Review on Machine Learning Approaches used for Stroke Prediction

Ms. Gagana M [1], Dr. Padma M C [2]

[1]*PG Student, Final year Computer Science and Engineering, PES College of Engineering, Mandya, Karnataka, India.*

[2]*Professor and Head of the Department, Computer Science and Engineering, PES College of Engineering, Mandya, Karnataka, India.*

---***---

Abstract - *Machine Learning is a type of AI that enables to predict outcomes without being explicitly programming the applications. This review aims to identify and analyze the Machine Learning techniques used for Stroke Prediction. We considered the previously published works to review the Machine Learning techniques used for Stroke Predictions. It's been found that the most research was at most on mortality rate and functional outcome as the predicted outcomes. The most commonly used techniques were random forest, support vector machines, decision trees and neural networks. However, a few models did basic reporting standards for clinical sector tools and none of them were available in a way which could be used in real time practically.*

***Key Words*: Stroke prediction, Machine learning approaches, Sensitivity and Specificity, Comparison Analysis.**

1. INTRODUCTION

According to the study of WHO, 15 million people suffer stroke world-wide every year. Among these 5 million die and another 5 million are permanently disabled. Europe has an average of 650,000 stroke deaths every year approximately. Many population based studies on stroke disorder have been conducted since the past decades to determine the types, prevalence, incidence and case fatality rates.

It is very difficult to predict the stroke symptoms and outbreaks taking note on the risk factors, since stroke is a complicated medical condition. This has enhanced the interests of people in technology sector to apply machine learning techniques to diagnose the stroke effectively by routinely collecting the datasets and delivering the accurate results for diagnose. Furthermore, many papers have been published frequently which explains machine learning techniques to address the issue. But there have been no reviews of studies on machine learning techniques used for stroke prediction.

The agenda of this review is to identify the better machine learning techniques used to predict stroke, which will also help to understand and resolve the problem in more effective ways.

2. RELATED MACHINE LEARNING APPROACHES

In this section, analysis and review is being done on the previously published papers related to work on prediction of stroke disease using different machine learning approaches and algorithms. At least, papers from the past decade have been considered for the review. They are explained below:

In 2010, Adithya Khosla and his team, proposed a novel automatic feature selection algorithm that selects robust features based on proposed heuristic; conservative mean. They combined it with SVM support vector machines, to achieve a greater area under the ROC Curve (AUC). In this study, they compared the cox proportional hazards model with a machine learning approach for stroke prediction on the cardiovascular Health Study (CHS) dataset. Further-more, they presented a margin-based censored regression algorithm that combined the margin-based classifiers with censored regression to attain a finer concordance index than the cox model. All- inclusively, this method outperformed the current state-of-the-art in both metrics of AUC and concordance index. This approach can be applied to clinical prediction of other diseases, where missing data are common & risk factors are not well understood. However, they realised that this feature selection algorithm may not work well in other datasets with highly correlated features as it evaluated the performance of each feature individually. To overcome this issue, they used an L1 regularised feature selection algorithm, to prune the features before applying conservative mean features selection for fine-tuning [10].

In 2012, Sudha. A under the guidance of her professors N.Jaisankar & P Gayathra, proposed a stroke predictive Model using classification techniques. They used classification algorithms-decision Tree, Naive Bayes & Neural Networks for predicting the stroke with related attributes. They utilised principle component analysis algorithm for dimension reduction. They studied & used sensitivity 7 accuracy indicators for evaluation. Decision tree achieved 95.29% of sensitivity & 98.01% of accuracy. Bayesian classifier achieved 8710% & 91.30% respectively. They compared these techniques and chose the decision tree as the best classification Method. The proposed Model takes the patient detail, & checks with reduced attributes (Glucose level, Blood pressure level, family history etc.,) and found out whether patient had stroke disease or not. The accuracy measured based on sensitivity & specificity. It was observed that Neural networks performance had more accuracy when compared with other two classification techniques [9].

In 2013, Esfan Al-Zahra and Mashhad Ghaem hospitals conducted the study during 2010-2011. They collected data on 807 healthy & sick subjects using a standard checklist that contains 50 risk factors for stroke such as history of cardiovascular disease, diabetes, hyperlipidemia, smoking and alcohol consumption. They used data mining techniques, k-nearest neighbour and C4.5 decision tree using WEKA tool to analyse the data. In this stud, 50 risk factors such as age, gender, sleep duration, hours of activity, hypertension, hyperlipidemia, smoking, alcohol, narcotics, stimulants and other risk factors that have not been considered previously, were extracted the C4.5 & k-nearest algorithms in WEKA 3.6 were used to analyse stroke data. Finally the best results were those pertaining to the C4.5 algorithm that outran the k-nearest neighbour algorithm with respect to accuracy, precision & specificity criteria by a small difference. So, the decision tree was selected as the stroke-predicting algorithm because it showed higher accuracy [8].

In 2014, Hamed Asadi, Richard Dowling, Bernard Yan, Peter Mitchell, conducted a look back study on a potential database of acute ischemic stroke. They compared different machine learning techniques, capable of predicting the outcome of cardiovascular intervention in acute anterior circulation ischemic stroke. They appended 107 conservative acute anterior circulation Ischemic stroke patients treated by cardiovascular technique. All the available information of demographic, procedural & clinical factors was included in the model. They used SPSS, MATLAB, Rapid miner, classical statistics as well as artificial neural network & support vector algorithms to design a supervised machine capable of classifying these predictors into potential outcomes & poor outcomes. Despite a small dataset used, there was promising accuracy, approaching 70% of predicting outcome, using supervised machine learning. They proposed a robust machine learning system that can potentially optimise the selection process for endovascular versus medical treatment in the management of acute stroke [7].

In 2015, Balar Khalid & Naji Abdelwahab, proposed a model for predicting Ischemic stroke using Data mining Techniques which were classification, logistic regression. They studied the risk factors of ischemic stroke. Then they used data software WEKA 3.6 & C4.5 algorithm & logistic regression for pre processing, cleaning & analysing the data. They used Microsoft "XLSTAT", for analysing the sample data they had collected. It was observed that the model of logistic regression in their case study allowed then to analyse the correlation between the occurrence of ischemic stroke & its risk factors. The XLSTAT software showed a very good sensitivity of 77.58% & specificity of 83% respectively. The ROC Curve is sensitivity according to specificity. However they concluded that prediction model achieved 19.7% error rate [6].

In 2016, Ahmet Kadir, Cemi Colak, Mehmet Ediz Sariham, collected dataset from Turget Ozal Medical Center, which contained the records of 112 healthy people and 80 patients and 2 target variable to predict Ischemic stroke using Data Mining approaches, They used support vector machine, Stochastic Gradient Boosing (SGB) and Penalized Logistic Regression (PLR). They included 10 fold cross validation re-sampling method. The performance evaluation metrices were accuracy, Area Under RoC Curve (AUC), sensitivity, specificity, positive predictive value and negative predictive value. The study showed that SVM produced the best performance compared to other models for predicting Ischemic stroke. The yield of accuracy values with 95% CI were 0.9789 for SVM, 0.9737 for SGB and 0.8947 for PLR. The yield of AUC values with 95% CI were 0.9783 for SVM, 0.9757 for SGB and 0.8953 for PLR. SVM and SGB showed remarkable predictive performance in identifying of Ischemic stroke [5].

In 2017, a team of medical university in Taiwan developed a model to automate the early detection of Ischemic stroke. They used CNN deep learning algorithm for this model. They collected CT images of the brains to analyze the possibility of stoke. The system preprocessed the CT images to remove the impossible area, which is not possible for occurrence of stroke. Then they selected the patch images and used data augmentation method to enhance the number of patch images. Later, they took patch images as input to the convolutional neural network for training and testing. Here they used 256 patch images to train and test the CNN module which has able to recognize the Ischemic stroke. It was observed that the proposed showed more than 90% result [4].

In 2018, A team of National Institute of Engineering, Karnataka, conducted a survey about AI applications in stroke and aimed to predict the accurate results of occurrence of stroke. They used predictive algorithms and parameters that include patients characteristics like gender, age, height, BMW, etc., they built a data model using decision tree algorithm to analyse these parameters. The result was analyzed using confusion matrix and the accuracy was 95%. To achieve this, they built the training model that helped to compare the newly fed data with the survey data. And the report was generated on the basis of this comparison [3].

In 2019, Department of Computer Architecture and Automation team of Universidad complustense de Madrid, Spain along with hospital Universitario de La Princesa, Madrid, Spain researched about the testing the hypothesis that state of art machine learning based modeling methods. They studied that non-invasive monitoring technologies could help on the diagnosis of stroke type. These techniques can even be employed for predicting future risks like the eventual death of the patient. They collected dataset comprised of the medical records of 119 patients with 7 predictors and 2 target variables which are prediction of stroke type and prediction of death. They used 7 different machine learning algorithms, which are Decision tree, KNN, logistic regression; Naïve Bayes, Neural Nework, Random Forest and Support Vector Machines and they evaluated over 6 different metrices using them. Also, they utilized 10-fold cross validation re-sampling technique, for guaranteed validation set from training one and the validation of the trained classifier against any unseen sample. The model performance evaluation metrices used to compare the algorithms used were; sensitivity, specificity, accuracy, F measure and areas under RoC as well as PRC. Among all of them, Random Forest models yielded the best performance in diagnosis of stroke prediction and death prediction compared to other algorithms, with values of 0.93+0.03 and 0.97±0.01 respectively [1].

In 2019, JoonNyung Heo, Hyungjong Park, young Dae Kim, Hyo Suk Nam & Ji Hoe Heo, build a model to predict the long-term outcomes of Ischemic stroke. They investigated the applications of Machine learning methods to predict outcomes of ischemic stroke. They did retrospective studying using a prospective cohort that had patients with ischemic stroke. They built three machine learning models which were deep neural network, random forest and logistic regression. After that they compared all of their predictability. Also, to analyze the accuracy of these models, they compared them to acute stroke registry and analysis of Lausanne score. They had included 2604 patients and 78% of them had favorable outcomes. Deep neural network model showed higher AUC curve than the ASTRAL score, while the logistic regression and random forest models' AUC curve were not higher than the ASTRAL score. It was observed that deep neural networks model performed better than any other models. It was more suitable for predicting outcomes [2].

Table -1: Comparison Analysis of Methods and Results

Sl. No	Paper Title	Method Used	Result
1	An Integrated Machine learning Approach to Stroke Prediction	Conservative mean feature selection, L1 regularized logistic regression novel prediction algorithm	Overall, this approach outperformed the current state-of-the art in both metrices of AUC and Concordance index
2	Prediction and Control of Stroke by Data Mining	Data Mining techniques, K-Nearest Neighbour and C4.5 Decision Tree Using WEKA	The accuracy of C4.5 decision tree algorithm and K-Nearest Neighbour were 95.42% and 94.18% respectively, in stroke prediction
3	Effective Analysis and Predictive Model of Stroke Disease using Classification Methods	Decision Tree, Bayesian Classifier, Neural Networks	Decision tree achieved 95.29% of sensitivity and 98.01% of accuracy. Bayesian Classifier achieved 87.10% and 91.30% respectively. Neural Networks achieved 94.82% and 97.87% respectively.
4	Machine learning for outcome prediction of Acute Ischemic Stroke Post Intra-Arterial Therapy	SPSS, MATLAB, Rapid Miner, ANN, Support Vector algorithm	It showed a promising accuracy up to 70% of predicting outcome.
5	A Model for Predicting Ischemic Stroke Using Data Mining Algorithms	Data Mining, Classification, Logistic Regression, WEKA 3.6	The results were obtained with the "XLSTAT" software. They showed the sensitivity of 77.58% and specificity of 83%.
6	Different Medical Data Mining Approaches based Prediction of Ischemic Stroke	Support Vector Machine, Stochastic Gradient Boosting (SGB) and penalized logistic regression (PLR)	The AUC values with 95% CI were 0.9783 for SVM, 0.9757 for SGB and 0.853 for PLR respectively.

7	An automated Early Ischemic Stroke Detection System using CNN Deep learning algorithm	Image pre-processing computer aided detection, Data augmentation, Convolutional Neural Network	It showed more than 90% accuracy
8	Stroke Prediction using Decision trees in AI	AI Decision Tree	The Decision tree algorithm showed the 95% accuracy in prediction of stroke
9	Comparison of different machine learning approaches to model stroke subtype classification and risk prediction	Decision tree, KNN, Logistic Regression, Naïve Bayes, Neural Network, Random Forest, Support vector machines	Random Forest Model showed best performance with average values of 0.93+0.03 and 0.9±0.01 respectively.
10	Machine learning based model for prediction of outcomes in acute stroke	Deep Neural Network, Random Forest. Logistic Regression	Deep neural network showed the highest accuracy.

3. REMARKS

The evaluation on each paper was done by carefully studying and analyzing them. Henceforth, the remarks with respect to each paper are mentioned below.

- Aditya Khosla et al. [10] showed that the machine learning methods significantly outperform the COX model in terms of both binary stroke prediction and stroke risk estimation. And it can be used for identifying potential risk factors for diseases without performing clinical trials.
- Leila Amini et al. [9] observed that the effect of data dimension reduction on classification accuracy and other algorithm performance criteria must be examined. Most investigations are presented theoretically. So this field is unfamiliar to medical specialists.
- A. Sudha et al. [8] showed that neural network performance was having more accuracy, compared to other two classification techniques. Here they predicted the presence of stroke based on a few parameters.
- Hamed Asadi et al. [7] proposed model inherent need for large training dataset which may affect the accuracy of machines in studies. No clear understanding of the true predictors, since an over corrected conservative.
- Balar Khalid et al. [6] built a model to predict the stroke disease. But to achieve medical data of higher quality all the necessary steps should be taken to build the medical information system which gives accurate information about patients medical history than their billing invoices.
- Ahmet Kadir Arslan et al. [5] observed; SVM produced the best predictive performance compared to other models. But for obtaining more accurate and robust comparison, comprehensive simulation is necessary.
- Chiun-Li-Chin et al. [4] used very less patch images (about only 256) to train the model, which decreases the efficiency of the system. But this proposed model can be effectively used by doctors to diagnose the diseases.
- Aishwarya Roy et al. [3] proposed a model that can assist the doctors in clinical trials. In this paper, they haven't mentioned clearly about the dataset they had used for prediction model and also about the method they have used.
- Luis Garcfa-Terriza et al. [1] have used different algorithms to predict stroke type (hemorrhagic v/s Ischemic) and to predict further complications of diseases. This will also allow doctors to use preventive treatments to avoid the adversity
- JoonNyung Heo et al. [2] observed that it was a single-center study and requires validation with respect to data from other sources. However, deep neural network itself may be more suitable for stroke prediction outcomes.

4. CONCLUSIONS

Many machine learning techniques have contributed to predict stroke in several different scenarios. Deciding to use a specific machine learning technique should be based on considerations of scenarios, data sets, parameters and other analysis. We cannot conclude on the best technique to use for stroke prediction. Each technique has its own advantages and disadvantages. It is wise to choose one among them based on the necessity of the individual problem statement. One must perform statistical analysis and initialisation to decide on the specific technique or model to use. However, random forest is one of the most popular and powerful technique for estimating a quantity from a data sample as it shows promising results.

REFERENCES

[1] Luis Garcfa-Terriza, Risco Martin, Ayala and Gemma Reig Rosello, “Comparison of different Machine Learning approaches to model stroke subtype classification and risk prediction”, Society for Modeling & Simulation International (SCS), 2019 April 29-May2.

[2] JoonNyung Heo, Jihoon G. Yoon, Hyungjong Park, Young Dae Kim, Hyo Suk Nam, Ji Hoe Heo, “Machine Learning-Based Model for Prediction of Outcomes in Acute Stroke”, 2019 February 1, doi:10.1161/strokeaha.118.024293

[3] Aishwarya Roy, Anwesh Kumar, Navin Kumar Singh and Shashank D, “Stroke Prediction using Decision Trees in Artificial Intelligence”, IJARIIT, Vol. 4, Issue 2, 2018, pp: 1636-1642.

[4] Chiun-Li-Chin, Guei-Ru Wu, Bing-Jhang Lin, Tzu-Chieh Weng, Cheng-Shiun Yang, Rui-Cih Su and Yu-Jen Pan, “An Automated Early Ischemic Stroke Detection System using CNN Deep Learning Algorithm”, IEEE 8th International Conference on Awareness Science and Technology, 2017.

[5] Ahmet Kadir Arslan, Cemil Colak, Mehmet Ediz Sarihan, “Different medical data mining approaches based prediction of ischemic stroke”, Elsevier, Computer Methods and Programs in Biomedicine 2016 March 18.

[6] Balar Khalid and Naji Abdelwahab, “A model for predicting Ischemic stroke using Data Mining algorithms”, IJISET, Vol. 2 Issue 11, Nov 2015, ISSN: 2348-7968.

[7] Hamed Asadi, Richard Dowling and Bernard Yan, “Machine Learning for outcome prediction of acute ischemic stroke”, PLOS ONE Vol. 9 Issue 2, Feb 2014.

[8] A. Sudha, P. Gayathri, “Effective analysis & predictive model of stroke disease using classification methods”, IJCA(0975-8887), Vol. 43-No. 14, April 2012.

[9] Leila Amini, Reza, Rasul Norouzi & Associates, “Prediction and Control of Stroke by Data Mining”, IJPM, 8th Iranian Neurology Congress, Vol. 4, 23 Feb 2013.

[10] Aditya Khosla, Yu cao, Honglak Lee & Associates, “An integrated machine learning approach to stroke prediction”, 25-28 July 2010, Washington, DC, USA.

AUTHORS

Gagana M
Final Year PG Student,
Computer Science and Engineering,
PES College of Engineering,
Mandya, Karnataka, India.

Dr. Padma M C
Ph.D. in Pattern Recognition & Image Processing.
Professor & Head of the Department,
Computer Science and Engineering,
PES College of Engineering,
Mandya, Karnataka, India