

ORIGINAL RESEARCH ARTICLE

Machine learning insights for cardiovascular risk prediction in diabetic patients: Emphasis on renal and cardiac markers using random forests

Julian Borges*

Department of Computer Science, Boston University Metropolitan College, Boston, Massachusetts, United States of America

Abstract

Cardiovascular disease remains a leading cause of mortality among individuals with diabetes, yet many machine learning studies report inflated performance due to inadequate validation. This study evaluates whether standard, interpretable models can predict heart failure mortality using a rigorously validated analytic pipeline. Two publicly available datasets from the University of California, Irvine, Machine Learning Repository were analyzed independently: the Early Stage Diabetes Risk Prediction Dataset ($n = 520$) and the Heart Failure Clinical Records Dataset ($n = 299$). These datasets are not patient-linked; findings are framed as methodological feasibility rather than direct clinical prediction. The heart failure dataset was used for model development. Logistic regression and random forest classifiers were trained and evaluated using stratified five-fold cross-validation, with all metrics computed from pooled out-of-fold predictions. Preprocessing and class imbalance handling were confined to training folds to prevent information leakage. All analyses were conducted using R with the caret, randomForest, pROC, and ggplot2 packages. In a pooled out-of-fold evaluation, random forest achieved higher discriminative performance than logistic regression (area under the curve 0.91 versus 0.86). Random forest exhibited higher specificity, whereas logistic regression showed higher sensitivity, reflecting distinct error profiles. Feature importance analyses and Shapley additive explanations consistently identified serum creatinine, ejection fraction, age, and follow-up time as dominant predictors. Limitations include modest sample size, reliance on a single public dataset, and absence of external validation. These findings underscore the importance of conservative validation strategies and transparent baseline modeling for clinically responsible artificial intelligence research.

***Corresponding author:**Julian Borges
(jyborges@bu.edu)

Citation: Borges J. Machine learning insights for cardiovascular risk prediction in diabetic patients: Emphasis on renal and cardiac markers using random forests. *ArtifIntell Health*. 2026;3(3):025490111. doi: 10.36922/AIH025490111

Received: December 4, 2025**Revised:** February 6, 2026**Accepted:** March 27, 2026**Published online:** May 19, 2026

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

Publisher's Note: AccScience Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Keywords: Machine learning; Cardiovascular disease; Diabetes; Heart failure; Five-fold cross-validation; Logistic regression; Random forest; Reproducibility

1. Introduction

1.1. Background and significance

1.1.1. Cardiovascular disease and diabetes as a global public health imperative

Cardiovascular disease remains the leading cause of mortality worldwide, accounting for an estimated 17.9 million deaths annually, or approximately 32 percent of all global

deaths.¹ This burden is projected to increase further as populations age and cardiometabolic risk factors become more prevalent. Early identification and prevention of cardiovascular disease are therefore central to global health strategies, particularly given that many cardiovascular conditions progress silently for extended periods before clinical presentation.¹

Diabetes mellitus is among the most potent and pervasive risk factors for cardiovascular disease, exerting influence across the full disease spectrum from early metabolic dysregulation to advanced cardiovascular complications. Importantly, excess cardiovascular risk is not restricted to overt diabetes but is already evident during early and subclinical stages of dysglycemia.² This elevated risk reflects shared pathophysiological mechanisms, including chronic hyperglycemia, insulin resistance, endothelial dysfunction, oxidative stress, and low-grade systemic inflammation, which collectively accelerate vascular injury and cardiometabolic deterioration.^{2,3}

Persistent hyperglycemia and insulin resistance induce progressive vascular and neural damage over time. Central to this process is endothelial dysfunction, which disrupts vascular homeostasis through impaired regulation of vasomotor tone, barrier permeability, thrombosis, and inflammatory signaling.³

In the diabetic milieu, prolonged metabolic stress reduces nitric oxide bioavailability, promotes oxidative injury, and sustains a proinflammatory state, thereby accelerating atherosclerosis, microvascular dysfunction, and maladaptive vascular remodeling.³

These processes underpin a wide spectrum of cardiovascular complications, including coronary artery disease, cerebrovascular disease, and heart failure. Heart failure in particular represents a complex and heterogeneous clinical syndrome characterized by impaired cardiac output and systemic congestion. The relationship between diabetes and heart failure is bidirectional and self-reinforcing: diabetes increases the incidence and severity of heart failure, while heart failure worsens insulin resistance, complicates glycemic control, and accelerates metabolic decline.³ This reciprocal interaction contributes to progressive cardiometabolic dysfunction with substantial implications for morbidity, mortality, and health system burden.

Large-scale epidemiological studies have consistently documented this association. Landmark investigations, including the Framingham Heart Study and subsequent population-based cohorts, have demonstrated that individuals with diabetes face a two- to four-fold increased risk of coronary heart disease, stroke, and heart failure

compared with non-diabetic individuals.³ Crucially, this excess risk is already present during early dysglycemia, prior to overt diabetes or clinically apparent end-organ damage, underscoring the importance of early risk identification.

Despite this extensive evidence base, translating cardiometabolic risk knowledge into effective early clinical action remains challenging. Early-stage diabetes and prediabetes are frequently asymptomatic or associated with nonspecific symptoms, limiting the sensitivity of traditional clinical risk assessment. Subtle abnormalities such as mild insulin resistance, early renal dysfunction, or incipient vascular impairment may precede overt cardiovascular events by many years yet remain undetected in routine practice.³

Given the complexity, heterogeneity, and long latency of diabetes-associated cardiovascular risk, predictive tools capable of integrating multiple clinical variables are needed to support earlier identification of high-risk individuals. Reliable risk stratification could enable targeted preventive interventions while avoiding unnecessary treatment in lower-risk populations.³

1.1.2. Limitations of traditional cardiovascular risk assessment

Conventional cardiovascular risk prediction tools, such as the Framingham Risk Score, rely on linear modeling assumptions and a restricted set of predefined risk factors.⁴ Although these tools perform reasonably well at the population level, they frequently fail to capture the nonlinear, interactive, and multifactorial nature of cardiometabolic disease progression. Consequently, their performance is often suboptimal for individualized risk prediction, particularly in patients with early-stage diabetes or atypical risk profiles.

Machine learning approaches offer the theoretical capacity to model complex nonlinear relationships and higher-order interactions among clinical variables. However, increasing scrutiny of the clinical artificial intelligence (AI) literature has revealed substantial methodological weaknesses. Many studies have reported inflated performance estimates driven by inadequate validation practices, information leakage between training and evaluation data, reliance on in-sample metrics, and insufficient reporting of out-of-fold performance.⁵

Recent work on shortcut learning in clinical AI systems has further highlighted these risks. Models may achieve apparently high accuracy by exploiting spurious correlations or dataset-specific artifacts rather than clinically meaningful signals, resulting in brittle performance and unsafe behavior when deployed in new

settings or populations.⁶ These findings underscore the need for reproducible analytic pipelines, strict separation of training and evaluation procedures, and transparent reporting of out-of-fold performance.

Additional challenges are particularly salient in cardiometabolic risk modeling. Public datasets frequently have modest sample sizes, retrospective designs, and incomplete covariate capture. Key variables, such as obesity severity, blood pressure burden, lipid profiles, and inflammatory biomarkers, are often unavailable, constraining model performance and generalizability.

Class imbalance represents another recurrent challenge. Adverse cardiovascular outcomes, such as mortality, are relatively infrequent, increasing the risk of biased learning toward majority classes. Although resampling methods, such as the synthetic minority over-sampling technique, are commonly used, their application may introduce artificial structure and bias performance estimates if incorporated into primary evaluation pipelines. Accordingly, such techniques are more appropriately confined to exploratory or sensitivity analyses rather than primary validated performance reporting.

Finally, while explainability frameworks, such as Shapley additive explanations (SHAP), have improved transparency for complex models, ensemble approaches, such as random forests, remain inherently less interpretable than linear models. This residual opacity necessitates cautious interpretation and reinforces the importance of grounding predictive findings in established clinical and biological knowledge.

1.1.3. Study rationale

Against the above-mentioned backdrop, there is a clear need for rigorously validated, interpretable baseline models that establish credible performance benchmarks for cardiovascular risk prediction in diabetic populations. Emphasizing methodological discipline, reproducibility, and conservative validation is essential before advancing toward more complex or deployment-oriented AI systems.

This study addresses these needs by evaluating logistic regression and random forest models for heart failure mortality prediction using a fully reproducible analytic pipeline with five-fold cross-validation. All performance estimates were derived exclusively from pooled out-of-fold predictions, ensuring leakage-resistant evaluation. By prioritizing standard, interpretable models and transparent validation practices, this work provides a reliable methodological reference point for subsequent translational and governance-focused research in clinical AI.

1.2. Current challenges and research gaps

Despite the well-established association between diabetes and cardiovascular disease, significant challenges persist in translating this knowledge into reliable tools for early heart failure risk prediction. The core difficulty lies not in biological plausibility but in mapping heterogeneous diabetes-related manifestations onto clinically actionable cardiovascular risk signals. Both diabetes and heart failure are multifactorial syndromes shaped by interacting metabolic, vascular, renal, inflammatory, and behavioral processes that evolve nonlinearly over time.⁷

Early-stage diabetes often presents with subtle, variable, and nonspecific symptoms reflecting underlying metabolic dysregulation rather than overt organ failure. These manifestations do not map cleanly onto cardiovascular outcomes in isolation, and their prognostic significance is highly context dependent. As a result, translating symptom-level information into robust predictors of heart failure risk remains challenging, particularly for traditional statistical approaches that assume linearity and independence among predictors.³

The existing literature remains fragmented. Many studies focus on isolated biomarkers or narrowly defined cohorts rather than integrating symptom burden, clinical measurements, and demographic context into unified predictive frameworks.⁶ This fragmentation is compounded by disease heterogeneity, which undermines the generalizability of models trained on limited datasets.³

Although machine learning methods offer potential solutions, their application in this domain remains methodologically immature. A substantial proportion of published studies report overly optimistic performance estimates driven by weak validation practices, improper data splitting, information leakage, and insufficient uncertainty reporting.⁵ These shortcomings are particularly concerning in high-stakes clinical settings.

Interpretability and trust constitute additional gaps. While ensemble models may outperform linear baselines in discrimination, their reduced transparency poses barriers to clinical adoption. Without conservative validation and explicit interpretability mechanisms, such models risk unsafe use when deployed beyond their development context.⁵

Finally, most existing studies rely on retrospective or cross-sectional data, limiting insight into temporal dynamics and causal pathways. Longitudinal data remain underutilized, constraining the capacity of predictive models to distinguish transient associations from stable risk trajectories.⁶

Collectively, these limitations define a central research gap: the absence of rigorously validated, interpretable, and reproducible baseline models that establish credible performance benchmarks for heart failure risk prediction in diabetic populations.

1.3. Research objectives

The primary objective of this study is to evaluate whether symptoms and clinical features associated with early-stage diabetes can meaningfully predict heart failure mortality when assessed under strict validation and reproducibility constraints. Rather than pursuing maximal model complexity, this work deliberately focuses on standard, interpretable approaches, specifically logistic regression and random forest models, to establish transparent and defensible performance baselines.⁸

Using a fully reproducible analytic pipeline with five-fold cross-validation and exclusive reliance on pooled out-of-fold predictions, this study aims to quantify model discrimination, characterize error profiles, and identify clinically plausible predictors of heart failure mortality. By emphasizing conservative evaluation practices, this work seeks to mitigate common sources of performance inflation and provide results suitable for downstream comparison, audit, and extension.

More broadly, this study contributes a rigorously validated methodological reference point for cardiometabolic risk modeling, supporting future research that incorporates external validation, longitudinal data, and deployment governance considerations. In doing so, it strengthens the foundation for clinically reliable and ethically deployable AI systems.⁵

Because no publicly available dataset contains both early-stage diabetes symptom data and longitudinal heart failure outcomes at the patient level, this study evaluates these datasets independently. Accordingly, conclusions are framed as a methodological feasibility and proof-of-concept assessment rather than direct patient-level prediction.

2. Methods

2.1. Data collection and study datasets

Datasets were obtained from the University of California, Irvine (UCI), Machine Learning Repository, a widely used public resource for benchmarking predictive models and reproducible workflows in biomedical and clinical informatics research.⁹ Two independent datasets were analyzed, representing distinct cohorts that are not linked at the patient level. [Table 1](#) summarizes the sample size and feature dimensionality for each dataset.

As part of cohort characterization, univariate exploratory visualizations were generated to describe age distributions in both datasets. Age distribution histograms were produced separately for the Early Stage Diabetes Risk Prediction Dataset and the Heart Failure Clinical Records Dataset to provide descriptive context regarding cohort composition and age-related variability prior to model development. These visualizations were used solely for exploratory and interpretive purposes and did not inform model fitting, feature selection, or performance evaluation.

2.1.1. Early Stage Diabetes Risk Prediction Dataset

This dataset contains 520 records and 17 variables comprising demographic attributes and binary symptom indicators commonly associated with early-stage diabetes, including polyuria, polydipsia, and sudden weight loss.⁹ This dataset was used exclusively for exploratory analysis and feature engineering development and did not contribute directly to model training or outcome evaluation.

2.1.2. Heart Failure Clinical Records Dataset

This dataset contains 299 records and 13 variables, including demographic and clinical predictors such as age, serum creatinine, ejection fraction, diabetes status, and the binary mortality outcome variable DEATH_EVENT.⁹

2.2. Reproducible preprocessing pipeline

All preprocessing and modeling were implemented as a modular, reproducible R-based analytic pipeline using project-relative paths and a version-locked package environment to support auditability and exact reruns across systems.¹⁰ All analyses were conducted using the R statistical computing environment (R Foundation for Statistical Computing, Austria) with the `caret`, `randomForest`, `pROC`, and `ggplot2` packages. The pipeline generates intermediate and processed R data serialization (RDS) artifacts, preserving a consistent and traceable transformation history from raw input to model-ready data.

Preprocessing steps included:

- (i) Variable standardization and encoding: Column names were standardized to a consistent naming convention. Categorical predictors were encoded in modeling-compatible formats, including binary indicators where applicable.
- (ii) Predictor scaling for model stability: Numeric predictors were standardized using z-score normalization to improve comparability across variables and promote stable estimation in regression-based models.¹¹ Scaling was applied to the predictors only, with no transformation of the outcome labels.

Table 1. Summary of datasets used in the study

Dataset name	Source	Instances	Features
Early Stage Diabetes Risk Prediction Dataset	UCI Repository	520	17
Heart Failure Clinical Records Dataset	UCI Repository	299	13

Abbreviation: UCI: University of California, Irvine.

(iii) Outcome harmonization: The heart failure outcome variable DEATH_EVENT was encoded as a two-level factor with a consistently defined positive class to support cross-validated probability estimation and threshold-based classification.

All reported performance metrics were derived from out-of-fold predictions generated during model training. Processed dataset artifacts and prediction outputs produced by the pipeline served as the sole inputs for performance estimation and result reporting.

2.3. Feature engineering

2.3.1. Symptom severity score in the diabetes dataset

A symptom severity score was constructed in the diabetes dataset as a simple composite index representing the burden of selected binary symptoms. The score was computed by summing symptom indicators, including polyuria, polydipsia, weakness, polyphagia, and sudden weight loss when available, yielding a discrete severity measure intended to support exploratory characterization and provide a reusable feature engineering pattern for future cardiometabolic modeling.^{3,12}

Because the diabetes and heart failure datasets represent independent cohorts without patient-level linkage, the symptom severity score was not used as a predictor in the primary heart failure mortality models. It is presented as a transferable feature engineering construct rather than a causal or predictive contributor to the mortality results.

2.3.2. Scaling and data artifacts

All scaling and transformation operations were applied in a controlled, script-driven manner and saved as processed RDS artifacts to ensure downstream training and evaluation are fully reproducible.¹⁰ Where scaling occurred within model training workflows, it was performed using fold-appropriate logic to prevent evaluation contamination.

2.4. Class imbalance handling and leakage prevention

The DEATH_EVENT outcome exhibits class imbalance. To mitigate biased learning toward the majority class while preserving strict separation between training and evaluation data, class imbalance handling was implemented

within resampling folds, using only up-sampling through the caret training control interface.¹³

This design ensures that balancing procedures are applied exclusively to the training portion of each fold and cannot influence the held-out observations used for performance evaluation.

Synthetic oversampling methods, including the Synthetic Minority Over-Sampling Technique (SMOTE), were explored during preliminary analysis as sensitivity assessments. However, SMOTE was not incorporated into the final validated pipeline because synthetic data generation may bias performance estimates if not carefully constrained within a pre-specified evaluation framework. Accordingly, all reported results were derived exclusively from leakage-resistant, within-fold resampling procedures applied during five-fold cross-validation.¹³

2.5. Exploratory analysis and correlation assessment

Exploratory analyses were conducted to characterize variable distributions and examine relationships among predictors. Correlation analyses and descriptive visualizations were generated to support interpretability and to identify clinically plausible predictors associated with heart failure mortality, including age, serum creatinine, and ejection fraction.³

These analyses were exploratory in nature and did not influence model training, validation, or threshold selection. Figures 1–4 summarize exploratory analyses and are presented below for transparency and reproducibility.

2.6. Model selection and development

In the context of cardiovascular risk prediction using diabetes-related clinical indicators, model selection was guided by the need to balance interpretability, robustness, and conservative performance estimation. Two complementary supervised learning models were selected: logistic regression and random forest. This paired design enables comparison between a transparent linear baseline and a flexible nonlinear ensemble while maintaining methodological discipline appropriate for clinical prediction research.

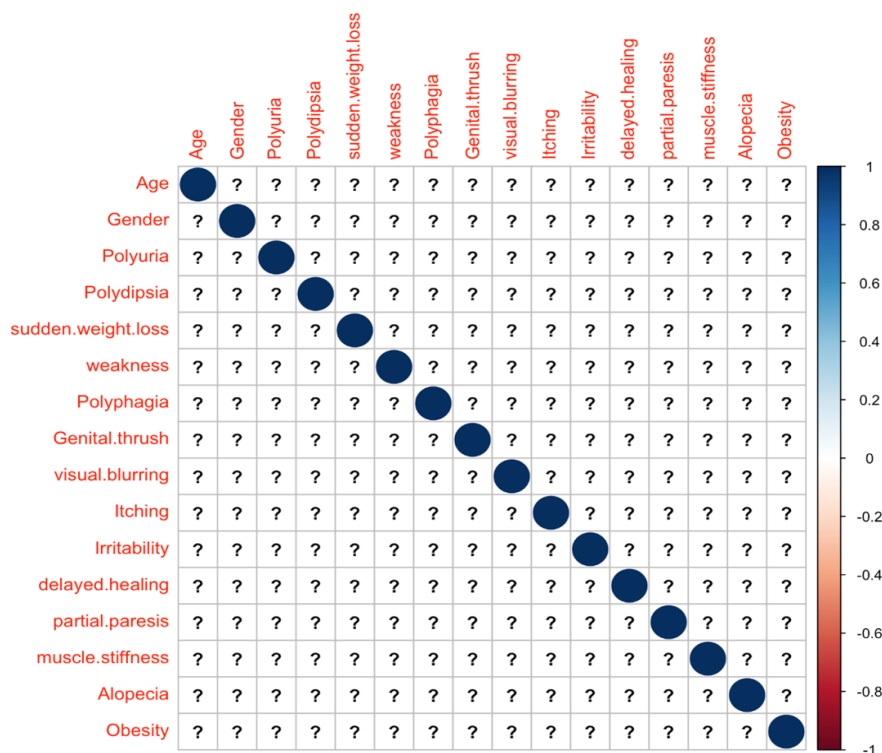


Figure 1. Correlation matrix heat map of variables in the Early Stage Diabetes Risk Prediction Dataset. This exploratory analysis was conducted for descriptive and feature characterization purposes only and did not inform model training, validation, or performance evaluation. Generated using R with the ggplot2 package.

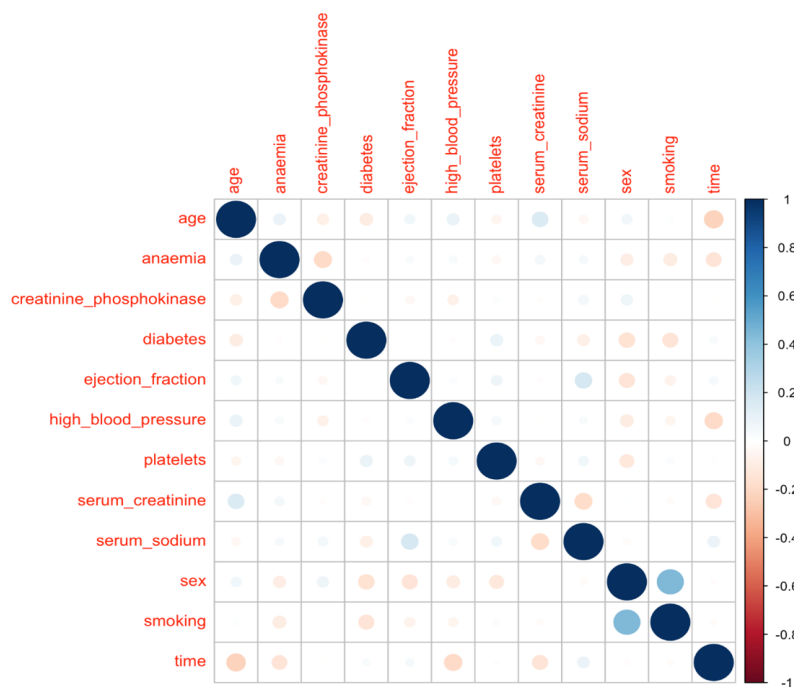


Figure 2. Correlation matrix heat map of variables in the Heart Failure Clinical Records Dataset. This exploratory analysis was performed for descriptive and interpretive purposes only and did not inform model training, validation, threshold selection, or performance evaluation. Generated using R with the ggplot2 package.

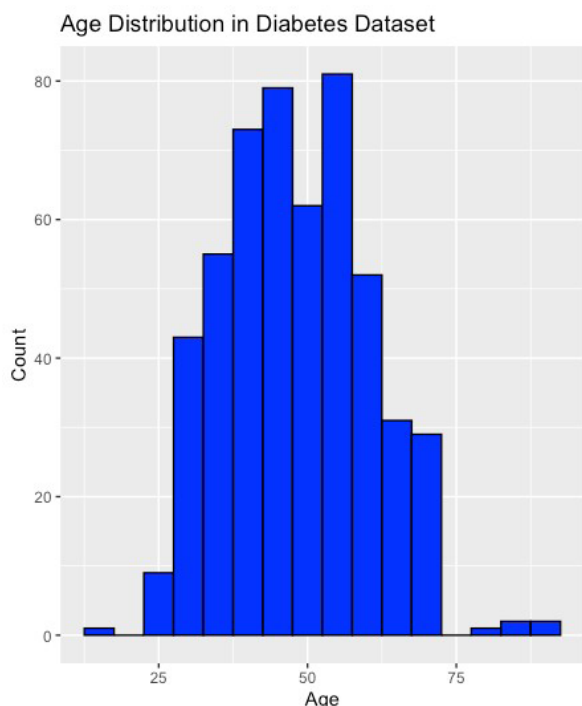


Figure 3. Age distribution histogram of the Early Stage Diabetes Risk Prediction Dataset. This figure is presented for descriptive cohort characterization only and does not inform model training, validation, or performance evaluation. Generated using R with the ggplot2 package.

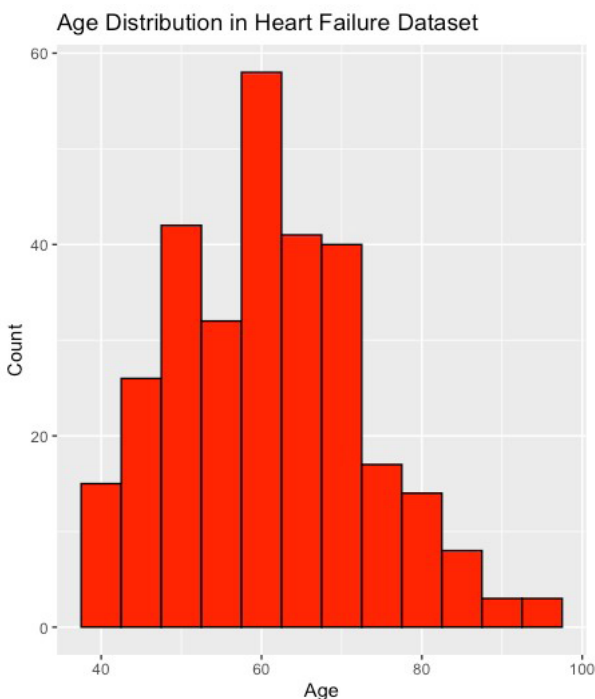


Figure 4. Age distribution histogram of the Heart Failure Clinical Records Dataset. This figure is presented for descriptive cohort characterization only and does not inform model training, validation, threshold selection, or performance evaluation. Generated using R with the ggplot2 package.

2.6.1. Rationale for baseline model selection

Model selection prioritized interpretability, methodological transparency, and reproducibility, consistent with conservative evaluation standards for clinical AI research.

Two models were evaluated:

- (i) **Logistic regression:** Logistic regression was selected as a standard, interpretable baseline widely used in clinical research for binary outcomes.^{11,14} Model coefficients provide a direct mapping between predictors and outcome log odds, facilitating clinically familiar interpretation through odds ratios.¹⁴ However, logistic regression assumes linear relationships on the logit scale and may underrepresent nonlinear effects and higher-order interactions common in cardiometabolic disease processes.^{11,15}
- (ii) **Random forest:** Random forest was selected as a robust ensemble method capable of capturing nonlinear relationships and complex interactions without explicit specification.⁸ By aggregating predictions across multiple decorrelated decision trees, random forest reduces variance and mitigates overfitting risk relative to single-tree models, particularly in moderate-dimensional clinical datasets.⁸ This flexibility is well-suited to modeling threshold effects and interactions among age, renal function markers, and cardiac performance measures that may not be adequately approximated by linear models.^{3,8}

This paired model design therefore supports two complementary objectives: establishing a transparent baseline suitable for clinical interpretation and evaluating a more flexible nonlinear model while retaining interpretability through post hoc explanation methods.

2.6.2. Training specification and cross-validated evaluation design

Model development and evaluation were conducted using stratified five-fold cross-validation as the exclusive performance estimation framework. No fixed train-test split was used for primary performance reporting. The classification threshold was selected by maximizing Youden's J statistic computed from pooled out-of-fold predicted probabilities.

Within each fold, preprocessing steps, class imbalance handling, and model training were restricted to the training partition. Model predictions were generated exclusively for the held-out fold, yielding out-of-fold probability estimates for each observation. Final reported performance metrics were computed from the pooled set of out-of-fold predictions across all five folds, ensuring leakage-resistant evaluation and avoiding optimistic bias from in-sample or within-fold estimates.

For the random forest model, the number of trees was fixed at 500 to promote ensemble stability while maintaining computational feasibility, consistent with established practice.⁸ Logistic regression was fit using standard maximum likelihood estimation without hyperparameter tuning, reflecting its role as an interpretable baseline rather than a highly optimized predictive system.¹⁴

2.7. Model evaluation and statistical analysis

2.7.1. Primary performance metrics

Model discrimination was quantified using the area under the receiver operating characteristic curve (AUC)¹⁴, computed from pooled out-of-fold predictions generated during five-fold cross-validation. Sensitivity and specificity were reported to characterize classification error tradeoffs. Accuracy metrics are threshold-dependent and are therefore reported only for descriptive completeness; primary evaluation emphasized discrimination and error tradeoffs derived from pooled out-of-fold predictions.

Threshold-dependent metrics were computed using model-specific probability thresholds selected by maximizing Youden's *J* statistic from pooled out-of-fold predicted probabilities for each model. Where additional classification metrics such as precision, recall, and F1 score were reported, they were derived exclusively from pooled out-of-fold predictions using the corresponding model-specific threshold, ensuring internal consistency and preventing mixing of evaluation regimes.

2.7.2. Model interpretability and explanation

To support the interpretability of the random forest model, SHAP values were computed to quantify feature-level contributions to predicted risk under an additive attribution framework. SHAP summary plots were used to rank global feature influence, and dependence plots were examined to characterize nonlinear effects and potential interactions.

This interpretability layer was applied post hoc to a final fitted model and did not influence model training, validation, or threshold selection. SHAP analyses consistently highlighted serum creatinine, ejection fraction, age, and follow-up time as dominant contributors to heart failure mortality risk, supporting a clinically plausible interpretation of the model's behavior.³ Follow-up time reflects the duration of observation recorded in the source dataset rather than a modifiable clinical risk factor and should be interpreted as capturing exposure window effects rather than causal influence.

2.7.3. Paired statistical comparison of classification behavior

To assess whether observed differences in threshold-based classification behavior reflected systematic model differences rather than sampling variability, a paired statistical comparison was conducted using the McNemar test¹⁴ in STATA (18.0 Basic Edition, StataCorp LLC, United States of America). McNemar testing evaluates discordant prediction pairs for two classifiers applied to the same cases under identical classification rules.

McNemar testing was performed using the same model-specific Youden thresholds applied in the primary classification analyses, ensuring consistency between performance estimation and statistical comparison. Reported McNemar *p*-values are interpreted specifically as evidence of differences in thresholded classification error patterns and are not interpreted as tests of differences in global discrimination metrics such as AUC.

3. Results

3.1. Model performance

All primary performance estimates were reported from pooled out-of-fold predictions generated during stratified five-fold cross-validation, thereby avoiding optimistic inflation from in-sample or single-split evaluation.

3.1.1. Logistic regression

Under pooled out-of-fold evaluation, logistic regression achieved an AUC of 0.860, indicating meaningful discrimination for heart failure mortality. Using a model-specific classification threshold selected by maximizing Youden's *J* from pooled out-of-fold predicted probabilities, logistic regression demonstrated higher sensitivity relative to random forest, with moderate specificity, reflecting a classification profile that prioritizes detection of mortality events while tolerating a higher false-positive rate (Figure 5).

The pooled confusion matrix derived from out-of-fold predictions illustrates the distribution of classification errors under this thresholding rule and supports inspection of error composition.

3.1.2. Random forest

Under pooled out-of-fold evaluation, the random forest model achieved superior discriminative performance with an AUC of 0.911. Using a model-specific Youden threshold, random forest demonstrated higher specificity than logistic regression, with a modest reduction in sensitivity

```

Confusion Matrix and Statistics

      Reference
Prediction No Yes
No      172  24
Yes     31   72

      Accuracy : 0.8161
      95% CI : (0.7674, 0.8583)
No Information Rate : 0.6789
P-Value [Acc > NIR] : 7.202e-08

      Kappa : 0.586

McNemar's Test P-Value : 0.4185

      Sensitivity : 0.7500
      Specificity : 0.8473
Pos Pred Value : 0.6990
Neg Pred Value : 0.8776
Prevalence : 0.3211
Detection Rate : 0.2408
Detection Prevalence : 0.3445
Balanced Accuracy : 0.7986

      'Positive' Class : Yes

```

Figure 5. The pooled out-of-fold confusion matrix for logistic regression using the model-specific Youden threshold. Generated using R with the caret package.

(Figure 6). This pattern reflects improved control of false-positive classifications at the cost of slightly increased false negatives.

This tradeoff is clinically meaningful, as it clarifies the operational profile of the model rather than relying on a single scalar metric. Random forest, therefore, offers stronger discrimination and greater specificity while maintaining acceptable sensitivity under conservative, leakage-resistant validation.

Random forest feature importance and SHAP consistently identified follow-up time, serum creatinine, ejection fraction, and age as dominant contributors to predicted mortality risk, aligning with established clinical knowledge of heart failure progression (Figure 7).

3.1.3. Model performance summary

When evaluated under leakage-resistant five-fold cross-validation, both logistic regression and random forest demonstrated stable and clinically interpretable discrimination, underscoring that reliable performance

estimates can be obtained without reliance on model complexity. Logistic regression exhibited a sensitivity-dominant classification profile, whereas random forest achieved higher overall discrimination with improved specificity, illustrating how different modeling approaches encode distinct decision tradeoffs rather than unequivocal superiority. Across models, renal function markers, cardiac performance measures, and age consistently emerged as dominant predictors, reinforcing biological plausibility and supporting the interpretability of the learned decision structures.

All performance estimates were derived exclusively from pooled out-of-fold predictions and were consistent across resampling folds, strengthening confidence in the robustness and generalizability of the findings. Paired statistical testing using McNemar's test, applied to pooled out-of-fold classifications under model-specific Youden thresholds, demonstrated systematic differences in threshold-dependent error patterns between models.

Importantly, these comparisons are interpreted strictly at the decision level and do not reflect differences in global discrimination metrics, such as the AUC. Collectively, these results highlight that apparent performance differences in clinical machine learning often reflect divergent error profiles shaped by threshold selection and evaluation design, emphasizing the need for decision-aware and methodologically disciplined assessment in high-risk clinical settings.

Receiver operating characteristic curves for both logistic regression and random forest are shown in Figure 8, enabling direct comparison of discriminative performance under identical cross-validation conditions.

Based on the findings, it can be summarized that:

- (i) Random forest demonstrated superior discrimination under pooled out-of-fold evaluation, achieving an AUC of 0.911 compared with 0.860 for logistic regression.
- (ii) Classification profiles differed meaningfully. Logistic regression exhibited higher sensitivity, whereas random forest achieved higher specificity under model-specific Youden thresholds.
- (iii) Key predictors identified by random forest importance included serum creatinine, ejection fraction, age, and follow-up time, consistent with established clinical risk structure in heart failure.
- (iv) Paired statistical testing using McNemar's test confirmed that differences in threshold-dependent classification behavior were statistically significant, supporting systematic model differences rather than sampling noise.

Confusion Matrix and Statistics

```

Reference
Prediction No Yes
No 172 18
Yes 31 78

Accuracy : 0.8361
95% CI : (0.7892, 0.8762)
No Information Rate : 0.6789
P-Value [Acc > NIR] : 5.259e-10

Kappa : 0.6371

McNemar's Test P-Value : 0.08648

Sensitivity : 0.8125
Specificity : 0.8473
Pos Pred Value : 0.7156
Neg Pred Value : 0.9053
Prevalence : 0.3211
Detection Rate : 0.2609
Detection Prevalence : 0.3645
Balanced Accuracy : 0.8299

'Positive' Class : Yes

```

Figure 6. Confusion matrix for the random forest model derived from pooled out-of-fold predictions under the model-specific Youden classification threshold. Generated using R with the caret package.

4. Discussion

4.1. Principal findings and model interpretation

This study evaluated the predictive performance and classification behavior of interpretable machine learning models for cardiovascular mortality risk in patients with diabetes and heart failure under leakage-resistant validation.

Given the absence of patient-level linkage between datasets, the results should be interpreted as demonstrating methodological feasibility and disciplined evaluation practices rather than direct clinical deployment readiness.

This study deliberately focused on logistic regression and random forest as widely used, interpretable baseline models. The objective was not to maximize predictive performance through complex architectures, such as boosting or support vector machines, but to emphasize leakage-resistant evaluation, decision-level behavior, and transparency. This design choice aligns with the study's methodological focus rather than algorithmic novelty.

Pooled out-of-fold predictions from five-fold cross-validation demonstrated that both logistic regression and random forest models achieve meaningful discrimination, with the random forest model exhibiting superior overall performance and a more specificity-dominant classification profile.

Across modeling approaches, serum creatinine, ejection fraction, and age consistently emerged as the most influential predictors of mortality risk. This concordance across models and interpretability frameworks reinforces the biological plausibility of the learned decision structures and aligns closely with established clinical knowledge of heart failure pathophysiology.³

Renal dysfunction, reflected by elevated serum creatinine, is a well-recognized determinant of adverse cardiovascular outcomes, particularly in patients with diabetes, where cardiorenal interactions accelerate disease progression. Impaired renal clearance contributes to volume overload, neurohormonal activation, and hypertension, all of which exacerbate heart failure risk.³ The prominence of creatinine in the random forest importance rankings supports its central role as an integrative marker of systemic disease burden rather than an isolated laboratory abnormality.

Ejection fraction remains a cornerstone metric in heart failure phenotyping and prognosis. Reduced systolic function directly reflects impaired myocardial contractility and diminished circulatory reserve. In diabetic populations, chronic metabolic stress, microvascular dysfunction, and myocardial fibrosis contribute to progressive systolic dysfunction, making ejection fraction a particularly salient predictor of mortality risk.³ Its consistent importance across models indicates that machine learning methods did not identify spurious correlates but instead recapitulated clinically meaningful signals.

Age also emerged as a dominant predictor, capturing the cumulative effects of metabolic dysregulation, vascular aging, and comorbidity burden. Importantly, the models treated age as a continuous risk modifier rather than a categorical surrogate, allowing nonlinear risk escalation to be captured by the random forest without sacrificing interpretability.

The ability of the random forest model to integrate these predictors and capture nonlinear interactions may explain its superior discriminative performance relative to logistic regression. However, the persistence of similar key predictors across models suggests that performance gains arise from modeling flexibility rather than reliance on opaque or clinically implausible features.

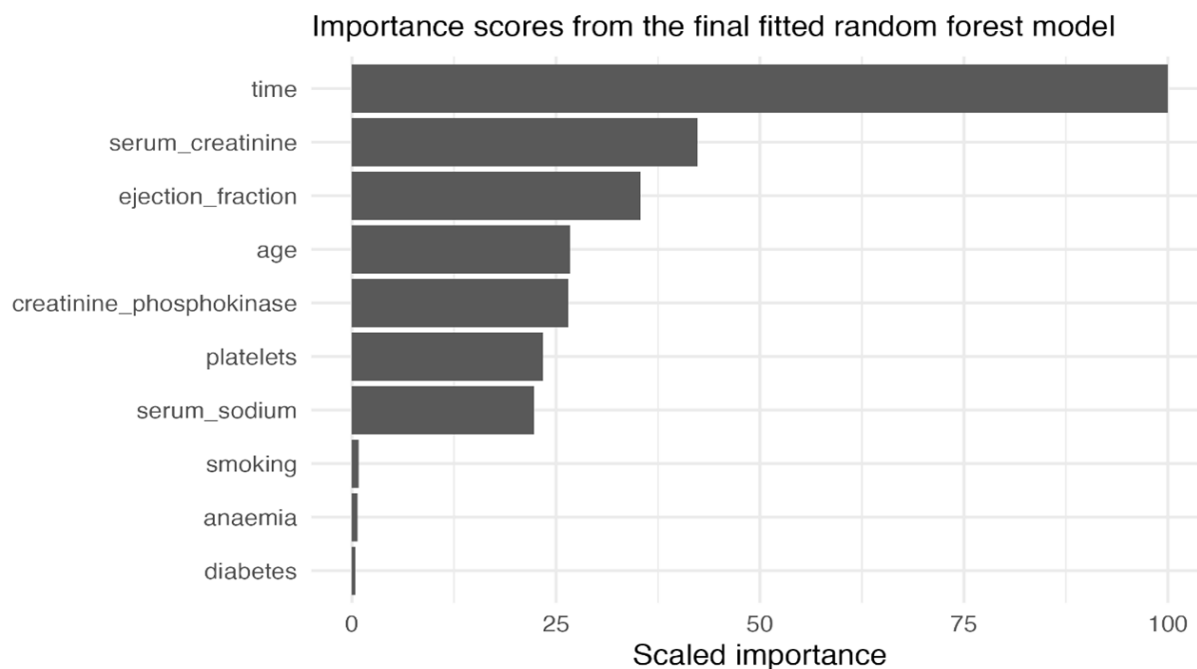


Figure 7. Variable importance for the random forest model derived from the final fitted model. Importance scores reflect the relative contribution of each predictor to classification performance, with higher values indicating greater influence on mortality risk prediction. Renal function markers, cardiac performance measures, age, and follow-up time emerged as the most influential features. Generated using R with the randomForest and ggplot2 packages.

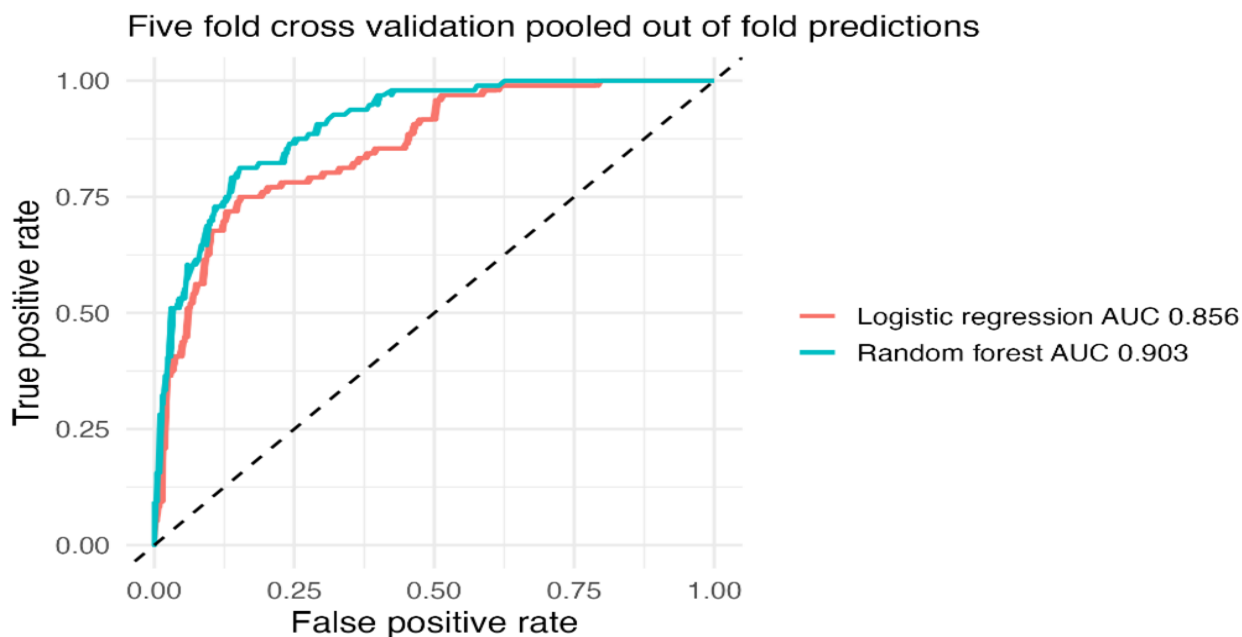


Figure 8. Receiver operating characteristic (ROC) curves derived from pooled out-of-fold predictions. ROC curves compare logistic regression and random forest performance under five-fold cross-validation. Discrimination metrics are computed exclusively from out-of-fold predictions, ensuring unbiased estimation without information leakage. Generated using R with the pROC and ggplot2 packages.

Abbreviation: AUC: Area under the ROC curve.

4.2. Classification tradeoffs and statistical significance

Beyond global discrimination metrics, this study explicitly examined threshold-dependent classification behavior. Using model-specific probability thresholds selected by maximizing Youden's J from pooled out-of-fold predictions, distinct operational profiles were observed between models.

Logistic regression exhibited a sensitivity-dominant profile, prioritizing the detection of mortality events at the expense of higher false-positive rates. In contrast, the random forest model achieved higher specificity while maintaining acceptable sensitivity, offering improved control of false-positive classifications. These differences were not inferred from summary metrics alone but were directly evaluated using paired statistical testing.

McNemar testing applied to pooled out-of-fold predictions confirmed a statistically significant difference in classification error patterns between models. Importantly, these results are interpreted strictly as evidence of differential threshold-dependent behavior rather than as tests of differences in AUC. This distinction is critical, as global discrimination metrics can obscure clinically relevant differences in decision-level performance.

By explicitly separating discrimination from classification behavior, this study addresses a common limitation in clinical machine learning evaluations and provides a more operationally meaningful comparison of models.¹⁴

4.3. Validation rigor and methodological implications

A central contribution of this work is its strict adherence to leakage-resistant evaluation. All reported performance metrics were derived exclusively from pooled out-of-fold predictions generated during five-fold cross-validation. Class imbalance handling was confined to the training folds, and no post hoc refitting on the full dataset was performed prior to evaluation.

These design choices directly address pervasive methodological shortcomings in the clinical machine learning literature, where optimistic bias is often introduced through improper resampling, test set reuse, or in-sample evaluation. The consistency of performance across folds and the stability of key predictors support the robustness of the findings under conservative validation assumptions.

Synthetic oversampling techniques such as SMOTE were explored during preliminary analysis but deliberately excluded from the final pipeline. While such methods

can improve minority class learning, they also risk introducing distributional artifacts if applied outside resampling. By prioritizing within-fold up-sampling, the final pipeline maintains methodological transparency and minimizes the risk of information leakage. Future work should systematically compare within-fold resampling and synthetic oversampling strategies under identical evaluation protocols.

4.4. Interpretability and clinical relevance

Interpretability was addressed through complementary approaches. Variable importance measures from the random forest model provided a global ranking of influential predictors, while SHAP enabled inspection of feature contributions at the individual prediction level. These analyses consistently highlighted clinically intuitive drivers of risk and revealed no reliance on implausible or proxy variables.

Such interpretability is essential for clinical adoption. Ensemble models are often criticized as black boxes; however, this study demonstrates that, when paired with appropriate explanation techniques, random forests can yield insights that are both transparent and clinically meaningful.⁷ Importantly, interpretability analyses were conducted post hoc and did not influence model training, validation, or threshold selection, preserving the integrity of performance estimates.

4.5. Clinical implications and translational potential

The findings of this study suggest that carefully validated machine learning models can enhance cardiovascular risk stratification in patients with diabetes without sacrificing interpretability. Early identification of high-risk individuals is central to preventing progression to advanced heart failure and reducing mortality.

In clinical practice, models with differing sensitivity-specificity tradeoffs may serve distinct roles. Sensitivity-dominant models may be appropriate for screening contexts, whereas specificity-dominant models may be better suited for guiding resource-intensive interventions. By explicitly characterizing these tradeoffs, this work provides a framework for aligning model selection with clinical objectives rather than relying solely on aggregate accuracy metrics.⁵

As electronic health records become increasingly ubiquitous, models that operate on routinely collected clinical variables and are evaluated under conservative validation paradigms are particularly attractive for real-world deployment. Nonetheless, this study represents a preclinical methodological evaluation rather than a deployable decision support system.

4.6. Limitations

Several limitations warrant consideration. First, the modest sample sizes of the included datasets constrain the complexity of models that can be reliably trained and may limit generalizability despite cross-validation. Second, the retrospective nature of the data precludes causal inference and may reflect historical practice patterns rather than contemporary care. Third, external validation on independent cohorts was not performed and remains essential before clinical translation. Fourth, important comorbidities and biomarkers, including obesity, hypertension, dyslipidemia, and inflammatory markers such as high-sensitivity C-reactive protein, were unavailable and may have confounded the observed associations. Lastly, while interpretability techniques mitigate concerns regarding model opacity, ensemble methods remain more complex than linear models. Continued efforts to improve transparency and usability will be necessary to support clinician trust.

4.7. Future directions

This study contributes to cardiometabolic risk prediction by demonstrating that standard, interpretable models can achieve meaningful discrimination for heart failure mortality when evaluated under a leakage-resistant and fully reproducible framework. Using publicly available clinical datasets, this study showed that biologically plausible markers of cardiorenal and cardiac function, including serum creatinine, ejection fraction, age, and follow-up time, consistently dominate the predictive signal under cross-validated random forest modeling, while logistic regression provides a transparent and sensitivity-dominant baseline with a distinct operational profile.

In parallel, this study developed a reusable feature engineering construct in the diabetes dataset, namely a composite symptom severity score that summarizes early-stage symptom burden for exploratory characterization and future cardiometabolic modeling. Because the diabetes and heart failure datasets are not patient-linked, this score was not used as a predictor in the primary mortality models. However, it establishes a modular and extensible pattern for symptom aggregation that may be leveraged in future integrated or longitudinal datasets.

Model interpretability was strengthened by applying SHAP post hoc to the random forest model, enabling clinicians to inspect feature attributions and supporting a governance-oriented understanding of model behavior beyond aggregate performance metrics. Future work should expand this interpretability layer to include stability analyses of feature attributions across resampling folds and under simulated dataset shift.

Class imbalance remains a key translational challenge. Although synthetic oversampling methods, such as SMOTE, were explored during preliminary sensitivity analyses, they were intentionally excluded from the final validated pipeline to prioritize leakage-resistant evaluation. Future studies should systematically compare imbalance-handling strategies, including within-fold resampling, cost-sensitive learning, calibrated threshold optimization, and synthetic sampling using pre-specified evaluation plans and strict separation between training and evaluation data.

The next steps toward clinical utility include external validation in independent cohorts, formal calibration assessment and decision curve analysis, and subgroup-level performance auditing to detect clinically meaningful disparities. Ultimately, prospective or shadow-deployment studies using electronic health record data will be required to evaluate model stability under dataset shift, integration into clinical workflows, and real-world error costs. Together, these extensions define a clear and responsible translational pathway from retrospective validation toward clinically trustworthy deployment.

5. Conclusion

This study demonstrates that standard, interpretable machine learning models can achieve meaningful and clinically plausible predictions of cardiovascular mortality risk in patients with diabetes and heart failure when evaluated under leakage-resistant validation. Using pooled out-of-fold predictions from five-fold cross-validation, the random forest model consistently outperformed logistic regression in discrimination while exhibiting a more specificity-dominant classification profile, whereas logistic regression favored higher sensitivity. These differences highlight the importance of evaluating models not only by global performance metrics but also by their operational behavior at clinically relevant decision thresholds.

Across modeling approaches, serum creatinine, ejection fraction, and age emerged as dominant predictors of mortality risk, reinforcing the biological plausibility of the learned models and aligning with established cardiovascular and cardiorenal pathophysiology. Importantly, these findings were obtained without reliance on complex or opaque feature representations, underscoring that rigorous evaluation and appropriate validation practices may be more consequential than model complexity for trustworthy clinical machine learning.

This work contributes methodologically by illustrating a reproducible framework for conservative model evaluation using pooled out-of-fold predictions, model-specific threshold selection, and paired statistical comparisons

of classification behavior. Substantively, it suggests that carefully validated machine learning models may support improved risk stratification in cardiometabolic populations, with potential applications in screening, triage, and clinical decision support.

However, this study represents a preclinical evaluation. External validation in independent and prospective cohorts, calibration assessment, subgroup analyses, and integration into clinical workflows are essential before real-world deployment. Future research should also explore longitudinal modeling and time-to-event approaches to better capture disease trajectories.

Taken together, these findings support a cautious but optimistic role for interpretable machine learning in cardiovascular risk prediction, emphasizing that methodological rigor and transparency are foundational prerequisites for clinical impact.

Acknowledgments

None.

Funding

None.

Conflict of interest

The author declares no conflict of interest.

Author contributions

This is a single-authored article.

Ethics approval and consent to participate

All analyses were performed using publicly available, fully de-identified datasets and did not involve human participants, clinical interventions, or identifiable personal data. Accordingly, institutional review board approval and informed consent were not required.

Consent for publication

Not applicable. This study used publicly available, anonymized datasets from the UCI Machine Learning Repository and did not involve identifiable human participants.

Availability of data

Model development and evaluation were implemented through a fully scripted and version-controlled analytic pipeline designed to ensure transparency, auditability, and reproducibility. All performance estimates were derived exclusively from pooled out-of-fold predictions generated under five-fold cross-validation, thereby avoiding

in-sample or single-split inflation. Class imbalance handling and preprocessing steps were confined strictly to training folds to prevent information leakage.

Interpretable baseline models were intentionally prioritized, and validation practices were aligned with principles articulated in Food and Drug Administration Good Machine Learning Practice guidance, emphasizing conservative evaluation, clear operating characteristics, and clinically inspectable behavior. All code, analytic procedures, and intermediate outputs are publicly available to support independent replication. The author welcomes correspondence regarding methodological details, reproducibility, and potential extensions of this work.

Further disclosure

A preliminary version of this work was previously made available as a preprint on SSRN under the title “Advancing Deep Learning Insights for Identifying Heart Disease in Diabetic Patients: A Data Mining Approach Using Logistic Regression and Random Forests” (doi:10.2139/ssrn.5091734). The version submitted here is the finalized and substantially revised manuscript, incorporating additional analyses, new modeling techniques (including SHAP explainability), reproducibility pipelines, and refined interpretation.

References

1. D’Agostino RB Sr, Vasan RS, Pencina MJ, *et al.* General cardiovascular risk profile for use in primary care: the Framingham Heart Study. *Circulation*. 2008;117(6):743-753. doi: 10.1161/CIRCULATIONAHA.107.699579
2. American Diabetes Association. Cardiovascular disease and risk management: standards of medical care in diabetes—2020. *Diabetes Care*. 2020;43(Suppl 1):S111-S134. doi: 10.2337/dc20-S010
3. Peters SAE, Huxley RR, Woodward M. Diabetes as risk factor for incident coronary heart disease in women compared with men: a systematic review and meta-analysis of 64 cohorts including 858,507 individuals and 28,203 coronary events. *Diabetologia*. 2014;57(8):1542-1551. doi: 10.1007/s00125-014-3260-6
4. World Health Organization. Cardiovascular diseases (CVDs). World Health Organization. 2025. Available from: [https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-\(cvds\)](https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)) [Last accessed on March 15, 2026].
5. Rajkomar A, Dean J, Kohane IS. Machine learning in medicine. *N Engl J Med*. 2019;380(14):1347-1358. doi: 10.1056/NEJMra1814259
6. Borges JYV. Auditing shortcut learning in AI-based breast

- cancer genomic subtyping. *JAMIA Open*. In press.
7. Tabák AG, Herder C, Rathmann W, Brunner EJ, Kivimäki M. Prediabetes: a high-risk state for diabetes development. *Lancet*. 2012;379(9833):2279-2290.
doi: 10.1016/S0140-6736(12)60283-9
8. Breiman L. Random forests. *Mach Learn*. 2001;45(1):5-32.
doi: 10.1023/A:1010933404324
9. Kuhn M, Johnson K. Applied Predictive Modeling. New York, NY: Springer. 2013.
10. Early Stage Diabetes Risk Prediction Dataset. UCI Machine Learning Repository. Available from: <https://archive.ics.uci.edu/dataset/529/early+stage+diabetes+risk+prediction+dataset> [Last accessed on March 15, 2026].
11. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. New York, NY: Springer. 2009.
doi: 10.1007/978-0-387-84858-7
12. Islam MMF, Ferdousi R, Rahman S, Bushra HY. Likelihood prediction of diabetes at early stage using data mining techniques. In: Gupta M, Konar D, Bhattacharyya S, Biswas S, eds. *Computer Vision and Machine Intelligence in Medical Image Analysis*. Singapore: Springer. 2020:113-125.
doi: 10.1007/978-981-13-8798-2_12
13. Kuhn M. Building predictive models in R using the caret package. *J Stat Softw*. 2008;28(5):1-26.
doi: 10.18637/jss.v028.i05
14. Hosmer DW, Lemeshow S, Sturdivant RX. *Applied Logistic Regression*. 3rd ed. Hoboken, NJ: Wiley. 2013.
doi: 10.1002/9781118548387
15. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc Ser B Stat Methodol*. 2005;67(2):301-320.
doi: 10.1111/j.1467-9868.2005.00503.x