Detection of Brain Stroke Using Machine Learning

Under The Supervision of
MR. RAVI SHARMA
(ravi.sharma@galgotiasuniversity.edu.in)
Prashant Kumar[1], Vinayak Kumar[2], Wazida Tabbasum[3]
Department of ComputerScience Engineering Galgotias University

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

I. Introduction

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 entropymethodologies, 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.

II. Literature Survey

Badriyah, Tessy et al. 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. N. Venketasubramanian et al.,Stroke is a major cause of death and disability in many countries. It was reported that, in 2013, globally, there were nearly 25.7 million stroke survivors, 6.5 million deaths due to stroke, 113 million disability-adjusted life-years (DALYs) lost because of stroke, and 10.3 million new cases of strokes
. A majority of the stroke burden was observed in developing countries, accounting for 75.2% of all stroke- relateddeaths and 81.0% of the associated DALYs lost.

G. A. P. Singh et al., Lung cancer is one of the most common causes of death among all cancer-related diseases (Cancer Research UK in Cancer mortality for common cancers. Automated classification of lung cancer is one of the difficult tasks, attributing to the varying mechanisms used for imaging patient's lungs.

C. L. Chin et al., 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 patientrecovers. The purpose of this paper is to develop an automated early ischemic stroke detection system using CNNdeep learning algorithm

III. Materials and methods

3.1. Materials

a) 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.

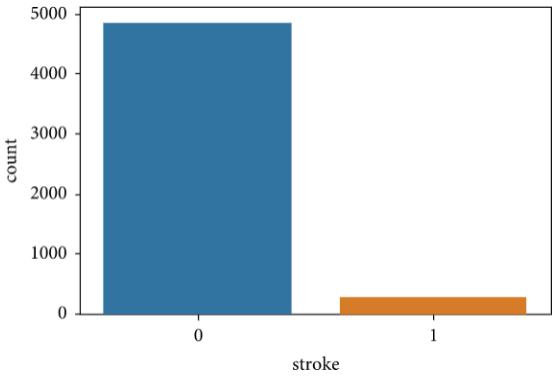


Fig 1: Total number of stroke and normal data

b) 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.

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.

C) Algorithms

i) 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.

ii) 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.

iii) K-Nearest Neighbor(KNN):

1. K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learningtechnique.
2. K-NN algorithm assumes the similarity between the new case/data and available cases and put the newcase into the category that is most similar to the available categories.
3. K-NN algorithm stores all the available data and classifies a new data point based on the similarity. Thismeans when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.
4. K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for theClassification problems.
5. KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifiesthat data into a category that is much similar to the new data.

iv) 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.

v) 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 Vector Machine.

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.

3.2. Methods

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.

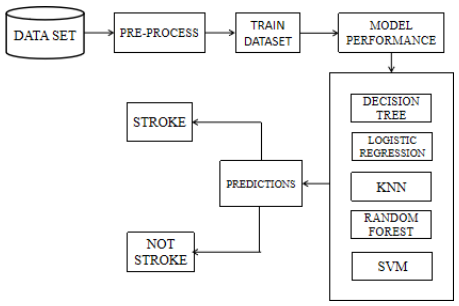


Fig 2: Block Diagram of Proposed System

vi)Hardware Requirements

The hardware requirements we are used in this project are:

4.1. Monitor:

A computer monitor is an output device that displays information in pictorial or text form. A monitor usually comprises a visual display, some circuitry, a casing, and a power supply. The display device in modern monitors is typically a thin film transistor liquid crystal display (TFT-LCD) with LED backlighting having replaced cold- cathode fluorescent lamp (CCFL) backlighting as shown in Fig 4.4.1.. Monitors are connected to the computer via VGA, HDMI, Display Port, USB-C, lowvoltage differential signaling (LVDS) or other proprietary connectors and signals.

4.2. Processor:

For the most part, you'll get faster CPU performance from the Core i5 parts over Core i3. Some Core i5 processors like in Fig 4.1.1 are dualcore and some are quad-core. Most of the time, a true quad-core CPU will perform better than a dualcore processor, especially on multimedia tasks like video transcoding or photo editing. All Core i3 processor are dual core. Occasionally, you'll find an older Ivy Bridge processor like the Intel Core i3- 3130M in a system that's the same price as system with a newer Haswell CPU like the Intel Corei3- 4012Y.

4.3. Hard Disk:

A computer's hard drive is a device consisting of several hard disks, read/write heads, a drive motor to spin the disks, and a small amount of circuitry, all sealed in a metal case to protect the disks from dust. In addition to referring to the disks themselves, the term hard disk is also used to refer to the whole of a computer's internal data storage. Beginning in the early 21st century, some personal computers and laptops were produced that used solid-state drives (SSDs) that relied on flash memory chips instead of hard disks to store information

4.4. RAM:

With 8 GB of RAM, you will have enough memory to run several programs at once. You can open lots of browser tabs at once, use photo or video editing programs, stream content, and play mid-tohigh-end games

v)Software Requirements

5.1 PYTHON:

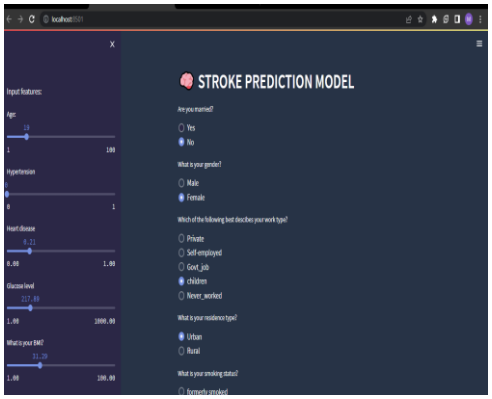
Python is a general purpose, dynamic, high level and interpreted programming language. It supports Object Oriented programming approach to develop applications. It is simple and easy to learn and provides lots of high- level data structures. It is easy to learn yet powerful and versatile scripting language which makes it attractive forApplication Development.

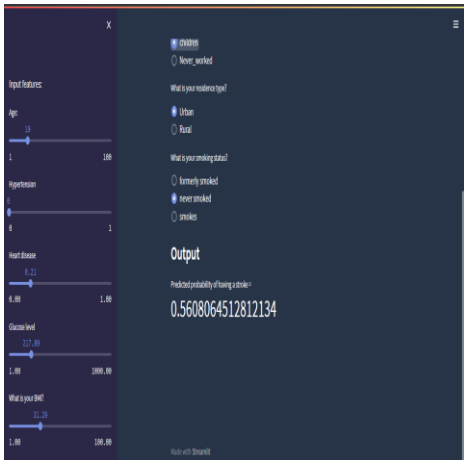
5.2 HTML:

The Hyper Text Markup Language or HTML is the standard markup language for documents designed to be displayed in a web browser. It can be assisted by technologies such as Cascading Style Sheets (CSS) and scripting languages such as Java Script. Web browsers receive HTML documents from a web server or from local storageand render the documents into multimedia web pages. HTML describes the structure of a web page semantically and originally included cues for the appearance of the document.

Result

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 patientcase sheets to detect the stroke based on the parameters (symptoms) and factors.





References

[1]. S. H. Pahus, A. T. Hansen, and A.-M. Hvas, “Thrombophilia testing in young patientswith ischemic stroke,” Thrombosis research, vol. 137, pp. 108–112, 2016.

[2]. 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.

[3]. L. T. Kohn, J. Corrigan, M. S. Donaldson, et al., To err is human: building a safer healthsystem, vol. 6. National academy press Washington, DC, 2000.

[4]. R. Jeena and S. Kumar, “Stroke prediction using svm,” in 2016 International Conferenceon Control, Instrumentation, Communication and Computational Technologies (ICCICCT), pp. 600–602, IEEE, 2016.

[5]. M. S. Singh and P. Choudhary, “Stroke prediction using artificial intelligence,” in 20178th Annual Industrial Automation and Electromechanical Engineering Conference (IEMECON), pp. 158–161, IEEE, 2017.

[6]. A. Sudha, P. Gayathri, and N. Jaisankar, “Effective analysis and predictive model ofstroke disease using classification methods,” International Journal of Computer Applications, vol. 43, no. 14, pp. 26– 31, 2012.