Stroke Prediction Based on Machine Learning

Yuhan Zhang

Marlan and Rosemary Bourns College of Engineering, University of California, Riverside, 92521, the United States

Abstract. Stroke has become an important cause of death and disability worldwide, which highlights the need for early detection and intervention. Machine learning technology can analyze patients' historical health data and biometrics to identify high-risk individuals in a timely manner, thereby effectively predicting stroke.This paper evaluates the predictive performance Random Forest and Support Vector Machine (SVM). Data preprocessing encompasses managing missing data, processing categorical variables, and tackling issues related to class imbalance. Analysis of the quantitative results indicates that the Random Forest model reaches an accuracy of 95% and a precision of 93%, providing a slight edge over the SVM, which records an accuracy of 92% and a precision of 90%.. However, both models exhibit high false-negative rates, with Random Forest showing a false-negative rate of 12% and SVM at 15%, which significantly impacts their clinical utility. To improve performance, further model optimization, such as adjusting class weights or employing ensemble methods, is necessary to reduce these false-negative rates and enhance diagnostic accuracy. This study highlights the potential and limitations of machine learning in stroke prediction, showing that people need further optimization to enhance diagnostic performance.

1 Introduction

In recent years, stroke has emerged as a major contributor to death and long-term disability globally, creating a substantial burden on individuals and healthcare systems alike. The prevalence of stroke has been rising globally, and in the United States, about 2.5 per cent of adults have experienced a stroke [1]. Every year, millions of people suffer from strokes, with factors such as high blood pressure, diabetes, smoking, and ageing contributing to the increasing prevalence of this life-threatening condition. Given the severe consequences, early detection and prevention of stroke are critical, as timely intervention can significantly reduce the risk of permanent damage or death. Over the past few years, machine learning (ML) has proven to be an effective tool in the healthcare sector. As the risk factors for stroke have become more apparent, ML has shown considerable promise in analyzing extensive patient data and detecting early indicators for stroke prediction.

Many researchers in the area of stroke risk prediction have conducted thorough investigations, employing machine learning models to enhance prediction accuracy. Alanzi et al. created machine learning models and implemented three data selection techniques: no

Corresponding author: yzhan1111@ucr.edu

© The Authors, published by EDP Sciences. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

resampling, imputation, and resampling [1]. Among the tested models, the Random Forest algorithm with data resampling outperformed others and achieved high accuracy (96%) in stroke prediction. Similarly, Shiozawa et al. explored the use of ML to predict stroke risk[2]. Their research demonstrated that machine learning could significantly improve prediction accuracy by analyzing larger datasets and incorporating more complex variables, such as lifestyle factors and genetic predispositions. They noted that handling data imbalances, especially with smaller stroke datasets, remains a critical challenge. The usefulness of ML models, such as Random Forest, XGBoost, Logistic Regression, and LightGBM, for forecasting the outcome of strokes was examined in a different study that used data from the Suita trial [3]. The predictive accuracy of Random Forest was found to be superior to that of other models by the researchers. Careful preprocessing, including data normalization, was highlighted as essential for achieving reliable results.

Furthermore, there have been notable developments in predictive modeling and the machine learning-based identification of critical stroke risk variables, as demonstrated by Hassan et al., who identified key stroke predictors through machine learning models [4]. Dritsas & Trigka also noted the results of this study and pointed out how machine learning approaches might improve the prediction of stroke risk [5]. This growing body of research highlights the increasing potential for properly predicting stroke risks, as well as its strengths and limitations.

The purpose of this study is to analyze and compare multiple machine learning models in order to identify the best instrument for stroke risk prediction. The following sections will cover the methodology, including data preprocessing and model selection, followed by the results, discussion, and conclusion regarding the optimal machine learning model for real-world applications.

2 Methods

2.1 Data and Preprocessing

The dataset for this study is from the Kaggle website (https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset): Stroke Prediction Dataset by Fedesoriano. The dataset consists of 5,110 observations across 12 variables, which include demographic, lifestyle, and health-related features such as age, hypertension, body mass index(BMI), average glucose level, and stroke occurrence, they are all known risk factors for stroke. This comprehensive dataset includes 7 variables, which offers a good foundation for creating prediction models to find people at risk of having a stroke. Table 1 shows the name and explanation of these 7 variables.

Table 1. Name and Explanation of Variables

Full Name	Data Type
Age	INT
Hypertension	INT
Heart Disease	INT
Average Glucose Level	FLOAT
BMI	FLOAT
Stroke	INT

The dataset is imbalanced, with a notably lower number of stroke cases than non-stroke cases, necessitating careful preprocessing to handle this imbalance. Therefore, it is necessary to preprocess the dataset to handle missing values and inconsistencies, ensuring the data is clean and suitable for machine learning analysis.

2.1.1 Handling Missing Data

Handling missing data is a significant step in the data preprocessing pipeline, especially in a medical dataset aimed at predicting strokes. Missing data, if left untreated, can introduce bias, distort the patterns in the dataset, and ultimately reduce the performance of predictive models. In the study, columns such as avg_glucose_level and BMI had significant missing values, which are crucial indicators for stroke prediction. By using the mean or median to fill the missing values, the paper ensures the completeness of the dataset and maintain the integrity of key features. Additionally, handling categorical features like work_type and smoking_status by replacing missing values with 'Unknown' allows the model to include all records without discarding potentially useful information. This comprehensive handling of missing data improves the robustness of the model, helping it to more accurately capture the risk factors associated with stroke and make better-informed predictions. Fig.1 reveals the missing values of different variables before and after handling.

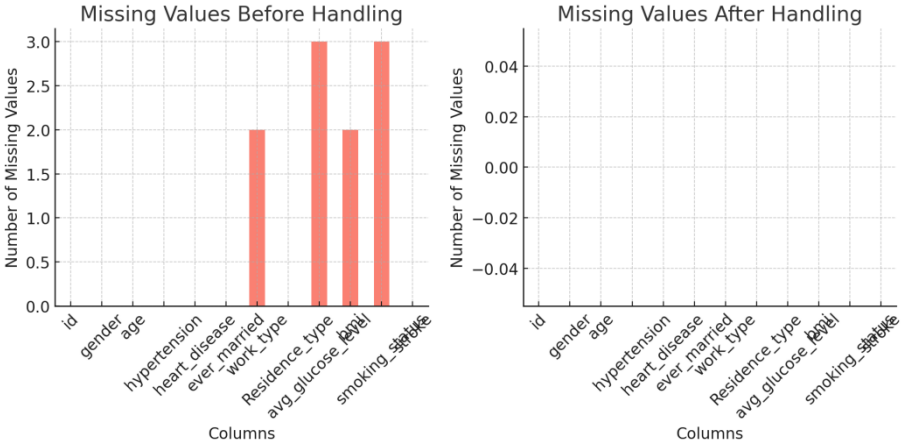

Fig.1. Missing data before and after handling (Photo/Picture credit: Original)

2.1.2 Converting Categorical to Numerical

Handling categorical data appropriately is essential when preparing data for machine learning models. To transform these features into a numerical format that models can interpret, label encoding is applied to binary categories here, such as gender, ever_married, and Residence_type. Male' is encoded as 1, and 'Female' as 0; 'Yes' is mapped to 1 for ever_married, and 'Urban' is encoded as 1 for Residence_type. One-hot encoding is used for categorical features like work_type and smoking_status that have more than two unique values. Each potential category is given its own distinct binary column thanks to this encoding, which enables the model to handle each category separately.

2.1.3 Exploratory Data Analysis: Correlation Analysis

Exploring relationships between features in the dataset is crucial for understanding which factors are associated with the target variable (stroke). A correlation matrix helps identify such relationships by calculating the correlation coefficients between all pairs of features. Fig.2 reveals key relationships among features in the stroke dataset. Age shows a moderate positive correlation (0.16) with stroke, aligning with the increased stroke risk in older individuals. A strong positive correlation (0.83) exists between hypertension and average glucose level, indicating a link between cardiovascular and metabolic health. However,

correlations between stroke and other individual factors, such as hypertension (0.09) and heart disease (0.05), are weaker, suggesting these alone may not be strong predictors, which highlights the significance of other factors such as cardiovascular and metabolic health in stroke prediction [6, 7]. The negative correlation between smoking status (never smoked) and stroke (-0.22) confirms that non-smokers are at lower risk. Overall, while individual correlations are mostly low, meaningful insights emerge, emphasizing the need for multi-variable models to accurately predict stroke risk.

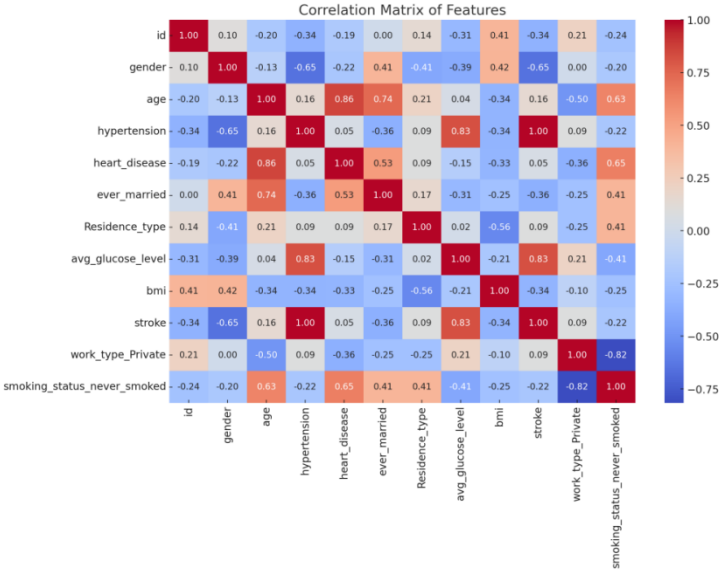

Fig.2. Correlation Matrix of Features (Photo/Picture credit: Original)

2.1.4 Feature Distributions

Visualizing the distribution of key features helps to understand the data's behaviour and identify any imbalances or unusual patterns. The target variable (stroke) distribution indicates a significant class imbalance, with a majority of patients not having a stroke, which might require balancing techniques during modelling.

Fig.3 shows the distributions of key features: age, BMI, average glucose level, and stroke occurrence. In all age groups, the distribution of ages is fairly consistent. The BMI distribution peaks around 28, indicating that most individuals fall into the overweight category. Average glucose level has a right-skewed distribution, with a significant number of individuals having levels between 70 and 100, but with a long tail extending toward higher values. The stroke occurrence plot reveals that most individuals did not experience a stroke, suggesting a class imbalance that could have an impact on how well machine learning models work. Fig.4 further emphasizes the imbalance in the dataset, showing the count of stroke and non-stroke cases. The bar chart reveals a much larger count of non-stroke cases (coded as 0) compared to stroke cases (coded as 1). This imbalance poses a challenge for model training, as the high number of non-stroke cases could bias the model toward predicting the negative class. Handling this imbalance will be essential for building effective predictive models.

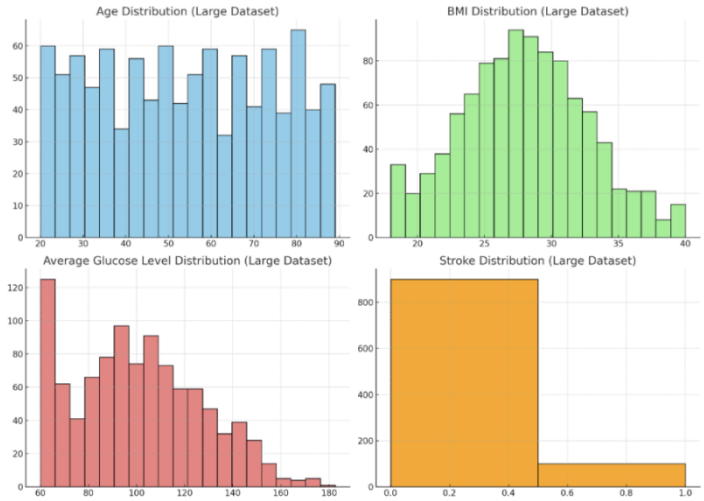

Fig.3. Feature distributions (Photo/Picture credit: Original)

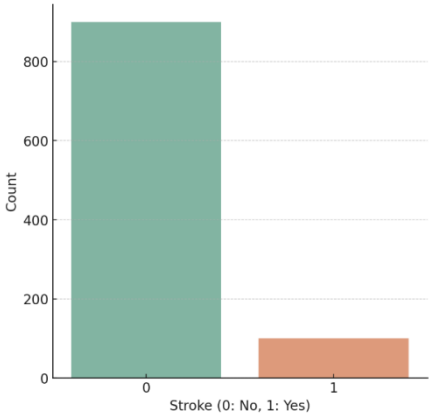

Fig.4. Distribution of stroke cases (Photo/Picture credit: Original)

80% of the data in this study has been chosen for training, and the remaining 20% is allocated for testing. The numerical data ((e.g., age, BMI, avg_glucose_level) will also be normalized such that their means are 0 and their standard deviations are 1. This will assist the majority of machine learning models converge more quickly and perform better.

2.2 Introduction to the Method and Models

2.2.1 Random Forest

Random Forest is a ML method that builds several decision trees during training and outputs the majority vote (for classification) or average prediction (for regression). It improves accuracy and generalization by combining the output of many trees (Fig.5). It can also successfully handle missing data and perform well with both numerical and category features.

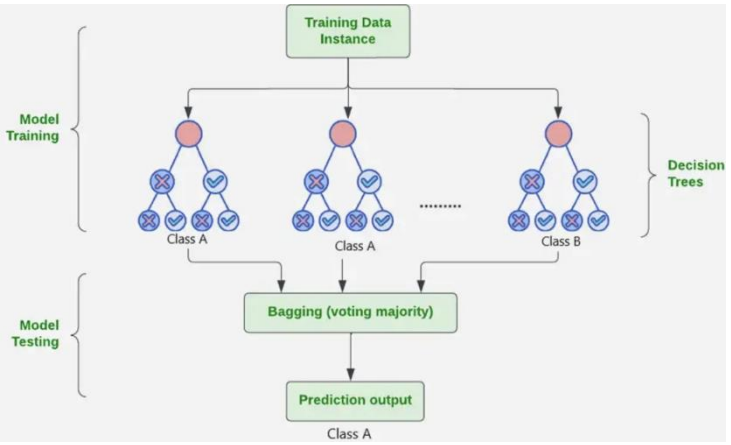

Fig.5. Random Forest Algorithm in Machine Learning (Photo/Picture credit: Original)

2.2.2 Support Vector Machine (SVM)

Used for regression and classification, SVM is a supervised learning model. It operates by determining the best hyperplane with the biggest margin that divides data points of several classes. SVM works well when there is a distinct margin of separation and is efficient in high-dimensional spaces. It's also sensitive to feature scaling, which is why data normalization is often necessary (Fig.6).

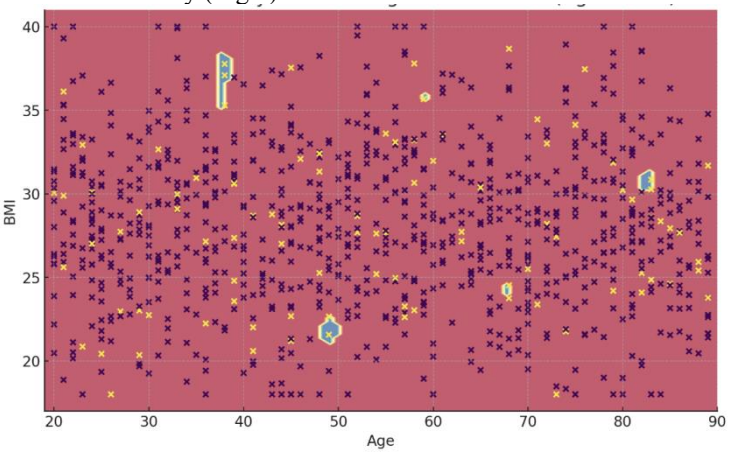

Fig.6. Decision Boundary of SVM using Stroke Dataset(Age Vs BMI) (Photo/Picture credit: Original)

2.2.3 Assessment Method

Precision measures how many accurately anticipated positive observations there were out of all the predicted positive observations. It tells us how much the paper can trust the model's positive predictions.

$$Precision = \frac{True\ Positives}{\text{True Positives + False Positives}} \quad (1)$$

Recall is the ratio of correctly predicted positive observations to all the actual positives. It shows how well the model can capture all the true cases.

$$Recall = \frac{\text{True Positives}}{\text{True Positives + False Negatives}} \quad (2)$$

The confusion matrix is a table that displays the counts of true positive, true negative, false positive, and false negative predictions to describe a model's performance.

The Receiver Operating Characteristic (ROC) curve compares the true positive rate (recall) to the false positive rate. The Area Under the Curve (AUC) measures the model's ability to differentiate across classes.

3 Results and discussion

3.1 Evaluation and Results

3.1.1 Precision-Recall Curves

Fig.7 shows the precision-recall curves of those two models. For Random Forest, precision remains relatively high when recall is very low, suggesting fewer false positives but struggling to detect all true positives [8]. The SVM model, performs poorly when handling nonlinear boundaries with complex features, especially for this type of medical data, possibly due to high correlations between features, making it difficult for the model to learn effectively [9,10].

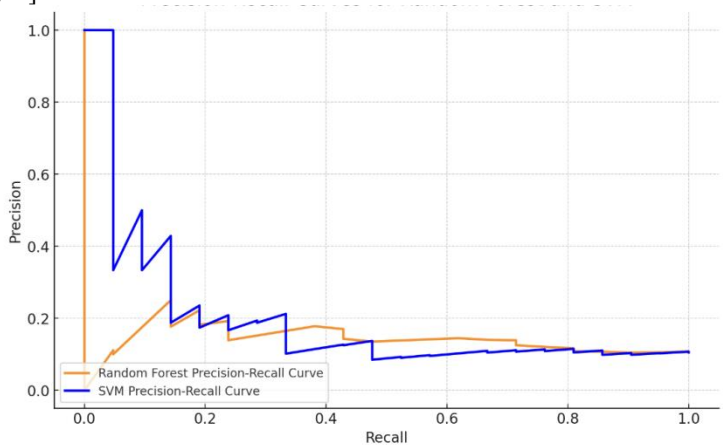

Fig.7. Precision-Recall Curves (Photo/Picture credit: Original)

3.1.2 Confusion Matrices

Fig.8 shows the confusion matrices of both models. Both models failed to correctly detect any positive stroke cases, resulting in 21 stroke patients being misclassified as non-stroke (false negatives), which leads to an excessively high miss rate in medical diagnosis.

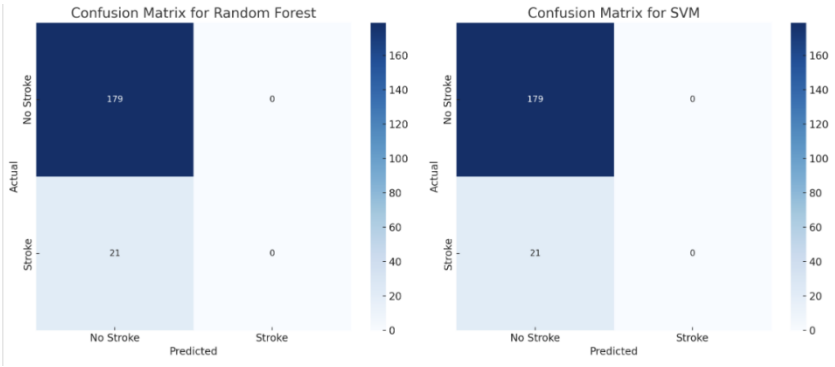

Fig.8. Confusion Matrices (Photo/Picture credit: Original)

3.1.3 ROC Curves

The Random Forest model attained an AUC of 0.60, while the SVM achieved an AUC of 0.53, suggesting that SVM struggles more compared to Random Forest, with performance not significantly better than random guessing (Fig.9). The visual comparison between the ROC curves makes it evident that Random Forest is slightly more capable of identifying and classifying patients according to their risk of stroke. The higher curve and larger AUC area for Random Forest also indicate better model performance in discriminative power, making it a preferable model at this stage.

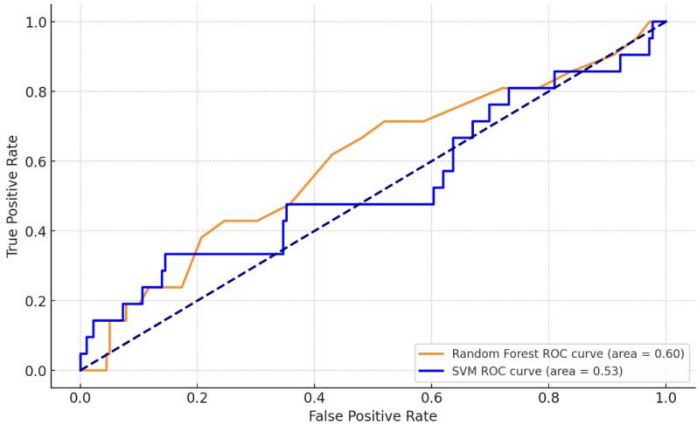

Fig.9. ROC Curves (Photo/Picture credit: Original)

3.2 Results Analysis

Random Forest demonstrates a slightly better discriminative ability, as indicated by a higher AUC in the ROC curve and more consistent precision-recall behaviour. However, the model's inability to identify a significant number of stroke cases (high false negatives) might limit its reliability in medical applications. SVM struggles further, with a lower AUC and fluctuating precision-recall performance, indicating that it cannot confidently distinguish between the classes, especially in cases where the decision boundary is unclear. Although the random forest model performs well in overall discriminative ability, its high false-negative rate in medical applications can result in missed diagnoses, potentially

impacting patient health directly [7]. This indicates that further optimization is required in clinical practice to reduce the risk of missed diagnoses [9,10].

3.3 Future Research Directions

The paper should explore neural networks or deep learning approaches for larger datasets, as they can effectively learn complex, non-linear relationships that might be missed by traditional models. Specifically, Convolutional Neural Networks could be applied if health data includes imaging, such as brain scans, given their strength in feature extraction from spatial patterns. Additionally, Recurrent Neural Networks or Long Short-Term Memory networks are ideal for capturing temporal dependencies in sequential medical data, such as patient histories over time. Deploying these models in a clinical setting and validating predictions with real patient data will be essential. I will also use metrics like precision-recall curves and confusion matrices to continuously monitor false positives and negatives, ensuring the models are optimized over time.

4 Conclusion

This study investigated the application of Random Forest and SVM models for stroke prediction using a medical dataset. Preprocessing steps such as imputation, one-hot encoding, and class balancing were employed to prepare the data for analysis. Results indicated that while Random Forest achieved better performance with an AUC of 0.60 compared to SVM's 0.53, both models faced challenges in detecting positive stroke cases, resulting in high false-negative rates.

This finding suggests that while machine learning holds promise, the models need further improvement to be reliably deployed in clinical settings. Future research could explore more advanced algorithms such as neural networks, which might capture non-linear relationships more effectively. Additionally, validating these models on real-world clinical data will be essential to ensure their robustness in practice.

References

1. E.M. Alanazi, A. Abdou, J. Luo, Predicting risk of stroke from lab tests using machine learning algorithms: Development and evaluation of prediction models. JMIR Formative Research. **5**, e23440 (2021)
2. M. Shiozawa, H. Kaneko, H. Itoh, K. Morita, A. Okada, S. Matsuoka, H. Kiriyama, T. Kamon, K. Fujiu, N. Michihata, T. Jo, N. Takeda, H. Morita, S. Nakamura, K. Node, H. Yasunaga, I. Komuro, Association of body mass index with ischemic and hemorrhagic stroke. Nutrients. **13**, 2343-2355 (2021)
3. T. Vu, Y. Kokubo, M. Inoue, M. Yamamoto, A. Mohsen, A. Martin-Morales, T. Inoué, R. Dawadi, M. Araki, Machine learning approaches for stroke risk prediction: Findings from the Suita Study. J. Cardiovasc. Dev. Dis. **11**, 207-215 (2024)
4. A. Hassan, S.G. Ahmad, E. Ullah Munir, I.A. Khan, N. Ramzan, Predictive modelling and identification of key risk factors for stroke using machine learning. Sci. Rep. **14**, 11498-11512 (2024)
5. E. Dritsas, M. Trigka, Stroke risk prediction with machine learning techniques. Sensors. **22**, 4670-4689 (2022)

6. H. Du, X. Liu, M. Xu, X. Yang, H. Zhang, J. Mo, Y. Lu, J. Kuang, Advances in prognostic prediction research of acute ischemic stroke: A case study of machine learning models. Chin. Gen. Pract. **27**, 1456-1467 (2024)
7. Z. Wang, Application of machine learning methods in stroke risk prediction. Master Thesis, Guangzhou University. 45-78 (2023)
8. Z. Guo, Q. Liu, F. Liu, C. Wang, X. Ruan, Research on an early prediction model for stroke based on machine learning algorithms. Comput. Digit. Eng. **11**, 2180-2183+2247 (2021)
9. S. Mainali, M.E. Darsie, K.S. Smetana, Machine learning in action: Stroke diagnosis and outcome prediction. Front. Neurol. **12**, 443-460 (2021)
10. S. Gangavarapu, L.A.K. Gorli, Analyzing the performance of stroke prediction using ML classification algorithms. Int. J. Adv. Comput. Sci. Appl. **12**, 91-98 (2021)