

Prediction of Brain Stroke Using Machine Learning

Korcha Ramu.

Assitant Professor Of EEE
Sanskriithi School Of
Engineering(JNTUA)
Ananthapur,India.

Gorantla Shanthi Priya

Department Of EEE
Sanskriithi School Of
Engineering (JNTUA)
Ananthapur,India

Konapalli Nandasree

Department Of EEE
Sanskriithi School Of
Engineering (JNTUA)
Ananthapur,India

Ediga Jyoshna

Department Of EEE
Sanskriithi School Of
Engineering (JNTUA)
Ananthapur,India

K.Sai Krishna Sree

Department Of EEE
Sanskriithi School Of
Engineering (JNTUA)
Ananthapur,India

Abstract—A stroke is a medical condition in which poor blood flow to the brain results in cell death. It is now a day a leading cause of death all over the world. Several risk factors believe to be related to the cause of stroke has been found by inspecting the affected individuals. Using these risk factors, a number of works have been carried out for predicting the stroke diseases. Most of the models are based on data mining and machine learning algorithms. In this work, we have used five machine learning algorithms to detect the stroke that can possibly occur or occurred form a person's physical state and medical report data. We have collected a good number of entries from the hospitals and use them to solve our problem. The classification result shows that the result is satisfactory and can be used in real time medical report. We believe that machine learning algorithms can help better understanding of diseases and can be a good healthcare companion.

Keywords: Brain Stroke, Machine learning, Algorithms

Date of Submission: 08-05-2022

Date of acceptance: 23-05-2022

I. INTRODUCTION

1.1 Domain Description

Health is considered as an essential aspect of everyone's life, and there is a need for a recording system which tracks data on diseases and the relationship between them. Most of the information pertaining to diseases could be found in the case summaries of patients, medical records found in clinics and other records that are maintained manually. The sentences in them could be deciphered through various methodologies of text mining and machine learning (ML). Machine learning is a tool which can disseminate the content as a part of information retrieval in which semantic and syntactic parts of the content are given prevalence. Various ML and text mining methodologies are proposed and implemented for feature extraction and classification. Stroke is a term used by most of the healthcare practitioners to describe injuries in the brain and spinal cord resulting from abnormalities in the supply of blood. Stroke projects its meaning based on different perspectives; however, globally, stroke evokes an explicit visceral response. Machine learning can be portrayed as a significant tracker in areas like surveillance, medicine, data management with the aid of suitably trained machine learning algorithms. Data mining techniques applied in this work give an overall review about the tracking of information with respect to semantic as well as syntactic perspectives. The proposed idea is to mine patient's symptoms from the case sheets and train the system with the acquired data. Next, the case sheets were mined using tagging and maximum entropy methodologies, and the proposed stemmer extracts the common and unique set of attributes to detects the stroke disease. Then, the processed data were fed into various machine learning algorithms such as, Decision tree, Logistic Regression, K-Nearest Neighbors, Random Forest, Support vector machine. Among these algorithms, Support vector Machine achieves high accuracy.

1.2 About The Project

A brain comprises 100 billion and a trillion neurons and glia, respectively, wrapped into more than three pounds of tissue, which contains every memory and encodes and stores them in a network. Brain activity supports each and every individual's breath and movement. The number of people who lose their life due to stroke is ten times greater in developing countries for more than the past five decades (i.e., from 1970), and it is projected to double globally by 2030. Generally, Stroke is classified into the following three types: ischemic stroke (IS), hemorrhagic stroke (HE), and transient ischemic attack (TIA). Ischemic stroke is the most common type of stroke. The American Heart Association (AHA) has predicted that 87% of strokes are ischemic stroke, which occur if a clot or an obstacle persist in a blood vessel of the brain. Ischemic stroke has two categories: embolic stroke and thrombotic stroke. Embolic stroke occurs if a block/clot forms in any part of the body and moves toward the brain and blocks blood flow. Thrombotic stroke is due to a clot that weakens blood flow in an artery, which carries blood to the brain. Hemorrhagic stroke occurs from a split/burst of weakened blood vessels.

Only 10–15% of strokes are predicted to be a hemorrhagic stroke, but the rate of mortality is high when compared with ischemic stroke.

II. LITERATURE SURVEY

2.1 Badriyah, Tessy et al. Improving stroke diagnosis accuracy using hyper parameter optimized deep learning. *International Journal of Advances in Intelligent Informatics*.

Data can be analyzed and used as consideration for the decision making. It can be carried out with a variety of approaches such as using the Deep Learning method which is increasingly being used today because it is proven to be powerful in solving various problems. The forerunner of Deep Learning itself began in 1980 when Kunihiko Fukushima made Neocognition, the first model of the Convolutional Neural Network before being refined by Yann LeCun, Leon Bottou, Joshua Bengio and Patrick Haffner .

Stroke may cause death for anyone, including youngsters. One of the early stroke detection techniques is a Computerized Tomography (CT) scan. This research aimed to optimize hyperparameter in Deep Learning, Random Search and Bayesian Optimization for determining the right hyperparameter. The CT scan images were processed by scaling, grayscale, smoothing, thresholding, and morphological operation. Then, the images feature was extracted by the Gray Level Co-occurrence Matrix (GLCM). This research was performed a feature selection to select relevant features for reducing computing expenses, while deep learning based on hyperparameter setting was used to the data classification process. The experiment results showed that the Random Search had the best accuracy, while Bayesian Optimization excelled in optimization time.

2.2 C. L. Chin et al., An automated early ischemic stroke detection system using CNN deep learning algorithm, vol. 2018-Janua, no. iCAST.

Over the past few years, stroke has been among the top ten causes of death in Taiwan. Stroke symptoms belong to an emergency condition, the sooner the patient is treated, the more chance the patient recovers. The purpose of this paper is to develop an automated early ischemic stroke detection system using CNN deep learning algorithm.

2.3 C. Y. Hung, W. C. Chen, P. T. Lai, C. H. Lin, and C. C. Lee, Comparing deep neural network and other machine learning algorithms for stroke prediction in a large-scale population-based electronic medical claims database.

Electronic medical claims (EMCs) can be used to accurately predict the occurrence of a variety of diseases, which can contribute to precise medical interventions. While there is a growing interest in the application of machine learning (ML) techniques to address clinical problems, the use of deep-learning in healthcare have just gained attention recently. Electronic medical claims (EMCs) can be used to accurately predict the occurrence of a variety of diseases, which can contribute to precise medical interventions. While there is a growing interest in the application of machine learning (ML) techniques to address clinical problems, the use of deep-learning in healthcare have just gained attention recently. Deep learning, such as deep neural network (DNN), has achieved impressive results in the areas of speech recognition, computer vision, and natural language processing in recent years. However, deep learning is often difficult to comprehend due to the complexities in its framework. Furthermore, this method has not yet been demonstrated to achieve a better performance comparing to other conventional ML algorithms in disease prediction tasks using EMCs. In this study, we utilize a large population-based EMC database of around 800,000 patients to compare DNN with three other ML approaches for predicting 5-year stroke occurrence. The result shows that DNN and gradient boosting decision tree (GBDT) can result in similarly high prediction accuracies that are better compared to logistic regression (LR) and support vector machine (SVM) approaches. Meanwhile, DNN achieves optimal results by using lesser amounts of patient data when comparing to GBDT method.

III. PROPOSED SYSTEM

3.1 Problem Statement

Stroke is the second leading cause of death worldwide and remains an important health burden. Every 4 minutes someone dies of stroke, but up to 80% of stroke can be prevented if we can identify or predict the occurrence of stroke in its early stage.

3.2 Problem Description

In the medical field, brain stroke is detected by using deep learning technique which is very time consuming and do not produce accurate results. Therefore, to overcome this problem, an alternative way is to design the system that will automatically identify the presence of brain stroke by using health condition of a person using

algorithms in machine learning techniques, which also provides faster and accurate solutions which is very useful to save life's.

3.3 Proposed System

In this proposed system, we are using different machine learning algorithms are Decision tree, Logistic Regression, K-Nearest Neighbors, Random Forest, Support vector machine. In our proposed system we test these many algorithms with each other and select maximum accuracy model. Based on the accuracy score we will select the best model for our dataset. This section will describe the detailed description of the proposed work done for the detection of Brain Stroke.

3.4 Advantages of Proposed System

- It takes less time to compute results.
- It will deal with a large size of data where existing system can't.
- More flexible compared to existing system.

3.5 Block Diagram

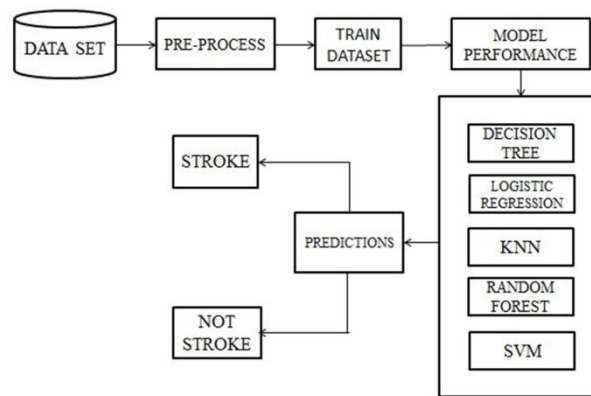


Fig 3.1 Block diagram of brain stroke detection

3.5.1 Data Set

A data set is a collection of data. In the case of tabular data, a data set corresponds to one or more database tables, where every column of a table represents a particular variable, and each row corresponds to a given record of the data set in question. The data set lists values for each of the variables, such as gender, age and bmi of the person. Data sets can also consist of a collection of documents or files.

The stroke prediction dataset was used to perform the study. There were 5110 rows and 12 columns in this dataset. The value of the output column stroke is either 1 or 0. The number 0 indicates that no stroke risk was identified, while the value 1 indicates that a stroke risk was detected. The probability of 0 in the output column (stroke) exceeds the possibility of 1 in the same column in this dataset. 249 rows alone in the stroke column have the value 1, whereas 4861 rows have the value 0. To improve accuracy, data pre processing is used to balance the data. Figure 3.2 shows the total number of stroke and non stroke records in the output column before pre processing.

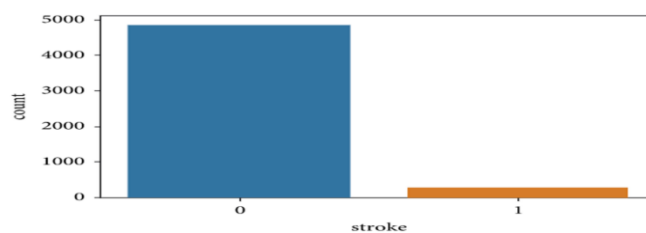


Fig 3.2 Total number of stroke and normal data

3.5.2 Pre Processing

Before building a model, data pre processing is required to remove unwanted noise and outliers 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. Figure 3.3 depicts the dataset's balance output column.

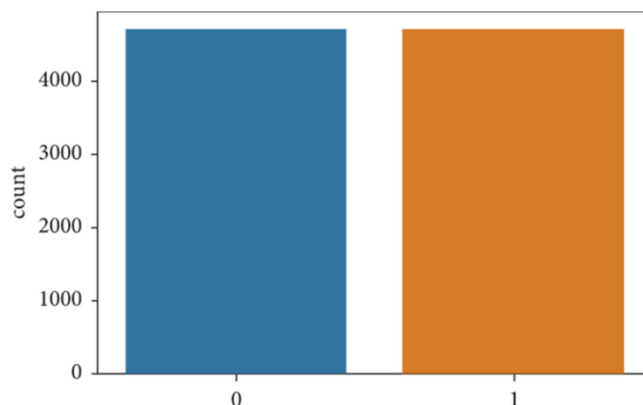


Fig 3.3 Output columns after processing

3.5.3 Train Dataset

- Training data (or a training dataset) is the initial data used to train machine learning models.
- Training datasets are fed to machine learning algorithms to teach them how to make predictions or perform a desired task.

3.6 Algorithms

3.6.1 Decision Tree

● Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. □

● In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.

- The decisions or the test are performed on the basis of features of the given dataset.

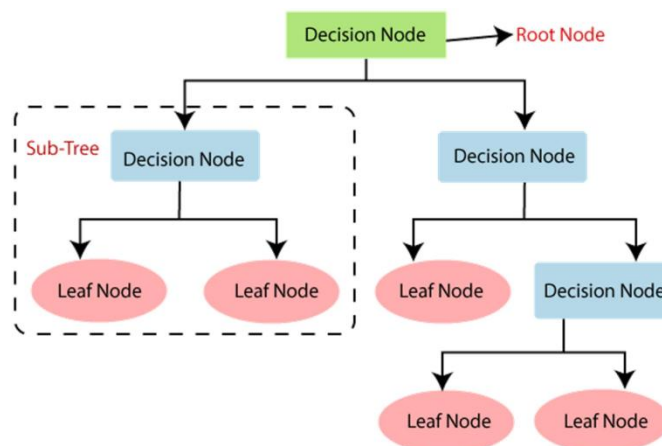


Fig 3.4 Structure of decision tree

3.6.2 Random Forest

Random forest is a supervised learning algorithm which is used for both classification as well as regression. But however, it is mainly used for classification problems. As we know that a forest is made up of trees and more trees means more robust forest. Similarly, random forest algorithm creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution by means of voting.

It is an ensemble method which is better than a single decision tree because it reduces the over-fitting by averaging the result.

Working of Random Forest Algorithm

We can understand the working of Random Forest algorithm with the help of following steps

1. First, start with the selection of random samples from a given dataset.
2. Next, this algorithm will construct a decision tree for every sample. Then it will get the prediction result from every decision tree.
3. In this step, voting will be performed for every predicted result.
4. At last, select the most voted prediction result as the final prediction result.

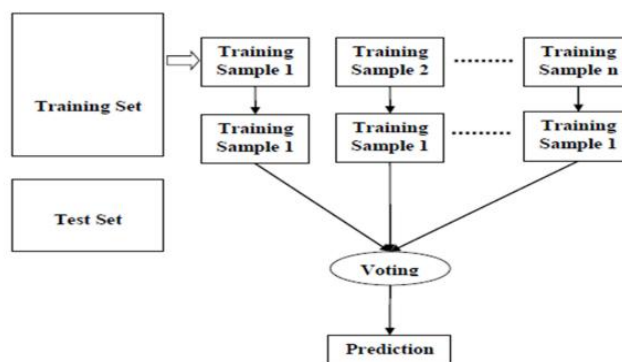


Fig 3.5 Example of Random Forest

3.6.3 K-Nearest Neighbor(KNN)

- K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique.
- K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.
- K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.

The K-NN working can be explained on the basis of the below algorithm:

1. Select the number K of the neighbours
2. Calculate the Euclidean distance of K number of neighbors
3. Take the K nearest neighbors as per the calculated Euclidean distance.
4. Among these k neighbors, count the number of the data points in each category.
5. Assign the new data points to that category for which the number of the neighbor is maximum.
6. Our model is ready.

Suppose we have a new data point and we need to put it in the required category. Consider the below image:

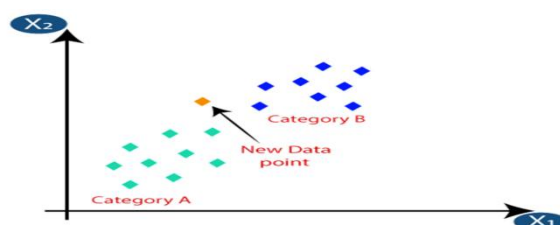


Fig 3.6 Example of KNN

- Firstly, we will choose the number of neighbors, so we will choose the k=5.
- Next, we will calculate the Euclidean distance between the data points. The Euclidean distance is the distance between two points, which we have already studied in geometry. It can be calculated as:

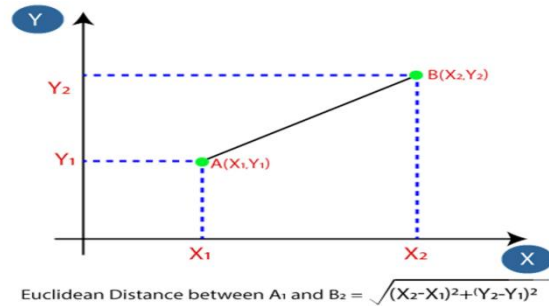


Fig 3.7 Finding Euclidean distance

- By calculating the Euclidean distance we got the nearest neighbors, as three nearest neighbors in category A and two nearest neighbors in category B.
-
- Consider the below image:

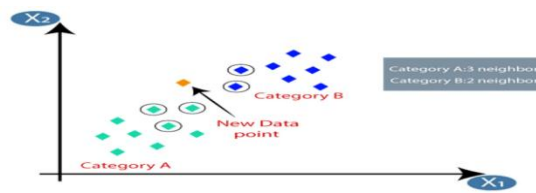


Fig 3.8 Finding near neighbors

- As we can see the 3 nearest neighbors are from category A, hence this new data point must belong to category A.

3.6.4. Logistic regression

Logistic regression is a Machine Learning classification algorithm that is used to predict the probability of a categorical dependent variable. In logistic regression, the dependent variable is a binary variable that contains data coded as 1 (yes, success, etc.) or 0 (no, failure, etc.). In other words, the logistic regression model predicts $P(Y=1)$ as a function of X .

Step 1: Logistic regression hypothesis

The logistic regression classifier can be derived by analogy to the logistic regression the function $g(z)$ is the logistic function also known as the *sigmoid function*. The logistic function has asymptotes at 0 and 1, and it crosses the y-axis at 0.5.

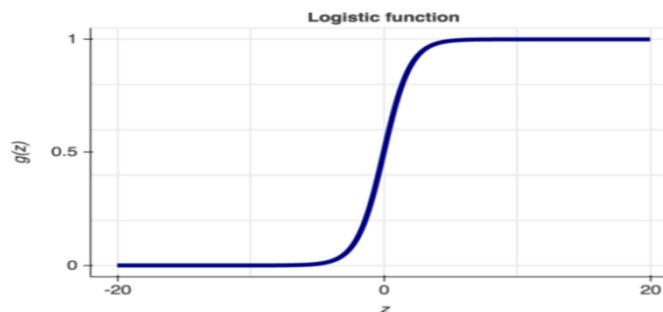


Fig 3.9 Logistic regression hypothesis

Step 2: Logistic regression decision boundary

Since our data set has two features: height and weight, the logistic regression hypothesis is the following:

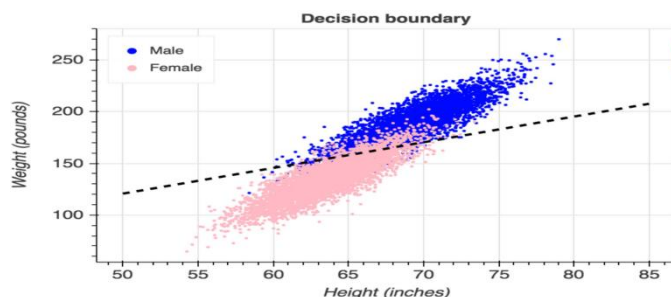


Fig 3.10 Logistic regression decision boundary

3.6.5 Support Vector Machine

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane. SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Machine. Consider the below diagram in which there are two different categories that are classified using a decision boundary or hyperplane:

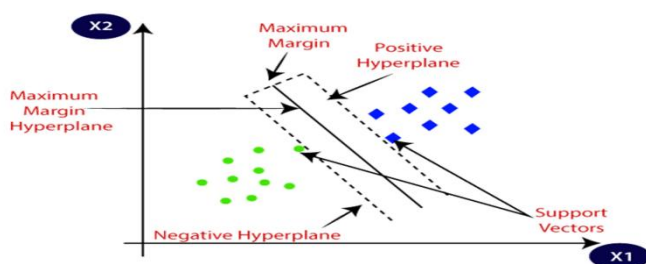


Fig 3.11 Boundaries of SVM

3.7 Steps for executing the Project

1. Install the required packages
2. Load the datasets.
3. Pre-process the data.
4. Split the dataset into train and test.
5. Use the train dataset to train the ml models.
6. Use the test data to test the model for prediction and accuracy generation.

IV. HARDWARE DESCRIPTION

- Processor : I3 or above processors
- Hard Disk : 160GB
- Key Board : Standard Windows Keyboard
- Mouse : Two or Three Button Mouse
- Monitor : SVGA
- RAM : 8Gb

V. SOFTWARE DESCRIPTION

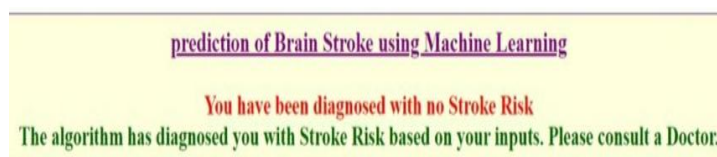
- Operating System : Windows 7/8/10
- Server side Script : Python
- IDE : PyCharm, Anaconda
- Libraries Used : SKlearn, pandas, numpy, Matplotlib, Collections
- Technology : Python 3.6
- Front End : HTML
- Dataset : Kaggle

VI. APPLICATIONS

1. Hospital applications
2. Medical appliances
3. Ayurveda treatments

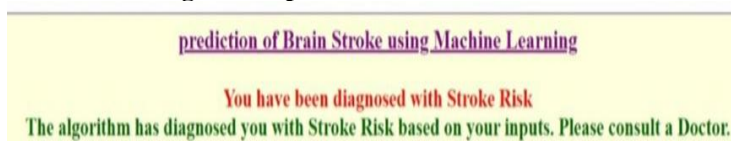
VII. RESULTS

- This study predicts the stroke for a patient based on classification methodologies.
- This study brings out the effectiveness of the classification methods for structured entities like patient case sheets to detect the stroke based on the parameters (symptoms) and factors.



The screenshot shows a yellow background with a purple header "prediction of Brain Stroke using Machine Learning". Below the header, there are two lines of text: "You have been diagnosed with no Stroke Risk" in red, and "The algorithm has diagnosed you with Stroke Risk based on your inputs. Please consult a Doctor." in green.

Fig 7.1 Output for no stroke detection



The screenshot shows a yellow background with a purple header "prediction of Brain Stroke using Machine Learning". Below the header, there are two lines of text: "You have been diagnosed with Stroke Risk" in red, and "The algorithm has diagnosed you with Stroke Risk based on your inputs. Please consult a Doctor." in green.

Fig 7.2 Output for stroke detection

REFERENCES

- [1]. P. Govindarajan, R. K. Soundarapandian, A. H. Gandomi, R. Patan, P. Jayaraman, and R. Manikandan, "Classification of stroke disease using machine learning algorithms," *Neural Computing and Applications*, pp. 1–12.
- [2]. R. Jeena and S. Kumar, "Stroke prediction using svm," in *2016 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT)*, pp. 600–602, IEEE, 2016.
- [3]. S. Y. Adam, A. Yousif, and M. B. Bashir, "Classification of ischemic stroke using machine learning algorithms," *Int J Comput Appl*, vol. 149, no. 10, pp. 26–31, 2016.
- [4]. A. Sudha, P. Gayathri, and N. Jaisankar, "Effective analysis and predictive model of stroke disease using . 26– 31, 2012.