International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE)
ISSN: 0976-1353 Volume 30 Issue 2 – March 2023.

BRAIN STROKE DETECTION USING MACHINE LEARNING

Dr.J.Mary Dallfin Bruxella.,Assistant professor,
Department of Computer Science and Information Technology,
KARE,
Krisnankovil,Tamilnadu.
Email ID: j.marydallfinbruxella@klu.ac.in

S.Rajayogha.,MSc Datascience,
Department of Computer Science and Information Technology,
KARE,
Krishnankovil,Tamilnadu.
Email ID: yogha961@gmail.com

.Abstract:A stroke is a medical disorder in which the brain's blood supply is disrupted, causing cell death. It is currently themain cause of death in many parts of the world. By examining the afflicted individuals, several risk variables believed to be connected to the etiology of stroke have been discovered. A lot of studies have been conducted using these risk variables to predict and diagnose stroke disorders. The majority of the models are built on machine learning and data mining technologies. In this work, we have used two machine learning algorithms like XG Boost algorithm and a Decision tree, to detect the type of stroke that can possibly occur or occurred from a person's physical state and medical report data. We have to collect a good number of entries from the hospitals and use them to solve our problem. Ml algorithms, we believe, can aid in improved illness knowledge and can be a useful healthcare companion.

Keywords:*Decision Tree, XGBoost, Machine Learning,Diagnosis, Stroke*

I. INTRODUCTION

Well-being is viewed as a significant part of everybody's life,and there is a requirement for a framework that keeps upwith information on sicknesses and their connections. Most infection-related data might be found in persistent case synopses, clinical records kept in facilities, and other physically kept up with information. Text mining and AI methods may be utilized to comprehend the texts in them(ML).

AI is a component for scattering material as the featured data recovery, with an emphasis on the semantic and syntactic parts of the substance. To include extraction and arrangement, a few ML and message mining approaches are introduced and applied. Most medical management professionals use the word "stroke" to describe damage tothe brainand spinal cordcaused by irregularities in blood flow.

In recent times, stress levels in individuals are at an all-time high. This increases the chances of strokes in individuals. In 2015, after coronary artery disease, stroke was the second most frequent cause of deathaccounting for 6.3 million deathsThe figures today are much higher than this. An exhaustive and easy-to-use tool is much needed for the detection of strokes. With the advancement of computer science in different research areas including medical sciences, this has been made possible. A machine-learning systemis trained rather than explicitly programmed, as it provides a better choice for achieving high accuracy in the detection of heart diseases. Medical organizations, all around the world, collect data on various health-related issues. These data can be exploited using various machine-learning techniques to gain useful insights. But the data collected is very massive and, many times, this data can be very noisy. These datasets, which are too overwhelming for human minds to comprehend, can be easily explored using various machine-learning techniques. Thus, these algorithms have become very useful, in recent times, to predict the presence or absence of heart-related diseases accurately.Worldwide, Stroke is the leading cause of death and it remains an important health burden both formales and females and for the national healthcare systems. The main aim of this project is to build an efficient prediction model and deploy it in the prediction of disease. Machine Learning is a faster-emerging technology of Artificial Intelligence that contributes various algorithms like Logistic Regression, SVM, Random Forests, and many more which is effective in making decisions and predictions from the large

quantity of data produced by the healthcare industry. Based on the proposed problem, ML provides different classification algorithms to divine the probability of a patient having a Brain Stroke.

Architecture

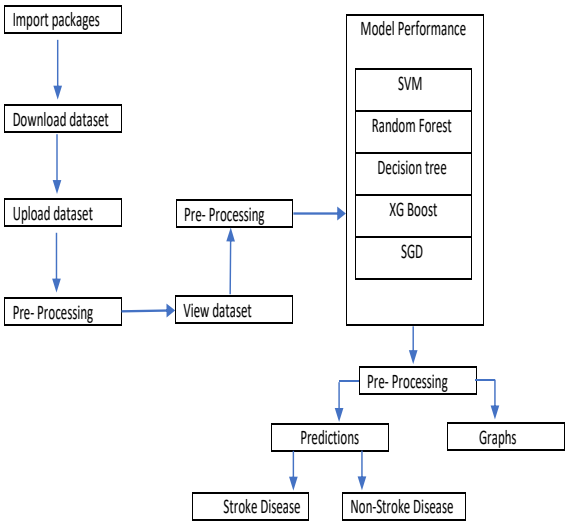

Fig1. Architecture

II. LITERATURE REVIEW

The standards mainly focus on the demands of the project that arisewhen scientific results when the machine learning model is applied to the construction of a software system. It is targeted to aid the clinical work of medical experts.This standard includes Machine Learning techniques by automated medical data processing and generation of graphs in order to analyze the stroke. When the pre-processed data is given as input, automatically themodel gets to divide the data by trained one in which it predicts the occurrence of a stroke. Finally, the region of graphs is segmented by the optimization techniques are compared with the standard trained data obtained by the neurologist to check the efficiency of the suggested methodologies and the segmented results can be reconstructed for further analysis to get better accuracy in the detection of stroke.

A brain stroke occurs when a brain artery bursts, resulting in bleeding, and can be devastating to brain function and performance. Medical professionals recommend using either an MRI or a CT scan to diagnose. Because of its ease of use, low cost, and rapid speed, CT imaging is employed in growing numbers. In the present system, the data mining techniques used in this study provide an overview of information tracking from both a semantic and syntactic standpoint. Malware detection techniques include Data Mining, Deep, Hypothesis Exploration, and others. In any case, one of the most generally-used procedures for recognizing malware is AI. Malware detection methods are divided into two groups. We utilize the XG Boost and Decision Tree AI calculations in the proposed framework. We give more exact results in our proposed framework.

C. L. Chin et al [1] considered that a brain hemorrhage occurs when a brain artery bursts, resulting in bleeding; and can be devastating to brain function and performance. Medical professionals recommend using either an MRI or a CT scan to diagnose a hemorrhage.As a result, we'll require a system that can segment CT scan images fast and automatically. The objective is to use the Deep Learning approach to swiftly and precisely segment the brain area that is damaged by hemorrhage. As a result, individuals with cerebral hemorrhage can receive medical care as soon as possible.

C. Y. Hung [2] Compared the deep neural network and other machine learning algorithms for stroke prediction in a large-scale population-based electronic medical claims database.The findings suggest that DNN and gradient boosting decision tree (GBDT) can achieve similar high prediction accuracies as logistic regression (LR) and support vector machine (SVM) techniques. Meanwhile, as compared to the GBDT approach, DNN produces ideal outcomes by utilizing fewer patient data.

N.VenketaSubramanian [3] used to say that Stroke is an indisputable justification behind mortality and failure in various countries. In 2013, there were over 6.5 million stroke fatalities, 113 million insufficiency-changed life-years (DALYs) lost in light of stroke, and 10.3 million new stroke cases, according to the World Health Organization.

Badriyah,Tessy, et al [4] processed the CT scan images by scaling, grayscale, smoothing, thresholding, and morphological operation. Then, the image feature was extracted by the Gray Level Co-occurrence Matrix (GLCM). This research performed a feature selection to select relevant features for reducing computing expenses, while deep learning based on a hyperparameter setting was used for the data classification process. The experiment results showed that the Random Search had the best accuracy, while Bayesian Optimization excelled in optimization time.

III. METHODOLOGY

EXISTING METHODOLOGY

In the present system, the data mining techniques used in this study provide an overview of information tracking from both a semantic and syntactic standpoint. Malware detection techniques include Data Mining, Deep, Hypothesis Exploration, and others. In any case, one of the most generally used procedures for recognizing malware is AI. Malware detection methods are divided into two groups. The first is the classic signature-based technique, in which malware is identified by its signature. The subsequent one and the new methodology utilized for malware location is the conduct-based methodology in which the malware is identified dependent on the exercises it is planned to perform on the framework it is attempting to assault.

Disadvantages:

1. Less accurate results.
2. Difficult to scale up.
3. Time-consuming.

PROPOSED METHODOLOGY

We utilize the XG Boost and Decision Tree AI calculations in the proposed framework. We give more exact results in our proposed framework. The main aim of this project is to build an efficient prediction model and deploy it forthe prediction of disease. Machine Learning is a faster-emerging technology of Artificial Intelligence that contributes various algorithms like Logistic Regression, SVM, Random Forests,and many more which is effective in making decisions and predictions from the large quantity of data produced by the healthcare industry. Based on the proposed problem, ML provides different classification algorithms to divine the probability of a patient having a Brain Stroke.

Advantages:

- It takes less time to compute results.
- More flexible compared to the existing system.

Decision tree classifier

Three machine-learning algorithms to filter spam from valid reviews with low error rates and high efficiency using a multilayer perception model. Several widely used techniques include decision tree classifier multilayer perceptron and Naïve Bayes classifier.

The Output of a C4.5 decision tree classifier is structural data in the form ofa binary tree. A C4.5 tree is modeled as follows. A training set is a set of base tuples to determine classes related to these tuples. A tuple X is represented by an adjective vector X = (x1, x2, …, xn). Assume that a tuple belongs to a predefined class that is determined by an adjective called a class label. The training set is randomly selected from the base; this step is called the learning step. This technique is very efficient and extensively uses classification. The structure of

the tree can be implemented with the following factors:

1. A node of the tree represents a test on an adjective;
2. A branch exiting from a node represents possible outputs of a test;
3. A leaf represents a class label.

A decision tree includes a rule set by which objective functions can be predicted. The algorithm used for this model uses greedy techniques

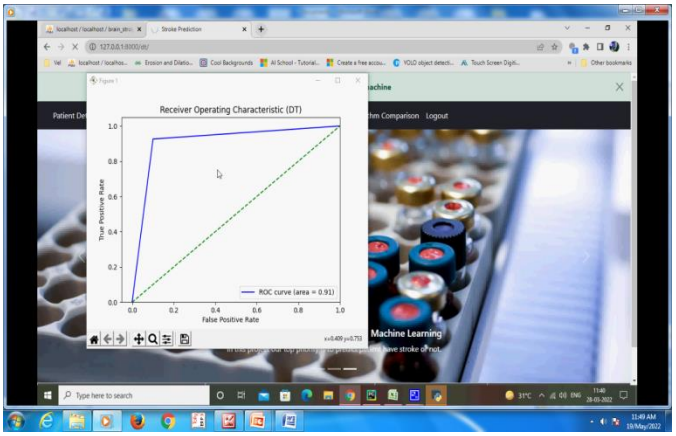

Fig2. Decision Tree

Logistic Regression:

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression is estimating the parameters of a logistic model. Mathematically, a binary logistic model has a dependent variable with two possible values, such as pass/fail which is represented by an indicator variable, where the two values are labeled "0" and "1". In the logistic model, the logodds (the logarithm of the odds) for the value labeled "1" is a linear combination of one or more independent variables ("predictors"); the independent variables can each be a binary variable (two classes, coded by an indicator variable) or a continuous variable (any real value).

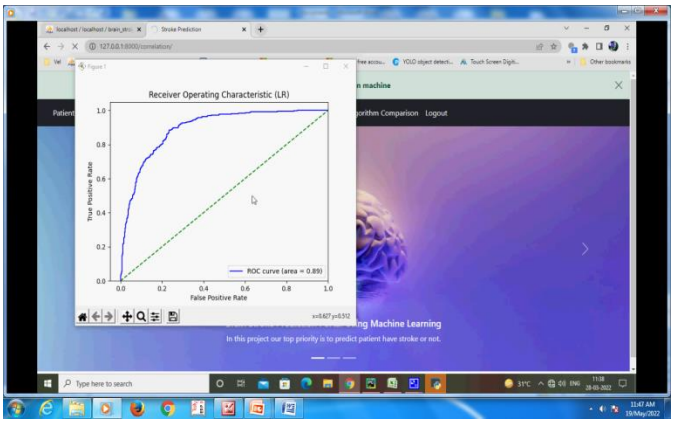

Fig3. Logistic Regression

Random Forest:

Random Forest is a mainstream ML calculation that has a place with the administered learning procedure. It tends to be utilized for both Arrangement and Relapse issues in ML. It depends on the idea of ensemble learning, which is a cycle of joining different classifiers to take care of a complex problem,

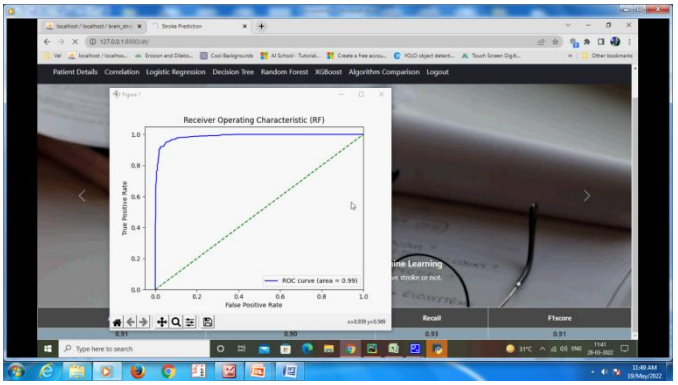

Fig4. Random forest

Extreme Gradient Boosting Classifier:

XGBoost is a decision-tree-based Machine Learning algorithm that uses a gradient-boosting framework. In prediction problems, unstructured data involves (images, text, etc.) artificial neural networks tend to perform all other algorithms and frameworks. However, decision tree-based algorithms are considered best-in-class right now,whenit comes to small-to-medium structured/tabulardata.

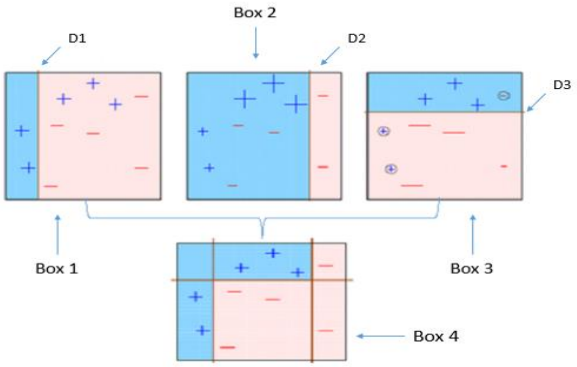

Fig5. XGBoost Process

Extreme Gradient Boosting, or XG support, is an adaptation of the Gradient Boosting strategy that utilizes more accurate approximations to decide the best tree model. It incorporates a lot of useful techniques that make it extremely successful, especially when dealing with structured data.

The most important are 1). computing second-order gradients: That is the second partial derivative of the loss function, which gives us further information about the gradient direction and how to get to the loss function's minimum. XG Boost utilizes the second request subsidiary as an estimation for decreasing the mistake of the general model, while standard inclination helping utilizes the misfortune capacity of our base model (for example choice tree) as an intermediary for limiting the mistake of the general model. 2). Advanced regularization: whichimproves modelgeneralization.Additional advantages: Training is quick and may be dispersed over several clusters.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

System Modules:

1) Upload data set (System takes dataset uploaded by the user).

2) View the data set (to View the data uploaded by the user).

3) Training (to train the model by the uploaded data).

4) Pre-Processing (minimizing the gaps to reduce the errors).

5) Model Performance (Selection of model to view the performance).

6) Prediction (To enter the entity Values in which the result would be calculated by the model).

7) Analyzation of Graphs (Graphs can be generated by the system and the user can view the graphs).

First, we have to upload the data set to the model so it would be trained by that after that we can enter the parameters of the particular patient whom we have to classify whether he/she would be attacked by the strokeor not, which is shown below infig i.

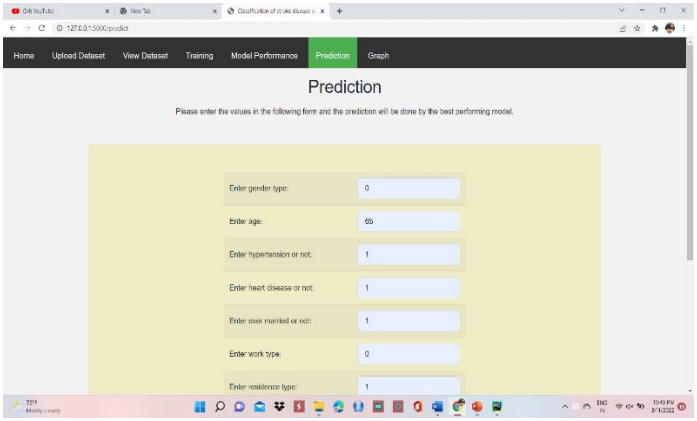

Fig6. Entering of Parameter Values

Here, fig ii shows the result obtained after the entering of parameters. It classifieswhether the particular patient would be affected by a brain stroke or not.

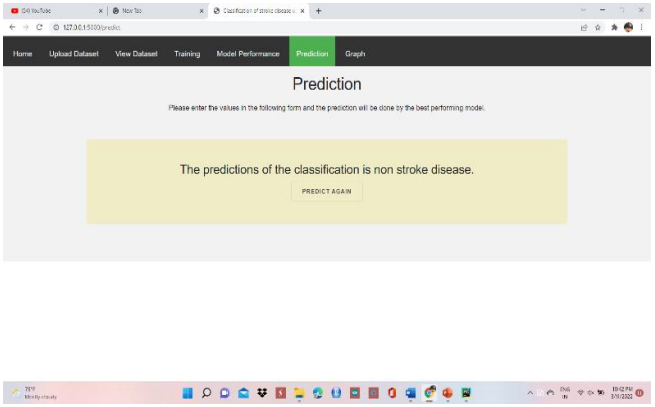

Fig7. Classified Result

And here the accuracy of the result would be shown by the graphs.

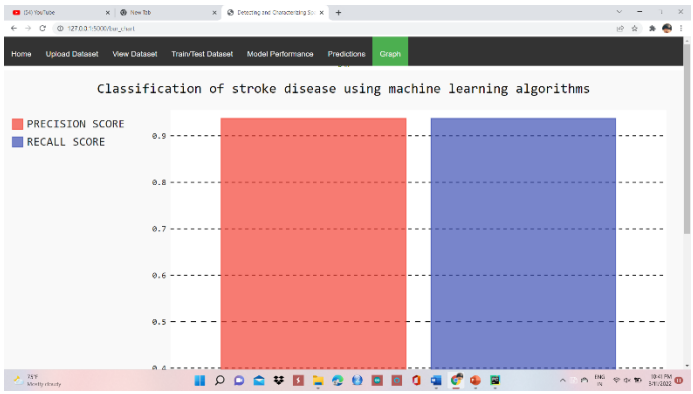

Fig8. Graphical Result

Balancing Dataset:

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 preprocessing is used to balance the data. It contains the total number of strokes and no stroke records in the output column before preprocessing.

Preprocessing

Before building a model, data preprocessing 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.

Correlation Matrix:

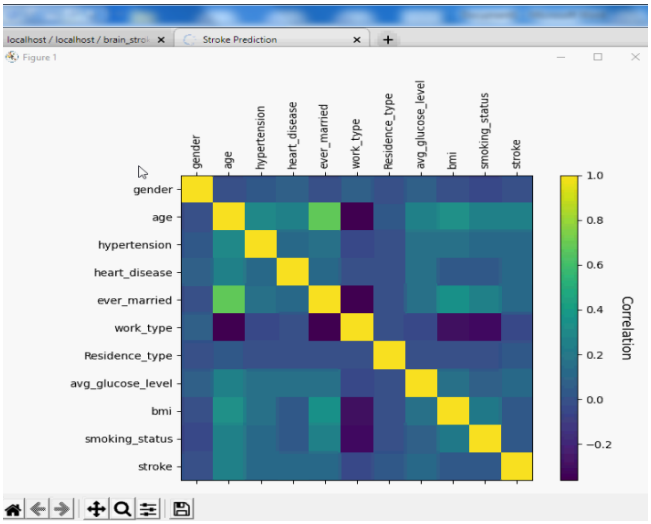

Fig9. Correlation Result

In the above heatmap, we can see that there is no multicollinearity present, 'Age' and 'Glucose Level' are some of the highest correlated features with 'Stroke'.

Best Features using Chi-Square Test

In the above table, we can see that Age, Average Glucose Level,and Hypertension are the top 3 features having the maximum impact on output 'Stroke'. Chi-Square Test is used to find out this result.

Evaluation Matrix

A Confusion matrix is a tool for evaluating the performance of machine learning classification algorithms. The confusion matrix has been used to test the efficiency of all models created. The confusion matrix illustrates how often our models forecast correctly and how often they estimate incorrectly. False positives and negatives have been allocated to badly predicted values, whereas true positives and negatives were assigned to properly anticipated values. The model's accuracy, precision-recall trade-off, and AUC were utilized to assess its performance after grouping all predicted values in the matrix

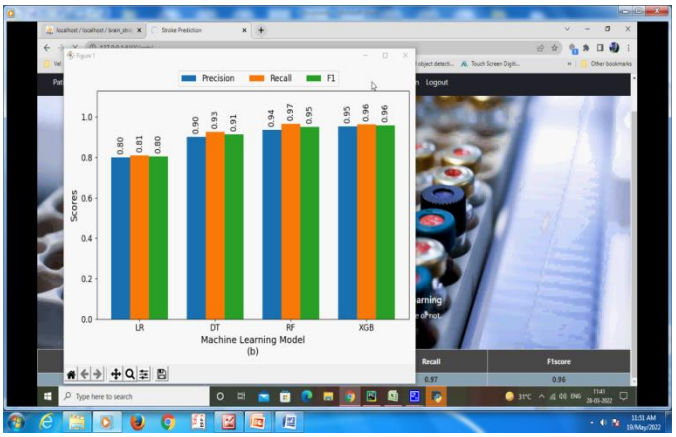

Fig10. Algorithm Comparison

V. CONCLUSION

The study brings out the effectiveness of the classification methods for structured entities like patient case sheets to classify strokes based on defined parameters (symptoms) and factors.

This study predicts the type of stroke for a patient based on classification methodologies This study indicates that stroke is more prevalent in men than in women and in the age group from 40 to 60 years old. Patients who suffered an ischemic stroke were greater in number than patients with hemorrhagic stroke. Determining the type of stroke depends not only on the impact of modifiable and non-modifiable risk factors of the patient but also on the individual patient's symptoms.The importance of knowing and understanding the risks of brain stroke is very much in these trying times. The model predicts the probability of brain stroke on the basis of very trivial day-to-day and known to all parameters. This makes this project highly relevant and of need to society. The objective of implementing the project on a web platform was to reach as many individuals as possible. The early warning can save someone's life who might have a probability of a stroke.

REFERENCES

[1] Roger VL, Go AS, Lloyd-©, Nichol G, Paynter NP, Soliman EZ, Sorlie PD, Sotoodehnia N, Turan TN, Virani SS, Wong ND, Woo D, Turner MB (2012) Executive summary: heart disease and stroke statistics—2012 update: a report. Circulation 125(1):188–197

[2] https://www.strokejournal.org/article/S1052-3057(19)30523-3/fulltext#seccesectitle0006

[3] Pahus SH, Hansen AT, Hvas AM (2016) Thrombophilia testing in young patients with Ischemic stroke. Thromb Res 137:108–112

[4] Dupont SA, Wijdicks EF, Lanzino G, Rabinstein AA (2010) Aneurysmal subarachnoid hemorrhage: an overview for the practicing neurologist. Semin Neurol 30(5):45–54

[5] Classification of stroke disease using machine learning algorithms Priya Govindarajan1 • Ravichandran Kattur Soundarapandian2 • Amir H. Gandomi3 • Rizwan Patan4 • Premaladha Jayaraman2 • Ramachandran Manikandan2

[6] "Computer Methods and Programs in Biomedicine" - Jae–woo Lee, Hyunsun Lim.

[7] "Strokepredictionusing artificial intelligence"- M. Sheetal Singh, Prakash Choudhary. (IEEE - 2017)

[8] "Deep learning algorithms for detection of critical findings in head CT scansa retrospectivestudy" - Rohit Ghosh, Swetha Tanamala, Mustafa Biviji, Norbert G Campeau, Vasantha Kumar Venugopal.

[9] "Prediction of stroke thrombolysis outcome using CT brain machine learning" - Paul Bentley, JebanGanesalingam, AnomaLalani, CarltonJones, KateMahady,SarahEpton,PaulRinne, Pankaj Sharma,OmidHalse, AmrishMehta, DanielRueckert

[10] "Probabilityof Stroke: A Risk Profile fromthe Framingham Study" - Philip A. Wolf, MD; Ralph B. D'Agostino, Ph.D., Albert J. Belanger, MA and William B. Kannel, MD.