**Predicting Stroke in Asian Patients with Atrial Fibrillation Using Machine Learning:**

A report from the KERALA-AF Registry, with external validation in the APHRS-AF Registry

Yang Chen MBChB MMSa; Ying Gue PhDa; Peter Calvert MBChBa; Dhiraj Gupta MDa; Garry McDowell PhDa,b; Jinbert Lordson Azariah MScc,d; Narayanan Namboodiri MD DMe;

Tommaso Bucci MDa,f; Jabir A. MD, DMg; Hung Fat Tse MD PhDh, Tze-Fan Chao MD PhDi;

Gregory Y. H. Lip MDa,j\*; Bahuleyan Charantharayil Gopalank\*;

on behalf of the KERALA-AF Registry & APHRS-AF Registry Investigators‡

[\*joint senior authors]

1. Liverpool Centre for Cardiovascular Science at University of Liverpool, Liverpool John Moores University and Liverpool Heart and Chest Hospital, Liverpool, United Kingdom
2. School of Pharmacy and Biomolecular Sciences, Liverpool John Moores University, Liverpool, United Kingdom
3. Department of Clinical Research, Ananthapuri Hospitals and Research Institute, Thiruvananthapuram, India
4. Department of Research, Global Institute of Public Health, Trivandrum, India
5. Sree Chitra Tirunal Institute for Medical Sciences and Technology, Trivandrum, India
6. Department of General and Specialized Surgery, Sapienza University of Rome, Rome, Italy
7. Lisie Heart Institute, Ernakulam, India
8. Division of Cardiology, Department of Medicine, School of Clinical Medicine; Queen Mary Hospital, the University of Hong Kong, Hong Kong SAR, China;
9. Institute of Clinical Medicine, and Cardiovascular Research Center, National Yang Ming Chiao Tung University, Taipei, Taiwan;
10. Division of Cardiology, Department of Medicine, Taipei Veterans General Hospital, Taipei, Taiwan; Danish Centre for Health Services Research, Department of Clinical Medicine, Aalborg University, Aalborg, DK-9220, Denmark
11. Department of Cardiology, Ananthapuri Hospitals and Research Institute, Thiruvananthapuram, India

**Corresponding authors:**

Prof. Bahuleyan Charantharayil Gopalan. E-mail: [bahuleyan2001@yahoo.co.uk](mailto:bahuleyan2001@yahoo.co.uk)

Prof. Gregory Y.H. Lip. E-mail: [gregory.lip@liverpool](mailto:gregory.lip@liverpool).ac.uk

**Keywords:** atrial fibrillation, stroke, machine learning, prediction, Kerala, South Asia

**Word Count:** 3351

**Abstract**

**Background** Atrial fibrillation (AF) is a significant risk factor for stroke. Predicting stroke in patients with AF is critical for risk stratification and treatment decision-making. Based on the higher stroke associated with AF in the South Asian population, we constructed a 1-year stroke prediction model using machine learning (ML) methods.

**Methods** Utilising KERALA-AF, a prospective registry, we included patients who had completed 1-year follow-up. Variables were first filtered before utilising five ML classifiers (light gradient boosting machine [LightGBM], random forest [RF], logistic regression, support vector machine, and multilayer perceptron) to predict 1-year stroke associated with AF, and applied 5-fold cross-validation to optimise hyperparameters. External validation was performed in another prospective AF cohort, the APHRS-AF registry.

**Results** We studied 2101 patients (median age 68.0 years; 46.2% female) with median CHA2DS2-VASc score 3.0 (IQR 2.0-4.0) in KERALA-AF registry. Twelve variables, including chronic kidney disease, age ≥ 75 years, and hypertension, were utilised as parameters in the model. Model performance was assessed using area under the receiver operator characteristic curve (AUC) and G-mean (specifically applied to the unbalanced data). An RF approach showed the best predictive performance in the internal validation (as compared to other classifiers and the CHA2DS2-VASc score) with AUC and G-mean of 0.821 (95%CI 0.816-0.825) and 0.427, respectively. In the external validation using the APHRS-AF registry, the LightGBM showed the best predictive performance with AUC and G-mean of 0.670 (95% CI 0.665-0.674) and 0.083, respectively.

**Conclusion** ML is a powerful non-invasive clinical tool for early identification of stroke risk in the South Asian AF population. We report the first demonstration of its applicability in an Indian prospective cohort, although the more modest prediction on external validation in a separate multinational Asian registry suggests the need for ethnic-specific ML models.

**Introduction**

Atrial fibrillation (AF) is the commonest cardiac arrhythmia and is associated with an increased risk of stroke and mortality. 1 A recent cross-national comparative study has shown that the age- and sex-standardised AF prevalence varies considerably by region, with South Asia having the lowest rate of approximately 30-60 cases per 100,000 people. 2 Despite South Asia's relatively low prevalence of AF, as a region with rapid population growth, the burden of AF will increase. 3

Recent studies have demonstrated ethnic and regional disparities in the occurrence of stroke associated with AF. In a multinational cohort study, the 1-year stroke incidence was approximately 4.3% in Asian AF (including India, China, and South-East Asian countries), which was higher than that (2.5%) in White AF (North America, Western Europe, and Australia). 4 In a multi-ethnic study, the risk of stroke as estimated using the CHA2DS2-VASc score was 1.67 times higher in South Asians than in Whites. 5 However, in a non-anticoagulated Asian AF cohort of over 180,000 people, the CHA2DS2-VASc score was not highly discriminatory for ischaemic stroke, with the area under curve (AUC) of 0.698. 6 Additionally, the CHA2DS2-VASc score does not include many other risk factors for stroke, such as chronic kidney disease (CKD) 7, type of AF 8, and electrocardiographic (ECG) features 9, nor does it distinguish between the relative importance of individual risk factors (for example, both hypertension and diabetes score +1, but may represent different levels of risk in reality). Also, stroke risk varies in different ethnicities, and many prior studies have focused on white Caucasian and east Asian cohorts, with limited data on stroke risk prediction in south Asian cohorts.

Thus, there is a need for better risk prediction models beyond the CHA2DS2-VASc score, particularly in South Asian AF. Generally, predictive models are developed using traditional logistic regression, which is based on the assumption of the linear relationship between variable and outcome, but the model may be unstable with increasing quantities of variables. However, machine learning (ML) enables the construction of predictive models by controlling variables' covariance through regularisation and thus inputting more factors, which is increasingly used in clinical studies for outcome prediction. 10 Current studies applying ML to predict stroke in the Asian AF cohort are limited, with only Korea and Japan, whose ML models with more input factors have been shown to perform better than the CHA2DS2-VASc score. 11, 12 As far as we are aware, no ML study predicting stroke associated with AF in South Asia has been previously reported.

In this ancillary analysis from a prospective Indian cohort of patients with AF, we aimed to build a novel model for predicting stroke associated with AF in a South Asian cohort based on ML, effectively incorporating various risk factors for more accurate stroke risk stratification.

**Methods**

**Study participants**

The KERALA-AF registry is an ongoing prospective, multicentre cohort study of AF patients in the Kerala region of India, and is the largest prospective AF study in South Asia. The proposal and results of this study with 1-year follow-up have been previously reported. 13, 14 During 2016-2017, 3401 AF patients were recruited from 53 independent centres. As an external validation cohort, we used the Asia-Pacific Heart Rhythm Society Atrial Fibrillation (APHRS-AF) registry which was also a prospective multinational multicentre cohort study of AF patients, with a total of 4,664 AF patients recruited from 52 independent centres in five Asian countries (but not India) from the end of 2015 to the beginning of 2017, and with 1-year follow-up. 15

**Inclusion and exclusion criteria**

In our KERALA-AF analysis, we included non-valvular AF (NVAF) patients and excluded those who were lost to follow-up prior to 1-year. The APHRS-AF registry for external validation included 1531 NVAF patients according to the same inclusion and exclusion criteria. (Figure 1).

**Data collection and outcome**

Demographic characteristics, lifestyle, disease history, comorbidities, pharmacological and surgical treatment, imaging features and laboratory parameters were collected at baseline. The primary outcome of interest was stroke at 1-year follow-up.

**Definitions**

Using categorical variables makes the interpretation of the model clearer and more intuitive, making it easier to share and explain how the model works to the applicants. We converted the continuous variables into categorical variables to include them as alternative variables in the model construction (their names and definitions are given in Supplementary Table S1).

**Feature selection and model construction**

The overall dataset was randomly split into a training and validation cohort on a 7:3 ratio. Based on the limited positive events in our analysis, incorporating too many variables in the predictive model may increase the risk of overfitting and cause the machine learning to over-memorise noise in the training cohort, making generalisation difficult. When explaining the model to non-technical applicants, fewer variables can make interpretation clearer and simpler. Also, reducing variables can decrease the computational and storage costs of the model, making it more practical and efficient. However, too few variables may lead to loss of information. Therefore, we selected 12 variables in order to strike a balance between information sufficiency and model simplicity, leading to the construction of more accurate, robust and explanatory predictive models. We filtered the 12 most critical features in the training cohort using the Chi-square test, which is most applicable to categorical variables. To avoid variable collinearity and multicollinearity, we performed Pearson correlation analysis and calculated the variance inflation factor (VIF).

We applied these features to 5 ML classifiers commonly used in medical binary problems to predict 1-year stroke associated with AF, including light gradient boosting machine (LightGBM), random forest (RF), logistic regression (LR), support vector machine (SVM), and multilayer perceptron (MLP). The ratio of positive to negative events in our dataset was approximately 1:26, so this is an unbalanced dataset, and we allocated sample weights to each category when constructing our model. Then, we used grid search and five-fold cross-validation on the training cohort to optimise and obtain the best hyperparameters for each ML classifier.

**Evaluation of parameters**

We plotted the receiver operating characteristic curve (ROC), obtained the mean area under the receiver operator characteristic curve (AUC) and 95% confidence interval (CI) of each classifier using 1000 bootstrapping iterations to assess their performances in the internal and external validation cohorts. When performing external validation, given the heterogeneity of the external cohort, we would attempt to retrain the model in 20% of the external cohort using the Fine-tuning technique to better adapt it to the external validation. Given that our data were unbalanced, we calculated the accuracy, specificity, sensitivity, precision, recall, F1-score and G-mean for each classifier, respectively, to assess the differences in their performance, and compared these with the CHA2DS2-VASc score.

**Online tool for the prediction model**

We developed a web-based tool utilising the predictive model with a simple user-friendly interface that allows clinicians to quickly and intuitively determine one-year risk of stroke in NVAF patients by collecting and inputting the appropriate features into the model to assist in making treatment decisions.

**Statistical analysis**

We used STATA (version 17) to clean the original dataset. Variables with more than 50% missing values were discarded because the values populated for these variables may not be sufficiently accurate or reliable, even when techniques such as multiple imputation are used. For other variables with missing values, we applied multiple imputation using the package 'miceforest' in Python (version 3.11.4). The variables were described using SPSS (version 27). For continuous variables, mean with standard deviations or median with interquartile range (IQR) were used based on the distribution, and t-tests or Mann-Whitney U tests were used to compare differences between stroke and non-stroke groups. For categorical variables, counts with percentages were used, and Fisher's exact test and Chi-square test were used to compare differences between groups. All statistical significance levels were set at two-tailed *P* < 0.05. Feature selection and model construction were implemented in Python (version 3.11.4), with the packages Scikit-learn (version 1.2.2) and lightgbm (version 3.3.5). Furthermore, in the best model, each sample produced a corresponding prediction value, and each feature in that sample was assigned a specific value, SHapley Additive exPlanation (SHAP) value, 16 to explain the importance of the feature to the model, and we used the Python's SHAP package (version 3.11.4).

**Results**

**Patient Characteristics**

From the KERALA-AF registry, 2,101 NVAF patients with completed follow-up were included in the analysis. The median age was 68.0 years (IQR: 60.0 to 76.0), and 979 (46.2%) were female, AF treatment was predominantly rate control (83.3%), and common comorbidities were hypertension (61.3%), diabetes (37.2%), dyslipidaemia (46.8%) and chronic kidney disease (CKD) (50.3%). 83 (4.0%) were in the stroke group (Table 1). Compared to the non-stroke group, the stroke group were older (median: 75.0 vs 68.0), more frequently female (61.4% vs 45.8%), and had higher rates of hypertension (79.5% vs. 60.5%) and CKD (69.9% vs. 49.7%), as well as higher CHA2DS2-VASc scores (median: 4.0 vs 3.0).

**Feature Selection**

We transformed continuous variables into categorical variables based on the corresponding definitions given in Supplementary Table S2. To ensure good predictive performance of the model, redundant variables needed to be removed. Two clinicians carefully assessed these variables, further filtered using the Chi-square test. Finally, the 12 best variables were applied to construct the prediction model.

**Collinearity Test**

To avoid serious collinearity among the variables within the model, we conducted a Pearson correlation analysis and plotted the heatmap. As shown in Figure 2A, the coefficients between all variables were less than 0.4, which implied no strong correlation.

The VIF was calculated to perform the multicollinearity test. As displayed in Figure 2B, the VIF values of all variables were less than 2.5, suggesting weak multicollinearity. Therefore, these variables could avoid the negative impact of variable collinearity to be effectively used for the predictive model.

**Model Construction**

Based on the 12 selected features, we built a predictive model for 1-year stroke. Using commonly applied classifiers for medical binary problems, including LightGBM, RF, LR, SVM, and MLP, the best hyperparameters of each classifier were obtained after 5-fold cross-validation in the training cohort (Supplementary Table S3).

**Model Evaluation**

In the internal validation cohort (Figure 3A), SVM obtained the highest AUC (0.835, 95% CI 0.831-0.839) and LR obtained the lowest AUC (0.723, 95% CI 0.720-0.727) among the five classifiers, but all classifiers had AUCs higher than the CHA2DS2-VASc (0.665, 95% CI 0.663-0.667). Further calculating the other metrics for the classifiers (Table 2), RF had the highest F1-score (0.417) and G-mean (0.427), thus RF had the best classification ability in that unbalanced data. So, RF was considered as the best classifier in internal validation.

According to the previous result of feature selection, the characteristics of the external validation cohort are presented in Supplementary Table S4. Since AST values were not collected from participants in the APHRS-AF registry, we used liver disease as a replacement. After using Fine-tuning, only LightGBM's performance was improved. Thus, we presented the ROCs and AUCs of LightGBM post Fine-tuning and the other classifiers without Fine-tuning (Figure 3B), with LightGBM obtaining the highest AUC (0.670, 95% CI 0.665-0.674) and MLP obtaining the lowest AUC (0.554, 95% CI 0.549-0.560) among the five classifiers. Only LightGBM and LR had AUCs higher than the CHA2DS2-VASc (0.615, 95% CI 0.611-0.619). For the other metrics for the classifiers (Table 2), LightGBM had the highest F1-score (0.039) and G-mean (0.083). Therefore, LightGBM was considered as the best classifier in external validation.

**Feature Importance**

To further identify the most influential features in the RF and LightGBM, we calculated and visualised the SHAP for each feature. According to Figure 4, the top-to-bottom position on the Y-axis indicates the order of importance of all variables. The top five risk features of RF were CKD, age ≥ 75, hypertension, diuretic use, and abnormal AST. The top five risk features of LightGBM were CKD, age ≥ 75, prior CVA/TIA/SE, enlarged LA size (≥ moderate), and AF treatment.

**Online Prediction Tools**

Based on the RF and LightGBM model, web-based tools were constructed with a simple and user-friendly interface containing options corresponding to the 12 features in the model. Using RF’s as an example, specific features can be input to obtain an intuitive score. the probability of output was used to assess outcome risk, so we set the optimal threshold of 0.438, with higher than 0.438 preferring 1-year stroke and lower than 0.438 preferring no stroke. Figure 5 demonstrates the utility of the online tool; in this case: female sex, age ≥ 75 years, hypertension, diabetes, CKD, abnormal AST, atrial fibrillation rate control, and diuretic use gave the patient a predictive function score of 0.74, so the predicted outcome was stroke.

**Discussion**

In our study, we developed a reasonably accurate ML model for personalised estimation of 1-year stroke associated with non-valvular AF, using readily obtained variables from a South Asian cohort. Our model is the first stroke prediction model constructed in a South Asian (Indian) AF cohort that incorporates potential risk factors not included in the CHA2DS2-VASc score with better predictive performance. The more modest prediction on external validation in a separate multinational Asian registry suggests the need for ethnic-specific ML models.

The incidence of 1-year stroke associated with AF in our study was approximately 4.0%. One multinational cohort study of 47 countries showed that the occurrence of stroke associated with AF in Southeast Asia was around 7% at 1 year, which was higher than our result. 4 This may be related to the AF cohort in Southeast Asia being older (median age 72), with more hypertension (64%), more previous stroke/TIA (22%), and less oral anticoagulation therapy use (50%); however, the North American, Western European, and Australian regional cohorts had similar baseline characteristics to our cohort, and our 1-year stroke rate in KERALA-AF was still approximately twice as high as theirs (2%). Additionally, a large systematic analysis of AF epidemiology in Asia reported that the annual stroke risk was approximately 3.0% in AF patients, which was broadly similar to our results. 17 Thus, the stroke rates in our study are consistent with existing evidence, and support the notion that South Asian people might have a higher stroke incidence than that seen in other regions.

How do our models compare with other prediction tools for stroke in single-centre Asian AF cohorts? Jung *et al.* constructed a prediction model using variables of demographic information, history of disease, and health screening to predict 5-year stroke in a Korean AF cohort of more than 750,000 participants. 11 Their best model was the deep neural network (AUC: 0.722, 95% CI: 0.718-0.726, F1-score: 0.223), which was lower than our RF and LightGBM, probably because they did not incorporate some of the known potential risk factors for stroke. Alternatively, as their model included larger numbers for significantly longer follow-up than in our study, it is possible that model accuracy over this timeframe deteriorates. For example, patient characteristics vary over time, and some may develop new co-morbidities such as hypertension or diabetes during follow-up, which may not be detected if only baseline characteristics are applied. Similarly, the longer follow-up continues for, the more competing mortality risks apply. The KERALA-AF registry is continuing follow-up and we hope to assess how our model performs in longer-term follow-up in the future.

The performance of our models in the external validation APHRS cohort was less satisfactory. Nishi *et al.* constructed a model for predicting stroke during the follow-up using the CatBoost algorithm in the Japanese non-anticoagulant AF cohort. 12 The model obtained an AUC of 0.82 but F1-score of only 0.26 for the internal cohort (where our models performed better), and AUC and F1-score of 0.72 and 0.18 for the external cohort (where our models performed worse). The better performance of our models in the internal validation cohort might result from methodological differences. For example, whilst Nishi *et al.* recognised that their cohorts were imbalanced with respect to outcome measures (7.9% stroke vs 92.1% non-stroke in the training cohort), they did not apply any methodology to manage this; however, it should be noted that the AUC and F1 scores for our best model in the external validation were only 0.67 and 0.08, probably because the 1-year stroke to non-stroke ratio in the APHRS-AF cohort was approximately 1:100, the more extremely imbalanced data may lead to worse performance. Also, although the APHRS-AF contained Asian AF patients (but not including India), there remains non-negligible heterogeneity with the South Asian AF cohort in KERALA-AF, ie. Asians are not homogeneous. Nonetheless, although the performance of our models in external validation is currently less satisfactory, we remain confident that ML can potentially be used as a precision prediction tool for stroke in South Asian AF population, with the caveat that ethnic-specific ML models may be needed for different ethnic groups.

There was no serious overfitting occurs in our model during the training process and the performance of models may improve a little as the amount of data are further increased. Although the accuracy of the models appears to be very high, the scarcity of positive samples may result in the difficulty predicting positive events. Our model overcomes this limitation by applying “class weights”, allowing higher weighting of the smaller number of positive events, thereby improving performance. This is one of the most valuable techniques for coping with modelling in unbalanced data.

SHAP explains the relative importance of each variable in the ML model. The variables in our best model were partially similar to the CHA2DS2-VASc score (age, female, prior CVA/TIA/SE, hypertension, diabetes). We also added widely accepted stroke risk factors, e.g., CKD, enlarged LA (≥ moderate), persistent AF, and mitral valve involvement. 8, 18, 19, 20 We also included several potential additional stroke risk factors, AF treatment strategy, abnormal AST and diuretic use. These may be relevant, as Weng *et al*. reported a lower incidence of IS (adjusted HR: 0.65, *P* = 0.002) in AF patients with rhythm control than with rate control. 21 Also, Choi *et al.* demonstrated that higher AST was significantly associated with the IS occurrence (adjusted HR: 1.04, 95% CI: 1.03-1.05, *P* < 0.001). 22 Green's group showed that lower serum potassium triggered by diuretic use was significantly associated with increased stroke risk (relative risk: 2.5, 95% CI: 1.7-3.5, *P* < 0.0001). 23 Alternatively, the use of diuretics may be reflective of heart failure, which is a well-described stroke risk factor. Despite these data supporting our results, whether they increase the stroke risk per se remains controversial. Overall, the predictions computed by our ML model are based on stroke-related variables, with significantly better overall performance than the CHA2DS2-VASc score.

*Limitations*

Several important limitations of this study must be emphasised. First, despite the performance of our models in the external validation was less satisfactory, most still outperformed the clinical factor based CHA2DS2-VASc score. Second, due to the imbalance of the dataset, although the several ML algorithms show good discrimination in this respect, it still needs to be treated with caution and further validation in larger cohorts is required. Third, the number of variables in the initial collection exceeded 100, and two stroke experts preselected the variables for inclusion in the feature screening before incorporation, which could not wholly exclude personal bias. Fourth, because of the small positive sample in our study, we included only the 12 best variables to prevent model overfitting, but we may have missed some potentially essential variables. Fifth, about 16% of patients were lost follow-up and excluding these patients may cause bias; however, this is a common exclusion criterion in real world observational cohorts. Finally, patients were enrolled due to previously diagnosed AF, however the duration of AF was not known and hence our model performance may underestimate stroke risk in those with longer-standing disease, or overestimate risk in those with new onset AF.

**Conclusion**

In this first demonstration of ML’s applicability in a South Asian cohort, we propose novel models based on the largest AF cohort in India using ML to predict 1-year stroke associated with AF, thereby enhancing monitoring and preventing stroke at an early stage. The poorer prediction on external validation in a separate multinational Asian (but non-Indian) registry suggests the need for ethnic-specific ML models. The results of internal and external validation showed that our ML models had better performance than the CHA2DS2-VASc score.

**Acknowledgments**

The authors thank all the participants and researchers in the KERALA registry.

**Contributions**

All Kerala AF Registry investigators contributed to data collection. Y.C. contributed to data processing and analysis. Y.G., P.C., D.G., G.M., J.L.A., N.N., J.A., G.Y.H.L., and B.C.G contributed to drafting of the manuscript. All authors contributed to critical revision of the manuscript and final approval.

**Declarations, and Conflicts of Interest**

G.Y.H.L. reports: Consultant and speaker for BMS/Pfizer, Boehringer Ingelheim, Daiichi-Sankyo, Anthos. No fees are received personally. G.Y.H.L. is a National Institute for Health and Care Research (NIHR) Senior Investigator and co-principal investigator of the AFFIRMO project on multimorbidity in AF, which has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 899871.

Other authors report no conflicts of interest.

**Funding**

No funding was received towards this work.

**Reference**

1. Lip GYH, Gue Y, Zhang J, Chao T-F, Calkins H, Potpara T. Stroke prevention in atrial fibrillation. Trends Cardiovasc Med. 2022;32(8):501-10.

2. Joseph PG, Healey JS, Raina P*, et al.* Global variations in the prevalence, treatment, and impact of atrial fibrillation in a multi-national cohort of 153 152 middle-aged individuals. Cardiovasc Res. 2021;117(6):1523-31.

3. Tse H-F, Wang Y-J, Ahmed Ai-Abdullah M*, et al.* Stroke prevention in atrial fibrillation--an Asian stroke perspective. Heart Rhythm. 2013;10(7):1082-8.

4. Healey JS, Oldgren J, Ezekowitz M*, et al.* Occurrence of death and stroke in patients in 47 countries 1 year after presenting with atrial fibrillation: a cohort study. Lancet. 2016;388(10050):1161-9.

5. Mathur R, Pollara E, Hull S, Schofield P, Ashworth M, Robson J. Ethnicity and stroke risk in patients with atrial fibrillation. Heart. 2013;99(15):1087-92.

6. Chao TF, Liu CJ, Tuan TC*, et al.* Comparisons of CHADS2 and CHA2DS2-VASc scores for stroke risk stratification in atrial fibrillation: Which scoring system should be used for Asians? Heart Rhythm. 2016;13(1):46-53.

7. Olesen JB, Lip GY, Kamper AL*, et al.* Stroke and bleeding in atrial fibrillation with chronic kidney disease. N Engl J Med. 2012;367(7):625-35.

8. Steinberg BA, Hellkamp AS, Lokhnygina Y*, et al.* Higher risk of death and stroke in patients with persistent vs. paroxysmal atrial fibrillation: results from the ROCKET-AF Trial. Eur Heart J. 2015;36(5):288-96.

9. O'Neal WT, Howard VJ, Kleindorfer D*, et al.* Interrelationship between electrocardiographic left ventricular hypertrophy, QT prolongation, and ischaemic stroke: the REasons for Geographic and Racial Differences in Stroke Study. Europace. 2016;18(5):767-72.

10. Lip GYH, Genaidy A, Tran G, Marroquin P, Estes C, Sloop S. Improving Stroke Risk Prediction in the General Population: A Comparative Assessment of Common Clinical Rules, a New Multimorbid Index, and Machine-Learning-Based Algorithms. Thrombosis and Haemostasis. 2022;122(1):142-50.

11. Jung S, Song MK, Lee E*, et al.* Predicting Ischemic Stroke in Patients with Atrial Fibrillation Using Machine Learning. Front Biosci (Landmark Ed). 2022;27(3):80.

12. Nishi H, Oishi N, Ogawa H*, et al.* Predicting cerebral infarction in patients with atrial fibrillation using machine learning: The Fushimi AF registry. J Cereb Blood Flow Metab. 2022;42(5):746-56.

13. Charantharayil Gopalan B, Namboodiri N, Abdullakutty J*, et al.* Kerala Atrial Fibrillation Registry: a prospective observational study on clinical characteristics, treatment pattern and outcome of atrial fibrillation in Kerala, India, cohort profile. BMJ Open. 2019;9(7):e025901.

14. Bahuleyan CG, Namboodiri N, Jabir A*, et al.* One-year clinical outcome of patients with nonvalvular atrial fibrillation: Insights from KERALA-AF registry. Indian Heart J. 2021;73(1):56-62.

15. Tse H-F, Teo W-S, Siu C-W*, et al.* Prognosis and treatment of atrial fibrillation in Asian cities: 1-year review of the Asia-Pacific Heart Rhythm Society Atrial Fibrillation Registry. Europace : European Pacing, Arrhythmias, and Cardiac Electrophysiology : Journal of the Working Groups On Cardiac Pacing, Arrhythmias, and Cardiac Cellular Electrophysiology of the European Society of Cardiology. 2022;24(12):1889-98.

16. Rodriguez-Perez R, Bajorath J. Interpretation of Compound Activity Predictions from Complex Machine Learning Models Using Local Approximations and Shapley Values. J Med Chem. 2020;63(16):8761-77.

17. Bai Y, Wang YL, Shantsila A, Lip GYH. The Global Burden of Atrial Fibrillation and Stroke: A Systematic Review of the Clinical Epidemiology of Atrial Fibrillation in Asia. Chest. 2017;152(4):810-20.

18. Ghoshal S, Freedman BI. Mechanisms of Stroke in Patients with Chronic Kidney Disease. Am J Nephrol. 2019;50(4):229-39.

19. Mayfield JJ, Otto CM. Stroke and Noninfective Native Valvular Disease. Curr Cardiol Rep. 2023;25(5):333-48.

20. Bouzas-Mosquera A, Broullón FJ, Álvarez-García N*, et al.* Left atrial size and risk for all-cause mortality and ischemic stroke. CMAJ. 2011;183(10):E657-E64.

21. Weng C-J, Li C-H, Liao Y-C*, et al.* Rhythm control better prevents stroke and mortality than rate control strategies in patients with atrial fibrillation - A nationwide cohort study. International Journal of Cardiology. 2018;270:154-9.

22. Choi KM, Han K, Park S*, et al.* Implication of liver enzymes on incident cardiovascular diseases and mortality: A nationwide population-based cohort study. Scientific Reports. 2018;8(1):3764.

23. Green DM, Ropper AH, Kronmal RA, Psaty BM, Burke GL. Serum potassium level and dietary potassium intake as risk factors for stroke. Neurology. 2002;59(3):314-20.

Table 1. Characteristics between stroke and non-stroke patients in the NVAF cohort.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Characteristic | All | Non-stroke | Stroke | *P*-value |
| N | 2101 | 2018 | 83 |  |
| Age, years | 68.0 (60.0, 76.0) | 68.0 (60.0, 76.0) | 75.0 (65.0, 80.0) | < 0.001 |
| Female, n (%) | 976 (46.5%) | 925 (45.8%) | 51 (61.4%) | 0.005 |
| BMI, kg/m2 | 24.5 (22.0, 26.8) | 24.5 (22.0, 26.8) | 24.4 (22.0, 26.9) | 0.931 |
| Heart Rate, beats/min | 90.0 (72.0, 115.0) | 90.0 (72.0, 115.0) | 100.0 (74.0, 120.0) | 0.278 |
| Systolic Blood Pressure, mmHg | 130.0 (120.0, 150.0) | 130.0 (120.0, 150.0) | 140.0 (120.0, 160.0) | 0.010 |
| Diastolic Blood Pressure, mmHg | 80.0 (70.0, 90.0) | 80.0 (70.0, 90.0) | 80.0 (70.0, 90.0) | 0.019 |
| CHA2DS2-VASc Score, n (%) |  |  |  | < 0.001 |
| 0 | 98 (4.7%) | 96 (4.6%) | 2 (2.4%) |  |
| 1 | 253 (12.0%) | 249 (12.3%) | 4 (4.8%) |  |
| 2 | 441 (21.0%) | 430 (21.3%) | 11 (13.3%) |  |
| 3 | 470 (22.4%) | 459 (22.7%) | 11 (13.3%) |  |
| 4 | 400 (19.0%) | 380 (18.8%) | 20 (24.1%) |  |
| 5 | 288 (13.7%) | 270 (13.4%) | 18 (21.7%) |  |
| 6 | 122 (5.8%) | 110 (5.5%) | 12 (14.5%) |  |
| 7 | 26 (1.2%) | 21 (1.0%) | 5 (6.0%) |  |
| 8 | 3 (0.1%) | 3 (0.1%) | 0 (0%) |  |
| Lifestyle, n (%) | | | | |
| Smoking Status |  |  |  | 0.050 |
| Never | 1570 (74.7%) | 1504 (74.5%) | 66 (79.5%) |  |
| Past | 461 (21.9%) | 444.0 (22.0%) | 17 (20.5%) |  |
| Current | 70 (3.3%) | 70 (3.5%) | 0 (0%) |  |
| Alcohol Consumption |  |  |  | 0.030 |
| Never | 1646 (78.3%) | 1572 (77.9%) | 73 (89.0%) |  |
| Past | 120 (5.7%) | 118 (5.8%) | 2 (2.4%) |  |
| Current | 335 (15.9%) | 328 (16.3%) | 7 (8.4%) |  |
| Disease History, n (%) | | | | |
| History of Rheumatic Fever | 136 (6.5%) | 132 (6.5%) | 4 (4.8%) | 0.532 |
| History of Valvular Disease | 345 (16.4%) | 326 (16.2%) | 19 (22.9%) | 0.104 |
| History of Congenital Heart Disease |  |  |  | 0.829 |
| None | 2054 (97.8%) | 1973 (97.8%) | 81 (97.6%) |  |
| Acyanotic | 43 (2.0%) | 41 (2.0%) | 2 (2.4%) |  |
| Cyanotic | 4 (0.2%) | 4 (0.2%) | 0 (0%) |  |
| Prior CVA/TIA/SE Event | 300 (14.3%) | 282 (14.0%) | 18 (21.7%) | 0.049 |
| Characteristic | All | Non-stroke | Stroke | *P*-value |
| Prior Bleeding Event | 131 (6.2%) | 125 (6.2%) | 6 (7.2%) | 0.702 |
| Coronary Disease | 214 (10.2%) | 205 (10.2%) | 9 (10.8%) | 0.840 |
| Aortic Involvement | 51 (2.4%) | 49 (2.4%) | 2 (2.4%) | 1.000 |
| Mitral Involvement | 212 (10.1%) | 198 (9.8%) | 14 (16.9%) | 0.036 |
| Comobidities, n (%) | | | | |
| Hypertension | 1287 (61.3%) | 1221 (60.5%) | 66 (79.5%) | < 0.001 |
| Diabetes | 782 (37.2%) | 745 (36.9%) | 37 (44.6%) | 0.157 |
| Dyslipidaemia | 983 (46.8%) | 943 (46.7%) | 40 (48.2%) | 0.793 |
| Thyroid Disease | 239 (11.4%) | 227 (11.2%) | 12 (14.5%) | 0.367 |
| Chronic Kidney Disease | 1060 (50.5%) | 1002 (49.7%) | 58 (69.9%) | < 0.001 |
| Respiratory Disease | 450 (21.4%) | 431 (21.4%) | 19 (22.9%) | 0.739 |
| Chronic Liver Disease | 40 (1.9%) | 37 (1.8%) | 3 (3.6%) | 0.208 |
| HF |  |  |  | 0.241 |
| None | 1560 (74.3%) | 1504 (74.5%) | 56 (67.5%) |  |
| HFrEF (< 50%) | 339 (16.1%) | 324 (16.1%) | 15 (18.1%) |  |
| HFpEF (≥ 50%) | 202 (9.6%) | 190 (9.4%) | 12 (14.5%) |  |
| NYHA Class |  |  |  | 0.423 |
| Class I & II | 1739 (82.8%) | 1673 (82.9%) | 66 (79.5%) |  |
| Class III & IV | 362 (17.2%) | 345 (17.1%) | 17 (20.5%) |  |
| Cardiomyopathy | 240 (11.4%) | 228 (11.3%) | 12 (14.5%) | 0.375 |
| History of therapeutic operation, n (%) | | | | |
| CABG | 108 (5.1%) | 107 (5.3%) | 1 (1.2%) | 0.125 |
| Valve Replacement |  |  |  | 0.564 |
| None | 2028 (96.5%) | 1946 (96.4%) | 82 (98.8%) |  |
| Aortic | 4 (0.2%) | 4 (0.2%) | 0 (0%) |  |
| Mitral | 63 (3.0%) | 62 (3.1%) | 1 (1.2%) |  |
| Combined | 6 (0.3%) | 6 (0.3%) | 0 (0%) |  |
| CHD Repair | 7 (0.3%) | 6 (0.3%) | 1 (1.2%) | 0.246 |
| Catheter Ablation | 10 (0.5%) | 10 (0.5%) | 0 (0%) | 1.000 |
| Pacemaker Implant | 120 (5.7%) | 115 (5.7%) | 5 (6.0%) | 0.810 |
| Surgery for AF | 3 (0.1%) | 3 (0.1%) | 0 (0%) | 1.000 |
| ICD Implant | 10 (0.5%) | 10 (0.5%) | 0 (0%) | 1.000 |
| LAAO | 6 (0.3%) | 5 (0.2%) | 1 (1.2%) | 0.215 |
| Bridged UFH | 257 (12.2%) | 250 (12.4%) | 7 (8.4%) | 0.281 |
| Bridged LMWH | 242 (11.5%) | 229 (11.3%) | 13 (15.7%) | 0.228 |
| Bridged Fondaparinux | 29 (1.4%) | 28 (1.4%) | 1 (1.2%) | 1.000 |
| Medicine, n (%) | | | | |
| Characteristic | All | Non-stroke | Stroke | *P*-value |
| ACEI Medicine | 194 (9.2%) | 189 (9.4%) | 5 (6.0%) | 0.303 |
| ARB Medicine | 352 (16.8%) | 339 (16.8%) | 13 (15.7%) | 0.786 |
| DHP CCB Medicine | 270 (12.9%) | 256 (12.7%) | 14 (16.9%) | 0.265 |
| Diuretics Medicine | 840 (40.0%) | 797 (39.5%) | 43 (51.8%) | 0.025 |
| Statin Medicine | 1189 (56.6%) | 1141 (56.5%) | 48 (57.8%) | 0.816 |
| Class I AAD | 36 (1.7%) | 33 (1.6%) | 3 (3.6%) | 0.168 |
| Class III AAD | 452 (21.5%) | 432 (21.4%) | 20 (24.1%) | 0.559 |
| VKA Medicine | 1228 (58.4%) | 1181 (58.5%) | 47 (56.6%) | 0.731 |
| NOAC Medicine | 174 (8.3%) | 168 (8.3%) | 6 (7.2%) | 0.723 |
| Non Anticoagulant Medicine | 703 (33.5%) | 673 (33.3%) | 30 (36.1%) | 0.597 |
| Antiplatelet Medicine | 1009 (48.0%) | 968 (48.0%) | 41 (49.4%) | 0.798 |
| AF-related variables, n (%) | | | | |
| AF Treatment Strategy |  |  |  | 0.080 |
| Rhythm | 350 (16.7%) | 342 (16.9%) | 8 (9.6%) |  |
| Rate | 1751 (83.3%) | 1676 (83.1%) | 75 (90.4%) |  |
| Persistent AF | 379 (18%) | 356 (17.6%) | 23 (27.7%) | 0.019 |
| AF Symptom |  |  |  |  |
| Palpitations | 990 (47.1%) | 953 (47.2%) | 37 (44.6%) | 0.636 |
| Breathlessness | 980 (46.6%) | 935 (46.3%) | 45 (54.2%) | 0.158 |
| Chest Pain | 453 (21.6%) | 431 (21.4%) | 22 (26.5%) | 0.264 |
| Syncope Presyncope | 178 (8.5%) | 171 (8.5%) | 7 (8.4%) | 0.990 |
| Fatigue | 331 (15.8%) | 313 (15.5%) | 18 (21.7%) | 0.130 |
| Imaging Feature, n (%) | | | | |
| ECG |  |  |  |  |
| Rhythm Enrollment |  |  |  | 0.748 |
| Normal Sinus Rhythm | 304 (14.5%) | 294 (14.6%) | 10 (12.0%) |  |
| AF | 1738 (82.7%) | 1668 (82.7%) | 70 (84.3%) |  |
| Paced Rhythm | 59 (2.8%) | 56 (2.8%) | 3 (3.6%) |  |
| Ischaemic Change | 290 (13.8%) | 275 (13.6%) | 15 (18.1%) | 0.250 |
| LBBB | 116 (5.5%) | 113 (5.6%) | 3 (3.6%) | 0.623 |
| RBBB | 101 (4.8%) | 100 (5.0%) | 1 (1.2%) | 0.183 |
| LVH | 352 (16.8%) | 332 (16.5%) | 20 (24.1%) | 0.068 |
| ST Change | 574 (27.3%) | 547 (27.1%) | 27 (32.5%) | 0.277 |
| Echocardiography |  |  |  |  |
| LVEF, % | 58.0 (50.0, 64.0) | 58.0 (50.0, 64.0) | 59.0 (50.0, 63.0) | 0.610 |
| LA Size, mm | 40.0 (36.0, 45.0) | 40.0 (36.0, 45.0) | 42.0 (38.0, 46.0) | 0.016 |
| LVH | 608 (28.9%) | 584 (28.9%) | 24 (28.9%) | 0.996 |
| Characteristic | All | Non-stroke | Stroke | *P*-value |
| RWMA | 374 (17.8%) | 362 (17.9%) | 12 (14.5%) | 0.417 |
| MS | 204 (9.7%) | 194 (9.6%) | 10 (12.0%) | 0.463 |
| AS | 97 (4.6%) | 91 (4.5%) | 6 (7.2%) | 0.276 |
| MR (≥ Moderate) | 632 (30.1%) | 606 (30.0%) | 26 (31.3%) | 0.801 |
| AR (≥ Moderate) | 171 (8.1%) | 163 (8.1%) | 8 (9.6%) | 0.610 |
| PAH (≥ Moderate) | 471 (22.4%) | 453 (22.4%) | 18 (21.7%) | 0.871 |
| Rheumatic Involvement | 225 (10.7%) | 215 (10.7%) | 10 (12.0%) | 0.687 |
| Labotary | | | | |
| Hemoglobin, g/dL | 12.5 (11.2, 13.6) | 12.5 (11.2, 13.6) | 12.3 (10.8, 13.1) | 0.098 |
| Total Cholesterol, mg/dL | 163.0 (134.0, 193.0) | 163.0 (134.0, 193.3) | 164.0 (136.0, 192.0) | 0.976 |
| LDL, mg/dL | 97.0 (72.0, 124.0) | 97.0 (71.0, 124.0) | 105.0 (75.0, 124.0) | 0.405 |
| HDL, mg/dL | 43.0 (37.0, 51.5) | 43.0 (37.0, 51.0) | 45.0 (36.0, 53.0) | 0.813 |
| AST, U/L | 31.0 (24.0, 42.0) | 31.0 (24.0, 42.0) | 32.0 (24.0, 61.0) | 0.107 |
| ALT, U/L | 121.0 (68.0, 121.0) | 121.0 (68.0, 121.0) | 121.0 (59.0, 121.0) | 0.837 |
| FBS, mg/dL | 110.0 (96.0, 138.0) | 110.0 (96.0, 136.0) | 113.0 (99.0, 165.0) | 0.108 |
| INR, sec | 1.4 (1.1, 2.0) | 1.4 (1.1, 2.0) | 1.4 (1.1, 2.0) | 0.793 |
| Serum Creatinie, mg/dL | 1.0 (0.9, 1.3) | 1.0 (0.9, 1.3) | 1.1 (0.9, 1.3) | 0.255 |
| Total Bilirubin, mg/dL | 0.9 (0.7, 1.2) | 0.9 (0.7, 1.2) | 0.9 (0.7, 1.0) | 0.700 |
| GFR, ml/min | 56.8 (41.7, 73.4) | 57.3 (41.9, 73.8) | 48.3 (34.3, 59.3) | < 0.001 |

Legend: AAD, antiarrhythmic drug; ACEI, angiotensin-converting enzyme inhibitor; AF, atrial fibrillation; ALT, alanine transaminase; AR, aortic regurgitation; ARB, angiotensin receptor blocker; AS, aortic stenosis; AST, aspartate Transaminase; BMI, body mass index; CABG, coronary artery bypass graft; CHD, congenital heart disease; CVA, cerebrovascular accident; DHP CCB, dihydropyridine calcium channel blocker; ECG, electrocardiography; ECHO, echocardiogram; FBS, fasting blood sugar; GFR, glomerular filtration rate; HDL, high density lipoprotein; HF, heart failure; HFrEF, heart failure with reduced ejection fraction; HFpEF, heart failure with preserved ejection fraction; ICD, implantable cardioverter defibrillator; INR, international normalized ratio; LA, left atrium; LAAO, left atrial appendage occlusion; LBBB, left bundle branch block; LDL, low density lipoprotein; LMWH, low molecular weight heparin; LVEF, left ventricular ejection fraction; LVH, left ventricular hypertrophy; MR, mitral regurgitation; MS, mitral stenosis; NOAC, non-vitamin K antagonist oral anticoagulant; NVAF, non-valvular atrial fibrillation; NYHA, New York Heart Association; PAH, pulmonary arterial hypertension; RBBB, right bundle branch block; RWMA, regional wall motion abnormality; UFH, unfractionated heparin; SE, systemic embolism; TIA, transient ischaemic attack; VKA, vitamin K antagonist.

Table 2. The Performance of classifiers and CHA2D2-VASc in the internal and external validation cohort.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Classifiers | AUC (95% CI) | Accuracy | Specificity | Sensitivity | Precision | Recall | F1-score | G-mean |
| Internal Validation Cohort | | | | | | | | |
| LightGBM | 0.815 (0.811, 0.818) | 0.824 | 0.824 | 0.833 | 0.118 | 0.708 | 0.202 | 0.288 |
| Random Forest | 0.821 (0.816, 0.825) | 0.945 | 0.950 | 0.824 | 0.340 | 0.560 | 0.417 | 0.427 |
| Logistic Regression | 0.723 (0.720, 0.727) | 0.741 | 0.744 | 0.684 | 0.077 | 0.682 | 0.137 | 0.219 |
| Support Vector Machine | 0.835 (0.831, 0.839) | 0.912 | 0.919 | 0.692 | 0.196 | 0.672 | 0.299 | 0.359 |
| Multilayer Perceptron | 0.822 (0.817, 0.826) | 0.943 | 0.992 | 0.276 | 0.494 | 0.370 | 0.410 | 0.421 |
| CHA2DS2-VASc | 0.665 (0.663, 0.667) | 0.957 | 1.000 | 0.000 | 0.008 | 0.048 | 0.007 | 0.219 |
| External Validation Cohort | | | | | | | | |
| LightGBM | 0.670 (0.665, 0.674) | 0.812 | 0.816 | 0.417 | 0.021 | 0.339 | 0.039 | 0.083 |
| Random Forest | 0.582 (0.577, 0.587) | 0.974 | 0.987 | 0.000 | 0.020 | 0.024 | 0.021 | 0.021 |
| Logistic Regression | 0.639 (0.635, 0.643) | 0.679 | 0.681 | 0.529 | 0.019 | 0.381 | 0.036 | 0.083 |
| Support Vector Machine | 0.590 (0.585, 0.595) | 0.918 | 0.928 | 0.190 | 0.014 | 0.091 | 0.024 | 0.035 |
| Multilayer Perceptron | 0.554 (0.549, 0.560) | 0.952 | 0.964 | 0.048 | 0.012 | 0.041 | 0.018 | 0.029 |
| CHA2DS2-VASc | 0.615 (0.611, 0.619) | 0.986 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |

Legend: AUC, area under curve; CI, confidence interval; LightGBM, light gradient boosting machine.