Informatics Bulletin, Faculty of Computers and Artificial Intelligence, Helwan University
Published Online ***Vol 4 Issue 2, July 2022***
(https://fcihib.journals.ekb.eg)

Machine Learning Models for Predicting Brain Strokes

Samaa Ahmed Mostafa
Information System department
Faculty of Computer Science and artificial intelligence, Helwan University, Cairo, Egypt
samaaahmed252@fci.helwan.edu.eg

Doaa Saad Elzanfaly
Information System department
Faculty of Computer Science and artificial intelligence, Helwan University, Cairo, Egypt
doaa.saad@fci.helwan.edu.eg

Ahmed El Sayed Yakoub
Information System department
Faculty of Computer Science and artificial intelligence, Helwan University, Cairo, Egypt
Eng_ahmedyakoup@yahoo.com

Abstract— Brain Strokes are one of the most serious and most common diseases in the world due to their sudden occurrence. It can be considered the death of brain cells as a result of the failure to provide these cells with the right amount of blood, which leads to the interruption of the brain's action and therefore leads to death within minutes. Therefore, predicting the incidence of strokes based on the risk factors of the patient is one of the most important reasons for preventing the occurrence of these strokes and providing early treatment for them. Machine learning techniques are widely used in building predictive models for strokes according to the patient's electronic health record, which contains the factors that lead to the occurrence of stroke. Ensemble methods are one of the most important concepts in machine learning as it works to collect more than one machine learning algorithm and combine them to produce a reliable predictive model with higher accuracy. The purpose of this paper is to present a survey on predictive models for Brain Strokes using a machine learning ensemble classifier.

Keywords — Brain Strokes, Machine Learning, Predictive Models, Ensemble methods.

1. INTRODUCTION

According to the World Health Organization, Brain Strokes are considered the second main reason for death in the world [1]. A stroke occurs at any time, which is why it is too dangerous [2]. The main two types of strokes are ischemic and hemorrhagic [3]. Ischemic stroke occurs when the amount of blood flowing into the brain cells decreases or stops as a result of the blockage, leading to the death of these cells within minutes and causing death. While hemorrhagic stroke occurs when weak blood vessels are ruptured by hypertension [4]. These two different types of stroke cause similar symptoms because they are affected by the amount of blood that flows in the brain. The most famous way to help find out about stroke is FAST, according to the National Stroke Association [5]. FAST is an abbreviation for four words: F is the sense of the face when a human laughs, does it hang part of the face or find it difficult to laugh. A refers to both arms when one raises his two arms, does anyone of them drift down? S is for speech if one is having trouble while talking. Finally, T refers to the time in which the person should go to the hospital immediately. Risk factors causing strokes can be classified as modifiable and non-modifiable risk factors. Age and sex are non-modifiable risk factors for both ischemic and hemorrhagic strokes, while hypertension, smoking, avg glucose level, and diet are considered modifiable risk factors [6]. Most of the causes of strokes are related to lifestyle factors. Therefore, you can reduce the risk of infection if you control them [7]. Moreover, building a predictive model for stroke is very important to provide early treatment for stroke patients.

This paper reviews the ensemble methods presented in the literature to build predictive models to predict whether the patient will have a stroke or not based on the risk factors of this disease [8]. Such models help to protect patients from the sudden occurrence of strokes. The most important two steps to reduce

the risk of stroke are: control risk factors and know the warning signs for strokes [9]. Most previous studies were based on just building a predictive model for brain strokes by using classical classification algorithms [10] [11] [12], but few studies employed the Ensemble Methods to improve the accuracy of the individual algorithms [13] [14] [15]. Also, studies that use ensemble methods have not highlighted all types of ensemble methods that affect the model accuracy. Another study focused on building a predictive model based on a few numbers of records of patients and did not cover all the possible attributes that might affect the model's accuracy [16]. Moreover, most previous studies did not rank the risk factors of the strokes.

The rest of this paper is organized as follows. Section 2 reviews predictive models for brain strokes. Section 3 investigates the ensemble methods presented in the literature to predict brain strokes. Finally, section 4 concludes the paper and sketches out the future work.

2. PREDICTIVE MODELS FOR BRAIN STROKES

A predictive model is a data mining technique in which we use mathematical tools to analyze current and historical datasets for patterns and calculate the probability of an outcome. Building a predictive model in the healthcare system help to identify the patients at the highest risk for a certain disease according to their health records and provide early treatment for them. Different classification algorithms have been used for the prediction of stroke by using several risk factors for this disease. Machine learning algorithms help us build predictive models to early predict a patient has a stroke or not according to the patient's health record. Early stroke prediction is useful for prevention and early treatment [17]. Some researchers used different classification algorithms to build the stroke predictive model such as decision tree, support vector machine, Naive Bayes, and k-nearest neighbors.

In [10], a support vector machine algorithm has been used to build a stroke prediction model. The model achieved an accuracy of 98.89%. Others developed a stroke predictive model to predict whether the patient will have a stroke or not [11]. They build the model using a support vector machine algorithm (SVM). They evaluated the performance of different kernel functions by using sensitivity, specificity, accuracy, precision, and f1 score. The result shows that the linear functions reached a greater accuracy of 91%. Others in [18] used logistic regression for building the predictive model. In [8], they developed a stroke predictive model using three machine learning classification algorithms: Decision Tree, Naïve Bayes, and Neural Network. It has been found that the Neural Network classifier is giving more accuracy when compared with two other two classifiers.

3. ENSEMBLE METHODS FOR BRAIN STROKE PREDICTION

Ensemble methods are a machine learning technique that is used to combine several machine learning classification algorithms as base models for this method to produce one optimal predictive model with higher accuracy [19]. These methods usually produce results at an accuracy better than any single classifier, which makes the prediction more accurate and reliable [20]. These methods are useful for classification and regression. There are many types of ensemble methods such as boosting, Stacking, bagging, and random forest [21]. It has been proven in [22] the effective use of the ensemble method to obtain more accurate prediction results than using classical classification algorithms independently, as shown in table1.

Rado et al. [22] concentrate on using ensembles methods for delivering results with high accuracy. The authors have shown the effectiveness of the ensemble methods for obtaining good predictions for stroke.

TABLE 1

COMPARISON BETWEEN USING SINGLE CLASSIFIERS AND ENSEMBLE METHODS ACCORDING TO THEIR ACCURACY [22]

Method	Accuracy %
SVM	77.29
KNN	84.58
Decision tree/C4.5	86.10
Random forest/Bagging	86.63
AdaBoost	82.43
Stacking	87.58

They used three ensemble methods which are Bagging, Boosting, and Stacking. Some Machine learning classification algorithms have been used as basic models for those ensemble methods such as SVM, DT, and KNN. They used a stroke dataset that contains 12 features such as gender, age, hypertension, heart disease. The result shows that the ensemble classifier produces higher prediction accuracy. They provide a comparison between single classifiers such as SVM and C4.5 and ensemble classifiers such as Bagging and Stacking where they reached from 86.63 to 87.58 accuracy with ensemble classifiers vs 77.29 to 86.10 in single classifiers. However, their model can predict stroke with acceptable accuracy, but it needs some improvements to achieve higher accuracy than they have reached.

MAHESH et al. [23] developed a prediction model for stroke by using three machine learning algorithms which are Decision Tree, Naïve Bayes, and Artificial Neural Network. They used a dataset containing risk factors of stroke-like age, gender, hypertension, heart disease, BMI, smoking status, Avg Glucose level. They evaluated the performance by using AUC (Area Under the Carve) and ROC (Receiver Operating Characteristics). The result shows that the used algorithms gave appropriate accuracy in predicting stroke patients. They also developed a web application to provide a warning about the level of stroke risk. Although their model can predict stroke, it needs to include more risk factors for stroke to help the model to provide an accurate warning for the possibility of stroke occurrence.

Nwosu et al. [24] authors in this study have applied several data mining techniques to develop a prediction model for stroke patients by using patients' medical records. They used three machine-learning algorithms Neural network, Decision tree, and Random forest. An electronic health records dataset with 12 attributes has been used for building a stroke prediction model.

They evaluated the performance according to the accuracy of the algorithm. The result shows that the neural network (multi-layer perception) model achieved the highest accuracy of 75.02%. They didn't build the model by using some features instead of using all features which are called feature selection, this may lead to an increase in the model accuracy.

Almadani et al. [25] this study focused on developing a model for stroke prediction using data mining techniques. They used J84 (C4.5), JRIP, and Neural Network (multilayer perception MLP) algorithms to make the prediction. A dataset that has been used is from the data governance department at King Abdulaziz Medical City. The result shows that c4.5 and JRIP have achieved the highest accuracy prediction of 95.25% and 94.42% respectively after applying Principal Component Analysis (PCA). This paper focuses on showing the effective use of PCA and how it helps in increasing the model prediction accuracy.

Singh et al. [26] presented a predictive model for stroke disease using classification algorithms. They used the cardiovascular health study dataset. To improve the performance of the prediction, they applied a decision tree for feature selection purposes. Then used PCA to reduce the dimension of the data. After that, they used Neural Network (NN) for making the classification process. In the final step, they analyzed the result by using a Confusion matrix. The result shows that the model gave a higher accuracy of 97.7%.

J. Lee et al. [27] have built a 10-year stroke prediction model and, they intended to develop a web application to provide a personalized warning on the user stroke level and provide a lifestyle correction message about the stroke risk factors. They built the prediction model by using a dataset obtained from national health in Korea. The result shows that the most influential risk factors for stroke were age, smoking, diabetes, and hypertension respectively. It was better for their model to provide the stroke occurrence probability.

A predictive model for stroke using an artificial neural network (ANN) is presented in [28]. They used a dataset containing patients' physiological data. It found that scaled conjugate gradient (SGD) and Levenberg Marquardt (LM) algorithms achieved an accuracy of 98%. They needed to do more data preprocessing to solve the problem of the imbalanced dataset.

P. Chantamit-o-pas et al. [29] have focused on building a stroke prediction model using a deep learning model. Predictive techniques have been applied to the heart disease dataset. They made a comparison between three models: Naïve Bayes, Support Vector Machine (SVM), and Deep Learning for the prediction of stroke disease. The result shows that the two algorithms of support vector machine and naïve Bayes help to predict if the patient has a stroke or not, while the deep learning technique shows the percentage of chance of having a stroke. It also shows that heart atrial fibrillation in patients is the main factor of stroke. The limitation of this paper is that they didn't use enough stroke risk factors that may enhance the result.

A stroke prediction model has been developed by using Cardiovascular Health Study (CHS) dataset in [17]. The authors addressed the problem of missing data imputation, feature selection, prediction in the heart dataset. They evaluated the performance by using the area under the ROC curve and concordance index. A stroke prediction model has been built with identified risk factors that have not been discovered before. The combined support vector machine with a conservative mean and feature selection in their model has achieved a higher area under the curve.

P. Govindarajan et al. [15] developed a model for stroke by using various classification algorithms such as Artificial Neural Network, Support Vector Machine, Boosting, Bagging, and Random Forest. Case sheets collected from Sugam Multi-specialty Hospital, India that contains information about stroke patients have been used to build the model. They evaluated the performance by using accuracy, sensitivity, specificity, recall, and precision. The result has shown that Neural Network achieved the highest accuracy about 95%. The accuracy they have reached is considered acceptable accuracy in the medical area.

Rakshit, T et al. [30] Proposed a machine learning prediction model for stroke. The model provides a prediction process for the early detection of stroke symptoms, enabling them to be prevented at an early stage. A dataset collected from Kaggle has been used to build the prediction model. Various machine learning classification algorithms have been used to make the prediction such as Random Forest, Naïve Bayes, Logistic Regression, K-Nearest Neighbor (KNN), Decision Tree, and Support Vector Machine (SVM). After building the model, it has been found that the decision tree gives the highest accuracy among the other used algorithms. This paper has some limitations, such as it removed null values in a specific column in the dataset, which considers a very informative and important feature. They were possible to fill the null values by taking the mean/median of the column rather than deleting these null values. They also built the model with an unbalanced dataset, rather than using some techniques which make the dataset more balanced. The unbalanced dataset results in predicting inaccurate results, and this isn't acceptable in medical prediction. Feature selection techniques don't exist in the model, which helps the results to be more accurate.

Rajora, M et al. [31] intended to develop a stroke prediction model. Various machine learning algorithms have been used for building the prediction model. The Receiver Operating Curve (ROC) is obtained for each algorithm. Dataset features have been analyzed by using univariate and multivariate plots which showed the correlation between the features. The analysis shows that age, gender, smoking status are considered important features and some features like residence type are of less importance. They made the implementation by using Apache Spark. The ROC for the Gradient Boosting algorithm gives the best results with a ROC area score of 0.90. They succeeded in fine-tuning the parameters, which made the values of Accuracy, Precision, Recall, and F1 score increased

4. CONCLUSION & FUTURE WORK

Brain Strokes is one of the most serious diseases that exist due to its sudden and fatal occurrence. Early prediction of this stroke helps prevent its occurrence and avoid exposure to its fatal dangers. Strokes occur as a result of the lack of blood flowing into the brain or there is a blockage in brain cells preventing from getting the right amount of oxygen and nutrients, causing the death of these cells within minutes. Early identification and

control of the risk factors that cause this stroke help in preventing this disease. Those factors involve controlling high blood pressure, stopping smoking, and controlling high cholesterol in the blood. Building a predictive model is necessary to predict the occurrence of brain strokes based on the patient's electronic record containing the risk factors causing this stroke.

Machine learning algorithms help us build predictive models to early predict a patient has a stroke or not according to the patient's health record. The use of ensemble methods helps build a more accurate prediction model than using machine learning algorithms individually, as it combines more algorithms to build a more accurate predictive model.

In the future, we plan to propose a machine learning model that exploits ensemble methods to predict brain strokes effectively. Specifically, we aim to use more classification algorithms as base models for more ensemble methods that have not been used before to improve the result.

5. REFERENCES

[1] "Global Health Estimates," World Health Organization, 06-Dec-2018. [Online]. Available: https://www.who.int/healthinfo/global_burden_disease/en/.

[2] "Stroke: Causes, symptoms, diagnosis, and treatment", Medicalnewstoday.com,2021. [Online]. Available: https://www.medicalnewstoday.com/articles/7624.

[3] Alrabghi, L., Alnemari, R., Aloteebi, R., Alshammari, H., Ayyad, M., Al Ibrahim, M., Alotayfi, M., Bugshan, T., Alfaifi, A., & Aljuwayd, H. (2018). Stroke types and management. International Journal of Community Medicine and Public Health, 5(9), 3715. https://doi.org/10.18203/2394-6040.ijcmph20183439.

[4] American Stroke Association Inc., "Types of Strokes," www.stroke.org. [Online]. Available: https://www.stroke.org/en/about-stroke/types-of-stroke.

[5] American Stroke Association Inc., "Stroke Symptoms," www.stroke.org. [Online]. Available: https://www.stroke.org/en/about-stroke/stroke-symptoms.

[6] Boehme, A. K., Esenwa, C., & Elkind, M. S. V. (2017). Stroke risk factors, genetics, and prevention. Circulation Research, 120(3), 472–495. https://doi.org/10.1161/circresaha.116.308398.

[7] "Stroke risk factors and prevention", Betterhealth.vic.gov.au, 2021. [Online]. Available: https://www.betterhealth.vic.gov.au/health/ConditionsAndTreatments/stroke-risk-factors-and-prevention.

[8] Sudha, A., Gayathri, P., & Jaisankar, N. (2012). Effective analysis and predictive model of stroke disease using classification methods. International Journal of Computer Applications, 43(14), 26–31. https://doi.org/10.5120/6172-8599.

[9] Cleveland Clinic. 2022. Stroke Risk Factors & Stroke Prevention. [online] Available at: <https://my.clevelandclinic.org/health/articles/13398-know-your-risk-factors-for-stroke>.

[10] Rosado, J. T., & Hernandez, A. A. (2019). Developing a predictive model of stroke using a support vector machine. 2019 IEEE 13th International Conference on Telecommunication Systems, Services, and Applications (TSSA). https://doi.org/10.1109/tssa48701.2019.8985498.

[11] Jeena, R. S., & Kumar, S. (2016). Stroke prediction using SVM. 2016 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT). https://doi.org/10.1109/iccicct.2016.7988020.

[12] Min, S. N., Park, S. J., Kim, D. J., Subramaniyam, M., & Lee, K.-S. (2018). Development of an algorithm for stroke prediction: A National Health Insurance Database Study in Korea. European Neurology, 79(3-4), 214–220. https://doi.org/10.1159/000488366.

[13] Rado, O., Al Fanah, M., & Taktek, E. (2019). Ensemble of multiple classification algorithms to predict stroke datasets. Advances in Intelligent Systems and Computing, 93–98. https://doi.org/10.1007/978-3-030-22868-2_7.

[14] Jeena, R. S., & Kumar, S. (2016). Stroke prediction using SVM. 2016 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT). https://doi.org/10.1109/iccicct.2016.7988020.

[15] Govindarajan, P., Soundarapandian, R. K., Gandomi, A. H., Patan, R., Jayaraman, P., & Manikandan, R. (2019). Classification of stroke disease using machine learning algorithms. Neural Computing and Applications, 32(3), 817–828. https://doi.org/10.1007/s00521-019-04041-y.

[16] Monteiro, M., Fonseca, A. C., Freitas, A. T., Pinho e Melo, T., Francisco, A. P., Ferro, J. M., & Oliveira, A. L. (2018). Using machine learning to improve the prediction of functional outcomes in ischemic stroke patients. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 15(6), 1953–1959. https://doi.org/10.1109/tcbb.2018.2811471.

[17] Khosla, A., Cao, Y., Lin, C. C.-Y., Chiu, H.-K., Hu, J., & Lee, H. (2010). An integrated machine learning approach to

stroke prediction. Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '10. https://doi.org/10.1145/1835804.1835830.

[18] S. Min, S. Park, D. Kim, M. Subramaniyam and K. Lee, "Development of an Algorithm for Stroke Prediction: A National Health Insurance Database Study in Korea", European Neurology, vol. 79, no. 3-4, pp. 214-220, 2018. Available: 10.1159/000488366.

[19] Chavan. A, "Improve Machine Learning Results with Ensemble Learning", AI, ML, Data Science Articles | Interviews | Insights | AI TIME JOURNAL, 2021. [Online]. Available: https://www.aitimejournal.com/@akshay.chavan/improve-machine-learning-results-with-ensemble-learning.

[20] Demir. N, "Ensemble Methods: Elegant Techniques to Produce Improved Machine Learning Results", Toptal Engineering Blog, 2021. [Online]. Available: https://www.toptal.com/machine-learning/ensemble-methods-machine-learning.

[21] E. Learning, "Ensemble Methods in Machine Learning | 4 Types of Ensemble Methods", EDUCBA, 2021. [Online]. Available: https://www.educba.com/ensemble-methods-in-machine-learning/.

[22] Rado, Omesaad, Muna Al Fanah, and Ebtesam Taktek, "Ensemble of Multiple Classification Algorithms to Predict Stroke Dataset". Intelligent Computing-Proceedings of the Computing Conference. Springer, Cham, 2019.

[23] KUNDER AKASH MAHESH, SHASHANK H N, SRIKANTH S, and THEJAS A M. "Prediction of Stroke Using Machine Learning." Researchgate, 2020.

[24] Nwosu, C. S., Dev, S., Bhardwaj, P., Veeravalli, B., & John, D. (2019). Predicting stroke from Electronic Health Records. 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). https://doi.org/10.1109/embc.2019.8857234.

[25] Almadani, O., & Alshammari, R. (2018). Prediction of stroke using data mining classification techniques. International Journal of Advanced Computer Science and Applications, 9(1). https://doi.org/10.14569/ijacsa.2018.090163.

[26] Singh, M. S., & Choudhary, P. (2017). Stroke prediction using Artificial Intelligence. 2017 8th Annual Industrial Automation and Electromechanical Engineering Conference (IEMECON). https://doi.org/10.1109/iemecon.2017.8079581.

[27] Lee, J.-woo, Lim, H.-sun, Kim, D.-wook, Shin, S.-ae, Kim, J., Yoo, B., & Cho, K.-hee. (2018). The development and implementation of stroke risk prediction model in National Health Insurance Service's Personal Health Record. Computer Methods and Programs in Biomedicine, 153, 253–257. https://doi.org/10.1016/j.cmpb.2017.10.007.

[28] Peng. C, Wang. S, Liu. S, Yang. Y and Liao. B, "Artificial Neural Network Application to the Stroke Prediction", 2020 IEEE 2nd Eurasia Conference on Biomedical Engineering, Healthcare and Sustainability (ECBIOS), 2020. Available: 10.1109/ecbios50299.2020.9203638.

[29] Chantamit-o-pas, P., & Goyal, M. (2017). Prediction of stroke using Deep Learning Model. Neural Information Processing, 774–781. https://doi.org/10.1007/978-3-319-70139-4_78.

[30] Rakshit, T. and Shrestha, A., 2022. Comparative Analysis and Implementation of Heart Stroke Prediction using Various Machine Learning Techniques. [online] Ijert.org. Available at: <https://www.ijert.org/comparative-analysis-and-implementation-of-heart-stroke-prediction-using-various-machine-learning-techniques>.

[31] Rajora, M., Rathod, M., and Naik, N., 2020. Stroke Prediction Using Machine Learning in a Distributed Environment. Distributed Computing and Internet Technology, pp.238-252.