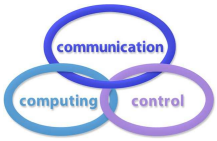
INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL
Online ISSN 1841-9844, ISSN-L 1841-9836, Volume: 16, Issue: 2, Month: April, Year: 2021
Article Number: 4108, https://doi.org/10.15837/ijccc.2021.2.4108




CCC Publications


# Prognosis Prediction of Stroke based on Machine Learning and Explanation Model

Q. Qin, X. Zhou, Y. Jiang

**Qiuli Qin*, Xuehan Zhou**
Department of Information Management
Beijing Jiaotong University, China
Beijing 100044, China
*Corresponding author: qlqin@bjtu.edu.cn
xhzhou@bjtu.edu.cn

**Yong Jiang**
China National Clinical Research Centre for Neurological Diseases
Beijing Tiantan Hospital,Capital Medical University , China
jiangyong@ncrcnd.org.cn

**Abstract**

The prognosis prediction of stroke is of great significance to its prevention and treatment. This paper used machine learning to predict stroke prognosis, and use SHAP method to make feature importance and single sample analysis. Firstly, feature engineering, use Borderline-SMOTE algorithm to deal with data imbalance, use Support Vector Machine(SVM) to build a prognostic prediction model, and use Random Forest(RF), Decision Tree(DT), Logistic Regression(LR) for comparative analysis, and find the performance of SVM after feature engineering better than other models, the accuracy, specificity, F1 score, AUC value reach 0.8306, 0.8356, 0.8415 and 0.9140. Then, the model was further analyzed for explainability, and it was found that the top three causes of the disease were Glasgow Coma Score, NIHSS and atrial fibrillation. Finally, try to analysis a single sample, which is performed to determine that the patient is a low-risk patient, and suffering from atrial fibrillation is the largest potential risk factor for the patient.

**Keywords:** machine learning; stroke; prognosis prediction; explanation model.

## 1 Introduction

Stroke is a global public health problem, and the prevention and treatment of stroke is also the primary task of public health in all countries. According to the joint statistics of the *American Heart Association and the National Institutes of Health* [30], stroke is the second leading cause of death and the third leading cause of disability in the world population. According to data published by the WHO[31], among the 56.9 million deaths worldwide in 2016, ischemic heart disease and stroke caused 15.2 million deaths, which was the main cause of death worldwide in 2001-2016.

With the increase in the incidence and recurrence rate of stroke, the burden of stroke has a serious impact on family and public medical care. In the absence of specific treatments, prevention and intervention are the best treatments. Therefore, it is of great significance for public health management to deeply understand the influencing factors of stroke disease and prognosis and to do a good job of secondary prevention of stroke.

Risk factors of stroke can be divided into modifiable and non-modifiable. WHO[32] shows, among the modifiable risk factors, hypertension, diabetes, obesity, and lack of physical activity are the important risk factors for stroke, among which hypertension is the most important risk factor for stroke. Among the non-modifiable factors, AHA/NIH[30] shows that age, gender and previous stroke history are important risk factors, and family stroke history and race are also listed as risk factors[33]. There is also [30], [5] evidence that recent use of antihypertensive drugs, rapid weight changes, and atrial fibrillation are also risk factors for stroke.

With the improvement of hospital information management system and the rapid development of machine learning, the study of stroke prognosis prediction based on machine learning has also developed. Prognosis prediction can help doctors make better medical decisions and achieve a better prognosis. Heo et al. [14] used deep neural networks, random forest and logistic regression algorithms to predict the long-term prognosis of ischemic stroke, and found that deep neural networks had the best performance. Machine learning can also analyze brain images and make prediction. Kuang et al.[18] used random forest to analyze brain NCCT images and realized automatic ASPECTS scoring. The sensitivity of the score reached 97.8%, and the AUC value was 0.89, which reflected good performance. Xie et al.[19] used GBM and XGB models for image segmentation of CT images and 90-day predicted mRS scores, and found that decision tree-based GBM performed better. Shameer[20] built a risk prediction model based on machine learning. By processing electronic medical record(EMR) data, predict the readmission of stroke patients with an accuracy rate of 83.19% and an AUC value of 0.78. Sung[29] used a supervised machine learning technique to mine EMR text to distinguish different stroke subtypes and found that binary classification based on decision tree and random forest is better than multi-classification. The use of machine learning to analyze medical data is a hot topic of current research. Most stroke prediction studies focus on brain imaging (CT) and electronic medical records(EMR). Afify H.M. et al.[1] automated diagnosis of different types of breast carcinoma histopathological images by machine learning algorithms, which was a multi-classification task. Romana C.H. et al.[7] used firefly algorithm combined with K-means clustering(KM-FA) in brain image segmentation and showed a great performance. Additionally, machine learning has been applied in electronic commerce, power system, education and other fields. Jing LI et al.[21] used machine learning in custom behavior prediction, include clustering, decision tree and naive bayesian algorithm, to explore the characteristics of target groups in which purchase behavior would occur.

In the study of stroke prediction based on machine learning, commonly used algorithms can be divided into interpretability models and difficult-to-interpret models. An interpretable model is a highly interpretable model structure, such as decision trees and logistic regression. The difficult-to-interpret model, which can also be called a complex model, usually has good predictive performance[11]. Therefore, although complex machine learning models show superior performance in the field of medical prediction, they usually lack the interpretability analysis of the results. At present, there are different interpretation methods for complex models. The more popular one is to construct an explanation model, namely an explanation, to assist in the interpretation of complex models. Ribeiro [26] et al. proposed a LIME (Local Interpretable Model-agnostic Explanation) method in 2016. The principle is to perturb the input data and observe the changes in the prediction results to fit an interpretable model and use the pseudo The combined model explains the prediction results. The LIME method can be independent of the model. It can explain a sample from the perspective of the model as a whole, which can explain the results prediction, image recognition, text classification and other issues. Inspired by the LIME method, Lundberg et al. [22] proposed the SHAP (SHapley Additive exPlanations) method combined with the Shapley value method. The Shapley value method is the benefit distribution plan of coalition members proposed by Shapley L.S. in 1953 [28], which distributes benefits according to the marginal contributions of the coalition members. The SHAP rule treats alliance members as features in the dataset to calculate the marginal contribution of each feature. Both LIME

and SHAP can interpret the model well and are independent of the model. In contrast, SHAP can provide local accuracy and consistency guarantees in the process of interpreting the model.

In this paper, we firstly proves that the complex machine learning method after feature engineering and parameter tuning has better performance than the simple machine learning method. With the development of machine learning, human not only pursue high performance, but also try to enhance the interpretability of machine learning, which help human can understand the prediction results. In the current research, many studies[[16], [9], [3]] showed the prediction results of machine learning are lack of interpretability. In the context of precision medicine, the interpretability model can effectively deal with the limitation of the machine learning prediction result, such as lack of robustness, lack of interpretability, and difficulty in application. In this paper, based on game theory, the interpretability research is divided into global interpretability and local interpretability. One part applies global interpretability to analyze the importance of different prognostic factors, and the other part applies local interpretability to analyze the risk factors of individual patients. This method can provide more timely, efficient and interpretable prediction results for clinical treatment decision-making, and use the analysis results to solve clinical practical problems.

Based on the above, in this paper, we use the dataset from the third international stroke trial(IST-3), which collected multi-dimensional dataset(266 variables). After feature selection, we use basic information, detection indicators, functional check, past medical history and other information. Then, we use different machine learning algorithms to build prediction models. Finally, we analyze the explanation of the models' results to improve the application further, which include feature important and single sample analysis. The purpose of this paper is to enhance the understanding of the influencing of stroke prognosis and to explore an explanation model applied to stroke diseases, hoping to help medical staff assist in diagnosis and treatment.

## 2 Materials

### 2.1 IST-3 Data

The Third International Stroke Trial (IST-3)[27] is a large-scale randomized trial and the largest randomized controlled trial in the history of acute ischemic stroke, which recorded stroke patients And the three-year follow-up data of the thrombolytic control group provided 266 variables including basic information, treatment plan, and follow-up information of the patient. The dataset included a total of 3035 patients older than 18 years old, of which 1617 (53%) patients were over 80 years old.

### 2.2 Input

Many risk factors affect the prognosis of stroke. In this paper, among the 266 variables in the original dataset, combined with the characteristics of the early stroke management guide[25] and the IST-3 dataset, we divide the independent variables into 5 categories, namely basic information, detection indicators, functional checks, and past The medical history, medication history, and stroke subtypes, and the variable assignment methods are shown in Table 1.

1) Basic information: The basic information of the patient includes age, gender, weight, and whether they live alone. Age and gender are important uncontrollable factors for stroke[30]. Obesity and lack of physical activity are also risk factors for stroke[32]. Therefore, weight and whether living alone before the stroke occurs can be considered as potential risk factors.

2) Testing indicators: The patient's testing indicators include systolic blood pressure, diastolic blood pressure, and blood sugar. The systolic and diastolic blood pressure reflect the patient's blood pressure level, and hypertension and diabetes are the most important risk factors[32]. 3) Functional check: The functional examination of the patient includes NIHSS, ADL scale, and Glasgow coma score. To assess the condition of patients, there are currently many scales that can quickly quantify the degree of functional impairment of patients, for example, NIHSS (NIH Stroke Scale) used to assess the degree of neurological impairment[25], ADL ( Activity of Daily Living Scale), GCS (Glasgow Coma Scale) to assess the degree of coma in patients, etc.

Table 1: Dataset features, feature description and value assignment methods

| Feature | Description | Assignment method |
|---|---|---|
| dead6mo | Died within 6 months | Yes=1,No=2 |
| **Basic Information** | | |
| Age | Age | Continuous variable |
| Gender | Gender | Female=1, male=2 |
| weight | Weight(kg) | Continuous variable |
| livealone_rand | Live alone | Yes=1,No=2 |
| **Detection Indicator** | | |
| sbprand | Systolic blood pressure(mm Hg) | Continuous variable |
| dbprand | Diastolic blood pressure(mm Hg) | Continuous variable |
| glucose | blood sugar (mmol/L) | Continuous variable |
| **Function check** | | |
| indepinadl_rand | Independent in ADL before stroke? | Yes=1,No=2 |
| gcs_score_rand | Total Glasgow Coma Scale score at randomisation | Continuous variable |
| nihss | Total NIH Stroke Score at randomisation | Continuous variable |
| **Medical history / Medication history** | | |
| antiplat_rand | Received antiplatelet drugs in last 48 hours? | Yes=1,No=2 |
| atrialfib_rand | Patient in atrial fibrillation at randomisation? | Yes=1,No=2 |
| stroke_pre | History of previous stroke or TIA? | Yes=1,No=2 |
| diabetes_pre | Treatment for diabetes before admission? | Yes=1,No=2 |
| hypertension_pre | Treatment for hypertension before admission? | Yes=1,No=2 |
| **Stroke subtype** | | |
| stroketype | Stroke subtype | 1='TACI', 2='PACI', 3='LACI', 4='POCI', 5='OTHER' |

4) Past medical history & medication history: The patient's past medical history and medication history include antiplatelet drugs, atrial fibrillation, past stroke history, and hypertension. Because stroke has a high recurrence rate, the history of past stroke is also an important factor. Antiplatelet drugs have been shown to have a preventive effect on ischemic stroke in previous studies[17], but at the same time, they may have a risk of hemorrhagic transformation[10]. There is a strong correlation between atrial fibrillation and stroke, and atrial fibrillation is also an important risk factor for stroke[24].

5) Stroke subtypes: In this dataset, stroke subtypes include five types, namely 1='TACI', 2='PACI', 3='LACI', 4='POCI', 5='OTHER'.

### 2.3 Output

In this paper, we selected the survival of patients within 6 months as the output.

## 3 Methods

### 3.1 Data Analytics Framework

The input of the model is the information when the patient is admitted to the hospital, and the output is the binary prediction result of whether the patient will die after 6 months. The data analysis framework is shown in Figure 1.

### 3.2 Data Preprocessing

1)Process Discrete Features

For discrete features that exist in the dataset, such as gender, stroke subtype, etc., the values of these features are not numerically differentiated, so one-hot encoding (one-hot) is used for processing. Taking gender as an example, the Gender feature is converted into two features gender_female and gender_male. If the sample value of the original gender feature is 1, the sample value of the converted

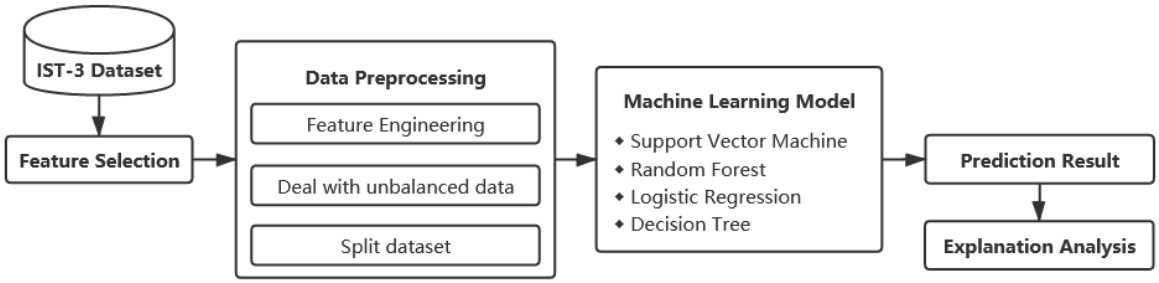



Figure 1: Data analysis framework for the prediction of stroke prognosis

[gender_female, gender_male] is [1, 0], thereby reducing The effect of meaning value on model training results.

2)Data Standardization

For continuous features in the dataset, such as systolic blood pressure, body weight and other features, because different features have different scales, features with a large scale have a greater impact on the results, so the dataset needs to be standardized. In this paper, z-score standardization is used to standardize the data by giving the mean and standard deviation of the original data. The processed data conforms to the standard normal distribution, that is, the mean is 0 and the standard deviation is 1. The conversion function is:

$$x^* = \frac{x - \bar{x}}{\sigma} \tag{1}$$

where $\bar{x}$ is the mean value of the feature, and $\sigma$ is the standard deviation of the feature.

3)Deal with Unbalanced Dataset

By analyzing the original dataset, it is found that the outcome feature "dead6mo" has 2220 positive samples and 815 negative samples, which means the dataset is unbalanced. Since machine learning tends to improve the accuracy of the majority class, the performance of the model will decrease, and the ability to judge minority samples will decrease. In the medical field, it is often necessary to keenly find high-risk patients. Therefore, it is necessary to deal with unbalanced dataset to improve the machine's ability to judge minority samples. This paper uses Borderline-SMOTE algorithm to deal with imbalanced dataset.

Borderline-SMOTE is an improved oversampling algorithm based on SMOTE, which uses only a few samples on the border to synthesize new samples, to improve the imbalanced distribution of the dataset[12]. The Borderline-SMOTE method divides minority samples into 3 categories, namely Safe, Danger, and Noise. Samples with more than half of the nearest neighbors are in the minority class are classified as Safe, samples with more than half of the nearest neighbors are in the majority class are classified as Danger, and all of the nearest neighbors are in the majority class are classified as Noise. Since the model usually has a weak distinguishing ability for Danger class, the Borderline-SMOTE method uses Danger class K nearest neighbors to randomly generate minority class samples. Based on SMOTE, the improved Borderline-SMOTE improves the ability to distinguish boundary samples.

4)Training and Test Set

Before training the model, the dataset needs to be divided into training set and test set. The training set is used to train and fit the model, and the test set is used to test the trained model. In this experiment, the dataset was randomly divided according to the ratio of 7:3, and the training set and the test set were constructed.

### 3.3 Hyper-Parameter optimization

Before model training, model parameters need to be determined. Such parameters are also called hyperparameters. Optimizing the hyperparameters and selecting a set of optimal hyperparameters for the machine learning model can improve the performance and effect of learning to a certain extent. This paper uses grid search algorithm to optimize the parameters.

The grid search algorithm (GridSearch, GS) is one of the most basic exhaustive parameter optimization algorithms, and the parameters obtained are relatively reliable[4]. The principle is to divide the parameters to be searched into grids in a specific spatial range, and then traverse all points in the grid to find the parameter values that optimize the performance of the model. The grid search algorithm is simple, convenient, easy to understand, and fast in finding the best.

### 3.4 Machine Learning model

In this paper, we chose 4 machine learning models to predict whether they would die after 6 months of stroke, include Support Vector Machine, Random Forest, Decision Tree and Logistic Regression. These 4 models were selected because they have great performance in binary classification. Among the 4 models, Support Vector Machines and Random Forests could process high-dimensional data well while maintaining great performance. But as a representative of complex models, it means that their results are always difficult to understand by human. Logistic Regression and Decision Tree are baseline algorithms for comparison, because they have interpretable advantages, people can usually understand the prediction results of Logistic Regression and Decision Tree models well.

1)Support Vector Machine

The support vector machine[8] divides all samples in the training set into two categories by finding an optimal hyperplane, and at the same time maximizes the classification interval between the two types of samples. The samples located on the optimal hyperplane are called support vectors. The solid dots and hollow dots respectively represent the samples of the two categories, H is the classification hyperplane, and the optimal classification plane not only ensures that the two types of samples are accurately separated, but also requires their classification Maximum interval. The former is to ensure that the empirical risk value is minimized, while the latter is to minimize the confidence range of the generalized community and ultimately lead to the smallest real risk.

2)Random Forest

Random forest[15] is a classifier based on the idea of ensemble learning. Multiple weak classifiers composed of decision trees are integrated into a strong classifier by voting. For each decision tree, the training set of each tree is randomly selected training samples with replacement. This method is called the bootstrap sample method.

3)Decision Tree

Decision tree is a classic machine learning algorithm, which is composed of internal nodes, directed edges, and leaf nodes. Through the decision tree, you can clearly see the choice of each root node when making a decision, so it has good interpretability.

4)Logistic Regression

Logistic regression methods are classic machine learning and statistical methods, which can not only solve regression problems, but also solve binary classification problems. Through the parameters of logistic regression fitting, the prediction results can be well understood, so it has good interpretability. For example, after a linear logistic regression model is fitted with $Y = X\beta + \varepsilon$ , the importance of each feature can be understood through $\beta$.

### 3.5 Explanation model

In the medical field, while pursuing a high rate of prediction accuracy, it is also necessary to explain the results of disease prediction to help doctors analyze the causes of patients and achieve the purpose of auxiliary diagnosis and treatment. This paper uses the SHAP method to analyze the predictive results. The SHAP method can either rank the overall feature importance or a single sample. The principle is as follows[22]:

Let $S \subseteq F$ ($F$ is the full set of features), $S$ represents the set of elements in the sequence before the feature $i$ , and $F\backslash S\backslash i$ is the set of elements after the feature $i$, then the elements of the $S$ set before $i$ and $F\backslash S\backslash i$ after the feature $i$ can form $|S|!\,(|F|-|S|-1)!$ sequence. In order to calculate the contribution value of the feature $i$ , it is necessary to train the model $f_{S\bigcup\{i\}}$ and the model $f_S$ first, and after the feature value $x_n$ is substituted into the model, the contribution value $f_{S\bigcup\{i\}}(x_{S\bigcup\{i\}})-f_S(x_S)$

is calculated. Finally, calculate the weighted average of each sequence, which is the SHAP value of each feature:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!\,(|F|-|S|-1)!}{|F|!}[f_{S\bigcup\{i\}}(x_{S\bigcup\{i\}}) - f_S(x_S)] \tag{2}$$

# 4 Experimental Results

To compare and intuitively see the model performance and experimental results, we first select the evaluation index and evaluate the model. Then, analyze the explanation for the problem that the complex model is difficult to explain.

## 4.1 Assessment measurement

The commonly used evaluation indicators of the binary classification model are accuracy, precision, specificity, F1-score, AUC (Area Under Curve) value, etc.

1) Confusion matrix

In the classification model, the prediction results can be recorded as TP (True Positive), FN (False Negative), FP (False Positive), TN (True Negative). And extend the accuracy rate, specificity, F1-score and other indicators, see Eq. (3)-(7).

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{3}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{4}$$

$$Specificity = \frac{TN}{FP + TN} \tag{5}$$

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

$$F_1 - score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{7}$$

2) ROC curve

The ROC curve refers to the receiver's operating characteristic curve. Each point on the curve reflects the susceptibility to the same signal stimulus, so it is also called the susceptibility curve. The horizontal axis of the ROC curve is the false positive rate, and the vertical axis is the true rate. The AUC value is the area under the ROC curve, and the value range is [0,1]. The larger the value, the better the classification effect.

3) Selection of evaluation indicators

In the two-class model, accuracy is the most intuitive and commonly used evaluation index to judge the performance of the model. In the scenario of stroke prognosis prediction, the model needs to prevent missed diagnosis as much as possible, so specificity is selected as one of the evaluation indicators. Also accuracy and specificity cannot fully evaluate the performance of the model, and F1-score can comprehensively evaluate the performance of the model. Therefore, this paper uses accuracy, specificity, F1-score and AUC value as the evaluation indicators of the model to comprehensively evaluate the model.

Table 2: Comparison of classification prediction results

| Model | Accuracy | Specificity | F1-score | AUC |
|---|---|---|---|---|
| SVM | 0.7747 | 0.6133 | 0.4279 | 0.7741 |
| SVM + Borderline-SMOTE | 0.8306 | 0.8356 | 0.8415 | 0.9140 |
| RF | 0.7802 | 0.6428 | 0.4285 | 0.7965 |
| RF + Borderline-SMOTE | 0.8045 | 0.8106 | 0.8172 | 0.8997 |
| LR | 0.7827 | 0.6351 | 0.5137 | 0.8056 |
| DT | 0.7131 | 0.4504 | 0.4739 | 0.6394 |

### 4.2 Assessment results and analysis

To compare the performance of the complex model and the baseline model, this paper uses the evaluation indicators mentioned above and obtains the performance results of different models (see Table 2).

1) The importance of Balancing the Dataset

In this experiment, by observing the specificity and F1-score, it can be found that after processing the imbalanced dataset, the model performance has been greatly improved. In terms of specificity, $0.8106_{(RF+Borderline-SMOTE)} > 0.6428_{(RF)}$, $0.8356_{(SVM+Borderline-SMOTE)} > 0.6133_{(SVM)}$. For F1-score, $0.8172_{(RF+Borderline-SMOTE)} > 0.4285_{(RF)}$, $0.8415_{(SVM+Borderline-SMOTE)} > 0.4279_{(SVM)}$. It can be seen that in the same machine learning algorithm, using the improved Borderline-SMOTE algorithm to process imbalanced dataset can greatly improve the prediction effect of the model in the minority class and improve the performance of the model. And it is suitable for different complex machine learning models, such as SVM and RF, has good portability. Therefore, the balance of sample categories is very important to the classification model.

2) Better performance of the Complex Models

Compared with the baseline model, the model after a series of processing has better performance under most evaluation indicators, and the complex model has strong predictive ability and significant advantages. In terms of accuracy, the accuracy of decision trees (0.7131) is lower than other models. Although the accuracy of the logistic regression model is slightly higher than that of the SVM and RF models after feature engineering, the accuracy of the complex model is significantly improved after data changes, which is much higher than LR. At the same time, the specificity and F1-score of DT and LR are much lower than SVM and RF models.

But it is undeniable that after a series of data processing, the complex model is far inferior to the baseline model in terms of interpretability. After data preprocessing, the dataset has lost part of information, and the dataset at this time has lost part of its interpretability. Then, after SVM and RF model training, the prediction results are more difficult to understand. Therefore, further analysis of the forecast results is needed.

### 4.3 Explanation Machine Learning Model

1) Feature Importance

Etiology research is an important part of modern medical research. Researching disease factors and risk factors can help people recognize and prevent diseases. The importance ranking of the overall sample can help people understand the importance of different features, so this paper uses importance ranking to help analyze the importance of different prognostic factors.

First, use the SHAP method to analyze the importance of features, calculate the SHAP value of the features of the overall sample, and sort the feature values. Through the overall feature importance ranking (see Figure 2), it can be found that the main features that affect the prognosis of stroke are the Glasgow coma score, the NIHSS, and whether the patient is atrial fibrillation. The patient's state at the time of admission has a great effect on the level of patient prognosis, which is basically consistent with the results of existing studies [[6],[2],[13]].

2) Single sample analysis

At present, most machine learning disease prediction models can achieve high prediction accuracy, but lack the interpretation of the prediction results of a single sample, and it is difficult to make targeted recommendations. Although the random forest and support vector machine constructed in

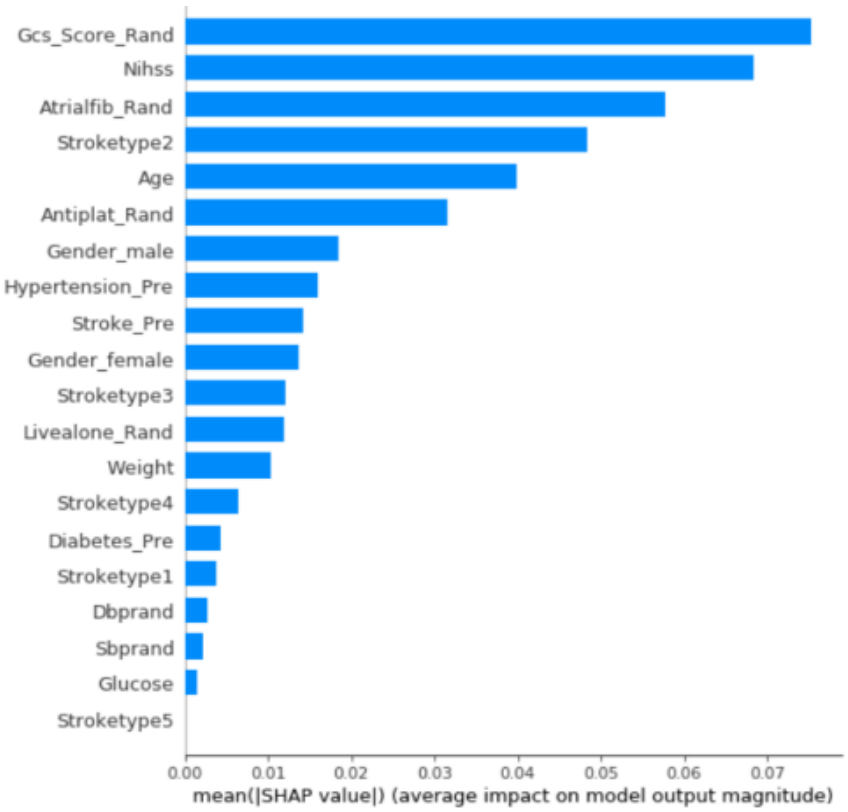



Figure 2: Feature Importance

this paper have more advantages in predictive performance, they are not as good as decision trees and logistic regression models in terms of interpretability. Especially after data preprocessing, the data becomes more difficult to understand.

When the SHAP method is used in the analysis of a single sample, it can enhance the interpretability of the model to a certain extent, help doctors understand the model, choose a better treatment plan, and improve the prognosis. This experiment randomly selects a sample for single-sample analysis (Figure 3). It can be seen from Figure 3 that the prediction result shows that the patient's risk value is 0.15, which is lower than the base value of 0.493, so the risk of death is low. Further analysis of this patient's low risk of death is due to the better performance of NIHSS and Glasgow coma scores, but since this patient is a patient with atrial fibrillation, which has a negative impact on the prognosis, standard anticoagulant therapy can be considered for this patient[23] to improve Prognosis.

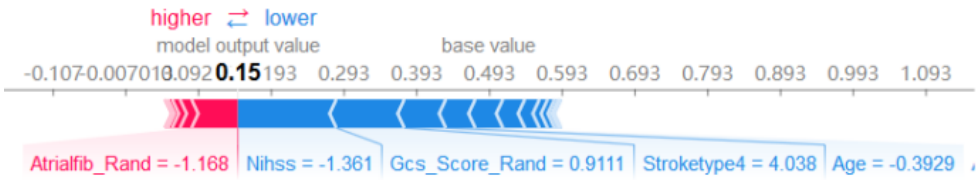



Figure 3: Single sample feature contribution value

## 5 Discussion

This paper mainly solves two problems. One is to study which machine learning method can improve the prediction effect of stroke as much as possible and to better serve public health; the other is how to explain the prediction results of the model, so that people can understand and use the results given by the model prediction.

1) Data preprocessing

Data preprocessing is one of the research focuses on this paper. The data preprocessing in this paper is mainly divided into three parts, namely discrete value processing, standardized processing and unbalanced dataset processing. The discrete value processing part, through the processing of one-hot encoding, will perform feature processing on features that do not have numerical values, such as gender. The standardized processing part avoids the problem of model bias caused by dimensional differences by standardizing all features. In terms of unbalanced dataset, by using the improved SMOTE method and the K-neighbor algorithm, new minority data are randomly generated at the edge of the classification, to avoid the problem of unbalanced data degrading model performance. Through the experimental results, it can be found that the performance of the model after data preprocessing is significantly superior to that of the unprocessed model. Especially after using the improved SMOTE algorithm, the specificity of the model is significantly improved. In practical applications, the probability of missed diagnosis can be reduced.

2) Machine learning algorithm selection

This paper considers different types of machine learning models in algorithm selection. Decision trees and logistic regression are baseline models with strong interpretability, while support vector machines and random forests are complex models with better prediction performance. It can be seen from the performance of the model that the complex model has obvious advantages when dealing with a large number of feature dataset, and the accuracy rate is higher than that of the baseline model. However, the results of complex models are difficult to interpret after data preprocessing and black-box model fitting, so they have not been widely used. This leads to the next research focus, interpretability analysis.

3) Interpretability analysis

This paper uses an advanced interpretability method, referring to the shapely value idea in game theory, and assists in explaining complex models by constructing an explanatory model. This method can not only rank the overall importance and analyze the important prognostic factors on the dataset, but also rank the feature importance of a single sample to analyze the main prognostic factors of a patient. Research on the interpretability of the model can help us understand the black box model, thereby better helping medical staff set up treatment plans and intervention measures.

4) Research limitations and future prospects

First of all, stroke is a complex disease, and current research usually focus on a kind of data structure, such as image data, electronic medical record data, streaming data, etc. In fact, we can try the method of data fusion, and make predictions after the fusion of multi-modal data. The prediction results may be more accurate and more stable. But the premise is that medical institutions and public health organizations consciously collect multi-modal data of stroke patients, and provide a good database platform for researchers who build models. Meantime, more data and more features will decrease classical machine learning performance, and we can try deep learning algorithm. Deep learning can deal with big data and high-dimensional features well and have a great performance.

Secondly, stroke disease may change dynamically over time. This paper uses death after 6 months as the outcome variable to predict, but this prediction method predicts the time point. In the future, we can try to use other methods for dynamic prediction, such as dynamically predicting the prognosis risk of patients in the form of a timeline.

Thirdly, this paper explores the explanation method and analyzes the overall feature importance and the feature importance ranking of individual samples. The ranking of overall feature importance can be confirmed by previous research. However, we don't have the dataset which is labeled, thus the interpretation of a single sample lacks expert verification, which is the limitation of the research process of this paper. Finally, although this paper uses the dataset of the International Stroke Trial, due to differences in race, geographic environment, and dietary habits, the results of studies in different regions, such as the importance of prognostic factors, may be different from this paper. However, the method in this paper is portable. Researchers in different regions can use the dataset and characteristics of local hospitals, regions, and countries to try the method proposed in this paper and verify the results.

# 6 Conclusion

In this paper, a representative IST-3 is used as the research dataset, and in response to the problems in the dataset, data preprocessing is used to process the original dataset. Then, different machine learning models were constructed, and the comparison found that the prediction effect of the support vector machine after data preprocessing was the best. Then the predictive results of the model were analyzed for interpretability, using feature importance to analysis the important prognostic risk factors of stroke. And we randomly selected a sample to assess prognostic risk, and try to give a treatment to improve the prognosis of patient.

### Author contributions

The authors contributed equally to this work.

### Conflict of interest

The authors declare no conflict of interest.

# References

[1] Afify, H.M.; Mohammed, K.K.; Hassanien, A.E.(2020). Multi-Images Recognition of Breast Cancer Histopathological via Probabilistic Neural Network Approach, *Journal of System and Management Sciences*, 1(2), 53-68, 2020.

[2] Asgedom, S. W. et al.(2020). Medical complications and mortality of hospitalized stroke patients, *Journal of stroke and cerebrovascular diseases*, 29(8), 104990, 2020.

[3] Azodi, C.B et al.(2020). Opening the Black Box: Interpretable Machine Learning for Geneticists, *Trends in genetics*, 36(6), 442-455, 2020.

[4] Bergstra, J.; Bengio, Y. (2012). Random Search for Hyper-Parameter Optimization, *The Journal of Machine Learning Research*, 13(1), 281-305, 2012.

[5] Boehme, Amelia K et al. (2017). Stroke Risk Factors, Genetics, and Prevention, *Circulation research*, 120(3), 472-495, 2017.

[6] Buonacera, Agata et al.(2019). Stroke and Hypertension: An Appraisal from Pathophysiology to Clinical Practice, *Current vascular pharmacology*, 17(1), 72-84, 2019.

[7] Capor Hrosik, R.; Tuba, E.; Dolicanin, E.; Jovanovic, R.; Tuba, M. (2019). Brain Image Segmentation Based on Firefly Algorithm Combined with K-means Clustering, Studies in Informatics and Control, *Studies in Informatics and Control*, 28(2), 167-176, 2019.

[8] Cortes, C; Vapnik, V. N. (1995). Support-Vector Networks, *Machine Learning*, 20(3), 273-297, 1995.

[9] Datta, A. et al.(2020). "Black Box" to "Conversational" Machine Learning: Ondansetron Reduces Risk of Hospital-Acquired Venous Thromboembolism, *IEEE journal of biomedical and health informatics*, 1-1, 2020.

[10] Esenwa, C.; Gutierrez, J. (2015). Secondary stroke prevention: challenges and solutions, *Vascular health and risk management*, 11, 437, 2015.

[11] Fellous, J.-M. et al.(2019). Explainable Artificial Intelligence for Neuroscience: Behavioral Neurostimulation, *Frontiers in neuroscience*, 13, 1346, 2019.

[12] Han, H.; Wang, W.; Mao, B. (2005). Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning, *International Conference on Intelligent Computing*, 2005.

[13] Hannon, N. et al.(2015). Antithrombotic treatment at onset of stroke with atrial fibrillation, functional outcome, and fatality: a systematic review and meta-analysis, *International journal of stroke*, 10(6), 808-814, 2015.

[14] Heo, J.N. et al. (2019). Machine Learning-Based Model for Prediction of Outcomes in Acute Stroke, *Stroke*, 50(5), 1263-1265, 2019.

[15] Ho, T.K. (1995). Random decision forests, *Proc. 3rd Int. Conf. Doc. Anal. Recognit.*, 278–282, 1995.

[16] Hofman, J. M. et al.(2017). Prediction and explanation in social systems, *Science*, 355(6324), 486-488, 2017.

[17] Kapil, N. et al. (2017). Antiplatelet and Anticoagulant Therapies for Prevention of Ischemic Stroke, *Clinical and applied thrombosis/hemostasis*, 23(4), 301-318, 2017.

[18] Kuang, H. et al. (2019). Automated ASPECTS on Noncontrast CT Scans in Patients with Acute Ischemic Stroke Using Machine Learning, *American journal of neuroradiology*, 40(1), 33-38, 2019.

[19] Kuang, H et al. (2019). JOURNAL CLUB: Use of Gradient Boosting Machine Learning to Predict Patient Outcome in Acute Ischemic Stroke on the Basis of Imaging, Demographic, and Clinical Information, *American journal of roentgenology*, 212(1), 44-51, 2019.

[20] Kuang, H. et al. (2017). PREDICTIVE MODELING OF HOSPITAL READMISSION RATES USING ELECTRONIC MEDICAL RECORD-WIDE MACHINE LEARNING: A CASE-STUDY USING MOUNT SINAI HEART FAILURE COHORT, *Pacific Symposium on Biocomputing*, 22, 276-287, 2017.

[21] Li, J; Pan, S.X.; Huang, L.; Zhu, X. (2019). A Machine Learning Based Method for Customer Behavior Prediction, *Tehnicki vjesnik-Technical Gazette*, 26(6), 1670-1676, 2019.

[22] Lundberg, S.; Lee, S. I. (2017). A Unified Approach to Interpreting Model Predictions, *NIPS*, 2017.

[23] Marijon, E., et al.(2013). Causes of Death and Influencing Factors in Patients With Atrial Fibrillation, *Circulation*, 128(20), 2192-2201, 2013.

[24] Pistoia, F. et al. (2016). The Epidemiology of Atrial Fibrillation and Stroke, *Cardiology clinics*, 34(2), 255-268, 2016.

[25] Powers, W. J. et al. (2019). Guidelines for the Early Management of Patients With Acute Ischemic Stroke: 2019 Update to the 2018 Guidelines for the Early Management of Acute Ischemic Stroke: A Guideline for Healthcare Professionals From the American Heart Association/American Stroke Association, *Stroke*, 50(12), e344-e418, 2019.

[26] Ribeiro, M. T.; Singh, S.; Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier, *International Conference on Knowledge Discovery and Data Mining*, 1135-1144, 2016.

[27] Sandercock, P.; Wardlaw, J.M.; Lindley, R.I. et al. (2012). The benefits and harms of intravenous thrombolysis with recombinant tissue plasminogen activator within 6 h of acute ischaemic stroke (the third international stroke trial [IST-3]): a randomised controlled trial, *Lancet*, 379, 2352-2363, 2012.

[28] Shapley, L. S.(1953). A value for n-person games, *Contributions to the Theory of Games*, 1953.

[29] Sung, S. F.; Lin, C. Y.; Hu, Y. H.(2020). EMR-based phenotyping of ischemic stroke using supervised machine learning and text mining techniques, *IEEE Journal of Biomedical and Health Informatics*, 24(10), 2922-2931, 2020.

[30] Virani, S.S. et al. (2020). Heart Disease and Stroke Statistics-2020 Update: A Report From the American Heart Association, *Circulation*, 141(9), e139-e596, 2020.

[31] [Online]. Available: https://www.who.int/zh/news-room/fact-sheets/detail/the-top-10-causes-of-death, Accesed on 10 January 2021.

[32] [Online]. Available: https://www.who.int/cardiovascular_diseases/en/cvd_atlas_03_risk_factors.pdf, Accesed on 10 January 2021.

[33] [Online]. Available: https://www.stroke.org/en/about-stroke/stroke-risk-factors/stroke-risk-factors-not-within-your-control?, Accesed on 10 January 2021.




This journal is a member of, and subscribes to the principles of,
the Committee on Publication Ethics (COPE).
https://publicationethics.org/members/international-journal-computers-communications-and-control

*Cite this paper as:*

Qin; Q.; Zhou, X.; Jiang, Y. (2021). Prognosis Prediction of Stroke Based on Machine Learning and Explanation Model, *International Journal of Computers Communications & Control*, 16(2), 4108, 2021.

https://doi.org/10.15837/ijccc.2021.2.4108