Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 7
TPCEE 2024

Research Status of Machine Learning in Stroke Treatment

Jiafeng Chen*
Maynooth International Engineering College, Fuzhou University, Fuzhou, China
*Corresponding Author: 832103114@fzu.edu.cn

Abstract. Stroke, characterized by high morbidity, mortality and disability, seriously jeopardizes human health. Timely and accurate prediction and diagnosis of the disease is of great significance. Machine learning can efficiently analyze large-scale datasets and shows great potential for application at various stages in the treatment of stroke disease. In this paper, we start from the common data processing, feature selection, model training and evaluation of the conventional application process of machine learning. This paper analyzes the application of Random Forest, Support Vector Machine, and Deep Learning, which are commonly used machine learning models, in the diagnosis, treatment, and prognosis of stroke. This paper summarizes their advantages over traditional statistical methods and the current shortcomings and points out the appropriate methods for improvement. In addition, this paper also looks forward to the future application of machine learning in the field of stroke prediction. It explores the possible path of machine learning technology in the future development of the fight against stroke. This study aims to promote the progress of machine learning in the field of stroke research by integrating and comparing related research approaches, and so on.

Keywords: Machine Learning; Random Forest; Support Vector Machine; Deep Learning.

1. Introduction

Stroke, also known as cerebrovascular accident (CVA) or brain attack, is a severe disease caused by a sudden interruption of the blood supply to the brain or rupture of a blood vessel. It is commonly seen in the elderly population, but also occurs in the younger population. The occurrence of stroke can cause serious consequences such as paralysis of the body, speech disorders, cognitive dysfunction, etc., which seriously affects the quality of life of the patients and even threatens the safety of their lives. According to statistics from the World Health Organization, stroke is one of the leading causes of disability and death in adults, causing a huge economic burden on society and families. Therefore, timely prediction and intervention for stroke is significant.

Traditional stroke prediction methods are mainly based on clinical indicators as well as epidemiological data. Still, these methods are limited by the limitations of statistical analysis methods, which usually fail to consider the specificity and diversity of stroke influencing factors fully. With the development of artificial intelligence and machine learning technologies, these advanced computer techniques have opened up new opportunities for the medical field. Using machine learning algorithms to process large-scale datasets can help healthcare professionals better understand disease risk factors, predict disease occurrence, and develop personalized treatment plans. Machine learning, as a powerful data-driven approach, is able to build prediction models based on a large number of data features and medical records to enable personalized stroke risk assessment and prediction. Particularly in stroke prediction, by using machine learning algorithms to process patients' clinical data, physiological parameters, and imaging features, it is possible to improve the accuracy and real-time performance of prediction. Mario Daidone et al.'s study on the application of machine learning algorithms such as in various aspects of stroke diagnosis, treatment and prognosis has also demonstrated the significant advantages they offer [1].

And machine learning, as an essential branch of artificial intelligence, has many different methods, such as Random Forest (RF), Support Vector Machine (SVM), Deep Learning (DL), Logistic Regression (LR), Artificial Neural Networks (ANN) and so on. The purpose of this paper is to explore the application of different types of commonly used machine learning algorithms in the fields of stroke prediction and analysis, and to analyze their advantages over traditional statistical methods as well as their limitations. By combining the corresponding machine learning research situation and results, the study is discussed and summarized to assess the effectiveness of machine learning in stroke prediction and other applications, and to put forward the outlook and suggestions for future research.

It is hoped to provide more accurate and personalized medical prediction tools for stroke patients, better decision support for healthcare professionals, scientific basis for stroke prevention and intervention, and reduce the impact of stroke on society. Machine learning, as a powerful tool, will bring new opportunities and challenges to the development and progress of the medical field.

2. Challenges in Stroke Prediction

Stroke, as an acute cerebrovascular disease, is a group of diseases that cause subsequent structural damage to brain tissue due to the sudden rupture of a blood vessel in the brain or the blockage of a blood vessel that prevents blood from flowing normally to the brain. It is mainly classified into two major types, including ischemic and hemorrhagic strokes. Stroke prediction is a challenging program that involves the patient's own health status as well as complex environmental factors.

The influencing factors involved in the development of stroke are complex and diverse, mainly involving congenital genetic factors and acquired influencing factors, of which the congenital genetic factors are mainly from individual genetic defects [2]. In contrast, the acquired influencing elements are related to the environment, age, dietary habits, lifestyle, etc., which are intertwined with each other, and the influence of the complexity of the impact of these factors. Congenital genetic factors are difficult to artificially intervene and adjust because they are mainly caused by DNA, including heredity and gene mutation. Acquired factors, however, are complex and variable and difficult to predict using specific data, as they come mainly from the individual's own situation and the environment, especially environmental factors.

However, in the current study, there are certain difficulties in obtaining the data cited and quality problems, such as the data set is too outdated to be used in the current predictive analysis program, or some of the main data are missing or incomplete, and there are imbalances in the data and other problems. Traditional predictive research is mainly analyzed through statistical models such as multivariate regression models and expert experience, but stroke prediction is a project with a wide variety of characteristic factors and a large amount of data. Its complex interactions make it difficult for traditional models to consider these factors in an all-round way, and they cannot effectively deal with the complex correlations and non-linear relationships between the data to carry out accurate simulations and assessments, resulting in poor prediction accuracy.

Machine learning methods have many advantages over traditional prediction models, and current research has confirmed that the use of machine learning methods for prediction and prognosis can achieve more accurate results.

3. Advantages of Machine Learning in Stroke Prediction

3.1. Overview of Machine Learning

Machine learning is a multidisciplinary subject first introduced by Samuel in 1959. In a broad sense, machine learning is an approach that gives a machine the ability to learn so that it can perform functions that cannot be accomplished by direct programming. In a practical sense, however, machine learning is an approach that utilizes existing datasets, etc., to train specific models that can be used to predict the desired outcome for a given set of data.

Machine learning, as an important branch in the field of artificial intelligence, has many unique advantages. Its advantages include learning patterns and regularities from data autonomously, automatically optimizing model performance based on new data, and handling large-scale data. In stroke machine learning algorithms research mainly focuses on Random Forest (RF), Support Vector Machine (SVM), Deep Learning (DL), Artificial Neural Network (ANN), etc., and is widely used in all stages of stroke diagnosis and classification diagnosis, prognosis, and treatment. These algorithms have high diagnostic accuracy by extracting and analyzing large-scale data for stroke diagnosis.

The stroke prediction problem is a typical binary prediction problem characterized by many influencing factors, various data types, large data volume, etc., which is suitable for machine learning methods to solve. Compared with traditional prediction models, machine learning-related prediction models often show higher accuracy, which can help healthcare professionals make a more accurate clinical diagnosis and strive for better treatment opportunities and time for patients.

3.2. Machine Learning Application Process for Stroke

In stroke-related research, the application process of machine learning models mainly includes the following steps: data set collection and preprocessing, feature selection, model training and evaluation.

First of all, the collection and preprocessing of the dataset need to take into account the type of data, the categorization of the data, and the special values, etc., and need to timely remove the outliers and other undesirable values that affect the model results to ensure the quality and reliability of the data. In order to ensure the integrity of the data and facilitate the intervention and learning of the later algorithms, it is necessary to fill the missing data with zeros and standardize the data content into numerical values. Methods such as Synthetic Minority Over-Sampling Technique (SMOTE) or Automatic Hyperparameter Optimization (AutoHPO) are used to do a good job of handling the imbalance of the data so as not to affect the model training in the later stage negatively.

Secondly, feature selection is a crucial part of machine learning, which aims to eliminate irrelevant factors from large amounts of data and extract key features that have an impact on stroke prediction, and needs to take into account whether the type of data has a real impact on the disease or not. Principal Component Analysis (PCA) can be used to go through the dimensionality reduction process on the original feature matrix, or to calculate the correlation between the features and the target variables, by using methods such as Filter model, Wrapper model, Embedded model, and so on, by evaluating the impact of the feature subset on the performance of the classifier, and so on. In turn, the most appropriate feature set is selected.

Finally, the model training and evaluation phase involves making full use of the processed dataset and selecting machine learning methods to train the model, as well as evaluating and attempting to optimize the model results by means of suitable evaluation methods. One of the more widely used methods is the cross-validation method, which involves dividing the dataset into training and validation sets and evaluating the model multiplefold to obtain more reliable performance metrics. This method can effectively limit overfitting.

3.3. Application Analysis of Common Machine Learning Methods

There are many research cases on machine learning in stroke prediction. These studies analyze a large amount of stroke patient data, apply different machine learning algorithms to construct prediction models, etc., and have achieved certain results and effects. Different machine learning algorithms have different advantages and limitations in stroke prediction, and researchers need to choose appropriate algorithms for experimentation and application according to specific situations to improve prediction accuracy and efficiency.

3.3.1. Random Forests

In general, RF is a further refinement of the Classification and Regression Tree (CART) model, which means that RF is an integrated algorithm of a CART model [3]. A single CART model usually does not meet the task of analyzing complex data using many interacting predictors. However, a collection of many such trees using different sets of predictors may produce very good predictions when aggregated. RF, developed by Brayman, use randomization to create many decision trees whose outputs are aggregated into a single output by voting on a classification problem or averaging on a regression problem. The biggest advantage of random forests is its ability to find meaningful interactions and nonlinear effects of predictors, the nature of the interactions or nonlinearities do not need to be specified in advance as in Cox models or other parametric survival models and are mostly used for classifying the factors influencing a problem, etc [4].

Desy Ika Puspitasari et al. used their study's decision tree and random forest methods to analyze and categorize stroke disease. The study used a stroke prediction dataset from the Kaggle database, which includes attributes such as gender, age, various diseases, and smoking status, and used decision trees and RF to model the associations between stroke risk factors and the occurrence of stroke, and to predict the occurrence of stroke. The decision tree method achieved 95.90% accuracy, while the RF, which is an integrated algorithm of the decision tree, achieved a high accuracy of 98.90%. Predicting and categorizing stroke risk factors is of vital importance, which can help in early prevention and treatment of stroke [5].

RF also have self-generated limitations. First, RF is only suitable for tabular data, not for image, audio and text data, where neural networks are more appropriate. Second, compared to neural networks, RF taper off as the input data increases. In addition, RF have difficulty capturing endogenous time effects, which need to be addressed by eliminating time effects or using Autoregressive Integrated Moving Average (ARIMA). Finally, RF is weak in dealing with imbalance in the dataset, and in compensating for this, methods such as Synthetic Minority Oversampling Technique (SMOTE), Balanced Random Forest (BRF) or Weighted Random Forest (WRF) can be used to cope with imbalanced data [6].

3.3.2. Support Vector Machine

SVM is a binary classification model used for data classification and regression analysis, and its basic idea is to map the data points in the high-dimensional space to the low-dimensional space, and maximize the classification interval between the two types of data by searching for an optimal separation hyperplane. This machine learning method is based on statistical learning theory, VC dimension theory and structural risk minimization principle [7]. By determining the decision function directly from the training data using SVM, the existing separation (boundaries) between decision boundaries can be made to be maximized in a high-dimensional space called feature space. This classification strategy minimizes the classification error of the data after it passes through the training and allows for better generalization, especially when the number of input data is small. SVMs are a powerful technique for data classification and regression analysis. A significant advantage is that they obtain a subset of support vectors during the learning phase, which is usually a small part of the original dataset. This set of support vectors represents a given classification task and consists of a small dataset.

The Jovel T. Rosado and Alexander A. Hernandez study used a dataset of 1,500 patients from Cavite, Philippines, and principal component analysis was used to downscale the dataset while ensuring data quality. The model was evaluated using assessments such as accuracy, precision, recall, f1 score and area under the curve. In this study, 60% of the dataset content was used for SVM model training, 30% for model testing, and the last 10% was used to validate the model.Finally the SVM model was found to have 99% accuracy, 98.89% testing accuracy and 97.33% validation accuracy after training using the data [8].

SVM consider both empirical risk and confidence intervals to minimize the actual risk and have good model generalization performance, but lack transparency of results. Since SVMs may produce high-dimensional features, they are difficult to represent in the form of parameterized functions. In addition, SVM is only suitable for binary classification problems; it is a binary classification algorithm, and multi-class problems need to be classified in pairs, which increases the complexity of the algorithm [9].

3.3.3. Deep Learning

DL is the most efficient, supervised, and time and cost effective machine learning method for learning descriptive and circumscribed features in a very hierarchical way. It is not a restricted learning method per se, but it follows various procedures and its high adaptability can be applied to a large number of multivariate problems to learn illustrative and circumscribed features in a very hierarchical way[10] .

After implementing and comparing different DL models (LSTM, bidirectional LSTM, CNN-LSTM, CNN-bidirectional LSTM), Yoon-A Choi et al. found that one of the CNN-bidirectional LSTM models achieved a final result of 94.0% accuracy, the highest among the above models, in stroke prediction using raw EEG data. Experimental ResultsThe system uses non-invasive electroencephalography (EEG) data as an input and can easily measure brainwaves individually in daily life, thus allowing real-time prediction and monitoring of stroke disorders in daily life. This helps to detect the risk of stroke in a timely manner before the patient is transported to the emergency room compared to other measurement techniques, and is expected to significantly improve early stroke detection, reduce costs and discomfort, and thus allow access to treatment within the golden period of salvation [11].

DL models usually have many parameters that fit the training data well. However, the features are automatically learned during the training process and are highly sensitive to the data, so if there are noisy or abnormal samples in the training data, the model can easily learn the features of these noisy or abnormal samples as well, leading to overfitting. To alleviate these problems, some regularization techniques, such as L1/L2 regularization, dropout, data augmentation, etc., are needed to limit the complexity of the model and reduce the risk of overfitting [11].

4. Recommendations and Outlook

Machine learning techniques have demonstrated excellent performance in stroke prediction, prognosis, diagnosis, and treatment. Machine learning algorithms such as deep learning, support vector machines, and random forests have significantly improved the accuracy of diagnosis and classification compared to traditional statistical prediction models in stroke-related fields. Different types of machine learning algorithms can extract different stroke subtype features based on input data, with higher classification accuracy and precision. For example, machine learning algorithms based on DWI and FLAIR imaging can be used to automatically identify the onset time of stroke to provide a basis for intravenous thrombolysis, to predict stroke severity and symptom onset, and to identify risk factors to provide a basis for clinical decision-making [12].

In addition, some machine learning algorithms are used in the automatic detection and diagnosis of intracranial hemorrhage, fracture, midline shift and other abnormalities in CT scan images, and their outstanding image analysis capabilities can help medical personnel to better diagnose and confirm stroke subtypes. Machine learning can help diagnose acute ischemic stroke earlier and more accurately. By analyzing a large amount of medical data, such as brain imaging scans, medical history, and clinical symptoms, machine learning algorithms can discover stroke's characteristic patterns, thus improving stroke diagnosis [13].

Machine learning has similarly demonstrated superior accuracy and great optimization potential in image analysis, diagnosing stroke subtypes, risk stratification, guiding medical treatment, and predicting patient prognosis. Furthermore, future research still needs to address issues such as standardized data collection, standardized model validation criteria, and ethical considerations regarding privacy and bias. Data imbalance can be discussed further and more sophisticated algorithms such as integrated learning can be explored to improve the accuracy and generalization of model predictions. For better promotion and application of relevant models among healthcare professionals with non-computer related specialties, combining machine learning models with web applications or Flask frameworks to provide more convenient prediction services for the healthcare field is also a development direction that deserves attention [13].

To summarize, machine learning algorithms have shown better application prospects in stroke diagnosis, but further promotion of medical data sharing, strengthening of clinical validation, and promotion of medical personnel's learning of related knowledge are still needed to promote the effective application of machine learning in clinical practice. Future research should focus on data quality, model optimization, and practicality in order to develop more accurate and reliable stroke prediction systems.

5. Conlusion

This paper summarizes the application examples of three widely used machine learning models, namely random forests, support vector machines and deep learning, in the fields of stroke prediction, treatment and prognosis, as well as their performances compared with traditional models, affirming their technical advantages and pointing out their current limitations. This paper concludes that the application of machine learning algorithms in the field of stroke research, various types of machine learning models are widely used in the prediction of stroke and stroke subtypes, classification and diagnosis, and post-treatment, as well as the identification of risk factors and other fields, and can have great potential for application and optimization.

This paper summarizes three widely used machine learning models, namely Random Forest, Support Vector Machines, and Deep Learning, in the fields of stroke prediction, treatment, and prognosis, as well as their performance compared to traditional statistical models. By analyzing and comparing these examples, this paper affirms the advantages of machine learning technology and at the same time points out its current limitations as well as related methods that can be applied to optimize the results. The study shows that by applying machine learning algorithms to stroke research, various types of models show great application advantages and optimization potential in predicting stroke and stroke subtypes, classification and diagnosis, prognostic treatment and risk factor identification. This application pathway has undoubtedly greatly promoted the research process in the field of stroke prediction and treatment, and can help clinicians better identify the cause of disease and formulate the treatment plan in the actual diagnosis and treatment.

However, one of the biggest obstacles to applying machine learning in the field of stroke is its dependence on computer languages such as Python and the complexity of the algorithms. This makes it difficult for healthcare professionals without knowledge of computer programming to understand and apply it in the actual treatment process, and indirectly leads to incomplete and outdated stroke-related datasets. In the further exploration of machine learning in the future, in addition to further solving the imbalance of data sources and improving the accuracy and generalization of the model, combining the model into a more convenient application can not only collect more standardized, more timely and more robust data sets for subsequent research, but also provide convenient services for medical personnel. In the future, the results and accuracy of machine learning in stroke diagnosis will certainly reach new heights.

Reference

[1] Daidone, Mario, Sergio Ferrantelli, and Antonino Tuttolomondo. "Machine learning applications in stroke medicine: Advancements, challenges, and future prospectives." Neural Regeneration Research 19.4 (2024): 769-773.

[2] Francis, J., Raghunathan, S., & Khanna, P. (2007). The role of genetics in stroke. Postgraduate medical journal, 83(983), 590-595.

[3] Shobayo, O., Zachariah, O., Odusami, M. O., & Ogunleye, B. (2023). Prediction of stroke disease with demographic and behavioural data using random forest algorithm. Analytics, 2(3), 604-617.

[4] Rigatti, S. J. (2017). Random forest. Journal of Insurance Medicine, 47(1), 31-39.

[5] Puspitasari, D. I., Dharmawati, A., Rosadi, M. E., & Dhuhita, W. M. P. (2021, November). Stroke disease analysis and classification using decision tree and random forest methods. In 2021 Sixth International Conference on Informatics and Computing (ICIC) (pp. 1-4). IEEE.

[6] Zhu, T. (2020, August). Analysis on the applicability of the random forest. In Journal of Physics: Conference Series (Vol. 1607, No. 1, p. 012123). IOP Publishing.

[7] Chang, W., Liu, Y., Xu, X., & Zhou, S. (2019, October). Prediction model of stroke recurrence based on support vector machine. In Journal of Physics: Conference Series (Vol. 1324, No. 1, p. 012095). IOP Publishing.

[8] Rosado, J. T., & Hernandez, A. A. (2019, October). Developing a predictive model of stroke using support vector machine. In 2019 IEEE 13th International Conference on Telecommunication Systems, Services, and Applications (TSSA) (pp. 35-40). IEEE.

[9] Karamizadeh, S., Abdullah, S. M., Halimi, M., Shayan, J., & javad Rajabi, M. (2014, September). Advantage and drawback of support vector machine functionality. In 2014 international conference on computer, communications, and control technology (I4CT) (pp. 63-65). IEEE.

[10] Dargan, S., Kumar, M., Ayyagari, M. R., & Kumar, G. (2020). A survey of deep learning and its applications: a new paradigm to machine learning. Archives of Computational Methods in Engineering, 27, 1071-1092.

[11] Choi, Y. A., Park, S. J., Jun, J. A., Pyo, C. S., Cho, K. H., Lee, H. S., & Yu, J. H. (2021). Deep learning-based stroke disease prediction system using real-time bio signals. Sensors, 21(13), 4269.

[12] Chang, S. H. U., Da, L. U. O., & Le, G. E. (2023). Advances in machine learning for the diagnosis and treatment of stroke. Chinese Journal of Contemporary Neurology & Neurosurgery, 23(1), 5.

[13] Sailasya, G., & Kumari, G. L. A. (2021). Analyzing the performance of stroke prediction using ML classification algorithms. International Journal of Advanced Computer Science and Applications, 12(6).