International Journal of Modern Developments in Engineering and Science
Volume 2, Issue 6, June 2023
https://www.ijmdes.com | ISSN (Online): 2583-3138

Brain Stroke Prediction System – A Review

Ramandeep Kaur[1*], Sirjana Dhillon[2]
[1,2]Student, Department of Computer Science Engineering, DAV Institute of Engineering and Technology, Jalandhar, India

***Abstract*: The primary organ of the human body that regulates all bodily activities is the brain. When the blood supply to a portion of the brain is cut off, brain tissue cannot receive oxygen and nutrients, which results in a stroke. In minutes, brain cells start to degenerate. fundamental organ of the body, in charge of all bodily processes. When the blood supply to a portion of the brain is cut off, brain tissue cannot receive oxygen and nutrients, which results in a stroke. In minutes, brain cells start to degenerate. The World Health Organization (WHO) reports that stroke is the second greatest cause of death in the world, accounting for about 11% of all fatalities. Subarachnoid haemorrhage affects 3% of the population, intracerebral haemorrhage affects 10%, and ischemic stroke affects 87% of the population. Brain stroke symptoms can appear suddenly and may include facial drooping, weakness, or paralysis. Those who have heart disease, high blood pressure, or other risk factors are also more likely to experience brain stroke. Based on input characteristics including gender, age, various diseases, and smoking status, our ML model uses a dataset to predict whether a patient is likely to have a stroke. A trustworthy dataset for stroke prediction was collected from the Kaggle website to increase the algorithm's efficacy. For precise prediction, we have employed Machine Learning techniques including Logistic Regression, Decision Tree Classification, Random Forest Classification, KNN, and SVM.**

***Keywords*: Machine Learning, Brain stroke, Logistic Regression, Decision Tree classification, Random Forest classification, KNN, SVM.**

1. Introduction

Stroke is a dangerous disorder that claims lives. The internal carotid and vertebral arteries are responsible for carrying blood to the brain. The internal jugular veins are then used to change it from the head to the heart [5].

The loss of blood may be seen in one of two ways: either an ischemic stroke, which occurs when the blood flow between the blood tissues decreases, or a haemorrhagic stroke, which occurs when internal bleeding within the brain tissues [6]. A clot in the blood vessel produced by atherosclerosis may also cause this blocking to occur. When blood leaks from the clot, it puts pressure on the brain because it spreads throughout the blood vessel.

Machine learning (ML) offers a quick and accurate prediction result, and it has developed into a potent tool in healthcare settings, providing stroke patients with individualised therapeutic care. This study aims to identify and forecast the risk of brain strokes using machine learning (ML) algorithms such logistic regression, SVM, KNN, decision trees, and random forests.

*Corresponding author: ramandeep0301@gmail.com

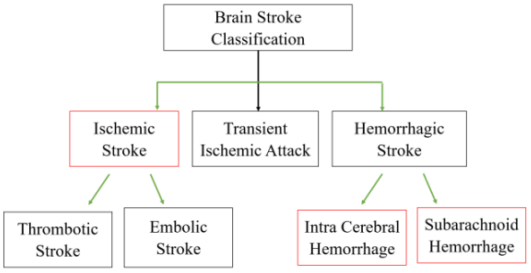

Fig. 1. Classification of brain stroke

In stroke, CT images are a widely used dataset. SVM and Random Forests are effective methods, too. For the diagnosis and prognosis of illnesses, ML is essential. Using a patient's medical history, the algorithm can determine which patients are most likely to contract the ailment. The device can forecast a person's likelihood of contracting an illness by considering their age, blood pressure, sugar levels, and other characteristics. The prediction of strokes using a feed-forward multi-layer artificial neural network was studied in [5].

The main objective of this research is to demonstrate how artificial neural networks (ANN), boosting methods, and machine learning algorithms can be utilised to forecast when a brain stroke would occur. The primary contribution of this study is the comparison and identification of the most effective method for stroke prediction by applying multiple algorithms to a dataset. In order to predict the presence of stroke disease and its various related characteristics, this study uses neural networks and machine learning algorithms as classification methods

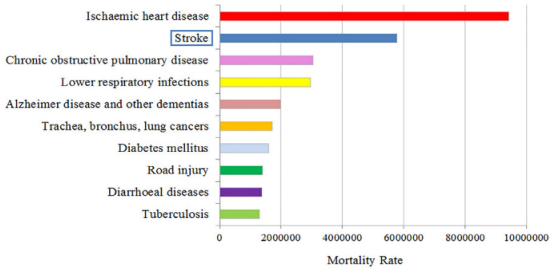

Fig. 2. Recording ranks of different diseases in 2016

2. Literature Survey

The section on stroke expectation has seen a lot of work completion. In order to understand the likelihood of stroke,

Jenna et al. investigate many factors [1]. The research also go into the accuracy and results of applying various Machine Learning models with text- and image-based datasets.

[2] Harish Kamal, Victor Lopez, Sunil A. Sheth, in their study discuss how Machine Learning (ML) through pattern recognition algorithms is currently becoming an essential aid for the diagnosis, treatment, and prediction of complications and patient outcomes in several neurological diseases. The evaluation and treatment of Acute Ischemic Stroke (AIS) requires the use of neuroimaging for decision-making. This study focuses on the recent developments and applications of ML in neuroimaging focusing on acute ischemic stroke. The ML is used for identification of imaging diagnostic findings, estimating time of onset, lesion segmentation, and fate of salvageable tissue, to the analysis of cerebral oedema, and predicting complications and patient outcomes after treatment.

[3] Govindarajan collected the information from the Sugam Multi-speciality Hospital. More than 500 patient records made up the dataset. In addition to Support Vector Machine (SVM), Artificial Neural Network (ANN), Logistic Regression, Decision Tree, Bagging, and Boosting, these techniques were used. With the ANN Algorithm, they achieved the highest accuracy (95%), out of the Machine Learning Algorithms.

[4] Sung et al. used clinical data that included information on the ischemic strokes suffered by 739 patients. There are 17 clinical factors in this information. They tested using four machine learning algorithms: Bootstrap choice forest, Logistic Regression, Boosted Trees, and Deep brain organisation. The exactness score obtained from the model is 0.966, 0.966, 0.966, and 0.946. The Boosted Tree calculation achieves the highest region under the curve with a value of 0.934 and precision of 0.966 out of all the calculations.

[5] Selma collected data from a few urgent care facilities and clinical Centres. The clinic report includes information on each patient admitted to the hospital as well as their CT scan, MRI analysis, age, gender, and other details. There were about 410 patients in the dataset., whose age is mostly somewhere in the range of 48 and 87 years. A couple of cases in the age of 32 years and most of them are male. The presentation of Decision tree characterization is superior to the exhibition of the KNN calculation. Clinical experts utilized a decision tree calculation to order and analyse ischemic stroke patients.

3. Methods Used

A. Overview of the Dataset

The stroke prediction dataset was used in the investigation. 5110 rows and 12 columns make up this dataset. The output column stroke has a value of either 1 or 0.s.

Table 1 contains a description of the dataset. The value 0 denotes that no stroke risk was found, but the value 1 denotes the discovery of a stroke risk. In this dataset, the potential of 0 in the output column (stroke) outweighs the possibility of 1. Only 249 rows in the stroke column alone have a value of 1, whereas 4861 rows have a value of 0. Data pre-processing is used to balance the data in order to increase accuracy.

Table 1
Description of the dataset used [3]

Attribute Name	Description
ID of Individuals	Unique identification number of 5110 patient
Gender	Male = '0', Female= '1'
Age (in years)	Age of the patient (1-82)
Hypertension	Indicating whether the patient has hypertension (1) problem or not (0)
Heart Disease	Demonstrating whether the unique id patient has heart disease problem (1) or not (0)
Ever married	Represents the marital status by yes (1) or no (0).
Work type	It indicates five category of work status of the patient. Children = '0' Government Job = '1' Never worked = '2' Private = '3' Self Employed ='4'
Residence type	It denotes the residential area type, whether Rural ='0' or Urban = '1'
Average Glucose Level	It gives the average glucose level which is represented in numeric form (55.12-271.74)
BMI	Body Mass Index is represented in numeric form (10.3-97.6).
Smoking Status	It indicates four categories of smokers. Formerly Smoked = '0' Never Smoked = '1' Smokes = '2' Unknown = '3' (There is no information/could not found about the unknown type)
Stroke	It indicates the target, whether have stroke (1) or non-stroke (0).

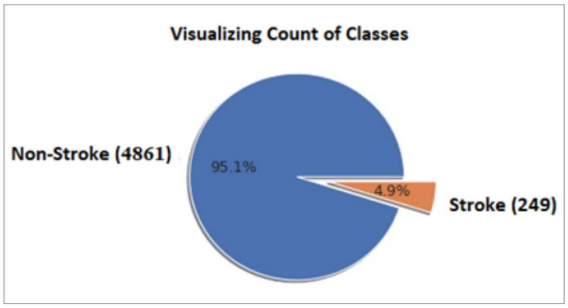

Fig. 3. Count of stroke vs. non-stroke

B. Visualization of Feature Selection

The correlation coefficient is a statistical measure of how strongly two variables are associated with one another's relative movements. The range of the values is from -1.0 to 1.0. If the projected value was greater than 1.0 or smaller than -1.0, there was a measurement error in the correlation. A correlation of 0.0 shows that there is no linear relationship between the movements of the two variables.

Age, high blood pressure, blood sugar levels on average, heart disease, having been married before, and BMI are all favourably connected with the objective characteristic, whereas gender is negatively correlated (-0.0069) with stroke.

4. Proposed Methodology

Data pre-processing is required to remove unwanted noise from the dataset that could lead the model to depart from its intended training. This stage addresses everything that prevents the model from functioning more efficiently. Following the

collection of the relevant dataset, the data must be cleaned and prepared for model development. As stated before, the dataset used has twelve characteristics. To begin with, the column id is omitted since its presence has no bearing on model construction.

The dataset is then inspected for null values and filled if any are detected. The null values in the column BMI are filled using the data column's mean in this case.

Label encoding converts the dataset's string literals to integer values that the computer can comprehend. As the computer is frequently trained on numbers, the strings must be converted to integers. The gathered dataset has five columns of the data type string. All strings are encoded during label encoding, and the whole dataset is transformed into a collection of numbers. The dataset used for stroke prediction is very imbalanced. The dataset has a total of 5110 rows, with 249 rows indicating the possibility of a stroke and 4861 rows confirming the lack of a stroke. While using such data to train a machine-level model may result in accuracy, other accuracy measures such as precision and recall are inadequate. If such an unbalanced data is not dealt with properly, the findings will be inaccurate, and the forecast will be ineffective. As a result, to obtain an efficient model, this unbalanced data must be dealt with first.

A MinMaxScaler was used to scale the features to between -1 and 1 to normalize them. After that, Principal component analysis (PCA) was utilized which chooses the minimum number of principal components such that 95% of the variance is retained.

The data is split into training and testing data with an 80/20 split, in order to increase the accuracy and efficiency of this job.

A. Algorithms

1) Logistic regression

The method of supervised learning known as logistic regression [2] is used to analyse the absolute dependent values by using the variable present in the necessary blocks of the independent values. The logistic regression of the absolute dependent values can be used to analyse the output values. As a result, the solutions can be expressed in terms of absolute or differential variables. It might take on any shape, including binary or numerical variables. The values are presented in a 0 or a format in computer-determined languages, however this model reflects a possible value that is between 0 and 1. While categorising the concerns was done using logistic regressions, the retrogradation of the issues can be resolved using linear regression is written as:

Sigmoid Function

$$\emptyset(z) = 1/1 + e - z \quad (1)$$

where,

z = b0 + b1 * age + b2 * systolic BP + b3 * diastolic BP + …… + b9 * cholesterol.

2) K-Means

combining the collection of unlabelled datasets into different configurations, which binds the datasets into a collection and configures the K-Means grouping algorithm. The number of existing groups that must be initialised in this technology is the variable K in this research work. This needed easy sorting of the datasets into different combination variables from among the different groups, without taking the training process into account, and by independently identifying different datasets in the unlabelled classes, as follows:

$$\sum k_j=1 \sum \| X_{i\,(j)}\ n_i=1 - C_j \| 2 \quad (2)$$

where, 'k' is number of clusters, 'n' is number of patients and 'C_j' defines risk factor can be low, medium, moderate, and risk.

3) Support Vector Machine

This model's primary objective is to create a precise linear or deterministic partition that divides n-proportional space into groups. Hyperplane is another name for this specific kind of accurate linear data sorting. With the use of demonstration graphs, which are separated into two different groups using either a hyperplane or a deterministic division. In order to linearly separate the information or data—which signifies separating the dataset into two distinct groups by a certain linear separation—the linear SVM is necessary. The classifier is denoted by the term "data," which is the linear differential:

Class 1 (Low risk) = (W * X + b) ≥ 1, ∀X
Class 2 (High risk) = (W * X + b) ≤ - 1 , ∀X

where, 'W' is a vector to Hyper plane, 'b' is a bias and 'x' is a matrix from dataset. Information is referred to as non-linear data and hence the classifier is written as:

$$K(X, Y) = (1 + X * Y)\ d \quad (3)$$

where, 'X' is data of low risk, 'Y' data of high risk and 'd' is degree of the polynomial. The RBF kernel represents a consequence that gives points relies upon the measured interval from the initial origination or from any point is written as:

$$K(X,X1) = \exp (-\|X - X1\| 2\ 2\ \sigma 2) \quad (4)$$

where, || X-X1 || defines the distances between the two risk vectors and

$$\text{let } \gamma = 1\ 2\ \sigma 2\ K(X,X1) = \exp(-\gamma \| X\text{-}X1 \| 2) \quad (5)$$

4) Decision Tree

The Decision apex and the Leaflet apex are the two apexes of the decision tree. In contrast, the leaflet apexes represent the results of these decision limbs and have no limbs linked to them. The Decision apex among these demonstrates the characteristics of the number of limbs attached to it and is essential in making decisions. Depending on the kind of dataset offered, a decision or implementation is made. The required scenarios are expressed in the form of a graph by drawing every potential result based on the choice or complexity. Since it originates at the source apex and develops in several directions,

the tree's structure is estimated as:

$$\text{Gini Index (G)} = \sum P\ c_i{=}1\ i\ (1 - P_i) \quad (6)$$

where, 'c' is number of classes, 'pi' defines the probability of class 'i', and 'G' becomes the root node which having more value.

5) Random Forest

The principle of collective learning is the foundation of random forest. It is a process that uses a variety of classifiers to improve how this technology is implemented in order to address repeated shortcomings. Using the random forest, which consists of numerous decision trees depending on various subsets within the provided dataset, to take into account the average in improving in detecting the speed of the dataset. As an alternative to relying solely on one decision tree, the random forest concludes every tree and uses the most possible identifications before displaying the identified result. The following formula can be used to determine the weight of each feature in a decision tree:

$$f_{ii} = \sum n_{ij}\ j{:}\ node\ j\ splits\ on\ feature\ i\ \sum n_{ik}\ k{\in}all\ nodes \quad (7)$$

where, 'f_{ii}' defines the 'i^{th}' feature importance, 'n_{ij}' defines the importance of node 'j'.

5. Conclusion

Following a review of the literature, we learned about the advantages and disadvantages of various research papers and consequently proposed a system that assists in the cost-effective and efficient prediction of brain strokes using a few user-provided inputs and trained machine learning algorithms. As a result, the system for predicting brain strokes has been created utilising five machine learning algorithms with a maximum accuracy of 98.56%. In order to provide a user interface that is both straightforward and effective while also showing empathy for both users and patients, the system was created in this manner. Future expansion of the system has the potential to produce better outcomes and an improved user experience. The user will benefit from this since they can save crucial time.

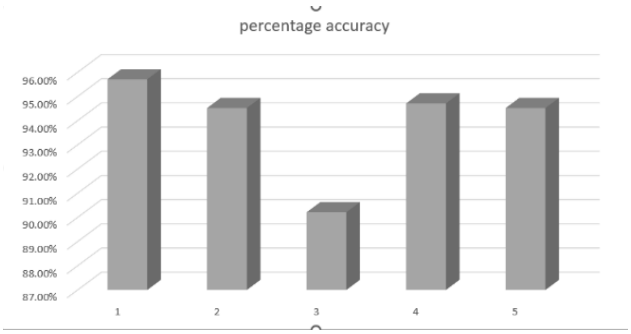

Fig. 4. Depicting accuracy of different algorithms used

6. Future Scope

Future goals of the work could include using a larger dataset or using the same model on multiple other datasets. In exchange for just providing some basic information, the artificial intelligence architecture may help the general public determine the likelihood that a stroke will occur in an adult patient, the associated risk level, and the determination of the likelihood that the condition will recur. It would help patients get early stroke therapy and recover from the incident in a perfect world. The implemented system's potential future range could include:

1. Improving the model's accuracy.
2. More details on brain strokes can be explained.
3. Providing people with the option to see their outputs based on their inputs.

References

[1] Neha Saxena, Arvind Choudhary, Deep Singh Bhamra, Preet Maru, "BrainOK: Brain Stroke Prediction using Machine Learning," April 2022.

[2] K. D. Mohana Sundaram, G. Haritha, A. Abhilash, K. Sona, E. Divya Sri, C. Bharath Kumar, "Detection of Brain Stroke Using Machine Learning Algorithm," 2022.

[3] Senjuti Rahman, Mehedi Hasan, and Ajay Krishno Sarkar, "Prediction of Brain Stroke using Machine Learning Algorithms and Deep Neural Network Techniques," 2023.

[4] Amol K. Kadam, Priyanka Agarwal, Nishtha, Mudit Khandelwal, "Brain Stroke Prediction Using Machine Learning Approach," 2022.

[5] Vamsi Bandi, Debnath Bhattacharyya, Divya Midhunchakkravarthy, "Prediction of Brain Stroke Severity Using Machine Learning," 2020.