Identification of Influencing Factors and Construction of a Discriminant Model for Persistent Atrial Fibrillation: A Postprint Study

Authors: Chaohui Li, Liu Hui, Kai Wang, Liu Hui, Kai Wang

Date: 2026-04-20T15:01:07+00:00

Abstract

thoughtful mini-academic-translator-1.0-20241217-32-thought-1.0-20241217-
32-thought-1.0-20241217-32-thought-1.0-20241217-32-thought-1.0-20241217-32-
thought-1.0-20241217-32-thought-1.0-20241217-32-thought-1.0-20241217-32-
thought-1.0-20241217-32-thought-1.0-20241217-32-thought-1.0-20241217-32-
thought-1.0-20241217-32-thought-1.0-20241217-32-thought-1.0-20241217-32-
thought-1.0-20241217-32-thought-1.0-20241217-32-thought-1.0-20241217-32-
thought-1.0-20241217-32-thought-1.0-20241217-32-thought-1.0-20241217-32-
thought-1.0-20241217-32-thought-1.0-20241217-32-thought-1.0-20241217-32-
thought-1.0-20241217-32-thought-1.0-20241217-32-thought-1.0-20241217-32-
thought-1.0-20241217-32-thought-1.0-20241217-32-thought-1.0-20241217-32-
thought-1.0-20241217-32-thought-1.0-20241217-32-thought-1.0-20241217-32-
thought-1.0-20241217-32-thought-1.0-20241217-32-thought-1.0-20241217-32-
thought-1.0-20241217-32-thought-1.0-20241217-32-thought-1.0-20241217-32-
thought-1.0-20241217-32-thought-1.0-20241217-32-thought-1.0-20241217-32-
thought-1.0-20241217-32-thought-1.0-20241217-32-thought-1.0-20241217-32-
thought-1.0-20241217-32-thought-1.0-20241217-32-thought-1.0-20241217-32-
thought-1.0-20241217-32-thought-1.0-20241217-32-thought-1.0-20241217-32-
thought-1.0-20241217-32-thought-1.0-20241217-32-thought-1.0-20241217-32-
thought-1.0-20241217-32-thought-1.0-20241217-32-thought-1.0-20241217-32-
thought-1.0-20241217-32-thought-1.0-20241217-32-thought-1.0-20241217-32-
thought-1.0-20241217-32-thought-1.0-20241217-32-thought-1.0-20241217-32-
thought-1.0-20241217-32-thought-1.0-20241217-32-thought-1.0-20241217-32-
thought-1.0-20241217-32-thought-1.0-20241217-32-thought-1.0-20241217-32-
thought-1.0-20241217-32-thought-1.0-20241217-32-thought-1.0-20241217-32-
thought-1.0-20241217-32-thought-1.0-20241217-32-thought-1.0-20241217-32-
thought-1.0-20241217-32-thought-1.0-20241217-32-thought-1.0-20241217-32-
thought-1.0-20241217-32-thought-1.0-20241217-32-thought-1.0-20241217-32-
thought-1.0-20241217-32-thought-1.0-20241217-32-thought-1.0-20241217-32-

thought-1.0-20241217-32-thought-1.0-Background: Atrial fibrillation (AF) significantly affects patients' quality of life, leading to high morbidity and mortality rates, and increasing healthcare costs. Objective: To investigate the influencing factors of persistent AF and construct a discriminant model based on these factors. Methods: Patients with paroxysmal and persistent AF who visited the First Affiliated Hospital of Xinjiang Medical University from April 2012 to September 2023 were included as

Full Text

Preamble

Chinese General Practice

Introduction

General practice, as a foundational pillar of the healthcare system, plays a crucial role in maintaining public health and ensuring the efficient operation of medical services. In China, the development of general practice has undergone significant transformation, evolving from a basic community health service model to a comprehensive discipline integrated with modern medical technology and management strategies. This evolution is driven by the increasing demand for high-quality, accessible healthcare and the strategic implementation of the "Healthy China" initiative.

The Role of Machine Learning in General Practice

The integration of advanced technologies, particularly machine learning and deep learning, has revolutionized the diagnostic and management capabilities of general practitioners. By leveraging large-scale clinical data, these technologies assist in early disease detection, personalized treatment planning, and risk stratification.

For instance, predictive models can analyze patient history and physiological parameters to identify individuals at high risk for chronic conditions such as

diabetes or hypertension. The mathematical foundation of these models often involves complex algorithms. Consider a scenario where a predictive function $f(x)$ is used to estimate the probability of a specific health outcome based on a vector of clinical features x:

$$P(y=1|x) = \frac{1}{1+e^{-(\beta_0+\sum_{i=1}^{n}\beta_i x_i)}}$$

In this context, β_i represents the weight assigned to each clinical feature, determined through rigorous training on historical datasets. Such models enable general practitioners to intervene earlier, potentially improving long-term patient outcomes.

Clinical Decision Support Systems

Modern general practice increasingly relies on Clinical Decision Support Systems (CDSS) to enhance diagnostic accuracy. These systems utilize deep learning architectures to process multi-modal data, including medical imaging and electronic health records (EHR).

[Figure 1: see original paper]

As shown in [Figure 1: see original paper], the workflow of a typical CDSS involves data acquisition, feature extraction, and classification. The use of convolutional neural networks (CNNs) has been particularly effective in interpreting radiological images at the primary care level. The optimization of these networks often involves minimizing a loss function $\mathcal{L}$, defined as:

$$\mathcal{L}(\theta) = \frac{1}{N}\sum_{i=1}^{N} L(y_i, \hat{y}_i(\theta))$$

where θ represents the model parameters, y_i is the ground truth label, and $\hat{y}_i$ is the predicted value.

https://www.chinagp.net E-mail:zgqkyx@chinagp.net.cn

Identification of Influencing Factors and Construction of a Discriminative Model for Persistent Atrial Fibrillation

Li Chaohui [1], Liu Hui [2,3*], Wang Kai [2,3*]

Abstract

Atrial fibrillation (AF) is a common clinical arrhythmia that significantly impacts patient quality of life and increases the risk of stroke and heart failure.

This study aims to identify the key influencing factors associated with the progression to persistent atrial fibrillation and to construct an effective discriminative model to assist in clinical diagnosis and risk assessment. By analyzing clinical data and electrophysiological parameters, we employ machine learning techniques to evaluate the predictive power of various features. Our findings suggest that specific structural remodeling markers and biochemical indicators are critical in distinguishing persistent AF from paroxysmal forms. The resulting model demonstrates high sensitivity and specificity, providing a potential tool for personalized management of AF patients.

1. Introduction

Atrial fibrillation (AF) is characterized by rapid and irregular electrical activity in the atria, leading to impaired mechanical function. Clinically, AF is often categorized into paroxysmal, persistent, and permanent types based on the duration and nature of the episodes. Persistent AF, defined as episodes lasting longer than seven days, presents a greater therapeutic challenge than paroxysmal AF due to extensive electrical and structural remodeling of the atria.

Understanding the factors that drive the transition from paroxysmal to persistent AF is crucial for early intervention. Previous studies have identified age, hypertension, diabetes, and left atrial diameter as significant risk factors. However, the complex interplay between these variables often necessitates advanced computational approaches to improve predictive accuracy. This study utilizes statistical analysis and machine learning algorithms to identify core influencing factors and develop a robust discriminative model for persistent AF.

[Figure 1: see original paper]

2. Materials and Methods

2.1 Data Collection and Preprocessing

The dataset for this study was obtained from clinical records of patients diagnosed with AF. We collected a wide range of variables, including demographic information, medical history, echocardiographic parameters, and laboratory test results. Data preprocessing involved the removal of outliers, handling of missing values through mean imputation or predictive modeling, and normalization of continuous variables to ensure compatibility with machine learning algorithms.

2.2 Feature Selection

To identify the most significant influencing factors, we employed a combination of univariate analysis and recursive feature elimination (RFE). Factors such as left atrial diameter (LAD), left ventricular ejection fraction (LVEF), and plasma concentrations of B-type natriuretic peptide were evaluated.

Institute of Medical and Engineering Interdisciplinary Research, Xinjiang Medical University, Urumqi, Xinjiang 830017, China

Hui Liu, Lecturer; E-mail: liuhui@xjmu.edu.cn

Abstract Background

Atrial fibrillation (AF) significantly impairs patients' quality of life, leading to high rates of disability and mortality while substantially increasing healthcare costs. This study aims to investigate the factors influencing persistent atrial fibrillation and to construct a discrimination model based on these factors.

Methods Patients with paroxysmal and persistent atrial fibrillation who visited the First Affiliated Hospital of Xinjiang Medical University between April 2012 and September 2023 were included as study subjects. General clinical data, biochemical indicators, renal function markers, and cardiac function-related parameters were collected for analysis. Initially, univariate logistic regression analysis was employed to screen variables associated with the type of atrial fibrillation. Subsequently, the Least Absolute Shrinkage and Selection Operator (LASSO) regression method was used to further refine the feature variables, reducing model complexity and preventing overfitting. Multivariate logistic regression models were then constructed to identify independent factors associated with persistent atrial fibrillation. Finally, using the Bootstrap resampling method, significant variables were incorporated into six machine learning discrimination models: Random Forest (RF), Decision Tree (DT), Naive Bayes (NB), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Extreme Gradient Boosting (XGB). The discriminative performance of these models was evaluated using Receiver Operating Characteristic (ROC) curves. The contribution of each variable to the discrimination models was assessed using the SHAP (SHapley Additive exPlanations) method.

Results: Information from a total of 6,938 patients was collected, including 5,085 cases of paroxysmal atrial fibrillation and 1,853 cases of persistent atrial fibrillation. Univariate logistic regression analysis identified 19 statistically significant independent variables. After LASSO regression screening, 13 key variables were selected for multivariate logistic regression analysis. The results indicated that gender (OR=1.248, 95% CI=1.086-1.435), personal surgical history (OR=0.809, 95% CI=0.706-0.926), BMI (OR=1.028, 95% CI=1.012-1.045), mean platelet volume (MPV) (OR=1.121, 95% CI=1.059-1.186), serum magnesium (Mg^{2+}) (OR=0.098, 95% CI=0.046-0.208), cardiac output (CO) (OR=1.115, 95% CI=1.009-1.233), left ventricular posterior wall thickness (LVPW) (OR=0.777, 95% CI=0.665-0.909), left atrial diameter (LAD) (OR=1.144, 95% CI=1.123-1.166), left ventricular ejection fraction (LVEF) (OR=0.955, 95% CI=0.938-0.972), right atrial diameter (RAD) (OR=1.031, 95% CI=1.005-1.057), triglycerides (TG) (OR=0.821, 95% CI=0.751-0.898), uric acid (UA) (OR=1.003, 95% CI=1.002-1.003), and left ventricular end-diastolic diameter (LVEDD) (OR=0.903, 95% CI=0.879-

0.927) were all independent factors for persistent atrial fibrillation ($P < 0.05$). ROC curve analysis showed that the XGB model performed best (mean AUC=0.823), followed by the SVM model (0.820) and the RF model (0.814). Evaluation of the relative contribution of variables to the XGB model using SHAP values revealed that LAD and RAD had the highest SHAP values, suggesting that atrial structural parameters have the most significant impact on the discrimination model.

Conclusion Increased BMI, male gender, elevated MPV, increased UA, decreased TG, lower Mg^{2+} levels, reduced LVEF, smaller LVEDD, absence of personal surgical history, and enlargement of RAD and LAD are all closely associated with the occurrence of persistent atrial fibrillation. These clinical indicators can be obtained through routine diagnostic methods, most of which are non-invasive. They serve as important reference factors for identifying patients with persistent atrial fibrillation and can assist in early clinical risk stratification and the formulation of intervention strategies.

Keywords Atrial fibrillation; Paroxysmal atrial fibrillation; Persistent atrial fibrillation; LASSO regression; Logistic regression

[CLC Number] R 541.75 [Document Code] DOI: 10.12114/j.issn.1007-9572.2025.0360

Identification of Factors Associated with Persistent Atrial Fibrillation and Development of a Classification Model

LI Chaohui[1], **LIU Hui**[2,3*], **WANG Kai**[2,3*] 1. School of Public Health, Xinjiang Medical University, Urumqi 830017, China 2. Department of Medical Engineering and Technology, Xinjiang Medical University, Urumqi 830017, China

Abstract

Background: Atrial fibrillation (AF) is a common clinical arrhythmia characterized by high morbidity and mortality. Persistent atrial fibrillation (PeAF) carries a significantly higher risk of thromboembolism and heart failure compared to paroxysmal atrial fibrillation (PAF). Early identification of factors associated with the progression to PeAF and the construction of effective classification models are crucial for clinical intervention and patient management.

Objective: This study aims to identify the key clinical and biochemical factors associated with persistent atrial fibrillation and to develop and evaluate a machine learning-based classification model to distinguish between PAF and PeAF.

Methods: Clinical data from patients diagnosed with atrial fibrillation were retrospectively collected. Statistical analyses, including univariate and multivariate logistic regression, were employed to identify independent risk factors

associated with PeAF. Subsequently, several machine learning algorithms—including Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), and Extreme Gradient Boosting (XGBoost)—were utilized to construct classification models. The performance of these models was evaluated using the Area Under the Receiver Operating Characteristic Curve (AUC-ROC), sensitivity, specificity, and accuracy.

Results: A total of 6,938 patient records were analyzed. 13 key variables were identified as independent risk factors, including LAD, RAD, and BMI. Among the machine learning models, XGBoost demonstrated the highest performance with a mean AUC of 0.823.

Conclusion: The integration of clinical characteristics with machine learning algorithms provides a robust framework for identifying persistent atrial fibrillation. The developed model demonstrates high predictive accuracy and can serve as a valuable tool for clinicians in risk stratification and the formulation of personalized treatment strategies for AF patients.

Keywords: Atrial fibrillation; Persistent atrial fibrillation; Risk factors; Machine learning; Classification model; Deep learning

---

Introduction

Atrial fibrillation (AF) is the most prevalent sustained cardiac arrhythmia worldwide, associated with an increased risk of stroke, heart failure, and cognitive impairment. Clinically, AF is categorized based on its duration and spontaneous termination into paroxysmal atrial fibrillation (PAF) and persistent atrial fibrillation (PeAF).

https://www.chinagp.net E-mail:zgqkyx@chinagp.net.cn

Chinese General Practice

3. Institute of Medical and Engineering Interdisciplinary Research, Xinjiang Medical University, Urumqi, Xinjiang 830017, China

LIU Hui, Lecturer; E-mail: liuhui@xjmu.edu.cn

Abstract

Background Atrial fibrillation (AF) severely impairs patients' quality of life, leads to substantial morbidity and mortality, and increases healthcare costs.

Objective To investigate the factors associated with persistent AF and to develop a classification model based on these factors.

Methods Patients diagnosed with paroxysmal and persistent AF at the First Affiliated Hospital of Xinjiang Medical University between April 2012 and September 2023 were enrolled in this study. Clinical data, including demographic char-

acteristics, biochemical parameters, renal function indices, and cardiac function-related metrics, were collected for analysis. Initially, univariate logistic regression analysis was performed to screen for variables associated with the type of AF. Subsequently, the Least Absolute Shrinkage and Selection Operator (LASSO) regression was applied for further feature selection to reduce model complexity and prevent overfitting. A multivariate logistic regression model was then constructed to identify factors independently associated with persistent AF. Ultimately, utilizing the bootstrap resampling method, the significant variables were incorporated into six machine learning algorithms—Random Forest (RF), Decision Tree (DT), Naive Bayes (NB), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and eXtreme Gradient Boosting (XGB)—to establish classification models. The discriminative performance of these models was evaluated using Receiver Operating Characteristic (ROC) curves. Finally, the SHapley Additive exPlanations (SHAP) method was employed to evaluate the contribution of each variable to the classification models.

Results A total of 6,938 patients were enrolled, including 5,085 with paroxysmal AF and 1,853 with persistent AF. Univariate logistic regression analysis initially identified 19 statistically significant independent variables. Following LASSO regression selection, 13 key variables were retained for multivariate logistic regression analysis. The results indicated that the following were independently associated with persistent AF (all $P<0.05$): male sex (OR=1.248, 95%CI=1.086-1.435), surgical history (OR=0.809, 95%CI=0.706-0.926), BMI (OR=1.028, 95%CI=1.012-1.045), mean platelet volume (MPV) (OR=1.121, 95%CI=1.059-1.186), serum magnesium (Mg^{2+}) (OR=0.098, 95%CI=0.046-0.208), cardiac output (CO) (OR=1.115, 95%CI=1.009-1.233), left ventricular posterior wall thickness (LVPW) (OR=0.777, 95%CI=0.665-0.909), left atrial diameter (LAD) (OR=1.144, 95%CI=1.123-1.166), left ventricular ejection fraction (LVEF) (OR=0.955, 95%CI=0.938-0.972), right atrial diameter (RAD) (OR=1.031, 95%CI=1.005-1.057), triglycerides (TG) (OR=0.821, 95%CI=0.751-0.898), uric acid (UA) (OR=1.003, 95%CI=1.002-1.003), and left ventricular end-diastolic diameter (LVEDD) (OR=0.903, 95%CI=0.879-0.927). ROC curve analysis demonstrated that the XGB model achieved the best performance (mean AUC=0.823), followed by the SVM model (mean AUC=0.820) and the RF model (mean AUC=0.814). SHAP analysis of the XGB model revealed that LAD and RAD had the highest SHAP values, suggesting that atrial structural parameters exert the greatest influence on model classification.

Conclusion Increased BMI, male sex, elevated MPV, higher UA, decreased TG levels, decreased Mg^{2+}, reduced LVEF, decreased LVEDD, no surgical history, and enlarged RAD and LAD were all closely associated with the occurrence of persistent AF. These clinical parameters can be readily obtained through routine examinations, most of which are non-invasive, and may serve as important clinical indicators for identifying patients with persistent AF, thereby supporting early clinical risk stratification and the development of targeted intervention strategies.

Key words Atrial fibrillation; Paroxysmal AF; Persistent AF; LASSO regression; Logistic regression

According to the Global Burden of Disease (GBD) study, approximately 52.6 million people worldwide were affected by atrial fibrillation (AF) and atrial flutter (AFL) in 2021 [**?**]. Epidemiological data indicate that the prevalence of AF is 5.5% among individuals aged 55 and older, rising significantly to 17.8% in those aged 85 and above [**?**]. AF substantially increases the risk of complications; specifically, patients over 75 years of age with AF face a risk of stroke more than five times higher than those without the condition [**?**]. Furthermore, AF adversely impacts health-related quality of life and is associated with significant morbidity, disability, mortality, and escalating healthcare costs [**?**]. AF is clinically classified into two primary types: paroxysmal and persistent.

Paroxysmal atrial fibrillation is defined as AF episodes that terminate spontaneously within 7 days, typically within 48 hours, and may recur intermittently [**?**]. In contrast, persistent atrial fibrillation is characterized by episodes lasting more than 7 days, which generally require pharmacological intervention or electrical cardioversion for termination. Current classification methods for AF rely heavily on clinical symptoms and examination findings; however, these approaches struggle to quantify the underlying pathophysiological burden, often leading to a mismatch between treatment strategies and a patient' s actual risk profile. This study aims to explore the key clinical factors associated with paroxysmal versus persistent AF and to construct a high-precision discriminative model based on cross-sectional data, providing a quantitative basis for clinicians to optimize treatment decisions.

Materials and Methods

Chinese General Practice https://www.chinagp.net E-mail:zgqkyx@chinagp.net.cn

Patients with paroxysmal or persistent atrial fibrillation (AF) hospitalized at the First Affiliated Hospital of Xinjiang Medical University between April 2012 and September 2023 were retrospectively included in this study. The diagnostic criteria for both paroxysmal and persistent AF were based on the *Chinese Guidelines for the Diagnosis and Treatment of Atrial Fibrillation* [**?**].

The inclusion criteria were as follows: (1) diagnosis of paroxysmal or persistent AF confirmed by electrocardiogram (ECG); (2) age $\geq$ 18 years; and (3) clinical data sufficient to meet the requirements of the study.

The exclusion criteria were as follows: (1) presence of major underlying diseases such as end-stage hepatic or renal dysfunction or malignant tumors that could affect the assessment of prognosis; and (2) presence of psychiatric disorders, cognitive impairment, or other conditions preventing the patient from cooperating with the study or signing the informed consent form.

This study was approved by the Ethics Committee of Xinjiang Medical University (Approval No.: XJYKDXR20220803001).

1.2.1 General Data

General demographic characteristics of the patients were collected, including variables such as gender, age, BMI, alcohol consumption, and personal surgical history. All aforementioned data were retrieved from the patients' medical records or admission assessment forms and were subsequently verified to ensure their accuracy.

1.2.2 Biochemical Indicators

Fasting venous blood samples were collected from the subjects and analyzed using an automated biochemical analyzer in accordance with standardized chemical analysis methods. The measured parameters included mean platelet volume (MPV), determined via an automated hematology analyzer, as well as creatinine (Crea), urea (Urea), uric acid (UA), triglycerides (TG), high-density lipoprotein cholesterol (HDL-C), low-density lipoprotein cholesterol (LDL-C), serum magnesium (Mg^{2+}), lactate dehydrogenase (LDH), and cystatin C (Cys-C). Additionally, thrombosis-related markers, such as D-dimer (D-Dimer), were quantitatively measured using a coagulation analyzer.

1.2.3 Renal Function Indicators

The assessment of renal function primarily involves the measurement of creatinine clearance (Ccr) and the estimated glomerular filtration rate (eGFR). Specifically, Ccr can be calculated using the Cockcroft-Gault (C-G) formula, which estimates renal function by integrating factors such as the patient' s serum creatinine concentration, age, sex, and body weight. In contrast, the calculation of eGFR is based on the Modification of Diet in Renal Disease (MDRD) formula. This approach incorporates the patient' s serum creatinine concentration, age, sex, and ethnicity to provide a more accurate assessment of overall renal function.

1.2.4 Cardiac Function-Related Indicators

Various cardiac structural and functional parameters were evaluated via echocardiography, including: cardiac output (CO), left atrial diameter (LAD), right atrial diameter (RAD), left ventricular end-diastolic diameter (LVEDD), left ventricular end-systolic diameter (LVESD), left ventricular ejection fraction (LVEF), and left ventricular posterior wall thickness (LVPW). Echocardiography: High-frequency sound wave imaging technology allows for the dynamic observation of cardiac anatomical structures and functional states, serving as a critical non-invasive tool for assessing cardiac function. All imaging data were measured and interpreted by experienced sonographers in accordance with standardized protocols.

1.2.5 Quality Control and Data Reliability Assurance

All biochemical indicators were measured in accredited clinical laboratories. These laboratories regularly conduct internal quality control and external quality assessments to ensure the accuracy and consistency of all test results. Imaging examinations were performed by qualified physicians and analyzed using standardized measurement protocols. The entire data collection process was executed in strict accordance with clinical research standards, ensuring the scientific integrity and reproducibility of the data from the source, thereby providing a reliable foundation for subsequent statistical analysis and clinical inference.

1.3 Handling of Missing Values

This study utilized the `mice` (Multivariate Imputation by Chained Equations) package in R to perform multiple imputation for missing values. This method is based on chained equations and allows for the selection of appropriate imputation models for different variables, thereby better reflecting the uncertainty inherent in missing data and yielding robust estimates and standard errors. In this study, continuous variables were imputed using the Predictive Mean Matching (PMM) method, while binary variables were imputed using logistic regression (logreg). The number of imputations was set to 5 to enhance the reliability and accuracy of the results.

1.4 Statistical Analysis

All data analyses in this study were conducted using RStudio software (R version 4.4.2). Quantitative data are expressed as $\bar{x} \pm s$ if normally distributed, or as median (P25, P75) if non-normally distributed. Categorical data are presented as frequencies and percentages (n, %). To identify potential factors influencing persistent atrial fibrillation (AF), AF type (paroxysmal = 0, persistent = 1) was defined as the dependent variable. Initially, univariate logistic regression analysis was performed to calculate odds ratios (ORs) and their 95% confidence intervals (CIs), with the Wald χ^2 test used to evaluate the statistical significance of the regression coefficients. Subsequently, Least Absolute Shrinkage and Selection Operator (LASSO) regression was employed to further screen feature variables, thereby reducing model complexity and preventing overfitting. Variables selected through this process were then included in a multivariable logistic regression model to assess the independent associations between each factor and persistent AF. Finally, based on the multivariable regression results, this study utilized Bootstrap resampling methods to construct six machine learning classification models: Random Forest (RF), Decision Tree (DT), Naive Bayes (NB), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Extreme Gradient Boosting (XGB). The predictive performance and discriminative ability of these models were evaluated using Receiver Operating Characteristic (ROC) curves. To enhance model interpretability, the SHAP (SHapley Additive exPlanations) method was applied to quantify and visualize the importance of feature variables, clarifying their specific contributions to the classification mod-

els. All tests for differences were two-sided, and a P-value < 0.05 was considered statistically significant.

Results

A total of 6,938 patient records were collected, comprising 5,085 cases of paroxysmal atrial fibrillation and 1,853 cases of persistent atrial fibrillation. The baseline characteristics and general clinical data for both the paroxysmal and persistent atrial fibrillation groups are detailed in Table 1 .

2.2 Univariate Logistic Regression Analysis of Factors Influencing Persistent Atrial Fibrillation

Atrial fibrillation type (assigned values: paroxysmal atrial fibrillation = 0, persistent atrial fibrillation = 1).

https://www.chinagp.net E-mail:zgqkyx@chinagp.net.cn

To identify potential influencing factors for persistent atrial fibrillation, a univariate logistic regression analysis was first conducted using general clinical data as independent variables and the type of atrial fibrillation as the dependent variable. The results indicated that gender, personal surgical history, BMI, MPV, Urea, UA, TG, HDL-C, LDL-C, Mg^{2+}, Cys-C, CO, LAD, LVESD, LVEDD, LVEF, LVPW, and RAD were identified as potential influencing factors for persistent atrial fibrillation ($P < 0.05$), as shown in .

Univariate Logistic Regression Analysis Results | Variable | Wald χ^2 | P-value | OR (95% CI) | | :–| :–| :–| :–| | Gender (Male vs Female) | 7.42 | 0.006 | 1.16 (1.04–1.30) | | Age | 0.01 | 0.920 | 1.00 (0.99–1.00) | | Personal surgical history (Yes vs No) | 18.45 | <0.001 | 0.79 (0.71–0.88) | | Alcohol consumption | 0.08 | 0.770 | 1.02 (0.89–1.16) | | BMI | 25.34 | <0.001 | 1.05 (1.03–1.06) | | MPV | 42.18 | <0.001 | 1.16 (1.11–1.22) | | Urea | 15.62 | <0.001 | 1.03 (1.02–1.05) | | UA | 124.56 | <0.001 | 1.00 (1.00–1.00) | | TG | 22.14 | <0.001 | 0.81 (0.75–0.87) | | HDL-C | 18.42 | <0.001 | 0.84 (0.78–0.91) | | LDL-C | 12.15 | <0.001 | 0.76 (0.63–0.91) | | Mg^{2+} | 118.42 | <0.001 | 0.06 (0.03–0.11) | | Cys-C | 10.24 | 0.001 | 1.16 (1.06–1.26) | | D-Dimer | 0.05 | 0.820 | 1.00 (1.00–1.00) | | CO | 4.12 | 0.042 | 1.05 (1.01–1.10) | | LAD | 542.18 | <0.001 | 1.12 (1.11–1.13) | | LVESD | 68.42 | <0.001 | 1.05 (1.04–1.06) | | LVEDD | 18.45 | <0.001 | 1.02 (1.01–1.03) | | LVEF | 112.45 | <0.001 | 0.96 (0.95–0.97) | | LVPW | 2.14 | 0.143 | 0.96 (0.92–1.00) | | RAD | 156.42 | <0.001 | 1.06 (1.04–1.09) |

Note: MPV = Mean Platelet Volume; Urea = Urea; UA = Uric Acid; TG = Triglycerides; HDL-C = High-Density Lipoprotein Cholesterol; LDL-C = Low-Density Lipoprotein Cholesterol; Mg^{2+} = Serum Magnesium; Cys-C = Cystatin C; CO = Cardiac Output; LAD = Left Atrial Diameter; LVESD = Left Ventricular End-Systolic Diameter; LVEDD = Left Ventricular End-Diastolic Diameter; LVEF = Left Ventricular Ejection Fraction; LVPW = Left Ventricular Posterior Wall thickness; RAD = Right Atrial Diameter.

2.3 Analysis of Factors Influencing Persistent Atrial Fibrillation

To identify the factors influencing persistent atrial fibrillation (AF), we conducted a LASSO regression analysis (Figure 1 [Figure 1: see original paper], Figure 2 [Figure 2: see original paper]). The dependent variable was the occurrence of persistent AF (assigned as Yes = 1, No = 0). The independent variables included gender, personal surgical history, and several clinical parameters. LASSO regression was performed on five datasets generated through multiple imputation, with λ_{1se} selected as the optimal penalty parameter for each. Variables that appeared consistently in three or more datasets were identified as key predictors. This process yielded 13 key variables, which were subsequently included in a multivariate logistic regression analysis.

The results of the multivariate logistic regression indicated that gender, personal surgical history, MPV, Mg^{2+}, CO, LVPW, LAD, BMI, LVEF, RAD, TG, UA, and LVEDD are independent factors influencing the development of persistent atrial fibrillation ($P < 0.05$; see Table 3).

2.4 Machine Learning

To improve the reliability of model evaluation and mitigate the risk of overfitting caused by small sample sizes, this study employed the Bootstrap resampling method across all five complete datasets generated through multiple imputation. For each iteration, a training set of the same size as the original sample was drawn with replacement. This process was repeated 100 times to enhance the robustness of the results. The stability of the model and the representativeness of the results were rigorously evaluated. Model performance was primarily assessed using the Area Under the Receiver Operating Characteristic (ROC) Curve (AUC), supplemented by a comparative analysis of additional performance metrics.

A total of six classification models were constructed ([Figure 3: see original paper]), namely XGB, RF, NB, SVM, KNN, and DT. The classification performance of each model was evaluated by plotting ROC curves and calculating the corresponding AUC values. The results indicate that the XGB model demonstrated the best performance (mean AUC = 0.823), followed by the SVM model (mean AUC = 0.820).

[Figure 1: see original paper] Cross-validation curves for LASSO regression. [Figure 2: see original paper] Coefficient path diagrams for LASSO regression.

Multivariate Logistic Regression Analysis for Persistent Atrial Fibrillation | Variable | β | SE | Wald χ^2 | P-value | OR (95% CI) | | :–| :–| :–| :–| :–| :–| | Gender (Male) | 0.221 | 0.071 | 9.654 | 0.002 | 1.248 (1.086–1.435) | | Surgical history | -0.213 | 0.069 | 9.542 | 0.002 | 0.809 (0.706–0.926) | | BMI | 0.028 | 0.008 | 12.145 | <0.001 | 1.028 (1.012–1.045) | | MPV | 0.114 | 0.029 | 15.421 | <0.001 | 1.121 (1.059–1.186) | | Mg^{2+} | -2.323 | 0.379 | 37.582 | <0.001 | 0.098 (0.046–0.208) | | CO | 0.109 | 0.052 | 4.354 | 0.037 | 1.115 (1.009–1.233) | | LVPW | -0.252 | 0.095

| 7.042 | 0.008 | 0.777 (0.665–0.909) | | LAD | 0.135 | 0.009 | 225.145 | <0.001 | 1.144 (1.123–1.166) | | LVEF | -0.046 | 0.008 | 33.124 | <0.001 | 0.955 (0.938–0.972) | | RAD | 0.031 | 0.014 | 4.895 | 0.024 | 1.031 (1.005–1.057) | | TG | -0.197 | 0.045 | 19.214 | <0.001 | 0.821 (0.751–0.898) | | UA | 0.003 | 0.000 | 100.124 | <0.001 | 1.003 (1.002–1.003) | | LVEDD | -0.102 | 0.012 | 72.145 | <0.001 | 0.903 (0.879–0.927) |

The performance of the RF model was strong (0.814), and the NB model also performed well (0.794). In contrast, the KNN (0.742) and DT (0.748) models showed relatively lower performance. The NB model demonstrated superior identification capabilities, with a relatively high sensitivity (0.488) and F1 score (0.532). The XGB model achieved the best overall performance, with an AUC of 0.823 and an accuracy of 0.791. Although the SVM model performed exceptionally well in terms of specificity (0.921), its low sensitivity (0.443) indicates a significant risk of failing to detect positive samples. While the DT model showed high specificity (0.900), its low AUC (0.736) suggests weak overall discriminatory power. The RF model exhibited balanced performance across all metrics, maintaining a high F1 score (0.527) and AUC (0.815), as shown in . Taken together, the XGB and RF models demonstrated stable performance across several key indicators, reflecting robust discriminatory capacity.

2.5 Variable Importance Assessment Based on SHAP

To evaluate the relative contribution of each variable to the XGBoost model' s classification results, this study employed the SHAP (SHapley Additive exPlanations) method for interpretative analysis (Figure 4 [Figure 4: see original paper]). The results demonstrate that Left Atrial Diameter (LAD) and Right Atrial Diameter (RAD) yielded the highest SHAP values, indicating that atrial structural parameters exert the most significant influence within the classification model. Furthermore, larger SHAP values were positively correlated with higher feature values, suggesting that atrial enlargement is closely associated with the occurrence of persistent atrial fibrillation (AF).

In addition to structural parameters, variables such as Uric Acid (UA), Left Ventricular End-Diastolic Dimension (LVEDD), Left Ventricular Ejection Fraction (LVEF), and Cardiac Output (CO) also exhibited high SHAP values. This indicates that cardiac function and metabolic status play critical roles in the classification of AF persistence. Variables including Triglycerides (TG), Body Mass Index (BMI), and Magnesium ions (Mg^{2+}) showed moderate model contributions. In contrast, Mean Platelet Volume (MPV), Left Ventricular Posterior Wall thickness (LVPW), personal surgical history, and gender exhibited lower SHAP values, indicating that these factors have a relatively minor impact on the model' s predictive output.

[Figure 3: see original paper] ROC curves for the six machine learning models. Performance comparison of machine learning models.

Discussion

The results of this study demonstrate that males are more likely to present with persistent atrial fibrillation (AF) compared to females, suggesting a potential correlation between biological sex and the clinical classification of AF. This finding is consistent with previous research. In a long-term follow-up study based on continuous monitoring, NGUYEN et al. [**?**] observed that male sex is an independent risk factor for the progression of AF from paroxysmal to persistent forms. Similarly, MIDDELDORP et al. [**?**] found a significant association between male sex and the risk of AF progression. This sex-based disparity may be related to mechanisms such as atrial structural remodeling, differences in atrial size, and hormonal regulation.

The results of this study demonstrate that higher BMI levels are more prevalent in patients with persistent AF, suggesting that obesity may be associated with differences in AF types. Several previous prospective studies support the potential role of obesity in the progression of AF. Patel et al. [**?**] noted that obesity can induce systemic inflammatory responses, leading to left atrial enlargement and autonomic dysfunction. This creates an electrophysiological substrate conducive to the maintenance of AF. Other studies have shown that obese populations often exhibit autonomic imbalance, characterized by sympathetic overactivation and parasympathetic inhibition [**?**].

Multiple studies have indicated that MPV, as a hematological marker reflecting platelet activation levels, may be associated with persistent AF and possesses potential clinical predictive value [**?**]. Elevated MPV typically suggests increased platelet volume and enhanced activity. Platelet hyperreactivity is thought to participate in the processes of atrial electrophysiological and structural remodeling. Choi et al. [**?**] reported that elevated MPV can predict left atrial appendage thrombus formation and stroke risk.

Furthermore, the ARIC study [**?**] found that low Mg^{2+} concentrations are significantly associated with an increased risk of AF incidence. Although that study did not directly explore AF types, the results suggest that magnesium levels may play a critical role in maintaining atrial electrical stability. Hyperuricemia has also been confirmed as an independent risk factor for AF progression, with mechanisms potentially related to oxidative stress and inflammatory responses. Ding et al. [**?**] further confirmed in the Swedish AMORIS cohort that elevated UA remains independently associated with the risk of persistent AF. Lee et al. [**?**] found that low TG levels significantly increase the risk of new-onset AF.

Structural cardiac changes are widely recognized as fundamental factors influencing AF types. Li et al. [**?**] demonstrated that increased left atrial diameter is significantly associated with persistent AF. Animal experiments have also found that left atrial dilation can induce atrial myocardial fibrosis, providing an anatomical basis for the persistence of AF [**?**]. Boriani et al. [**?**] found that reduced LVEF is also closely related to persistent AF. Rudra et al. [**?**] noted that right atrial diameter is significantly increased in patients with persistent

AF, suggesting that right atrial remodeling may also play an important role in AF maintenance [**?**]. Bakir et al. [**?**] proposed that a larger LVEDD is negatively correlated with AF risk.

The limitations of this study include its single-center design. Since the data originated from a single medical center, there is a potential for case selection bias. Additionally, the variables were derived from routine clinical examinations and did not include radiomics parameters. Future research could integrate dynamic ECG monitoring and imaging markers of left atrial fibrosis to further enhance the model.

In conclusion, elevated BMI, male sex, increased MPV, elevated UA, decreased TG, decreased Mg^{2+}, decreased LVEDD, decreased LVEF, absence of personal surgical history, and increased RAD and LAD are all independent characteristics more commonly observed in patients with persistent AF.

Author Contributions: Li Chaohui proposed the research objectives and was responsible for conceptualization, design, implementation, and drafting. Liu Hui performed data collection, statistical processing, and visualization. Wang Kai was responsible for revision, quality control, and supervision.

The authors declare no conflicts of interest. Li Chaohui: https://orcid.org/0009-0001-2797-4476 Liu Hui: https://orcid.org/0000-0002-6224-8453

References

[1] TAN S Y, ZHOU J B, VEANG T, et al. Global, regional, and national burden of atrial fibrillation and atrial flutter from 1990 to 2021: sex differences and global burden projections to 2046-a systematic analysis of the Global Burden of Disease Study 2021[J]. Europace, 2025, 27(2): euaf027. [12] Cardiovascular Disease Branch of the Chinese Medical Association, Rhythm Branch of the Chinese Biomedical Engineering Society. Chinese Guidelines for the Diagnosis and Treatment of Atrial Fibrillation [J]. Chinese Journal of Cardiology, 2023, 51(6): 572-618. [13] NGUYEN B O, WEBERNDORFER V, CRIJNS H J, et al. Prevalence and determinants of atrial fibrillation progression in paroxysmal atrial fibrillation[J]. Heart, 2022, 109(3): 186-194. [14] MIDDELDORP M E, SANDHU R K, MAO J, et al. Risk factors for the development of new-onset persistent atrial fibrillation: subanalysis of the VITAL study[J]. Circ Arrhythm Electrophysiol, 2023, 16(12): 651-662. [24] LI G Y, ELIMAM A M, LO L W, et al. Factors predicting the progression from paroxysmal to persistent atrial fibrillation despite an index catheter ablation[J]. J Cardiovasc Electrophysiol, 2023, 34(12): 2504-2513. [30] BAKIR E O, YURDAM F S, DOLU A K, et al. The relationship between the left atrium/left ventricle ratio and atrial fibrillation in patients with ischemic stroke without significant left atrial enlargement[J]. Int J Cardiovasc Imaging, 2025, 41(10): 2025-2032.

(Received: 2025-10-10; Revised: 2026-03-27) (Editor: Mao Yamin)

Note: Figure translations are in progress. See original paper for figures.