Intelligent Prediction of Cryptogenic Stroke Using Patent Foramen Ovale from TEE Imaging Data and Machine Learning Methods

Jiao Bai¹ · Jia Yang¹ · Wanwan Song¹ · Yumin Liu² · Haibo Xu³ · Yang Liu^{4,5}

Received: 26 December 2021 / Accepted: 28 January 2022
© The Author(s) 2022

Abstract

In spite of the popularity of random forests (RF) as an efficient machine learning algorithm, methods for constructing the potential association for between patent foramen ovale (PFO) and cryptogenic stroke (CS) using this technique are still barely. For the vital regional study areas (atrial septum), RF was used to predict CS in patients with PFO using partial clinical data of patients and remotely sensed imaging examination data obtained from Tee imaging. We validated our method on a dataset of 151 consecutive patients with detected PFO at a large grade A hospital in China from November 2018 to December 2020, we obtained an area under the relative operating characteristic curve of 0.816, with 65% specificity at 73% sensitivity. The RF models accurately represented the relationship between the CS and remotely sensed predictor variables. Therein, maximum mobility, large right-to-left shunt during Valsalva maneuver, size of PFO in diastole and systole, and diastolic length of the tunnel present higher predictive value in CS. Our findings suggest that multi-Doppler sensor data by transesophageal echocardiography (TEE)-detected morphologic and functional characteristics of PFO may play important roles in the occurrence of CS. These results indicate that the established random forest model has the potential to predict CS in patients with PFO and great promise for application to clinical practice.

Keywords Multi-Doppler sensor · Random forest · Patent foramen ovale · Cryptogenic stroke · Prediction diagnosis

Abbreviations

RF	Random forests
PFO	Patent foramen ovale
CS	Cryptogenic stroke
TEE	Transesophageal echocardiography
ASA	Atrial septal aneurysm
RL	Right-to-left
SVM	Support vector machine

ANN	Artificial neural network
c-TEE	Contrast transesophageal echocardiography
OBB	Out-of-bag data

1 Introduction

Stroke is the second leading cause of death worldwide and first leading cause of death in China. The diverse variety of ischemic stroke due to its etiology is an important research

Jiao Bai and Jia Yang contributed equally to the work.

✉ Haibo Xu
xuhaibo1120@hotmail.com

✉ Yang Liu
dabailiu@whu.edu.cn

Jiao Bai
baijiao20211001@163.com

Jia Yang
jessieyang1992@163.com

Wanwan Song
675887254@qq.com

Yumin Liu
wb001792@whu.edu.cn

¹ Department of Ultrasonic Medicine, Zhongnan Hospital of Wuhan University, Wuhan 430071, People's Republic of China

² Department of Neurology, Zhongnan Hospital of Wuhan University, Wuhan 430071, People's Republic of China

³ Department of Medical Imaging, Zhongnan Hospital of Wuhan University, Wuhan 430071, People's Republic of China

⁴ Wuhan University, Zhongnan Hospital of Wuhan University, Wuhan 430071, People's Republic of China

⁵ School of Economics and Management, Wuhan University, Wuhan 430072, People's Republic of China

part of stroke, with about 30% of ischemic stroke being cryptic stroke [1]. Patent foramen ovale (PFO) is present in approximately 25% of the general population, and approximately 40–50% of cryptogenic stroke (CS) presents with PFO [2]. However, the potential association between PFO and CS has been a controversial issue for several decades. TEE is the gold standard for the diagnosis of PFO, and it is also the first choice for diagnosis. The front end of the TEE is the Doppler sensor, where the end is a multiplanar probe with the most widely used rotating sensor, and a series of horizontal, vertical sectional images can arbitrarily “cut” the heart from an angle of 0–180°, with less bending stretch, you can get easier and more detailed heart imaging data [3]. Therefore, remote sensing technologies provide valuable assisted information that can be used to enhance the precision and timeliness of PFO parameter acquisition. Machine Learning algorithms are crucial for prediction the association between PFO and CS [4]. In this article, we propose a random forest algorithm based on TEE imaging data, and compare it with other machine learning algorithms.

2 Literature Review

Previous studies have stated that large-size PFO, long-tunnel PFO, the presence of atrial septal aneurysm (ASA), and severe right-to-left (RL) shunt are high-risk factors related to CS [5]; conversely, other studies have shown that there are no morphological differences in PFO between those with and without CS, and that only interatrial shunt is significantly related to possible embolism [6]. However, both sets of studies have ignored the fact that PFO shape, interatrial hemodynamics, and atrial pressures change with the cardiac cycle. Regardless of physiological condition, each normal cardiac cycle during right ventricular early diastole and isovolumetric contraction presents a transient spontaneous reversal of left and right atrial pressures [7], and this reversal gradient may increase substantially with certain physiological conditions, such as a cough, inspiration, and Valsalva maneuver. Transcatheter PFO closure has been suggested as one therapeutic option of CS; while no strong data from randomized clinical trials have been provided to support primary preventive PFO closure, many results have pointed to the benefits of preventative PFO closure in certain high-risk patients [8]. Moreover, evidence from recent trials has demonstrated that PFO device closure is more effective at preventing the recurrence of stroke compared to medical therapy, especially in patients with high-risk PFO [9, 10].

Other studies have found that logistic analysis to propose a predictive model of high-risk PFO related to CS [11]. Machine learning, as a branch of artificial intelligence, has been widely used in the medical and biological sectors. However, due to the complex distribution of PFO-related CS

data collected from ultrasound images with many features, high noise, and many discrete attribute variables. Although these machine learning algorithms can achieve better prediction results than empirical statistical models. However, these traditional intelligent prediction models have inherent defects such as being highly dependent on the accuracy of the database, being inefficient, time-consuming, and easily falling into local optimum, including support vector machine (SVM) and artificial neural network (ANN) approaches [12]. The random forest approach, on the other hand, has good generalization ability and is insensitive to noise, so it is suitable to establish a prediction model of high-risk features of PFO. Random forest is a supervised integrated learning classification technology. As an emerging machine learning ensemble algorithm, random forest can solve complex problems of nonparametric and nonlinear classification [13]. Moreover, it can reduce the computing complexity under the premise of improving the accuracy, with the advantages of few parameters, strong generalization ability and strong resistance to overfitting [14]. For high-dimensional data, the comprehensive performance index of the random forest approach, such as classification accuracy and algorithm efficiency, are clearly superior to other single classifiers and integrated classifiers.

Therefore, random forest technology has been applied widely in the fields of biological information, text mining, and image classification in recent years, and has become a frequent research topic in the fields of data mining, machine learning, and pattern recognition [15]. Based on aforementioned discussion, our study investigates whether PFO morphological features in different phases of the cardiac cycle can predict CS, and evaluates which of these features show the highest predictive value of high-risk PFO related to CS by establishing a random forest model. All data are taken from Zhongnan Hospital of Wuhan University, and the results are realized using MATLAB.

3 Materials and Methods

3.1 Study Population

From November 2018 to December 2020, we retrospectively enrolled 151 consecutive patients with detected PFO at our institution. All patients were subject to brain and carotid imaging, 12-lead electrocardiography, echocardiography, and a hypercoagulability panel. The TEE and contrast transesophageal echocardiography (c-TEE) findings were reviewed in patients with and without CS. CS was defined to present with a transient or permanent neurological deficit on magnetic resonance imaging after excluding all other identifiable causes of ischemic lesion. Patients without CS presented with migraine and dizziness, without cerebral

infarction lesion (confirmed by magnetic resonance imaging). CS was evaluated by an experienced neurologist. Patients with poor visualization of PFO, an atrial septal defect, atrial fibrillation, valvular heart disease, congestive heart failure with ejection fraction < 50%, or an inability to perform the Valsalva maneuver because of impaired cognition or coordination, and having other causes of CS were excluded. Patients with observed PFO on TEE and underwent a saline contrast study were ultimately included. This study was approved by the ethics committee of Zhongnan Hospital of Wuhan University (No. 2020060 K), and all enrolled patients provided their written informed consent.

3.2 Assessment of Characteristics of PFO by TEE and c-TEE

Echocardiographic studies were performed using the GE Vivid E95 cardiovascular ultrasound system equipped with an M5S probe and 6VT-D probe (GE Vingmed Ultrasound AS, Horten, Norway). The principle of the Doppler sensor is transesophageal ultrasound imaging, including 2-D, M-mode color ultrasound and spectral Doppler, which can be transesophageal 20–25 cm, middle 35–40 cm, stomach 40–45 cm, multi-angle 0–180°, long- and short-axis, multi-section to observe the dynamic structure and function of the heart and blood vessel, using the caliper and tracker to calculate the area, volume, diameter, length of each anatomical part of the heart and blood vessel. Understand myocardial systolic and diastolic functions, detect the degree of valve disease, and the direction of blood vessels. Understand the function of the heart, regulate the circulation in real time, and maintain the stability of the circulation. A saline-contrast study was administered both during normal respiration

and during the Valsalva maneuver. The presence of PFO was confirmed based on either (1) direct visualization of microbubbles passing through the atrial septum to the left atrium within three consecutive cardiac cycles after entire right atrium opacification, or (2) visualization of color Doppler flow through the atrial septum (Fig. 1) [16].

The following parameters on the anatomical and functional characteristics of PFO were studied: PFO size in ventricular systole and diastole, length of PFO tunnel in ventricular systole and diastole, presence of ASA, maximum mobility, presence of prominent Eustachian valve or Chiari's network, and degree of RL shunt at rest and during Valsalva maneuver. PFO size was defined as the maximum separation between the septum primum and septum secundum at the point of entry to the left atrium (Fig. 2), and a size of ≥ 2 mm was defined as large-size PFO [17]. The length of the PFO tunnel was measured according to the maximum overlap between the septum primum and septum secundum (Fig. 2), and a length of ≥ 10 mm was defined as long-tunnel PFO [18]. Prominent Eustachian valve was defined as a ≥ 10 mm protrusion within the right atrium. Chiari's network was defined as a network of threads in the right atrium with attachments to the upper wall of the right atrium or the interatrial septum (Fig. 3A) [19]. ASA was defined as ≥ 10 mm of septal excursion from the midline into the right or left atrium, or ≥ 15 mm of the total excursion between the right and left atrium (Fig. 3B) [20]. Maximum mobility was equal to the sum of the excursions (the greatest leftward and rightward deflections of the septum with respect to a line perpendicular to the fossa ovalis plane; Fig. 3C). The degree of RL shunt was assessed either at rest or during the Valsalva maneuver using agitated saline contrast. According to the amount of microbubbles appearing

Fig. 1 PFO diagnosis. **A** Visualization of color Doppler flow through the atrial septum. **B** Microbubbles passing through the atrial septum; the number of microbubbles in the left atrium was counted

Fig. 2 PFO size and tunnel length. The PFO size (white arrow) was measured via the maximum separation in the ventricular systole and diastole, the length of the PFO tunnel (yellow arrow) was measured via the maximum overlap in the ventricular systole and diastole

Fig. 3 PFO characteristics. **A** A network of threads in the right atrium with attachments to the upper wall of the right atrium were defined as Chiari's network (arrow). **B** ASA was defined as ≥ 10 mm of septal

excursion from the midline into the right or left atrium. **C** A moving and floppy septum defined the maximum mobility of the interatrial septum

in the left atrium, the degree of shunt was defined as mild (3–9), moderate (10–30), or severe (> 30). Each transthecal echocardiographic study was reviewed and analyzed on the EchoPAC system (GE Vingmed Ultrasound AS, Horten, Norway) by two independent cardiologists who were blinded to the CS status of the patients.

3.3 Random Forest

Random forest is a statistical learning theory based on a classification tree. Random forests utilize the bootstrap resampling method to draw multiple sample sets back and forth from the original sample set, and model each sample set separately for decision trees. Each decision tree

randomly selects features during modeling to split attributes on the internal nodes, and constitutes part of the random forest; the final prediction result is synthesized from the resulting vote of each decision tree [21] (Fig. 4).

3.3.1 Generalization Error of Random Forest

When taking a sample set, about 36.8% of the samples in each original sample set are not selected. These samples are called out-of-bag data (OOB), and can be used to calculate the generalization error of the model [22]. The generalization error of a random forest can be expressed as follows:

Fig. 4 Flow diagram of the random forest regression

$$E^* = P_{X,Y}(M(X, Y) < 0), \quad (1)$$

where subscripts X and Y indicate that probability P covers X and Y spaces. In a random forest, when the number of decision trees is large enough, E^* converges to

$$P_{X,Y}(P_{\theta}(h(X, \theta) = Y) - \max_{j \neq Y} P_{\theta}(h(X, \theta) = j) < 0). \quad (2)$$

This shows that the generalization error does not cause overfitting as the number of decision trees increases, but approaches a finite upper bound.

3.3.2 Evaluation of the Significance of Features

There two primary methods by which to judge the importance of random forest features: one is to rank each feature according to the impurity of the Gini coefficient, and the other is to calculate the influence of each feature on the accuracy of the model. This paper chooses the latter method to evaluate the importance of feature variables.

When determining the importance of the model, the OOB error R1 was calculated using the corresponding OOB data, and then the order of a feature in the OOB data was randomly transformed to calculate the OOB error R2 again. Assuming that there are N decision trees in a random forest, the importance I of a certain feature is

$$I = \frac{\sum_{i=1}^N (r_1 - r_2)}{N}. \quad (3)$$

After obtaining the importance degree of each feature, recursive feature elimination was used to sequentially reject the features with the least importance until the optimal number of features was reached, thus enabling feature selection.

3.4 Risk Assessment of CS Risk Based on the Random Forest Approach

A technical flowchart of the PFO-related CS risk prediction model based on the random forest approach is shown in

Fig. 5. First, the CS risk-assessment system was constructed, and the training set was established by random sampling. Next, the parameters were optimized and a random forest training model was established. The importance of the evaluation indicators was then determined, the test set data were input into the training model, and each regression tree in the model obtained a set of predicted values based on the test set data. The mean value was the final prediction result, and error analysis and variable sensitivity analysis were performed thereon.

3.5 Construction of PFO-Related CS Risk Assessment System and Establishment of Training Set

Establish a risk-assessment system. The mechanism of CS induced by PFO was analyzed, and relevant influencing factors were obtained. According to a large amount of practical experience and relevant references, the PFO-related CS risk evaluation index system was constructed and the risk grade was determined.

Set up the original training set sample. Each index of the index system was taken as the random forest variable, and the index-related data were taken as the original training set, of which 80% of the original data were trained.

Generate random self-help sample set. The original training set was recorded as $T = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$. The bootstrap sampling method was used to extract k times from sample T with n sample size, so as to form a k mutually independent training set $\{T_i, i = 1, 2, \dots, k\}$.

3.6 Determination of Optimal Parameters and Establishment of the Training Model

Choose the best branch via k -fold cross-validation. The k -fold cross-validation method was adopted to divide the initial sample into k subsamples, and a single subsample was retained as the data of the validation mode. The other $k-1$ samples were used for training. Finally, the average prediction accuracy of the k models was used as the final estimate of the prediction accuracy of the model, and the

Fig. 5 Flow diagram of the random forest regression model on cryptogenic stroke risk

splitting mode with the highest prediction accuracy was selected as the optimal branch. Tenfold cross-validation was adopted in this study [23].

Select the optimal parameters. In the process of tree generation, for each node, M features were randomly selected from all feature sets, and then an optimal eigenvalue $mtry$ was selected as the split variable value according to the criterion that the information gain ratio reaches the maximum. A random forest model was established, the trend of $ntree$ and mean square error was observed, and the decision tree corresponding to the minimum mean square error was selected as the best $ntree$ value—that is, the number of regression trees.

Establish a training model. Taking the optimal branch as the random forest input, the node was divided into two branches according to the characteristics, and then the best features were determined from the remaining features. In this way, the branches of the classification tree were constructed recursively to maximize the growth of the regression tree without any clipping, and a decision tree was generated. The process was repeated to establish the random forest training model.

3.7 Evaluation of Variable Importance and Model Fit Prediction

Evaluate variable importance. For each tree in the random forest, the corresponding OOB was used to calculate its OOB data error, which was recorded as err_{OOB1} . Noise interference was randomly added to the characteristic X of all samples of OOB data, and its OOB data error was calculated again and recorded as err_{OOB2} . Thus, the importance of feature X is

$$\text{Importance} = \sum (\text{err}_{OOB2} - \text{err}_{OOB1}) / N_{tree}. \quad (4)$$

Evaluate model fit prediction. The remaining 20% of the data test set was input into the training model. The random forest model was used to predict the test set data and establish the prediction model. The average of the output values of all decision trees were taken as the prediction value of the random forest; the random forest training model fit by the training set and the random forest prediction model predicted by the test set were drawn visually, and the model fit diagram and prediction diagram were

thereby obtained. The prediction result of the random forest regression model is

$$f_r(x) = \frac{1}{k} \sum_{i=1}^k h_i(x). \quad (5)$$

In Eq. (5), $f_r(x)$ represents the predicted value of the random forest regression model, and $h_i(x)$ represents the predicted value of a single regression tree model.

3.8 Evaluation of Model

Once the model had been established, it had to be evaluated to determine whether it is suitable for disease prediction. In this study, the true-positive rate and false-positive rate of the test set were calculated using R language, and the relative operating characteristic curve and the area under the curve were drawn using the pROC package within R to evaluate the random forest model.

3.9 Statistical Analysis

All statistical analyses were performed using SPSS19.0 (IBM Corporation, Armonk, NY, USA). Data were described as mean $\pm$ SD for continuous variables and as

number (percentage) for categorical variables. Student's t test, Mann–Whitney U test, χ^2 test and Fisher's exact test were used to compare baseline characteristics and echocardiographic PFO characteristics between patients with and without CS where appropriate. The interobserver agreements were analyzed using the intraclass correlation coefficient or Cohen's κ statistics based on data from 15 randomly selected patients recorded by observers A and B. P values of < 0.05 were considered significant.

4 Results

4.1 Patient Characteristics

Of the 151 patients recruited, the mean age was 50.9 ± 13.9 years and 80 (53%) were male. CS was found in 66 (43.7%) patients, hypertension in 90 (59.6%) patients, diabetes in 45 (29.8%) patients, and hyperlipidemia in 58 (38.4%). In addition, 46 (30.4%) patients reported being a current or prior smoker and 47 (31.1%) had a body mass index ≥ 25 . The baseline characteristics did not differ significantly between patients with and without CS, as shown in Table 1.

Table 1 Comparison of basic information between CS and non-CS groups

Indicators	Total patients ($n = 151$)	Patients with CS ($n = 66$)	Patients without CS ($n = 85$)	p value
Age, years	50.9 ± 13.9	50.2 ± 14.3	51.4 ± 13.6	0.622
Gender, male	80 (53.0%)	38 (57.6%)	42 (49.4%)	0.319
Hypertension	90(59.6%)	42 (63.6%)	48(56.4%)	0.373
Hyperlipidemia	58(38.4%)	27(40.9%)	31(36.5%)	0.578
Diabetes	45(29.8%)	21(31.8%)	24(28.2%)	0.633
BMI (≥ 25)	47(31.1%)	24(36.4%)	23(27.1%)	0.221
Current or prior smoker	46(30.4%)	24(36.4%)	22(25.9%)	0.165
TEE findings of PFO				
Blood flow via atrial septum	136(90.1%)	64(97.0%)	72(84.7%)	0.012
Size of PFO in systole [M (P25, P75), mm]	1.8(1.3, 2.3)	2.0(1.5, 2.9)	1.6(1.1, 2.0)	< 0.001
Large PFO in systole (Size ≥ 2 mm)	60(39.7%)	34(51.5%)	26(30.6%)	0.009
Size of PFO in diastole [M (P25, P75), mm]	1.5(1.2, 2.0)	1.7(1.4, 2.2)	1.3(1.1, 1.8)	< 0.001
Large PFO in diastole (Size ≥ 2 mm)	39(25.8%)	23(34.8%)	16(18.8%)	0.026
Systolic length of the tunnel [M (P25, P75), mm]	7.1(5.1, 11.2)	7.8(5.3, 11.4)	7.1(5.0, 11.0)	0.363
Long tunnel in systole (≥ 10 mm)	47(31.1%)	22(33.3%)	25(29.4%)	0.606
Diastolic length of the tunnel [M (P25, P75), mm]	6.9(5.0, 11.0)	8.0 (5.0, 11.6)	6.7(5.1,10.6)	0.319
Long tunnel in diastole (≥ 10 mm)	47(31.1%)	22(33.3%)	25(29.4%)	0.606
Atrial septal aneurysm	19(12.6%)	18(27.3%)	1(1.2%)	< 0.001
Maximum mobility (mm) [M (P25, P75), mm]	3.9(2.7, 5.9)	5.9(3.3, 7.6)	3.2(2.3, 4.6)	< 0.001
Prominent Eustachian valve or Chiari's network	21(13.9%)	11(16.7%)	10(11.8%)	0.388
Large RL shunt during Valsalva maneuver	59(39.1%)	42(63.6%)	17(20.0%)	< 0.001
Large RL shunt at rest	18(11.9%)	18(27.3%)	0	< 0.001

4.2 Echocardiographic Characteristics of PFOs

As shown in Table 1, the size of the PFOs was significantly greater in patients with CS compared to those without CS in systole and diastole [systole 2.0 (1.5, 2.9) mm versus 1.6 (1.1, 2.0) mm, $p < 0.001$; diastole 1.7 (1.4, 2.2) mm versus 1.3 (1.1, 1.8) mm, $p < 0.001$). Large PFOs in systole and diastole were more common in patients with CS (systole 51.5% versus 30.6%, $p = 0.009$; diastole 34.8% versus 18.8%, $p < 0.001$), while long-tunnel PFO and length of tunnel showed no significant difference between the two groups in either systole or diastole. Patients with CS had a greater maximum mobility [5.9 (3.3, 7.6) mm versus 3.2 (2.3, 4.6) mm, $p < 0.001$]. ASA was present in 27.3% of patients with CS, compared with 1.2% of patients without CS ($p < 0.001$).

4.3 Random Forest Model Results

The test results of the random forest model are summarized in Fig. 6. Among 30 samples, 21 were correctly predicted, with an accuracy rate of 70%. Moreover, the relative operating characteristic curve diverged from the 45° line near the coordinates (0,0) and (1,1) and yielded an area under the curve value of 0.816, which also indicates that the model has acceptable discrimination to diagnose patients with low and high risk.

4.4 Discovery of High-Risk Factors

The order of importance of variables in the random forest model is shown in Fig. 7, in which maximum mobility, large RL shunt during Valsalva maneuver, size of PFO in diastole and systole, and diastolic length of the tunnel are the top five most important variables of the random forest.

4.5 Reproducibility

Data from 15 randomly selected patients were used to assess interobserver agreement. The interobserver intra-class correlation coefficient between two reviewers for size of PFO in diastole was 0.91 (0.76–0.97), and was 0.85 (0.62–0.95) for maximum mobility. There was 100% agreement between reviewer 1 and reviewer 2 for the presence of ASA and the classification of the severe RL shunt at rest and Valsalva maneuver.

4.6 Prediction Model Accuracy Comparison

Using the same dataset, we choose ANN [24, 25] and SVM [26] to predict the CS based on PFO from TEE Imaging Data. The prediction results are compared with the RF model results, and the root mean square error and goodness of fit are used to measure the prediction effect of the model. The error comparison of the prediction results of different models is shown in Table 2:

(1) The RF model has the highest fitting degree of resistance to permeability. The goodness of fit of the RF prediction results in the training set and test set is 0.968 and 0.951, and the coefficient of certainty is closest to 1 compared with the other two models. (2) The prediction error of RF resistance to permeability is the smallest. The RMSEs of the RF prediction results in the training set and test set are 0.036 and 0.095, the error is very close to 0, and lower than the other two prediction models. To sum up, the random forest prediction model has strong adaptability and superiority in A prediction, and can obtain prediction results with high accuracy and reliability [27, 28].

Fig. 6 Classification results visualized by the confusion matrix and relative operating characteristic curve

Fig. 7 The importance of the random forest model

Table 2 Error comparison

Model	RMSE		R2	
	Training set	Test set	Training set	Test set
RF	0.036	0.095	0.968	0.951
ANN	0.048	0.146	0.853	0.841
SVM	0.134	0.279	0.836	0.825

5 Discussion

This study developed a random forest model for high-risk PFO associated with CS, in which 21 variables were included. The model's predictive ability was found to be acceptable. The random forest approach has been shown to have high efficiency in processing medical data. It has been widely used in the fields of genes, proteins, drugs, diseases, and so on. However, investigations of the morphologic characteristics of PFO by TEE and ischemic lesions based on random forest models have not been conducted to date. The accuracy of the final test set in this study was 70%; in addition, the area under the curve value of the prediction ability of the model was 0.816 with high sensitivity of 73% and high specificity of 65%, indicating that the established random forest model performed well in identifying the risk factors for CS in patients with PFO.

In this study, random forest was implemented to quantify feature importance; it was found that maximum mobility, large RL shunt during Valsalva maneuver, size of PFO in diastole and systole, and diastolic length of the tunnel are closely related to CS. In approximately 40% of patients with ischemic stroke, the origin of cerebral ischemic events remains unknown [29, 30]. Multiple studies have shown that PFO can be implicated in the pathogenesis of CS [31, 32]. Other studies have determined that PFO size is closely related to CS. Nevertheless, some studies have shown that there is no significant association between the anatomy of PFO and paradoxical cerebral embolism [33]. Schuchlenz et al. [34] concluded that PFO size measured at exit location (left atrial side) is an independent risk factor for ischemic events, and that patients with a PFO size > 4 mm have a substantial risk of recurrent strokes. In contrast, Nakayama et al. observed that PFO size measured in the end-systolic frame is not related to CS [10]. However, PFO size changes during the cardiac cycle and differs depending on the location in the tunnel. PFO size in the systole has been found to be greater than that in the diastole at the entrance, mid-, and exit location, and PFO size at the entrance (right atrial side) has been shown to be greater than that of exit (left atrial side). The current study investigated the relationship between PFO size at the exit location (left atrial side) in both systole and diastole and CS; our findings reveal that the sizes of the PFO in the diastole and systole are both related to CS.

PFO is generally considered to be the anatomic means by which paradoxical cerebral embolism develops [35]. The saline contrast TEE test is a widely accepted noninvasive standard for diagnosing PFO, enabling the RL flow to be noted along with semi-quantification of RL shunt size according to the bubble count in the left atrial. In the present study, we observed that patients with CS had a higher frequency of severe RL shunt compared to those without CS ($p < 0.001$). This finding is unexpected, because the larger RL shunt leads to greater potential for thrombus to pass directly from venous to arterial circulation when the pressure in the right exceeds that in the left cardiac chamber, which increases the likelihood of paradoxical embolic stroke. Nevertheless, the finding is consistent with previous reports.

In the present study, maximum mobility was found to be associated with CS in patients with PFO. In fact, De Castro et al. investigated the morphological and functional characteristics of PFO and their embolic implications in patients with a median follow-up period of 31 months [6]. They found that greater interatrial septum mobility was more common in patients with CS, and RL shunt at rest with a hypermobile interatrial septum seemed to identify PFO patients who were at high risk of paradoxical cerebral embolism recurrence. Such findings have also been supported in other research [36]. It is believed that increased interatrial septum mobility may be an indicator of a larger PFO and is able to strengthen the preferential orientation of blood flow from the inferior vena cava via the PFO into the left atrium, leading to an increase in the potential thrombus passage and the occurrence of paradoxical embolism [37]. Nevertheless, in this study, a moderate correlation was found between the presence of ASA and CS; this differs from prior PFO–ASA studies [37], where ASA accompanied by PFO was demonstrated to be more frequent in patients with CS and recurrent stroke. Indeed, this discrepancy could partly be explained by the different definitions of ASA used in these studies. In the study by [18], an ASA was diagnosed when the atrial septum extended ≥ 11 mm into the right or left atrium or if the sum of the excursion into the left and right atria was ≥ 11 mm. In [38], ASA was defined as a septum primum excursion ≥ 10 mm from the atrial septum into the left or right atrium. However, in our study, we considered septal excursion from the midline into the right or left atrium ≥ 10 mm, or total excursion between the right and left atrium ≥ 15 mm, as diagnostic criteria. Therefore, non-invasive “gold standard” criteria are needed to normalize the identification of atrial septal aneurysm.

Similar to a previous study [11], we found that the diastolic length of the tunnel was highly associated with CS. This indicates that the long tunnel may tend to serve as a conduit for paradoxical emboli, or produce stroke via in situ thrombus formation [39].

6 Conclusions

Large PFO in diastole, the presence of hypermobile interatrial septum, severe right-to-left shunt and Eustachian valve or Chiari’s network were independently associated with CS suggesting that TEE-detected morphologic and functional characteristics of PFO may play important roles in the occurrence of CS.

Our model’s credibility is supported by the fact that the importance of the influencing factors used can be traced, meaning that we can use it to effectively evaluate patients with high-risk PFO in clinical practice. Nevertheless, the study has some limitations. First, there are many variables involved, which is the best number of variables obtained using the CARET package in the language, but these may not be practical in clinical settings. The number of variables may need to be optimized in future studies using other methods. Second, this was a single-center study, which included a small sample, and some variables had to be deleted due to missing values. Future research should increase the sample size and the number of variables to provide robust data. In addition, this study was exploratory, and needs to be verified by samples that include more populations. Finally, the study only considered the presence/absence of cerebral infarction lesion, not the severity of neurological events such as cerebral infarct distribution, location, number, and so on.

Acknowledgements The authors are also grateful for the constructive comments and suggestions of the manuscript reviewers.

Author Contributions JB, writing—original draft; JY, data curation; WS, software; YL, supervision; HX, validation; YL, writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Funding This research was funded by the Philosophy and Social Science research Project in Department of Education of Hubei Province (Grant No. 21G001), the medical Sci-Tech innovation platform of Zhongnan Hospital, Wuhan University, Grant Number PTXM2021008, and the Construction of Science and Technology Planning Project of Hubei Province in 2020 (Grant No. 2020041).

Data Availability Not applicable.

Declarations

Conflict of interest The authors declare no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not

permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Zhou, M., Wang, H., Zeng, X.: Mortality, morbidity, and risk factors in China and its provinces, 1990–2017: a systematic analysis for the global burden of disease study 2017 (vol 394, pg 1145, 2019). *Lancet* **396**(10243), 26–26 (2020)
- Kent, D.M., Thaler, D.E.: Is patent foramen ovale a modifiable risk factor for stroke recurrence? *Stroke* **41**(10), S26–S30 (2010)
- Gennarelli, G., Ludeno, G., Soldovieri, F.: Real-time through-wall situation awareness using a microwave Doppler radar sensor. *Remote Sens* **8**(8), 621 (2016)
- Ntaios, G., Weng, S.F., Perlepe, K., Akyea, R., Condon, L., Lambrou, D., Sirimarco, G., Strambo, D., Eskandari, A., Karagkiozi, E., Vemmou, A., Korompoki, E., Manios, E., Makaritsis, K., Vemmos, K., Michel, P.: Data-driven machine-learning analysis of potential embolic sources in embolic stroke of undetermined source. *Eur J Neurol* **28**, 192–201 (2021)
- Zhu, Y., Zhang, J., Huang, B., Liu, Y., Deng, Y., Weng, Y., Sun, R.: Impact of patent foramen ovale anatomic features on right-to-left shunt in patients with cryptogenic stroke. *Ultrasound Med Biol* **47**, 1289–1298 (2021)
- De Castro, S., Cartoni, D., Fiorelli, M., Rasura, M., Anzini, A., Zanette, E.M., Beccia, M., Colonnese, C., Fedele, F., Fieschi, C., Pandian, N.G.: Morphological and functional characteristics of patent foramen ovale and their embolic implications. *Stroke* **31**, 2407–2413 (2000)
- Vitarelli, A.: Patent foramen ovale: pivotal role of transesophageal echocardiography in the indications for closure, assessment of varying anatomies and post-procedure follow-up. *Ultrasound Med Biol* **45**, 1882–1895 (2019)
- Nietlispach, F., Meier, B.: Percutaneous closure of patent foramen ovale: an underutilized prevention? *Eur Heart J* **37**, 2023–2028 (2016)
- Mas, J.L., Derumeaux, G., Guillon, B., et al.: Patent foramen ovale closure or anticoagulation vs. antiplatelets after stroke. *N Engl J Med* **377**(11), 1011–1021 (2017)
- Saver, J.L., Carroll, J.D., Thaler, D.E., Smalling, R.W., MacDonald, L.A., Marks, D.S., Tirschwell, D.L., Investigators, R.: Long-term outcomes of patent foramen ovale closure or medical therapy after stroke. *N Engl J Med* **377**, 1022–1032 (2017)
- Nakayama, R., Takaya, Y., Akagi, T., Watanabe, N., Ikeda, M., Nakagawa, K., Toh, N., Ito, H.: Identification of high-risk patent foramen ovale associated with cryptogenic stroke: development of a scoring system. *J Am Soc Echocardiogr* **32**, 811–816 (2019)
- Shan, J., Alam, S.K., Garra, B., Zhang, Y.T., Ahmed, T.: Computer-aided diagnosis for breast ultrasound using computerized BI-RADS features and machine learning methods. *Ultrasound Med Biol* **42**(4), 980–988 (2016)
- Liu, Z., Wen, T., Sun, W., Zhang, Q.: Feature-weighting and clustering random forest. *Int J Comput Intell Syst* **14**(1), 257–265 (2021)
- Liu, Y., Chen, H.Y., Zhang, L.M., Feng, Z.B.: Enhancing building energy efficiency using a random forest model: a hybrid prediction approach. *Energy Rep* **7**, 5003–5012 (2021)
- Molpeceres Barrientos, G., Alaiz-Rodriguez, R., Gonzalez-Castro, V., Parnell, A.C.: Machine learning techniques for the detection of inappropriate erotic content in text. *Int J Comput Intell Syst* **13**(1), 591–603 (2020)
- Tanaka, J., Izumo, M., Fukuoka, Y., Saitoh, T., Harada, K., Harada, K., Gurudevan, S.V., Tolstrup, K., Siegel, R.J., Shiota, T.: Comparison of two-dimensional versus real-time three-dimensional transesophageal echocardiography for evaluation of patent foramen ovale morphology. *Am J Cardiol* **111**, 1052–1056 (2013)
- Lee, P.H., Song, J.K., Kim, J.S., Heo, R., Lee, S., Kim, D.H., Song, J.M., Kang, D.H., Kwon, S.U., Kang, D.W., Lee, D., Kwon, H.S., Yun, S.C., Sun, B.J., Park, J.H., Lee, J.H., Jeong, H.S., Song, H.J., Kim, J., Park, S.J.: Cryptogenic stroke and high-risk patent foramen ovale: the DEFENSE-PFO trial. *J Am Coll Cardiol* **71**, 2335–2342 (2018)
- Goel, S.S., Tuzcu, E.M., Shishebor, M.H., de Oliveira, E.I., Borek, P.P., Krasuski, R.A., Rodriguez, L.L., Kapadia, S.R.: Morphology of the patent foramen ovale in asymptomatic versus symptomatic (stroke or transient ischemic attack) patients. *Am J Cardiol* **103**, 124–129 (2009)
- Schneider, B., Hofmann, T., Justen, M.H., Meinertz, T.: Chiari's network: normal anatomic variant or risk factor for arterial embolic events? *J Am Coll Cardiol* **26**, 203–210 (1995)
- Schnieder, M., Siddiqui, T., Karch, A., Bahr, M., Hasenfuss, G., Liman, J., Schroeter, M.R.: Clinical relevance of patent foramen ovale and atrial septum aneurysm in stroke: findings of a single-center cross-sectional study. *Eur Neurol* **78**, 264–269 (2017)
- Namvar, A., Siami, M., Rabhi, F., Naderpour, M.: Credit risk prediction in an imbalanced social lending environment. *Int J Comput Intell Syst* **11**(1), 925–935 (2018)
- Liu, Y., Chen, H.Y., Zhang, L.M., Wang, X.J.: Risk prediction and diagnosis of water seepage in operational shield tunnels based on random forest. *J Civ Eng Manag* **27**(7), 539–552 (2021)
- Chen, Z.S., Yang, L.L., Chin, K.S., Yang, Y., Pedrycz, W., Chang, J.P., Skibniewski, M.J.: Sustainable building material selection: an integrated multi-criteria large group decision making framework. *Appl Soft Comput* **113**, 107903 (2021)
- Liu, Y., Wang, X.-J., Zhou, S., Chen, H.: Enhancing public building energy efficiency using the response surface method: an optimal design approach. *Environ Impact Assess Rev* **87**, 106548 (2021)
- Hussain, H.I., Kamarudin, F., Thaker, H.M.T., Salem, M.A.: Artificial neural network to model managerial timing decision: non-linear evidence of deviation from target leverage. *Int J Comput Intell Syst* **12**, 1282–1294 (2019)
- Liu, Y., Chen, H.Y., Zhang, L.M., Wu, X.G., Wang, X.J.: Energy consumption prediction and diagnosis of public buildings based on support vector machine learning: a case study in China. *J Clean Prod* **272**, 122542 (2020)
- Xiao, L., Chen, Z.S., Zhang, X., Chang, J.P., Pedrycz, W., Chin, K.S.: Bid evaluation for major construction projects under large-scale group decision-making environment and characterized expertise levels. *Int J Comput Intell Syst* **13**(1), 1227–1242 (2020)
- Chen, Z.S., Martinez, L., Chang, J.P., Wang, X.J., Xiong, S.H., Chin, K.S.: Sustainable building material selection: a QFD-and ELECTRE III-embedded hybrid MCGDM approach with consensus building. *Eng Appl Artif Intell* **85**, 783–807 (2019)
- Fonseca, A.C., Ferro, J.M.: Cryptogenic stroke. *Eur J Neurol* **22**, 618–623 (2015)
- Boutet, C., Rouffange-Leclair, L., Garnier, P., Quenet, S., Delsart, D., Varvat, J., Epinat, M., Schneider, F., Antoine, J.C., Mismetti, P., Barral, F.G.: Brain magnetic resonance imaging findings in cryptogenic stroke patients under 60 years with patent foramen ovale. *Eur J Radiol* **83**, 824–828 (2014)
- Bonati, L.H., Kessel-Schaefer, A., Linka, A.Z., Buser, P., Wetzel, S.G., Radue, E.W., Lyrer, P.A., Engelter, S.T.: Diffusion-weighted imaging in stroke attributable to patent foramen ovale: significance of concomitant atrial septum aneurysm. *Stroke* **37**, 2030–2034 (2006)

32. Natanzon, A., Goldman, M.E.: Patent foramen ovale: anatomy versus pathophysiology—which determines stroke risk? *J Am Soc Echocardiogr* **16**, 71–76 (2003)
33. Kumar, P., Rusheen, J., Tobis, J.M.: A comparison of methods to determine patent foramen ovale size. *Catheter Cardiovasc Interv* **96**, E621–E629 (2020)
34. Schuchlenz, H.W., Weihs, W., Horner, S., Quehenberger, F.: The association between the diameter of a patent foramen ovale and the risk of embolic cerebrovascular events. *Am J Med.* **109**(6), 456–62 (2000)
35. De Castro, S., Cartoni, D., Conti, G., Beni, S.: Continuous monitoring by biplane transesophageal echocardiography of pulmonary and paradoxical embolism. *J Am Soc Echocardiogr* **8**(2), 217–220 (1995)
36. Holda, M.K., Koziej, M.: morphometric features of patent foramen ovale as a risk factor of cerebrovascular accidents: a systematic review and meta-analysis. *Cerebrovasc Dis* **49**, 1–9 (2020)
37. Overell, J.R., Bone, I., Lees, K.R.: Interatrial septal abnormalities and stroke: a meta-analysis of case-control studies. *Neurology* **55**, 1172–1179 (2000)
38. Turc, G., Lee, J.Y., Brochet, E., Kim, J.S., Song, J.K., Mas, J.L.: Atrial septal aneurysm, shunt size, and recurrent stroke risk in patients with patent foramen ovale. *J Am Coll Cardiol* **75**, 2312–2320 (2020)
39. Yan, C., Li, H.: Preliminary investigation of in situ thrombus within patent foramen ovale in patients with and without stroke. *JAMA* **325**, 2116–2118 (2021)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.