International Journal of General Medicine

Dovepress
Taylor & Francis Group

Open Access Full Text Article

ORIGINAL RESEARCH

# Development and Validation of an Explainable Machine Learning Model for Predicting Repeat Catheter Ablation for Atrial Fibrillation: A Single-Center Retrospective Cohort Study

Shuai Shang[1,2,*], Huasheng Lv[1,2,*], Guoxiang Ma[3], Meng Wei[1,2], Kai Wang[3], Yanmei Lu[1,2], Baopeng Tang[1,2]

[1]Department of Cardiac Pacing and Electrophysiology, The First Affiliated Hospital of Xinjiang Medical University, Urumqi, People's Republic of China; [2]Xinjiang Key Laboratory of Cardiac Electrophysiology and Remodeling, The First Affiliated Hospital of Xinjiang Medical University, Urumqi, People's Republic of China; [3]Department of Medical Engineering and Technology, Xinjiang Medical University, Urumqi, People's Republic of China

*These authors contributed equally to this work

Correspondence: Baopeng Tang; Yanmei Lu, Email tangbaopeng1111 @163.com; gracy @189.cn



**Background:** Atrial fibrillation (AF) is the most prevalent sustained cardiac arrhythmia worldwide. Catheter ablation is the first-line therapy for symptomatic/refractory AF, yet post-procedural recurrence remains extremely common, driving a high rate of repeat ablation procedures. Repeat ablation is associated with elevated medical costs, incremental procedural risks, and impaired quality of life and clinical outcomes in affected patients. Existing clinical risk scores for predicting repeat AF ablation have limited discriminative ability, poor interpretability, and suboptimal clinical utility. This study aimed to develop and validate an explainable machine learning model, using routine clinical and echocardiographic features, to predict the risk of requiring repeat catheter ablation for AF.

**Methods:** A retrospective cohort of 1073 patients undergoing AF ablation from 2012 to 2023 was analyzed, with data split into training (70%) and testing (30%) sets. Feature selection was performed using LASSO regression and the Boruta algorithm, followed by the construction of eight machine learning models. Model performance was evaluated using area under the receiver operating characteristic curve (AUC), sensitivity, specificity, F1 score, balanced accuracy, Brier score, and clinical utility via decision curve analysis. Interpretability was enhanced using Shapley Additive Explanations (SHAP).

**Results:** Among 1073 patients undergoing AF ablation, 352 (32.8%) required a second procedure. LASSO regression combined with the Boruta algorithm identified nine predictive features: NT-proBNP, age, globulin (GLO), direct bilirubin (DBIL), left ventricular ejection fraction (LVEF), cystatin C (Cys-C), smoking history, creatine kinase (CK), and urea. Among the eight models evaluated, XGBoost demonstrated the best overall performance, achieving an AUC of 0.811 (95% CI: 0.762–0.859) in the testing cohort, with a sensitivity of 0.748, specificity of 0.726, and Brier score of 0.1682. It also outperformed alternative models in terms of F1 score and clinical net benefit. SHAP analysis confirmed NT-proBNP and age as the most influential predictors, alongside non-linear contributions from the remaining variables.

**Conclusion:** The XGBoost model may provide a useful and interpretable tool for predicting repeat AF ablation, providing clinical insights to guide patient management and optimize procedural outcomes.

**Keywords:** atrial fibrillation, catheter ablation, machine learning, repeat ablation, XGBoost, SHAP, prediction model, interpretability

## Introduction

Atrial fibrillation (AF) remains the most prevalent cardiac arrhythmia, affecting millions worldwide and contributing to significant morbidity, including stroke and heart failure, as well as substantial healthcare costs.[1] Catheter ablation, particularly pulmonary vein isolation, is a cornerstone treatment for patients with symptomatic or drug-refractory

https://doi.org/10.2147/IJGM.S574069
Received: 6 November 2025
Accepted: 16 March 2026
Published: 23 March 2026

AF, offering improved rhythm control compared to antiarrhythmic drugs.[2] However, AF recurrence remains a critical challenge, with 20–40% of patients requiring repeat ablation due to factors such as pulmonary vein reconnection, incomplete lesion formation, or progressive atrial remodeling.[3] Accurate identification of patients at risk for repeat ablation is essential to optimize treatment strategies, enhance patient outcomes, and reduce healthcare burdens.

Conventional risk stratification for AF recurrence often relies on clinical variables such as age, left atrial diameter, and comorbidities, but these models frequently lack precision due to their inability to capture complex, non-linear interactions.[4,5] Machine learning (ML) approaches have emerged as powerful tools for cardiovascular risk prediction, leveraging high-dimensional data to uncover intricate patterns.[6] Despite their potential, many ML models are criticized for their lack of interpretability, which hinders clinical adoption.[7] Explainable AI techniques, such as Shapley Additive Explanations (SHAP), have addressed this limitation by providing transparent insights into feature contributions, thereby enhancing trust and applicability in clinical settings.[8] The SHAP method, based on game theory, calculates the marginal contribution of each feature to the prediction for an individual patient, allowing for both global and local model interpretation.

Recent studies have demonstrated the utility of ML in predicting AF-related outcomes. For instance, predictive models using Random Forest and gradient boosting have shown promise in identifying patients at risk of AF recurrence post-ablation, though often without sufficient focus on interpretability.[9] The integration of explainable ML frameworks, such as those combining XGBoost with SHAP, has been shown to improve both predictive accuracy and clinical utility in cardiovascular applications.[10,11] However, few studies have specifically targeted the prediction of repeat AF ablation while prioritizing model transparency, a critical gap in the era of personalized medicine.[12,13] This study aims to develop and validate an explainable ML model to predict the need for repeat AF ablation, utilizing a comprehensive set of clinical, laboratory, and echocardiographic features. By employing advanced feature selection and interpretable ML techniques, we seek to deliver a clinically actionable tool to guide patient management and optimize procedural outcomes.

# Methods

## Study Design and Ethical Approval

This was a single-center, retrospective cohort study conducted at the First Affiliated Hospital of Xinjiang Medical University. We enrolled consecutive patients who underwent catheter ablation for AF between June 2012 and September 2023.This study was approved by the Ethics Committee of The First Affiliated Hospital of Xinjiang Medical University (Approval No. 231124–05) and performed in strict accordance with the principles of the Declaration of Helsinki. Given the retrospective, non-interventional design, the requirement for written informed consent from enrolled patients was waived by the ethics committee.

## Study Population

Inclusion criteria: 1) Aged ≥18 years at the time of the index ablation procedure; 2) Confirmed AF diagnosis, verified by ≥30 s single-lead electrocardiogram (ECG) or ≥10 s 12-lead ECG showing absent P waves, irregular fibrillatory waves, and irregular RR intervals; 3) Underwent first-time, successful index catheter ablation for AF; 4) Complete clinical, echocardiographic and follow-up data available.

Exclusion criteria: 1) Valvular AF; 2) Unsuccessful index ablation procedure; 3) Underwent early touch-up ablation within the 3-month post-ablation blanking period; 4) Incomplete clinical or follow-up data.

## Outcome Definition

The primary endpoint of this study was the occurrence of repeat catheter ablation for AF (defined as the second ablation procedure). Specifically, repeat ablation was defined as a second planned catheter ablation performed for recurrent AF, or AF-related atrial flutter/atrial tachycardia after the 3-month blanking period following the index first-time ablation. Repeat procedures for non-AF-related arrhythmias or non-recurrent clinical symptoms were not counted as the primary endpoint. Early touch-up procedures within the 3-month blanking period were also not counted as the primary endpoint,

and corresponding patients were excluded from the final cohort. A total of 1073 eligible patients were finally included, categorized into a single-ablation group (n = 721, no repeat ablation meeting the endpoint definition after the blanking period) and a repeat-ablation group (n = 352, met the primary endpoint definition). Patient inclusion and exclusion criteria are summarized in Figure 1.

## Data Collection

Demographic characteristics, medical history, laboratory results, and echocardiographic parameters were extracted from the hospital's electronic medical records, yielding 46 feature variables (Supplementary Table 1). These included sex, age, number of AF ablations, complete blood count, coagulation profile, biochemical markers, lipid profile, thyroid function, cardiac injury biomarkers, and echocardiographic data. Categorical variables, such as smoking history, were converted into a numerical format suitable for machine learning models using one-hot encoding. Missing data were handled using listwise deletion to ensure complete case analysis. Variable correlations were analyzed using the R packages caret (version 4.3.3) and DataExplorer (version 4.3.3).

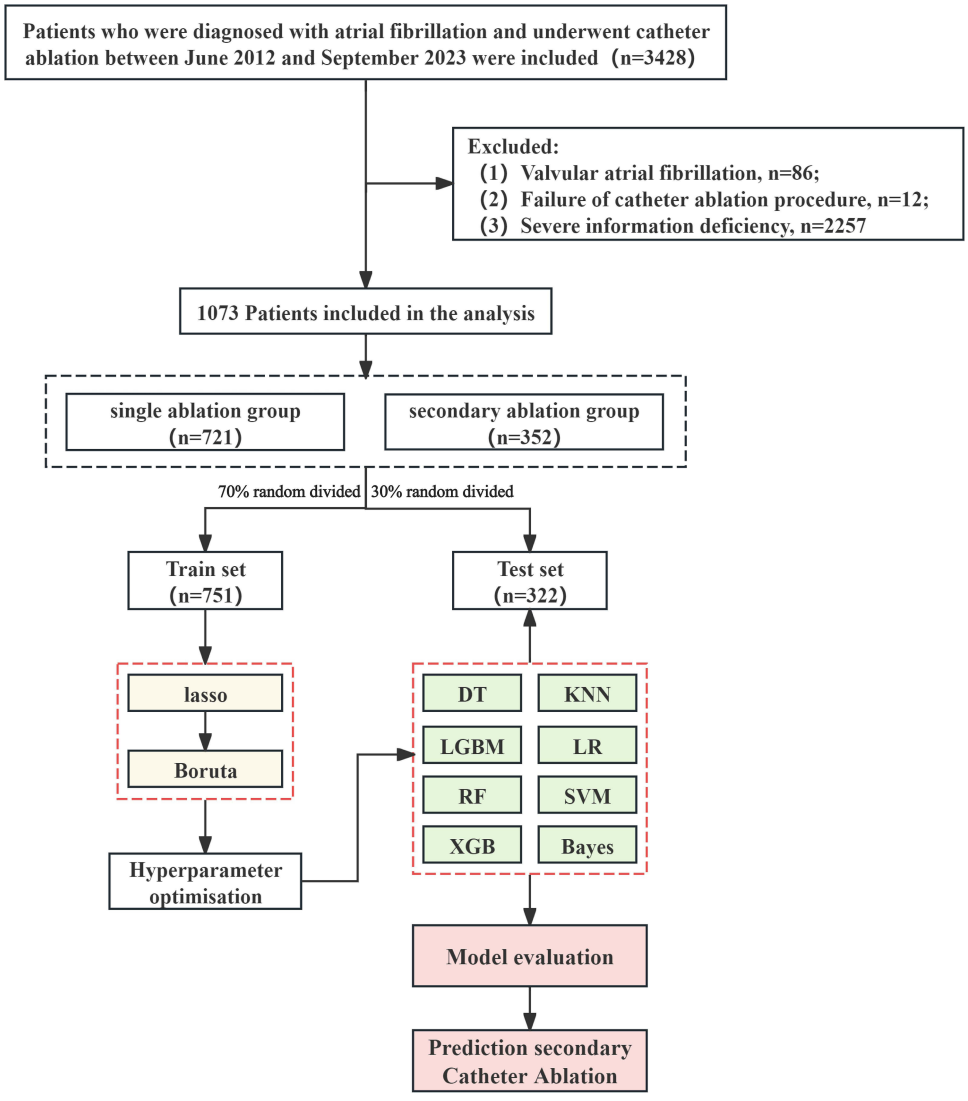


**Figure 1** Flowchart of patient selection.

**Abbreviations**: DT, Decision Tree; KNN, K-Nearest Neighbors; LGBM, Light Gradient Boosting Machine; LR, Logistic Regression; RF, Random Forest; SVM, Support Vector Machine; XGB, eXtreme Gradient Boosting; Bayes, Naive Bayes.


https://doi.org/10.2147/IJGM.S574069

## Feature Selection

To assess potential multicollinearity, the correlation structure of all variables was visualized prior to feature selection. The dataset was split using stratified sampling based on the primary outcome (repeat ablation) into training (70%) and testing (30%) sets, with a random seed of 12345, to ensure a balanced distribution of cases in both sets. Least Absolute Shrinkage and Selection Operator (LASSO) regression was applied to the training set for preliminary variable selection. LASSO employs L1 regularization to shrink coefficients of non-significant variables to zero, identifying variables with the greatest predictive impact. All variables were included in the LASSO model, with the optimal lambda value (lambda.1se) selected via 10-fold cross-validation to enhance model robustness and generalizability.

Subsequently, the LASSO-selected variables were further refined using the Boruta algorithm, with two core justifications for this sequential two-stage approach: first, LASSO regression enables efficient linear screening of the initial high-dimensional feature set to eliminate redundant noise variables, but cannot fully capture non-linear relationships between features and the endpoint; the Boruta algorithm, based on random forest principles, can robustly validate the non-linear predictive value of candidate variables, avoiding the exclusion of biologically meaningful features with non-linear effects. Second, this two-step approach combines the efficiency of LASSO-based dimensionality reduction with the strict false-positive control of the Boruta algorithm, which validates feature importance against randomly permuted "shadow features" over 100 iterations, ensuring only variables with statistically robust predictive power are retained.

After feature selection, we performed post-selection validation to confirm the robustness of the final feature set: we re-assessed multicollinearity using the variance inflation factor (VIF), confirming all final selected features had a VIF $< 5$ with no significant residual multicollinearity; we also examined feature interaction effects using SHAP interaction values, confirming no strong confounding interactions between final features, while the model effectively captured meaningful non-linear interactions between key predictors. Only variables retained by this two-stage selection process were used for model construction.

## Model Construction and Validation

Eight machine learning algorithms were employed to develop predictive models based on the selected features: Decision Tree (DT), Random Forest (RF), Extreme Gradient Boosting (XGBoost), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Light Gradient Boosting Machine (LGBM), Logistic Regression (LR), and Naive Bayes (NB). Models were trained on the training set, and performance was evaluated using Receiver Operating Characteristic (ROC) curves, Calibration Curves, and Decision Curve Analysis (DCA) to assess discrimination, calibration, and clinical utility, respectively. These curves were also generated for the testing set to further evaluate model performance.

## Model Evaluation

Model performance was comprehensively assessed using the following metrics: (1) Area Under the ROC Curve (AUC) to evaluate discrimination for repeat ablation; (2) sensitivity, specificity, F1 score, Balanced Accuracy (BA), and recall to measure predictive performance across classification thresholds; (3) Brier Score to assess calibration performance. The Shapley Additive Explanations (SHAP) method was applied to enhance model interpretability, quantifying the contribution of each variable to the prediction of repeat AF ablation and improving clinical transparency.

## Statistical Analysis

Data analysis was performed using SPSS 26.0 and R (version 4.2.0). Continuous variables were tested for normality. Normally distributed variables were expressed as mean ± standard deviation (SD) and compared using independent t-tests. Non-normally distributed variables were reported as medians (25th–75th percentiles) and compared using Mann–Whitney U or Kruskal–Wallis tests. Categorical variables were presented as frequencies (%) and compared using chi-square tests. A $P < 0.05$ was considered statistically significant.

# Results

## Baseline Characteristics of the Study Population

Overall, 1073 patients who underwent AF ablation were included, with 721 (67.2%) in the single-ablation group and 352 (32.8%) in the repeat-ablation group. Table 1 presents the baseline characteristics comparison between the two groups. Significant differences were observed in age (P < 0.001), gamma-glutamyl transferase (GGT; P = 0.006), glycosylated hemoglobin (HbA1c; P = 0.003), right atrial diameter (RA; P = 0.004), triglycerides (TG; P = 0.003), urea (P = 0.011), creatinine (Crea; P < 0.001), thyroid-stimulating hormone (TSH; P = 0.024), hemoglobin (Hb; P = 0.006), cystatin C (Cys-C; P < 0.001), N-terminal pro-B-type natriuretic peptide (NT-ProBNP; P < 0.001), direct bilirubin (DBIL; P = 0.004), creatine

**Table 1** Comparison of Baseline Data Between Groups with Different Numbers of Ablations

| Variables | Total (n = 1073) | Single Ablation Group (n = 721) | Secondary Ablation Group (n = 352) | *P* |
|---|---|---|---|---|
| Neut, % | 56.00 ± 8.49 | 55.92 ± 8.52 | 56.16 ± 8.44 | 0.659 |
| RBC, 10^12/L | 4.61 ± 0.55 | 4.59 ± 0.56 | 4.65 ± 0.52 | 0.102 |
| Age, year | 63.00 (55.00, 70.00) | 64.00 (55.00, 72.00) | 61.00 (54.00, 68.00) | <0.001 |
| D-Dimer, ng/mL | 69.00 (38.00, 123.00) | 71.00 (41.00, 128.00) | 68.67 (37.00, 111.00) | 0.219 |
| GGT, U/L | 23.00 (16.90, 36.60) | 22.00 (16.10, 35.60) | 25.35 (17.00, 38.78) | 0.006 |
| AST, U/L | 19.10 (15.90, 23.70) | 19.00 (15.70, 23.50) | 19.48 (16.30, 24.10) | 0.191 |
| HbA1c, % | 6.00 (5.70, 6.48) | 6.00 (5.61, 6.40) | 6.10 (5.80, 6.56) | 0.003 |
| RA, mm | 36.00 (33.00, 40.50) | 36.00 (33.00, 39.00) | 36.00 (34.00, 41.00) | 0.004 |
| TG, mmol/L | 1.11 (0.83, 1.56) | 1.07 (0.81, 1.51) | 1.17 (0.89, 1.69) | 0.003 |
| LVEF, % | 61.90 (58.96, 63.69) | 61.90 (58.91, 63.37) | 61.90 (59.00, 63.80) | 0.302 |
| Urea, mmol/L | 5.60 (4.60, 6.80) | 5.70 (4.70, 6.90) | 5.48 (4.50, 6.50) | 0.011 |
| FFA, mmol/L | 0.41 (0.30, 0.54) | 0.41 (0.30, 0.54) | 0.41 (0.30, 0.55) | 0.670 |
| GLO, g/L | 23.60 (21.00, 26.65) | 23.60 (20.80, 26.70) | 23.60 (21.10, 26.61) | 0.236 |
| Crea, umol/L | 74.90 (64.70, 88.00) | 77.00 (65.00, 90.00) | 72.37 (63.86, 83.00) | <0.001 |
| TT3, nmol/L | 1.63 (1.40, 1.82) | 1.63 (1.39, 1.84) | 1.62 (1.43, 1.81) | 0.949 |
| TSH, mIU/L | 2.44 (1.61, 3.83) | 2.37 (1.55, 3.80) | 2.61 (1.76, 3.86) | 0.024 |
| ApoB, g/L | 0.75 (0.60, 0.94) | 0.74 (0.59, 0.93) | 0.77 (0.62, 0.96) | 0.191 |
| PLT, 10^9/L | 197.00 (165.00, 231.00) | 197.00 (165.00, 233.00) | 197.00 (165.00, 226.00) | 0.561 |
| ApoA1, g/L | 1.13 (1.01, 1.27) | 1.13 (1.01, 1.27) | 1.13 (1.00, 1.28) | 0.785 |
| ALT, U/L | 18.01 (13.34, 24.87) | 17.00 (12.90, 24.10) | 19.35 (14.50, 26.90) | 0.001 |
| LVEDd, mm | 49.00 (47.00, 51.00) | 49.00 (47.00, 51.00) | 49.00 (47.00, 51.00) | 0.395 |
| UA, umol/L | 328.03 (270.00, 395.90) | 329.60 (267.00, 397.00) | 328.00 (276.77, 392.19) | 0.864 |
| TT4, nmol/L | 89.44 (78.64, 101.00) | 88.93 (78.14, 100.70) | 90.19 (80.57, 102.06) | 0.234 |
| TP, g/L | 65.30 (62.10, 68.70) | 65.10 (62.20, 68.70) | 65.45 (61.98, 68.70) | 0.671 |
| HDL C, mmol/L | 1.01 (0.86, 1.19) | 1.01 (0.86, 1.17) | 1.02 (0.87, 1.21) | 0.227 |
| Hb, g/L | 141.00 (130.00, 152.00) | 140.00 (129.00, 151.00) | 144.00 (132.00, 154.00) | 0.006 |
| LPa, mg/L | 114.60 (58.00, 225.86) | 115.50 (56.30, 232.06) | 110.76 (66.70, 211.02) | 0.978 |
| LDL-C, mmol/L | 2.26 (1.78, 2.79) | 2.24 (1.74, 2.77) | 2.32 (1.82, 2.83) | 0.133 |
| Cys-C, mg/L | 0.97 (0.82, 1.15) | 1.00 (0.87, 1.18) | 0.92 (0.76, 1.08) | <0.001 |
| NT-ProBNP, ng/L | 474.00 (152.00, 1055.50) | 377.00 (115.00, 975.00) | 635.02 (335.10, 1167.28) | <0.001 |
| WBC, 10^12/L | 5.79 (4.89, 6.79) | 5.75 (4.80, 6.68) | 5.84 (5.00, 6.86) | 0.136 |
| TC, mmol/L | 3.56 (2.95, 4.18) | 3.54 (2.93, 4.17) | 3.62 (3.01, 4.18) | 0.229 |
| DBIL, umol/L | 4.40 (3.12, 6.15) | 4.25 (3.00, 6.00) | 4.70 (3.50, 6.50) | 0.004 |
| ALP, U/L | 66.00 (55.00, 78.00) | 65.50 (55.00, 78.00) | 66.75 (55.00, 79.20) | 0.589 |
| BMI, $Kg/m^2$ | 25.72 (23.53, 28.07) | 25.71 (23.53, 28.01) | 25.76 (23.56, 28.13) | 0.537 |
| CK, IU/L | 73.43 (54.00, 100.00) | 70.80 (52.10, 97.76) | 80.00 (60.00, 106.15) | <0.001 |
| IBIL, umol/L | 9.84 (6.86, 13.58) | 9.90 (6.90, 13.59) | 9.70 (6.72, 13.54) | 0.907 |
| LA, mm | 40.00 (35.00, 44.00) | 40.00 (35.00, 43.00) | 41.00 (37.00, 44.00) | 0.003 |

(*Continued*)

**Table 1** (Continued).

| Variables | Total (n = 1073) | Single Ablation Group (n = 721) | Secondary Ablation Group (n = 352) | *P* |
|---|---|---|---|---|
| ALB, g/L | 41.60 (39.40, 44.10) | 41.60 (39.50, 44.20) | 41.40 (39.00, 43.90) | 0.427 |
| Train set, n (%) | 751 (69.99) | 506 (70.18) | 245 (69.60) | 0.846 |
| Male, n (%) | 648 (60.39) | 419 (58.11) | 229 (65.06) | 0.029 |
| Smoke, n (%) | 231 (21.53) | 133 (18.45) | 98 (27.84) | <0.001 |
| Hypertension, n (%) | 489 (45.57) | 337 (46.74) | 152 (43.18) | 0.272 |
| Diabetes, n (%) | 161 (15.00) | 110 (15.26) | 51 (14.49) | 0.741 |
| Alcohol, n (%) | 189 (17.61) | 119 (16.50) | 70 (19.89) | 0.172 |
| Stroke, n (%) | 100 (9.32) | 67 (9.29) | 33 (9.38) | 0.965 |

kinase (CK; P < 0ouwd), left atrial diameter (LA; P = 0.003), sex (P = 0.029), and smoking history (P < 0.001). Comparison of baseline characteristics between the training set (n = 751) and testing set (n = 322) (Supplementary Table 2) showed no significant differences across all variables (P > 0.05), indicating consistent data partitioning.

## Feature Selection Outcomes

Feature selection was performed on 46 initial variables using LASSO regression followed by the Boruta algorithm (Figure 2). LASSO regression identified variables with potential predictive value, which were further refined by Boruta. Supplementary Table 3 details the Boruta feature selection results, confirming nine significant variables (age, left ventricular ejection fraction [LVEF], urea, globulin [GLO], smoking history, Cys-C, NT-ProBNP, DBIL, CK) and two tentative variables (TG, Hb), with the remaining variables excluded. Supplementary Figure 1 illustrates variable correlations, supporting the feature selection process by highlighting potential multicollinearity.

## Model Construction and Performance Evaluation

Evaluation Eight machine learning models (LR, DT, RF, XGBoost, SVM, KNN, LGBM, NB) were constructed using the selected features and evaluated on both training and testing sets (Figure 3) and Table 2 summarizes the performance metrics for each model. On the training set, RF and KNN achieved the highest AUC (1.000; 95% CI: 1.000–1.000), which is a definitive marker of severe overfitting, as confirmed by a dramatic deterioration in their performance on the independent testing set. This substantial performance gap between the training and testing sets means these models cannot reliably generalize to unseen patient data — the core requirement for clinical predictive models — and is the key reason these models were deemed unsuitable for clinical application.

The XGBoost model was selected as the final model for three core, evidence-based reasons, despite its lower training set performance relative to the overfitted RF and KNN models. First, it exhibited excellent generalizability: it achieved an AUC of 0.843 (95% CI: 0.814–0.872) on the training set and 0.811 (95% CI: 0.762–0.859) on the testing set, with a minimal performance drop of only 0.032 between the two sets, in stark contrast to the severe performance degradation seen in the overfitted RF and KNN models. Second, it outperformed all other models on the independent testing set (the gold standard for real-world predictive performance) across all key metrics, with the highest AUC, F1 score, balanced accuracy, and the lowest Brier score (0.1682), indicating superior discrimination and calibration. Third, decision curve analysis confirmed that the XGBoost model provided the highest net clinical benefit across most clinically relevant risk thresholds, making it the most suitable model for clinical translation. The XGBoost model's robust generalizability, optimal testing set performance, and favorable clinical utility underscore its superiority for predicting repeat ablation risk (Supplementary Figure 2).

## Model Interpretability Analysis

SHAP analysis was applied to the XGBoost model to enhance interpretability (Figure 4). The importance ranking based on mean absolute SHAP values indicated that NT-proBNP was the most influential predictor, followed by age, GLO,

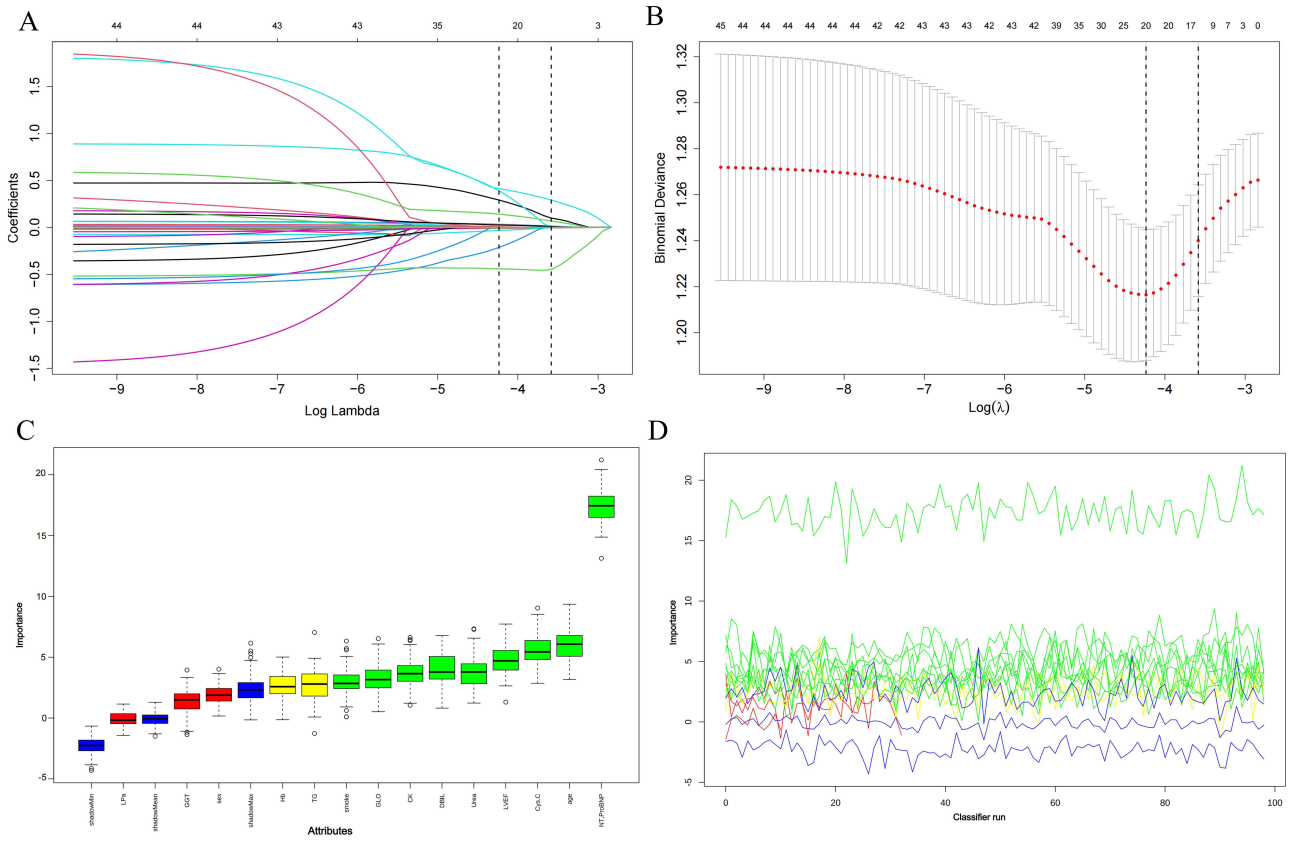


**Figure 2** Feature selection process using LASSO and Boruta algorithms (**A**) Solution path of the Least Absolute Shrinkage and Selection Operator (LASSO) regression showing the coefficient trajectories of all variables as the penalty parameter (log λ) changes.(**B**) Cross-validation error curve for LASSO, with the optimal λ value (lambda.1se) indicated by the dotted vertical line.(**C**) Variable importance plot generated from the Boruta algorithm, highlighting selected features (green), tentative features (yellow), and rejected features (red and blue).(**D**) Detailed run stability of the Boruta feature selection process across 100 iterations.

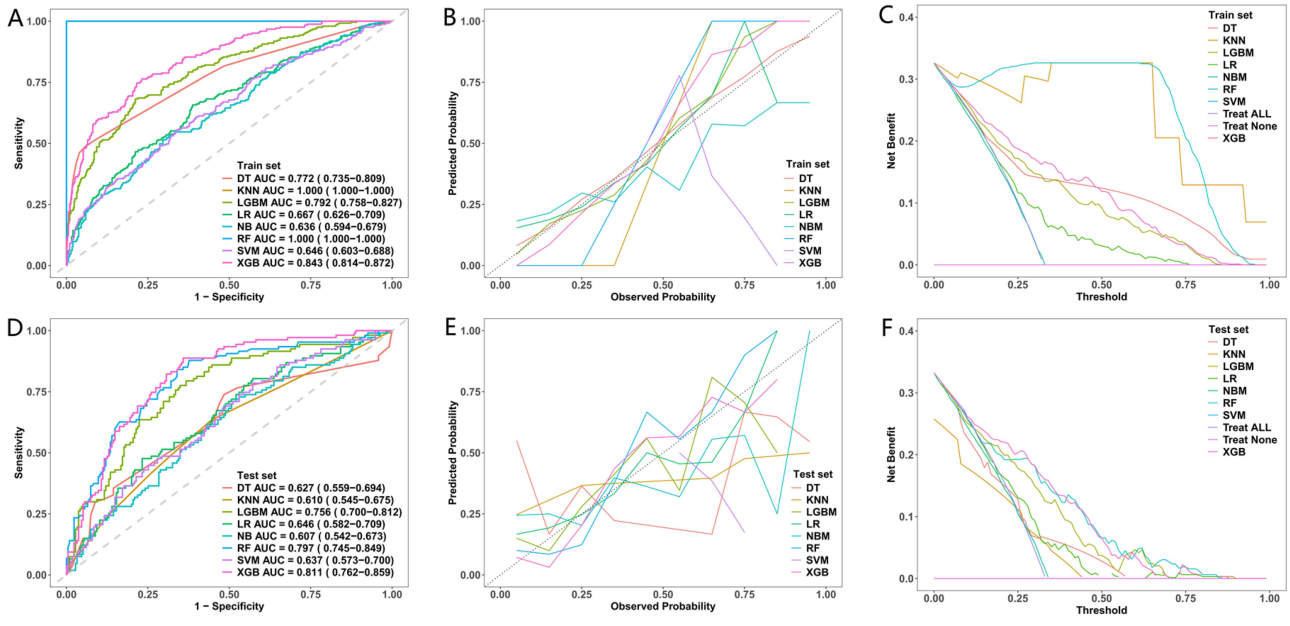


**Figure 3** Performance comparison of eight machine learning models in both training and test sets. (**A** and **D**) display the ROC curves and corresponding AUC values for each model in the training and test sets, respectively. (**B** and **E**) show calibration plots comparing predicted probabilities with observed outcomes. (**C** and **F**) present DCA to evaluate the net clinical benefit across different threshold probabilities.


https://doi.org/10.2147/IJGM.S574069

**Table 2** Comparison Results of Eight Machine Learning Models in Training and Testing Datasets

| model | Data Split | Sensitivity (Recall) | Specificity | F1 | BA | AUC (95% CI) | PRAUC | BS score |
|---|---|---|---|---|---|---|---|---|
| LR | Train | 0.657 | 0.613 | 0.535 | 0.635 | 0.667 (0.626–0.709) | 0.542 | 0.2005 |
| DT | Train | 0.85 | 0.786 | 0.598 | 0.818 | 0.772 (0.735–0.809) | 0.543 | 0.1555 |
| RF | Train | 1 | 1 | 1 | 1 | 1 (1–1) | 0.542 | 0.0268 |
| XGB | Train | 0.763 | 0.769 | 0.681 | 0.766 | 0.843 (0.814–0.872) | 0.542 | 0.1512 |
| SVM | Train | 0.445 | 0.34 | 0.317 | 0.392 | 0.646 (0.603–0.688) | 0.542 | 0.3564 |
| KNN | Train | 1 | 1 | 1 | 1 | 1 (1–1) | 0.496 | 0.0364 |
| LGBM | Train | 0.682 | 0.789 | 0.644 | 0.735 | 0.792 (0.758–0.827) | 0.542 | 0.1654 |
| NBM | Train | 0.514 | 0.704 | 0.484 | 0.609 | 0.636 (0.594–0.679) | 0.557 | 0.2163 |
| LR | Test | 0.645 | 0.553 | 0.507 | 0.599 | 0.646 (0.582–0.709) | 0.539 | 0.2092 |
| DT | Test | 0.579 | 0.721 | 0.402 | 0.65 | 0.627 (0.559–0.694) | 0.539 | 0.2264 |
| RF | Test | 0.505 | 0.865 | 0.568 | 0.685 | 0.797 (0.745–0.849) | 0.539 | 0.1709 |
| XGB | Test | 0.748 | 0.726 | 0.65 | 0.737 | 0.811 (0.762–0.859) | 0.539 | 0.1682 |
| SVM | Test | 0.458 | 0.391 | 0.341 | 0.424 | 0.637 (0.573–0.7) | 0.539 | 0.3505 |
| KNN | Test | 0.411 | 0.744 | 0.427 | 0.578 | 0.61 (0.545–0.675) | 0.496 | 0.2719 |
| LGBM | Test | 0.607 | 0.781 | 0.594 | 0.694 | 0.756 (0.7–0.812) | 0.539 | 0.1836 |
| NBM | Test | 0.43 | 0.698 | 0.422 | 0.564 | 0.607 (0.542–0.673) | 0.551 | 0.2315 |

DBIL, LVEF, Cys-C, smoking history, CK, and urea. These nine features were all used in model construction and demonstrated varying contributions to repeat ablation risk. SHAP dependency plots revealed complex, non-linear relationships between key predictors and the model's output probability (Figure 5). NT-proBNP levels above 1000 ng/L were associated with a sharp increase in recurrence risk, particularly beyond 5000 ng/L. Age exhibited an inverse contribution, with elevated risk in younger patients (<60 years), peaking between 40–50 years. GLO and DBIL displayed non-monotonic patterns, with peak contributions around 30 g/L and 20 μmol/L, respectively. LVEF showed a protective gradient, with lower values (<50%) increasing predicted risk. Cys-C was positively associated with recurrence above 1.5 mg/L, while smoking history and elevated CK (>200 IU/L) contributed discretely to risk elevation. Urea exerted a modest positive effect above 10 mmol/L. Importantly, these SHAP findings represent statistical associations between features and the model's predicted risk, rather than causal relationships between variables and repeat ablation, particularly given potential correlations between the included clinical features.

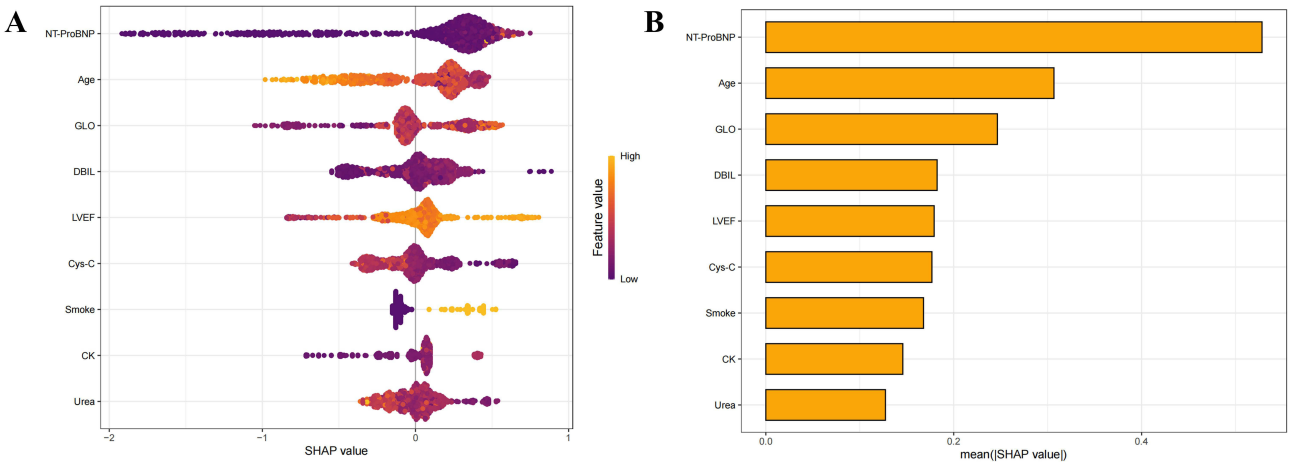


**Figure 4** SHAP summary plots of the XGBoost model.(**A**) Mean absolute SHAP values for each selected feature, representing their overall importance in predicting repeat atrial fibrillation ablation. (**B**) SHAP value distribution colored by feature value, illustrating the direction and magnitude of each variable's contribution to model output. NT-proBNP, age, GLO, and DBIL showed the greatest influence.


https://doi.org/10.2147/IJGM.S574069

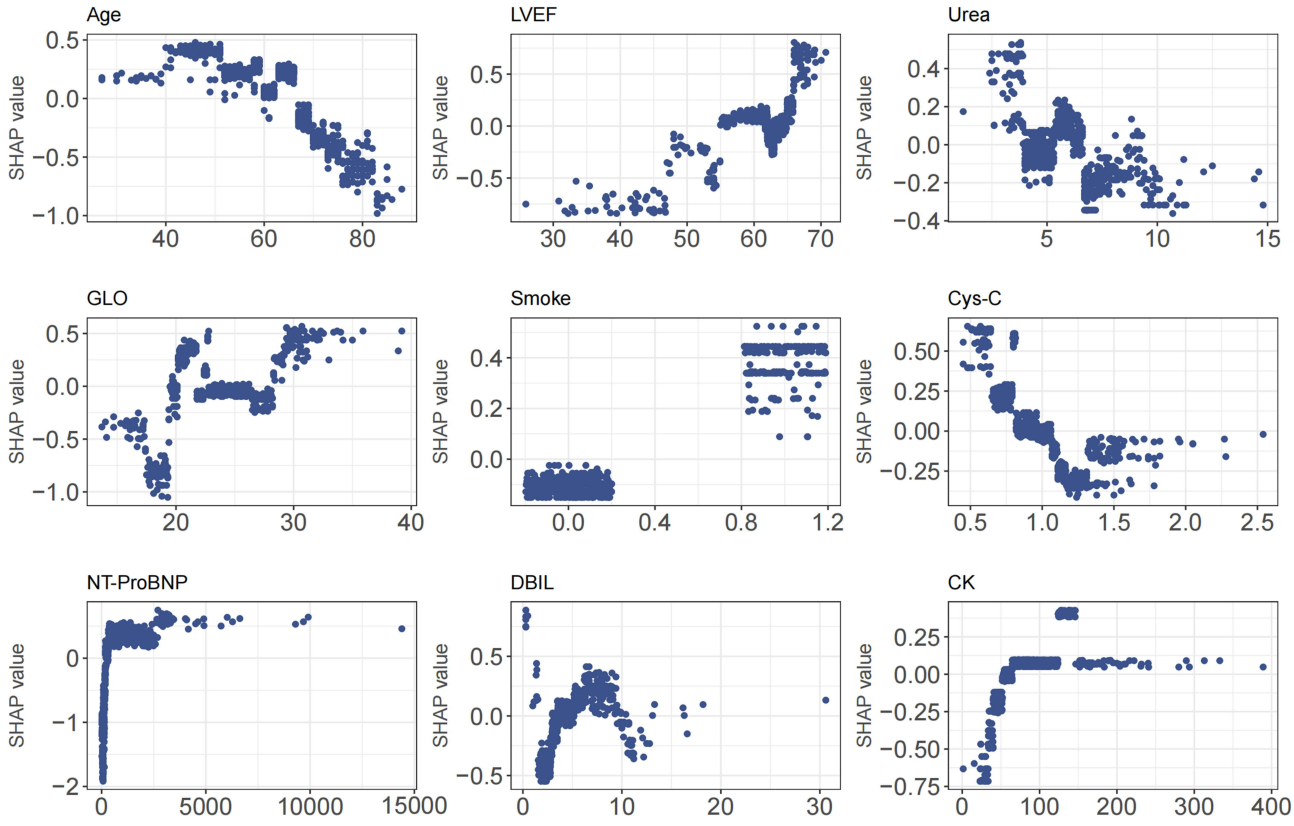


**Figure 5** SHAP dependence plots for top features in the XGBoost model.SHAP dependence plots demonstrate the relationship between each feature and its SHAP value, highlighting nonlinear effects and interaction patterns. NT-proBNP and Cys-C exhibited sharp threshold-like behaviors, whereas variables such as age, GLO, and LVEF showed more gradual gradients.

## Discussion

In this study, we developed and validated an explainable ML model to predict the need for repeat catheter ablation in patients with AF. Using nine readily available clinical and biochemical variables, our XGBoost model achieved strong discriminatory performance (AUC ~0.81). Direct comparison with established scores like APPLE and CAAP-AF is inherently challenging due to fundamental differences in primary endpoints: these scores were developed to predict any post-ablation AF recurrence (with reported AUCs of ~0.55–0.60), while our model specifically identifies patients with clinically significant recurrence that warrants a repeat ablation procedure.[14,15] Rather than replacing these existing clinical scores, our model serves a complementary role in clinical practice. Conventional recurrence risk scores are suited for universal post-ablation risk stratification to guide routine follow-up intensity, while our model is optimized to identify patients at high risk of needing repeat invasive intervention, which can inform pre-procedural patient counseling, personalized ablation strategies, and post-procedural risk factor management. This complementary positioning strengthens the clinical relevance of our model, as it can be integrated into existing workflows alongside established risk scores. Our findings are consistent with prior work indicating that ML algorithms can enhance outcome prediction in electrophysiology. For instance, the AFA-Recur random forest model and other XGBoost-based approaches reported AUCs in the range of 0.72–0.75.[16,17] Given that long-term single-procedure success rates remain suboptimal (~50% at 4–5 years),[18,19] our model may provide clinically meaningful improvements in patient risk stratification and selection for repeat ablation. Notably, this endpoint definition has inherent limitations that may introduce classification bias. Specifically, two groups of patients may be misclassified in our cohort: first, patients with documented post-blanking period AF recurrence who opted for optimized antiarrhythmic medical therapy rather than repeat ablation; second, patients who underwent repeat ablation for atypical atrial flutter or tachycardia unrelated to the index AF substrate. This misclassification may lead to underestimation of the model's true predictive performance for underlying AF recurrence,


https://doi.org/10.2147/IJGM.S574069

and may affect the generalizability of the model to clinical settings with different practice patterns for recommending repeat ablation, or different patient preferences for invasive vs. medical management.

Multiple risk stratification tools have been developed to predict AF recurrence or repeat ablation after catheter ablation, including traditional clinical scores (eg., HATCH, APPLE scores) and machine learning models. However, traditional scores generally have only moderate discriminative ability and fail to capture complex non-linear correlations between variables; most existing machine learning models are "black-box" frameworks with poor interpretability, and many incorporate intraoperative/postoperative parameters that cannot be used for preoperative risk stratification. The model developed in this study addresses these limitations with two core improvements: first, we adopted an explainable machine learning framework, which maintains excellent predictive performance while quantifying the contribution of each feature to the risk of repeat ablation, enabling transparent and clinically interpretable risk prediction; second, all features included in the model are routine preoperative clinical, laboratory and echocardiographic indicators, which can be used for risk stratification before the index ablation to guide personalized clinical decision-making.

The top predictive features identified by our SHAP analysis align closely with established pathophysiological mechanisms of AF recurrence post-ablation, supporting the biological plausibility of our model. NT-proBNP, the most influential feature, reflects atrial pressure and stretch, with elevated levels consistently associated with higher recurrence risk.[20] NT-proBNP, a top-ranked feature in our model, is mainly secreted by atrial cardiomyocytes in response to atrial wall stretch, a direct biomarker of left atrial hemodynamic overload and structural remodeling. Long-term atrial stretch induces RAAS activation, atrial fibrosis and electrical remodeling, forming a stable AF substrate that drives post-ablation recurrence and repeat ablation. Impaired LVEF further elevates left atrial filling pressure, aggravating atrial stretch and remodeling, and forming a vicious cycle between ventricular dysfunction and AF progression. Renal function impairment is associated with volume overload, uremic toxin-induced myocardial fibrosis and systemic inflammation, all of which aggravate AF substrate formation and increase the risk of repeat ablation. In addition, systemic inflammation is a key driver of atrial fibrosis and electrical remodeling before ablation, and also predicts poor scar formation and pulmonary vein reconnection after ablation, which explains its independent predictive value for repeat ablation in our study. These well-established pathophysiological pathways from prior literature provide a mechanistic rationale for the strong predictive associations of these features observed in our model. Age, another key variable, likely captures cumulative atrial remodeling and fibrosis.[21] Reduced LVEF, a marker of structural heart disease, has also been linked to poorer ablation outcomes.[22] Cystatin C and urea, indicators of renal function, support the growing recognition that chronic kidney disease promotes atrial arrhythmogenic remodeling through inflammatory and fibrotic pathways.[23,24] The inclusion of liver-related biomarkers such as globulin and direct bilirubin may reflect systemic inflammation and hepatic congestion, which have recently been implicated in AF persistence.[22,24] Smoking history, a modifiable risk factor, was also predictive and has been associated with atrial oxidative stress and adverse remodeling.[25] Together, these variables span cardiac function, end-organ status, and lifestyle, underscoring the multifactorial nature of AF recurrence.

Compared to conventional scores, our ML model capitalized on non-linear interactions and high-dimensional data, explaining its superior performance. Notably, risk scores like APPLE and CAAP-AF rely on limited and largely linear predictors, which may limit precision.[14,15] Previous ML studies integrating biomarkers and echocardiographic features have shown similar gains in accuracy (AUC ~0.70–0.75),[17,26] although performance remains moderate due to AF's inherent heterogeneity. Nonetheless, improvements in model discrimination are clinically meaningful. High-risk individuals may benefit from closer follow-up or early reintervention, while low-risk patients might avoid unnecessary procedures.

Our results also reinforce growing evidence that ML-based models can enhance AF outcome prediction. Studies employing gradient boosting, random forest, and logistic regression algorithms have shown reproducible performance across different settings.[16] For instance, Budzianowski et al achieved AUC 0.75 with only 12 features,[17] mirroring our parsimonious nine-variable model. Some groups have explored advanced modalities such as MRI-derived atrial scar or CT anatomy,[27] but such inputs may limit practicality. In contrast, our model's reliance on routine data may promote wider clinical adoption. Furthermore, our use of blood biomarkers (eg., NT-proBNP, cystatin C) adds value, highlighting the importance of systemic pathophysiology in arrhythmia persistence.

A distinct advantage of our study is the incorporation of model interpretability through SHAP. Lack of transparency is a major barrier to ML integration in clinical practice.[28] SHAP analysis enabled case-level insight into prediction drivers, with major contributions from NT-proBNP, age, and LVEF aligning with established clinical knowledge.[29] This interpretability facilitates clinical trust and allows for shared decision-making. For example, a patient flagged as high-risk due to elevated renal and cardiac biomarkers may be counseled differently than one with isolated risk factors. Moreover, our model highlights modifiable factors such as smoking and renal dysfunction, which may be targets for upstream intervention.[25,29] Importantly, we emphasize that SHAP analysis quantifies the contribution of each feature to the model's predictive output, which reflects statistical associations rather than causal inference. This is particularly relevant given the inherent correlations between the included clinical features, as SHAP attribution cannot disentangle independent causal effects from correlated variable associations. While our SHAP findings align with established mechanistic understanding of AF recurrence from prior studies, this alignment is based on existing clinical evidence, not causal inference from our model. All feature attributions should therefore be interpreted as predictive associations, not definitive causal drivers of repeat ablation.

## Limitations

This study has several caveats. First, its retrospective, single-centre design introduces selection bias and limits generalisability; external validation in large, multicentre cohorts is therefore warranted. Second, we used repeat catheter ablation as a surrogate for clinically significant AF recurrence, which introduces inherent classification bias and may impair the model's generalizability. Two key sources of misclassification were identified in our study: (1) Patients with confirmed post-blanking period AF recurrence who chose optimized antiarrhythmic medical management instead of repeat ablation were misclassified into the single-ablation group, which may lead to underestimation of the model's predictive ability for underlying AF recurrence, and skew the model toward identifying patients with more severe, symptomatic recurrence who are more likely to opt for invasive reintervention. (2) A small proportion of patients who underwent repeat ablation for non-AF-related arrhythmias or non-recurrent symptoms may have been misclassified into the repeat-ablation group, which could introduce noise to the endpoint definition and reduce the model's specificity for AF-related recurrence.The impact of this bias on model generalizability is twofold: first, the model's performance may vary across medical centers with different clinical decision-making thresholds for offering repeat ablation vs. medical management for recurrent AF; second, the model may have reduced accuracy in patient populations with a higher preference for non-invasive therapy after arrhythmia recurrence. To address this limitation, we have explicitly discussed the impact of endpoint bias on model performance in the Discussion section, and we will conduct sensitivity analyses using alternative endpoint definitions (eg., documented post-blanking period AF recurrence regardless of reintervention) in future work to validate the model's robustness. Third, our model was evaluated using a single stratified train-test split. While this approach is widely used in clinical machine learning-based predictive model studies, it cannot fully capture the stability and generalizability of the model across different data distributions, and may lead to over- or under-estimation of the model's true predictive performance. More robust validation strategies, such as k-fold cross-validation or bootstrapping, can provide more stable and reliable performance estimates by repeatedly partitioning the dataset and evaluating the model across multiple non-overlapping test sets, which can better characterize the model's generalizability. We explicitly acknowledge that the use of a single train-test split in this study may limit the robustness of our reported performance metrics, which is an important limitation of this work.Finally, despite rigorous feature selection, the modest sample size of this single-center cohort raises potential over-fitting concerns that can only be definitively addressed through independent, large-scale multicenter external validation and prospective implementation in future work.

## Conclusion

We developed and validated an interpretable XGBoost model to predict repeat AF ablation, achieving superior discrimination compared to existing clinical scores. By integrating routine clinical and biochemical features with SHAP-based interpretability, the model offers a transparent and practical tool for risk stratification. Future work should focus on external multicenter validation, prospective implementation, and sensitivity analyses with alternative endpoint

definitions (eg., documented AF recurrence regardless of reintervention) to further validate model robustness and refine its clinical utility.

## Disclosure

The authors report no conflicts of interest in this work.

## References


1. Li H, Song X, Liang Y, et al. Global, regional, and national burden of disease study of atrial fibrillation/flutter, 1990-2019: results from a global burden of disease study, 2019. *BMC Public Health*. 2022;22(1):2015. doi:10.1186/s12889-022-14403-2
2. Hindricks G, Potpara T, Dagres N, et al. 2020 ESC Guidelines for the diagnosis and management of atrial fibrillation developed in collaboration with the European Association for Cardio-Thoracic Surgery (EACTS). *Eur Heart J*. 2021;42(5):373–498. doi:10.1093/eurheartj/ehaa612
3. Ganesan AN, Shipp NJ, Brooks AG, et al. Long-term outcomes of catheter ablation of atrial fibrillation: a systematic review and meta-analysis. *J. Am. Heart Assoc*. 2013;2(2):e004549. doi:10.1161/JAHA.112.004549
4. Karanikola AE, Tzortzi M, Kordalis A, et al. Clinical, electrocardiographic and echocardiographic predictors of atrial fibrillation recurrence after pulmonary vein isolation. *J Clin Med*. 2025;14(3):809. doi:10.3390/jcm14030809
5. Yin Y, Li Y, Wang L, et al. Left atrial size and echocardiographic diastolic parameters as predictors of incident atrial fibrillation in older hospitalized patients. *Aging Clin Exp Res*. 2025;37(1):38. doi:10.1007/s40520-025-02936-6
6. D'Ascenzo F, De Filippo O, Gallone G, et al. Machine learning-based prediction of adverse events following an acute coronary syndrome (PRAISE): a modelling study of pooled datasets. *Lancet*. 2021;397(10270):199–207. doi:10.1016/S0140-6736(20)32519-8
7. Sanchez-Martinez S, Camara O, Piella G, et al. Machine learning for clinical decision-making: challenges and opportunities in cardiovascular imaging. *Front. Cardiovasc. Med*. 2022;8:765693. doi:10.3389/fcvm.2021.765693
8. Lundberg SM, Erion G, Chen H, et al. From local explanations to global understanding with explainable AI for trees. *Nature Mach Intell*. 2020;2(1):56–67. doi:10.1038/s42256-019-0138-9
9. Fan X, Li Y, He Q, et al. Predictive value of machine learning for recurrence of atrial fibrillation after catheter ablation: a systematic review and meta-analysis. *Rev cardiovasc med*. 2023;24(11):315. doi:10.31083/j.rcm2411315
10. Salah H, Srinivas S. Explainable machine learning framework for predicting long-term cardiovascular disease risk among adolescents. *Sci Rep*. 2022;12(1):21905. doi:10.1038/s41598-022-25933-5
11. Luo H, Xiang C, Zeng L, et al. SHAP based predictive modeling for 1 year all-cause readmission risk in elderly heart failure patients: feature selection and model interpretation. *Sci Rep*. 2024;14(1):17728. doi:10.1038/s41598-024-67844-7
12. Ma Y, Zhang D, Xu J, et al. Explainable machine learning model reveals its decision-making process in identifying patients with paroxysmal atrial fibrillation at high risk for recurrence after catheter ablation. *BMC Cardiovasc. Disord*. 2023;23(1):91. doi:10.1186/s12872-023-03087-0
13. F BS, Macheret F, D SG, et al. Explainable machine learning to predict anchored reentry substrate created by persistent atrial fibrillation ablation in computational models. *J. Am. Heart Assoc*. 2023;12(16):e030500. doi:10.1161/JAHA.123.030500
14. Karlo F, Daniel S, Arian S, et al. Validation of seven risk scores in an independent cohort: the challenge of predicting recurrence after atrial fibrillation ablation. *Int J Arrhythm*. 2022;23(1):29. doi:10.1186/s42444-022-00080-0
15. J MM, B KMJ, A HLHG, et al. Comparison of the predictive value of ten risk scores for outcomes of atrial fibrillation patients undergoing radiofrequency pulmonary vein isolation. *Int J Cardiol*. 2021;344:103–110. doi:10.1016/j.ijcard.2021.09.029
16. Saglietto A, Gaita F, Blomstrom-Lundqvist C, et al. AFA-recur: an ESC EORP AFA-LT registry machine-learning web calculator predicting atrial fibrillation recurrence after ablation. *Europace*. 2023;25(1):92–100. doi:10.1093/europace/euac145
17. Budzianowski J, Kaczmarek-Majer K, Rzeźniczak J, et al. Machine learning model for predicting late recurrence of atrial fibrillation after catheter ablation. *Sci Rep*. 2023;13(1):15213. doi:10.1038/s41598-023-42542-y
18. T MB, Bilbrough J, Eranki A, et al. Mid-to-long-term recurrence of atrial fibrillation in surgical treatment vs. catheter ablation: a meta-analysis using aggregated survival data. *Ann. Cardiothorac. Surg*. 2024;13(1):.
19. A EM, A QJ, R MJ, et al. Recurrence after atrial fibrillation ablation and investigational biomarkers of cardiac remodeling. *J. Am. Heart Assoc*. 2024;13(6):e031029. doi:10.1161/JAHA.123.031029
20. Yuan Y, Nie B, Gao B, et al. Natriuretic peptides as predictors for atrial fibrillation recurrence after catheter ablation: a meta-analysis. *Medicine*. 2023;102(19):e33704. doi:10.1097/MD.0000000000033704
21. Bannehr M, Georgi C, Edlinger C, et al. Myeloperoxidase and N-terminal proatrial natriuretic peptide as predictors for atrial fibrillation recurrence in patients undergoing redo ablation. *Heart Rhythm O2*. 2024;5(11):770–777. doi:10.1016/j.hroo.2024.09.003
22. Donnellan E, G CT, M WO, et al. Impact of nonalcoholic fatty liver disease on arrhythmia recurrence following atrial fibrillation ablation. *JACC Clin Electrophysiol*. 2020;6(10):1278–1287. doi:10.1016/j.jacep.2020.05.023
23. Chung I, Khan Y, Warrens H, et al. Catheter ablation for atrial fibrillation in patients with chronic kidney disease and on dialysis: a meta-analysis and review. *Cardiorenal Med*. 2022;12(4):155–172. doi:10.1159/000525388
24. Vempati R, Garg A, Shah M, et al. Predictors of atrial fibrillation recurrence after catheter ablation: a state-of-the-art review. *Hearts*. 2025;6(2):12. doi:10.3390/hearts6020012
25. Giomi A, Bernardini A, P PA, et al. Clinical impact of smoking on atrial fibrillation recurrence after pulmonary vein isolation. *Int J Cardiol*. 2024;413.
26. Dretzke J, Chuchu N, Agarwal R, et al. Predicting recurrent atrial fibrillation after catheter ablation: a systematic review of prognostic models. *EP Europace*. 2020;22(5):748–760. doi:10.1093/europace/euaa041
27. Liu CM, Chen WS, Chang SL, et al. Use of artificial intelligence and I-score for prediction of recurrence before catheter ablation of atrial fibrillation. *Int J Cardiol*. 2024;402:131851. doi:10.1016/j.ijcard.2024.131851
28. Otaki Y, Singh A, Kavanagh P, et al. Clinical deployment of explainable artificial intelligence of SPECT for diagnosis of coronary artery disease. *JACC Cardiovasc Imaging*. 2022;15(6):1091–1102. doi:10.1016/j.jcmg.2021.04.030

29. Guo C, Gao B, Han X, et al. Interpretable artificial intelligence model for predicting heart failure severity after acute myocardial infarction. *BMC Cardiovasc. Disord*. 2025;25:362. doi:10.1186/s12872-025-04818-1