reliability. As reflected in Table 2, the full pipeline maintains strong discrimination while improving reliability compared to the uncalibrated monotone model, indicating that reliability gains are achieved without sacrificing the ability to separate higher-risk from lower-risk patients.

Table 1. Main test-set performance. Discrimination is reported by AUROC/AUPRC/F1/Sensitivity/Specificity; reliability is reported by ECE and Brier score.Hybrid + adaptive conformal (FULL) results on ETTh1 test set (target OT).

Logistik	AUROC	AUPRC	F1	Sens	Spec	ECE	Brier
Logistic Regression	0.949	0.938	0.905	0.870	0.920	0.166	0.089
XGBoost (no constraints)	0.933	0.930	0.864	0.870	0.840	0.169	0.104
XGBoost + monotone	0.935	0.933	0.837	0.857	0.840	0.168	0.106
XGBoost + monotone + calibration (FULL)	0.938	0.921	0.870	0.952	0.840	0.102	0.105

Table 2. Ablation study on the test set, isolating the effect of monotone constraints and calibration.

Logistik	AUROC	AUPRC	F1	Sens	Spec	ECE	Brier
LRbaseline (uncalibrated)	0.949	0.934	0.905	0.870	0.920	0.166	0.089
XGBno-constraints (uncalibrated)	0.933	0.930	0.864	0.870	0.840	0.169	0.104
XGBoost + monotone	0.935	0.933	0.837	0.857	0.840	0.168	0.106
XGB+monotone + calibration (FULL)	0.938	0.921	0.870	0.952	0.800	0.102	0.105

Calibration Analysis

Since calibrated probabilities are required for threshold-based decision support, we inspect reliability beyond scalar summaries. Figure 2 shows the reliability diagram on the test set. Points closer to the diagonal indicate that predicted probabilities better match observed event frequencies, while deviations from the diagonal reflect overconfidence or

underconfidence. After isotonic calibration, the curve moves closer to the diagonal, consistent with reduced miscalibration.

Weinterpret the reliability diagram jointly with ECE and the Brier score. ECE provides an interpretable summary of the accuracy–confidence gap aggregated across bins (Guo et al., 2017; Nixon et al., 2019), while the Brier score is a proper scoring rule that captures probabilistic accuracy (Brier, 1950). In small benchmarks, both ECE and Brier can vary depending on the split and binning choices; thus, we view them as complementary indicators rather than definitive proof of calibration quality (Van Calster et al., 2019). Overall, the calibrated model exhibits improved probability reliability, which is the property most directly required when using predicted risk to trigger interventions or allocate limited resources.

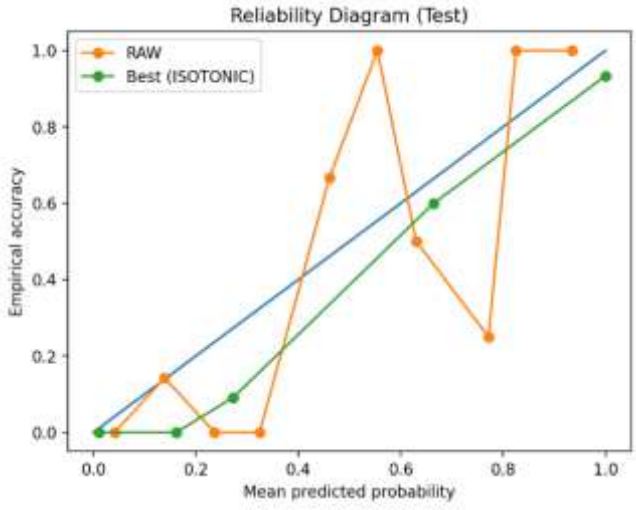


Figure 2. Reliability diagram on the test set. Points closer to the diagonal indicate better calibration.

Discrimination Curves

Scalar discrimination metrics can hide important operating-point behavior. To complement AUROC and AUPRC values, Figure 3 reports ROC and precision–recall curves on the test set. ROC curves summarize ranking performance across thresholds, while precision–recall curves are particularly informative under class imbalance because precision directly reflects the expected positive predictive value at different recall levels. In clinical triage scenarios, decision-makers often care about operating points where recall is high (to avoid missed cases) while precision remains acceptable (to avoid overwhelming downstream resources). Therefore, the precision–recall curve helps contextualize whether performance remains adequate in clinically relevant regions, rather than relying solely on AUROC.

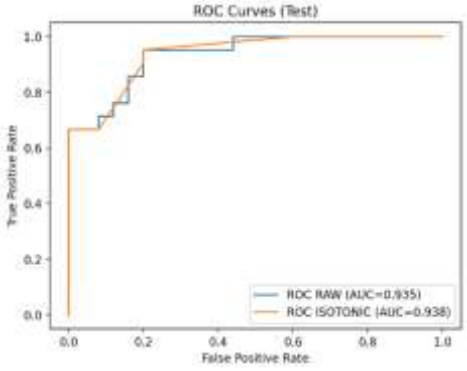


Figure 3. False Positive Rate.

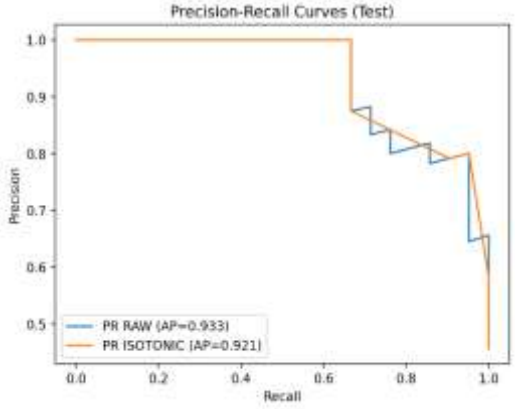


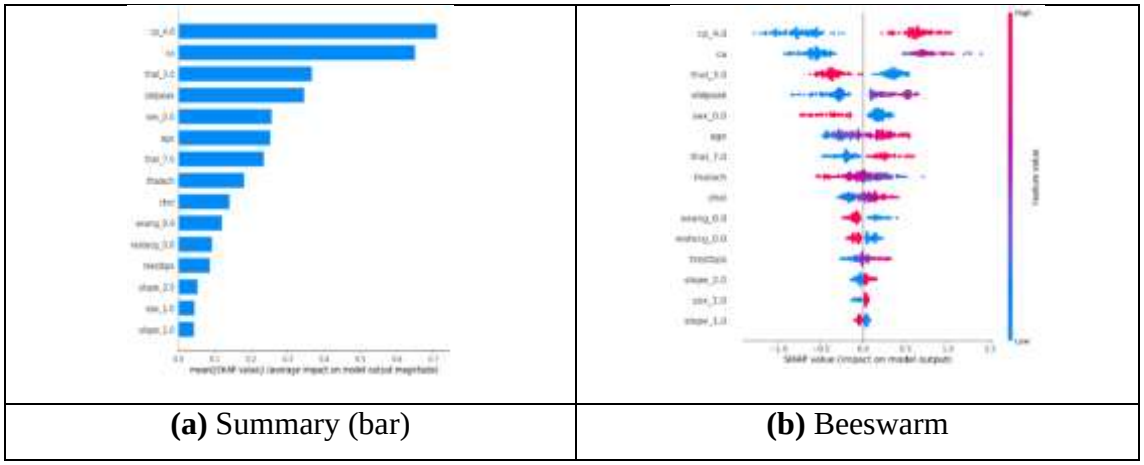
Figure 4. Recall.

Explainability with SHAP

We analyze model behavior using SHAP for the proposed monotone XGBoost model. SHAP provides additive feature attributions that can be aggregated globally to summarize dominant predictors and inspected locally to explain individual predictions (Lundberg & Lee, 2017; Lundberg et al., 2020). Figure 4 reports global explanations, highlighting which features most influence predicted risk across the cohort and how feature values relate to increased or decreased model output. This global view supports sanity checks: it can confirm whether known clinical risk factors are used in a broadly plausible manner and can reveal whether the model relies excessively on a small set of predictors.

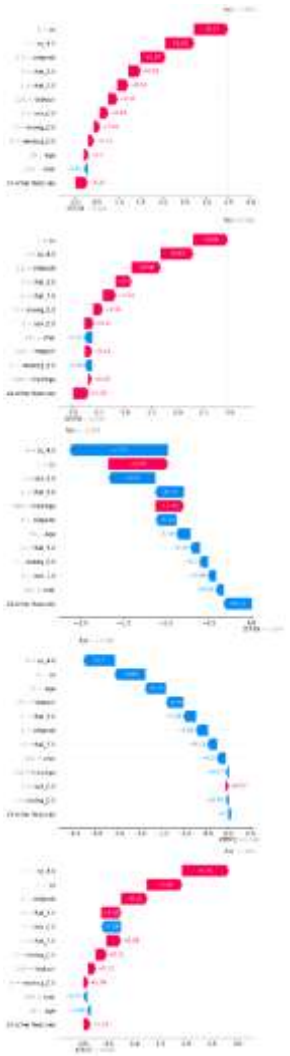
Figure 5 provides patient-level SHAP waterfall explanations for representative test cases, illustrating how different combinations of feature contributions push a prediction toward higher or lower risk. In deployment, such local explanations can support clinician communication and auditing, but they should be interpreted as descriptions of the model's internal logic rather than causal mechanisms. In healthcare, there is substantial concern that post-hoc explanations may increase trust even when the model is biased or invalid, and explanations should not be treated as evidence of clinical causality (Arrieta et al., 2020; Ghassemi et al., 2021). More broadly, some argue that high-stakes decisions should prioritize

inherently interpretable models rather than relying on explaining complex predictors after the fact (Rudin, 2019). For these reasons, we position SHAP as an auditing and communication tool that complements calibration diagnostics and monotonicity verification rather than replacing rigorous validation.



(a) Summary (bar) | **(b)** Beeswarm

Figure 5. Global SHAP explanations for the monotone XGBoost model.



Monotonicity Audit

Monotonic constraints are intended to encode defensible directional behavior, but it remains good practice to empirically verify that the learned model behaves monotonically along constrained dimensions under controlled perturbations (Pei et al., 2016; Wang et al., 2022). We perform a counterfactual sweep audit on constrained features by selecting instances and sweeping one feature across a grid while holding others fixed, then evaluating whether predicted risk is non-decreasing. Figure 6 visualizes representative sweep curves; ideally, curves should be flat or increasing as the swept feature increases, and systematic decreases would indicate monotonicity violations.

Table 3 reports violation rates per constrained feature, where lower values indicate better compliance. In our benchmark, the violation rates are low (and in audited cases may reach zero), supporting that the constraint mapping and training procedure effectively enforce the intended ordering. This audit is valuable for governance because it provides a concrete behavioral check that complements standard performance reporting: a model can be highly discriminative yet fail basic directional expectations, which can impede deployment even if calibration is improved.

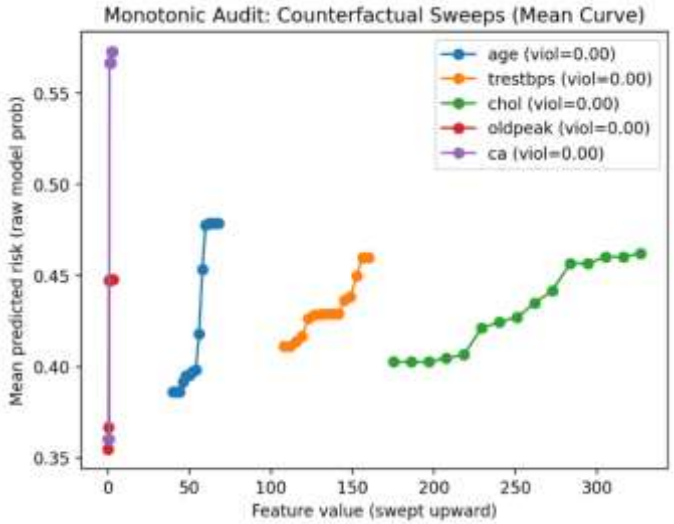


Figure 3. Counterfactual sweeps for constrained features; predicted risk should be non-decreasing as the swept feature.

feature	grid$_{min}$	grid$_{max}$	grid$_{points}$	n$_{samples}$	Violation$_{rate}$
age	40.00	68.00	15	120	0.00
trestbps	108.00	160.00	15	120	0.00
chol	175.10	326.90	15	120	0.00
oldpeak	0.00	3.40	15	120	0.00
ca	0.00	3.00	15	120	0.00

Risk Stratification

To connect calibrated probabilities to operational decision support, we report risk stratification into low, medium, and high probability bands. Figure 7 summarizes the empirical event rate per band on the test set. A useful risk model should concentrate observed events in higher-risk strata and provide band-level summaries that are easy to interpret for triage planning. When calibration is improved, the mean predicted probability within each band should align more closely with the corresponding observed event rate, strengthening the argument that predicted risk can be used as an actionable signal rather than merely a ranking score (Van Calster et al., 2019).

While risk bands provide a practical view, clinical decision-making often requires evaluating the trade-off between true positives and false positives across thresholds. Decision curve analysis is a standard tool to quantify net benefit over threshold ranges (Vickers & Elkin, 2006; Vickers et al., 2019). Although our primary analysis focuses on stratification and reliability, the calibrated probability outputs produced by our pipeline are compatible with decision-analytic evaluation and can be assessed under utility-based criteria in future work or external validations.

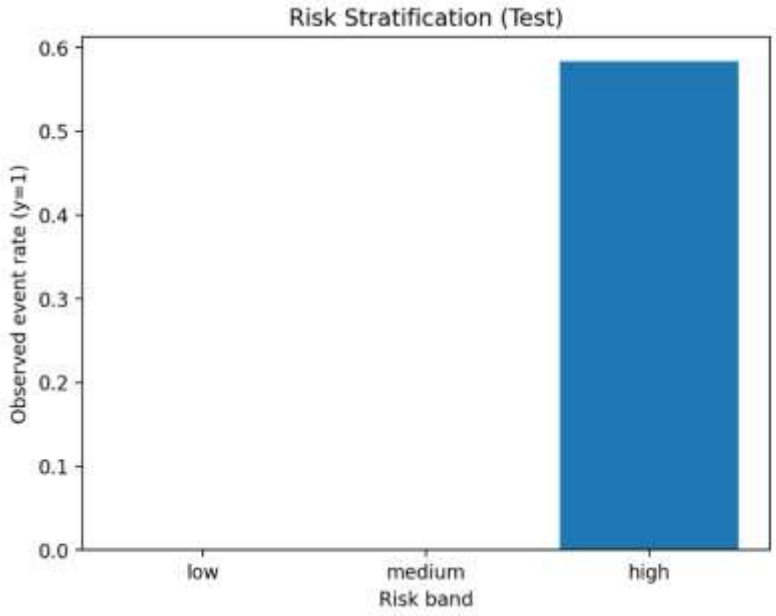


Figure 6. Event rate per risk band on the test set using calibrated probabilities.

Comparison

On small public benchmarks, claiming decisive superiority over diverse prior work is rarely meaningful because results are highly sensitive to preprocessing choices, missing-value handling, categorical encoding, class balance, split strategies, and hyperparameter tuning. These issues are amplified in clinical prediction, where reporting gaps and subtle sources of bias can materially affect conclusions (Moons et al., 2019; Stevenson et al., 2021). For these reasons, rather than presenting the study as a leaderboard-style comparison, we frame our comparison around trust-relevant properties that are directly needed for safe and maintainable clinical deployment: (1) reliable probabilities, (2) clinically coherent directional behavior for ordered variables, and (3) auditable explanations that support governance and communication. This deployment framing also aligns with monitoring-aware operational frameworks in which auditability and rapid response are treated as first-class system requirements.

Tree ensembles are a commonandoften strong choice for structured clinical prediction due to performance and practical deployability (Chen & Guestrin, 2016; Ke et al., 2017; Prokhorenkova et al., 2018). However, two shortcomings regularly limit their decision-support readiness. First, raw predictive scores are frequently miscalibrated, so high discrimination (e.g., AUROC) does not guarantee that predicted probabilities can be trusted for threshold-based actions (Guo et al., 2017; Silva Filho et al., 2023; Van Calster et al., 2019). Second, unconstrained models may exhibit non-intuitive local reversals along variables that clinicians expect to be directionally ordered, reducing face validity and complicating adoption. Prior methodological work shows that monotonicity constraints can encode defensible domain structure within tree-based learners (Pei et al., 2016; Wang et al., 2022), while calibration methods can correct probability scale distortions without materially changing ranking (Guo et al., 2017; Platt, 1999).

Our empirical comparison therefore emphasizes these two deployment-relevant improvements. The ablation study in Table 2 isolates the effect of adding calibration and monotone constraints on top of a strong tree baseline. The results indicate that post-hoc calibration (selected on the validation set) improves probability reliability, as reflected in lower calibration error and improved proper scoring behavior, while retaining discriminative performance. This is consistent with the broader literature noting that calibration is often the “Achilles heel” of predictive analytics in medicine and must be explicitly addressed when probabilities drive interventions (Van Calster et al., 2019). Complementarily, monotone constraints target behavioral coherence rather than pure accuracy: the monotonicity audit (Table 3 and Figure 6) verifies that predicted risk is non-decreasing along constrained

dimensions under counterfactual sweeps, providing an explicit behavioral guarantee that is rarely reported in baseline tree models.

We also note that explainability is commonly requested in healthcare, but explanations should be treated as descriptions of model behavior rather than causal evidence (Arrieta et al., 2020; Ghassemi et al., 2021). In high-stakes decisions, some argue for inherently interpretable models when feasible (Rudin, 2019). In our setting, SHAP is used primarily as an auditing and communication tool for a tree-based pipeline, complementing calibration diagnostics and monotonicity verification (Lundberg & Lee, 2017; Lundberg et al., 2020). This positioning avoids over-claiming that post-hoc explanations by themselves resolve clinical validity concerns.

Finally, while risk stratification into bands provides an operational summary, decision-support evaluation can be expanded beyond fixed bands by decision curve analysis (DCA), which quantifies net benefit across threshold probabilities and is recommended when a model is intended to trigger interventions (Vickers & Elkin, 2006; Vickers et al., 2019). Incorporating DCA is feasible using the exported calibrated probabilities produced by our pipeline. However, DCA is most meaningful when external validation cohorts and clinically grounded utility assumptions are available; therefore, we treat DCA as a natural next step once broader validation data and intervention costs/benefits are defined.

5. CONCLUSIONS

This paper presented a reproducible and CPU-friendly pipeline for explainable clinical risk prediction from public tabular EHR-style data. Using a strong tree-ensemble baseline, we explicitly targeted deployment relevant trust properties rather than maximizing discrimination alone. The proposed approach combines monotone-constrained gradient boosting with validation-driven post-hoc probability calibration, producing risk estimates that are more suitable for threshold-based decision support. Across our benchmark, the full pipeline maintains strong discriminative performance while improving probability reliability, as reflected by reduced calibration error and improved proper scoring behavior. In addition, the monotonicity audit provides a concrete behavioral verification step: under counterfactual sweeps, predicted risk remains non-decreasing along the selected constrained clinical variables, aligning model behavior with basic directional expectations that are often required for governance and clinician acceptance. Finally, SHAP-based global and local explanations provide complementary transparency into how the model forms predictions, supporting

auditing and communication for a tree-based predictor (Lundberg & Lee, 2017; Lundberg et al., 2020).

Limitations. Several limitations constrain the interpretation and generalizability of these findings. First, the study is conducted on a small public benchmark (the processed Cleveland subset), and results are reported under a single stratified split. In such settings, both discrimination and especially calibration estimates can exhibit high variance; therefore, reported reliability improvements should be interpreted cautiously and not treated as sufficient evidence for clinical deployment (Van Calster et al., 2019). Second, no external validation cohort is used, and EHR-based prediction models are known to be sensitive to dataset shift, site-specific coding practices, and selection mechanisms; these factors can introduce bias and degrade real-world performance even when internal test results appear strong (Stevenson et al., 2021). Third, monotone constraints encode a limited form of domain knowledge and are appropriate only when directionality is clinically defensible over the modeled range. If a predictor has a non-monotone or U-shaped relationship with risk, constraining it may distort the learned function; in such cases, the variable should be transformed into a monotone proxy or left unconstrained, and the choice should be justified and reported. Finally, while SHAP explanations improve transparency, they describe model behavior rather than causal effects and may still encourage over-trust if used without rigorous validation and bias assessment (Arrieta et al., 2020; Ghassemi et al., 2021; Rudin, 2019)

Future work. Future work should evaluate this pipeline on larger and more diverse cohorts, ideally across multiple sites, to assess generalization and robustness. Beyond internal validation, external validation and subgroup analyses are essential for understanding performance heterogeneity and potential fairness or applicability issues, in line with established clinical prediction reporting and risk-of-bias guidance (Moonsetal., 2019). In addition, the calibrated probabilities produced by our approach enable decision-analytic evaluation: decision curve analysis can quantify net benefit across clinically relevant threshold probabilities and can better connect prediction quality to real intervention trade-offs (Vickers &Elkin, 2006; Vickers et al., 2019). Finally, robustness studies should investigate sensitivity to missingness mechanisms, measurement noise, and dataset shift, and explore calibration maintenance strategies when operational distributions drift. Together, these steps would move the proposed method from a trustworthy benchmark demonstration toward a clinically actionable and well-governed prediction model.

REFERENCES

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., et al. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review, 78*(1), 1–3. https://doi.org/10.1175/1520-0493(1950)078

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). https://doi.org/10.1145/2939672.2939785

Dua, D., & Graff, C. (2019). *UCI machine learning repository*.

Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health, 3*(11), e745–e750. https://doi.org/10.1016/S2589-7500(21)00208-9

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning* (Vol. 70, pp. 1321–1330). https://proceedings.mlr.press/v70/guo17a.html

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). *LightGBM: A highly efficient gradient boosting decision tree*. In *Advances in Neural Information Processing Systems*. https://arxiv.org/abs/1712.01034

Kull, M., Perello-Nieto, M., Kängsepp, M., Filho, T. S., Song, H., & Flach, P. (2019). *Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with Dirichlet calibration*. In *Advances in Neural Information Processing Systems*. https://arxiv.org/abs/1910.12656

Lundberg, S. M., & Lee, S.-I. (2017). *A unified approach to interpreting model predictions*. In *Advances in Neural Information Processing Systems*. https://arxiv.org/abs/1705.07874

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence, 2*(1), 56–67. https://doi.org/10.1038/s42256-019-0138-9

Moons, K. G. M., Wolff, R. F., Riley, R. D., Whiting, P. F., Westwood, M., Collins, G. S., Reitsma, J. B., Kleijnen, J., & Mallett, S. (2019). PROBAST: A tool to assess risk of bias and applicability of prediction model studies: Explanation and elaboration. *Annals of Internal Medicine, 170*(1), W1–W33. https://doi.org/10.7326/M18-1377

Niculescu-Mizil, A., & Caruana, R. (2006). *Knowledge discovery in the Cleveland heart disease data*. In *Proceedings of the AAAI Workshop on Evaluation Methods for Machine Learning*.

Nixon, J., Dusenberry, M. W., Zhang, L., Jerfel, G., & Tran, D. (2019). *Measuring calibration in deep learning*. *arXiv Preprint*. https://arxiv.org/abs/1904.01685

Pei, S., Xue, B., Liu, H., & Wang, X. (2016). Multivariate decision trees with monotonicity constraints. *Information Sciences, 369*, 178–198. https://doi.org/10.1016/j.ins.2016.06.019

Platt, J. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers* (pp. 61–74). https://doi.org/10.7551/mitpress/1113.003.0008

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). *CatBoost: Unbiased boosting with categorical features*. In *Advances in Neural Information Processing Systems*. https://arxiv.org/abs/1706.09516

Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., Liu, P. J., Liu, X., Marcus, J., Sun, M., et al. (2018). Scalable and accurate deep learning with electronic health records. *npj Digital Medicine, 1*, 18. https://doi.org/10.1038/s41746-018-0029-1

Rudin, C. (2019). Stop explaining black box machine learning models for high-stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*, 206–215. https://doi.org/10.1038/s42256-019-0048-x

Silva Filho, T. M., Song, H., Kull, M., & Flach, P. (2023). Classifier calibration: A review of probabilistic outputs in classification models. *Machine Learning*. https://doi.org/10.1007/s10994-023-06336-7

Stevenson, M. D., et al. (2021). EHR-based clinical prediction models: A systematic review of risks of bias and reporting. *Journal of the American Medical Informatics Association, 28*(8), 1759–1771.

Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L., & Steyerberg, E. W. (2019). Calibration: The Achilles heel of predictive analytics. *BMC Medicine, 17*(1), 230. https://doi.org/10.1186/s12916-019-1466-7

Vickers, A. J., & Elkin, E. B. (2006). Decision curve analysis: A novel method for evaluating prediction models. *Medical Decision Making, 26*(6), 565–574. https://doi.org/10.1177/0272989X06295361

Vickers, A. J., Van Calster, B., & Steyerberg, E. W. (2019). Decision curve analysis for evaluating prediction models: A tutorial. *Medical Decision Making, 39*(5), 583–594. https://doi.org/10.1177/0272989X19855449

Wang, Y., et al. (2022). Monotonic gradient boosting for risk prediction with domain constraints. *IEEE Journal of Biomedical and Health Informatics, 26*(8), 3890–3901.

Yang, W., et al. (2022). Modified Brier score for evaluating prediction accuracy in binary outcomes. *Statistics in Medicine*. https://pmc.ncbi.nlm.nih.gov/articles/PMC9691523/