iJRASET
International Journal For Research in Applied Science and Engineering Technology

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

**Volume:** 11 **Issue:** III **Month of publication:** March 2023

**DOI:** https://doi.org/10.22214/ijraset.2023.49651

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
*ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538*
*Volume 11 Issue III Mar 2023- Available at www.ijraset.com*

# Early Prediction of Brain Stroke Using Logistic Regression

S. Rajayogha[1], Dr. J. Mary Dallfin Bruxella[2]
[2]*Assistant Professor, Department of Computer Science and Information Technology Kalasalingam Academy of Research and Education, 626126, Tamil Nadu, India*

***Abstract: A stroke is a condition where there is an interruption in the blood flow to the brain, which results in cell death. In numerous regions of the world, it is today a major cause of death.***
***Many risk factors thought to be related to the etiology of stroke have been identified by studying the affected individuals. These risk factors have been used in numerous research to forecast and identify stroke problems.***
***The vast majority of the models are constructed using data mining and machine learning techniques. In this study, we used two machine learning algorithms, the XG Boost algorithm, and the Decision tree, to identify the type of stroke that would have occurred or had already occurred based on a person's physical condition and data from medical***
***Keywords: Decision Tree, Xgboost, Machine Learning, Diagnosis, Stroke***

## I. INTRODUCTION

Everyone's well-being is an important aspect of life, and there is a need for a framework that is updated with knowledge about illnesses and their links. Most of the infection-related information can be found in ongoing case summaries, clinical records preserved at facilities, and other physically updated data.

The texts included therein may be understood using text mining and AI technique AI is a component for recovering data from scattered sources with a focus on the semantic and syntactic components of the data. Several ML and message mining techniques are presented and used for include extraction and arrangement.

The term "stroke" is most used by medical management experts to denote damage to the brain and spinal cord brought on by anomalies in blood flow.

This has been made possible by the development of computer science in numerous scientific fields, including the medical sciences. To achieve high accuracy in the identification of heart disorders, a machine-learning system is trained rather than explicitly designed.

Worldwide, medical organizations compile information on a range of health-related topics. To extract insightful knowledge from these data, multiple machine-learning approaches can be used. Yet, the amount of data gathered is enormous, and it is frequently highly noisy.

These datasets, which are too enormous for human minds to comprehend, may be easily investigated utilizing various machine-learning techniques.

Hence, in recent years, these algorithms have greatly improved at predicting the presence or absence of heart-related disorders. Stroke continues to be a significant health burden for individuals and national healthcare systems and is the second highest cause of death globally.

The major goal of this project is to create and implement an effective disease prediction model. With the use of numerous algorithms like Logistic Regression, SVM, Random Forests, and others, Machine Learning, a rapidly developing field of artificial intelligence, can make judgments and predictions from the vast amounts of data generated by the healthcare sector. Several classification algorithms are offered by ML based on the presented problem to determine the likelihood that a patient will experience a brain stroke.

*A. Architecture*

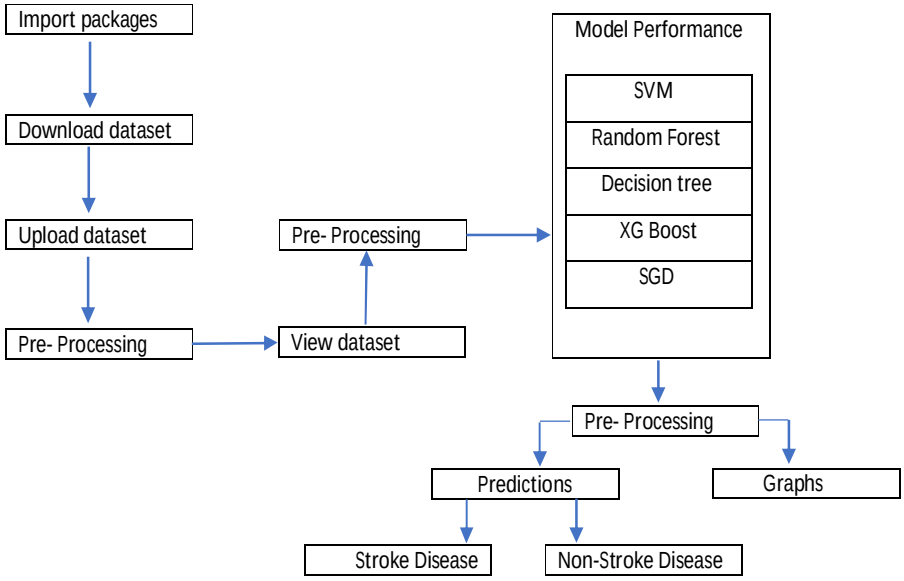



Fig1. Architecture

## II. LITERATURE REVIEW

The guidelines primarily concentrate on the demands that project face when applying scientific findings to the development of software systems. It is intended to support medical professionals' clinical activities. This standard contains machine learning methods for analyzing medical data automatically and creating graphs to assess strokes. When pre-processed data is used as input, the trained model automatically divides the data by that to forecast the occurrence of a stroke. In order to determine the effectiveness of the offered approaches, the region of graphs segmented by optimization techniques is compared with the standard training data collected by the neurologist. The segmented results can then be reassembled for additional analysis to get higher accuracy in.

A brain stroke can be fatal to brain function and performance and is caused when a brain artery ruptures, resulting in a hemorrhage. Experts in medicine advise utilizing either a CT scan or an MRI to make a diagnosis. CT imaging is being used more frequently due to its simplicity, affordability, and quickness. The data mining method utilized in this study gives an overview of information tracking in the current system from both a semantic and a syntactic perspective. Techniques for detecting malware include hypothesis exploration, deep learning, and data mining. In any event, AI is one of the most often utilized methods for identifying malware. There are two categories of malware detection techniques. In the suggested framework, we employ the XG Boost and Decision Tree AI calculations. With the paradigm we suggest, we provide more precise results.

According to C. L. Chin et al. [1], a brain hemorrhage happens when a brain artery breaks, causing bleeding that can have a disastrous impact on brain performance. To diagnose a hemorrhage, medical specialists advise utilizing either an MRI or a CT scan. We will therefore need a system that can quickly and automatically segment CT scan pictures. The goal is to separate the hemorrhaged brain area using the Deep Learning technique quickly and accurately. As a result, persons with brain hemorrhage can receive medical attention as quickly as feasible.

In a sizable population-based electronic medical claims database, C. Y. Hung [2] compared the deep neural network and various machine learning techniques for stroke prediction. According to the results, logistic regression (LR) and support vector machine (SVM) techniques can be matched in high prediction accuracy by DNN and gradient boosting decision tree (GBDT) methods. DNN, on the other hand, achieves optimum results while using fewer patient data than the GBDT method.

Stroke is a proven cause of death and failure in a number of nations, according to N. Venketasubramanian [3]. According to the World Health Organization, there were 10.3 million new stroke cases, 113 million DALYs lost due to stroke, over 25.7 million stroke survivors, 6.5 million stroke fatalities, and over 6.5 million insufficiency-altered life-years (DALYs) in 2013.

The CT scan pictures were processed by scaling, grayscale, smoothing, thresholding, and morphological operation by Badriyah, Tessy, and colleagues [4]. Subsequently, the Gray Level Co-occurrence Matrix was used to extract the pictures feature (GLCM). In this study, feature selection was utilized to choose pertinent features and lower processing costs, and deep learning based on a hyperparameter setting was used to classify the data. The experiment's findings demonstrated that while Bayesian Optimization was superior in terms of optimization time, Random Search had the best accuracy.

## III. METHODOLOGY

### *A. A Current Methodology*

The data mining methods employed in this study give an overview of information tracking in the current system from both a syntactic and semantic perspective. Data mining, deep learning, hypothesis exploration, and other methods are used to detect malware. In any event, AI is one of the techniques that are most frequently used to identify malware. There are two categories of malware detection techniques. First up is the traditional signature-based approach, in which malware is recognized by its signature. The next technique, which is also a novel one, is called a conduct-based methodology, and it is used to locate malware. With this technique, the malware is discovered based on the actions it intends to take against the target system.

Disadvantages

*1)* Inaccurate outcomes.
*2)* Tough to scale up.
*3)* It takes time.

### *B. Methodology Proposed*

In the suggested framework, we employ the AI calculations from the XG Boost and Decision Tree. Under our suggested framework, we provide more precise results. Building a reliable prediction model and applying it to disease prediction is the major goal of this project. Machine Learning, a faster-emerging branch of artificial intelligence, contributes several algorithms, including Logistic Regression, SVM, Random Forests, and many others, which are useful for drawing conclusions and making predictions from the vast amounts of data generated by the healthcare sector. Several classification algorithms are offered by ML based on the presented problem to determine the likelihood that a patient will experience a brain stroke.

Compared to the current system, it is more versatile and results are computed faster.

#### *1) A Decision Tree Classifier*

Using a multilayer perception model and three machine learning algorithms, valid reviews may be separated from spam with low mistake rates and high efficiency. Multilayer perceptron, decision tree classifier, and Naive Bayes classifier are a few often used methods. Structured data in the form of a binary tree is the output of a C4.5 decision tree classifier. The following is a model of a C4.5 tree. A training set is a collection of base tuples used to identify classes associated with the basis tuples. An adjective vector with the form X = (x1, x2,..., xn) represents a tuple X. Suppose that a tuple is a member of a predefined class, which is identified by the adjective class label. The learning stage is where the training set is randomly chosen from the basis. This method makes considerable use of classification and is quite effective. The following elements can be used to implement the tree's structure:

A test on an adjective is represented by a node in the tree, and potential test results are represented by a branch that leaves a node. A class label is represented by a leaf. A decision tree has a collection of rules that can be used to forecast objective functions. This model's algorithm employs greedy methods.

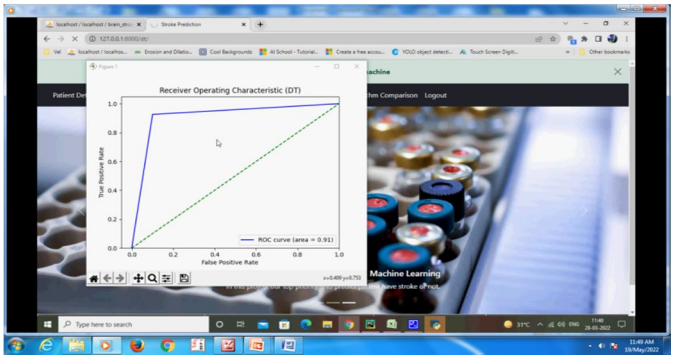



Fig 2: decision tree classifier

*2) Logistic Regression*

A logistic function is used to represent a binary dependent variable in the simplest form of logistic regression, though there are many more intricate variants. Using logistic regression, a logistic model's parameters are estimated in regression analysis. A binary logistic model has a dependent variable that can have two alternative values, such as pass/fail, which is represented by an indicator variable, and the two values are denoted by the letters "0" and "1". A linear combination of one or more independent variables makes up the log odds (the logarithm of the odds) for the value designated "1" in the logistic model. Both binary variables (two classes, each coded by an indicator variable) and continuous variables can be used as independent variables (any real value).

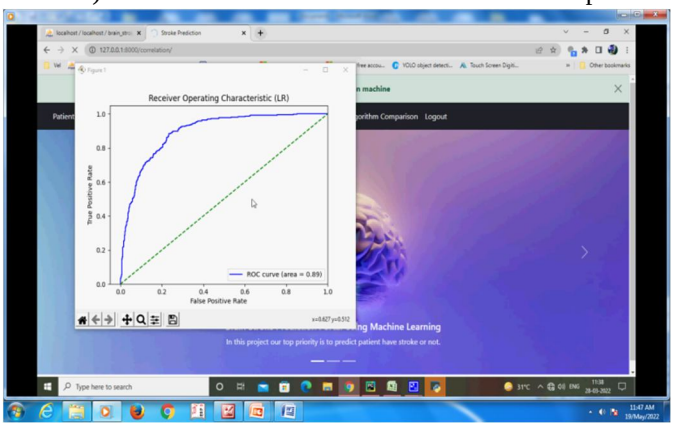



Fig3. Logistic Regression

*3) Random Forest*

Random Forest is a widely used machine learning calculation that fits nicely with the guided learning process. It is used in ML for both Arrangement and Relapse difficulties.

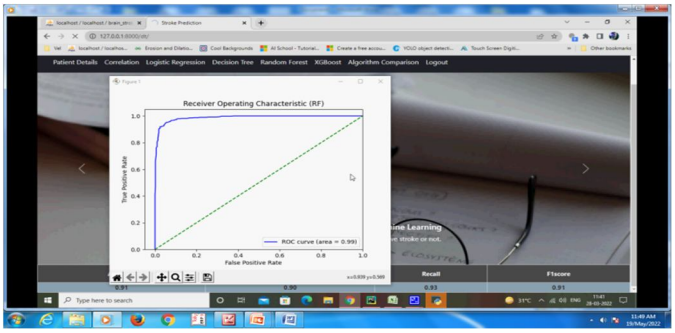



Fig4. Random forest

*4) Extreme Gradient Boosting Classifier*

The Xgboost ensemble machine learning algorithm is based on decision trees and employs a gradient-boosting framework. Artificial neural networks frequently outperform all other algorithms or frameworks in prediction issues involving unstructured data (pictures, text, etc.). Nonetheless, decision tree-based algorithms are now regarded as best-in-class when it comes to small- to medium-sized structured/tabular data.

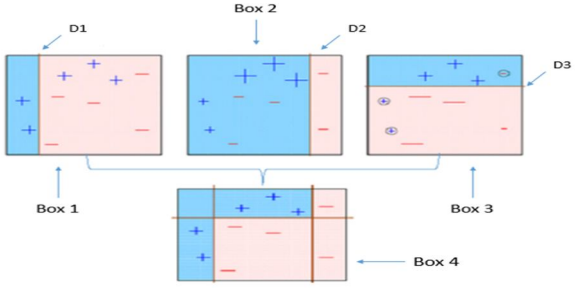



Fig5. Xgboost Process

Extreme Gradient Boosting, often known as XG support, is a modification of the Gradient Boosting approach that chooses the optimal tree model using more precise approximations. It combines many helpful strategies that greatly enhance its success, especially when working with structured data. 1) Calculating second-order gradients, or the second partial derivatives of the loss function is crucial because it reveals more details about the gradient's direction and shows us how to reach the loss function's minimum. Standard inclination aiding uses the loss capacity of our base model (for example, decision tree) as a middleman for limiting the mistake of the general model, whereas XG Boost uses the second request subsidiary as an estimator for decreasing the mistake of the general model. 2) Improved regularization: This enhances the generalizability of the model. Other benefits include quick training that can be spread across multiple clusters.

## IV. EXPERIMENTAL RESULTS AND DISCUSSION

### A. *System Modules*

*1)* Upload the data set (The system takes the dataset uploaded by the user).
*2)* Examine the data set (to View the data uploaded by the user).
*3)* Instruction (to train the model by the uploaded data).
*4)* Pre-processing (minimizing the gaps to reduce errors).
*5)* Model Performance (Selection of model to view the performance).
*6)* Projection (To enter the entity Values in which the result would be calculated by the model).
*7)* Analysis of Graphs (Graphs can be generated by the system and the user can view those graphs).

The parameters of the specific patient for whom we need to determine whether he or she will be affected by a stroke or not must first be entered into the model so that it may be trained. These parameters are illustrated below in Fig. I.

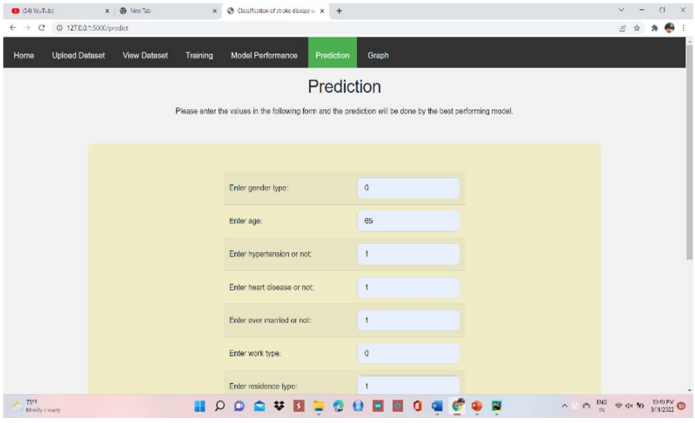



Fig6. Entering of Parameter Values

The outcome achieved after entering the settings is shown in this case in Fig. II. It determines whether the specific patient will experience a brain stroke.

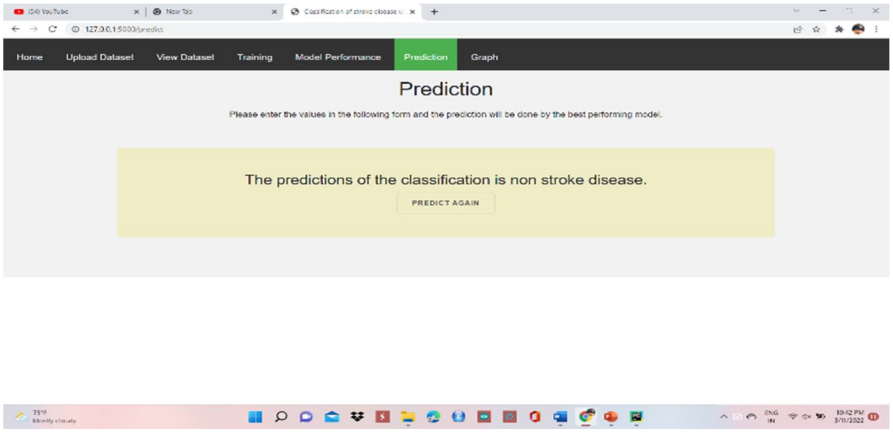



Fig7. Classified Result

And in this case, the graphs would demonstrate the precision of the outcome.


Fig8. Graphical Result

*B. Balancing Dataset*

12 columns and 5110 rows made up this dataset. In this dataset, the likelihood of the output column (stroke) being 0 is greater than the likelihood of the same column being 1. Only 249 rows in the stroke column alone have a value of 1, whereas 4861 rows have a value of 0. Data preparation balances the data to increase accuracy. Before preprocessing, it has no stroke records and the total number of strokes in the output column.

*1) Preprocessing*

Data preprocessing is crucial when creating a model since it helps to get rid of unwanted noise and outliers from the dataset that can cause the model to deviate from the training it was meant for. This step deals with anything preventing the model from working more successfully. Prior to starting to develop a model, the necessary dataset must be obtained, cleansed, and prepared. As indicated before, the dataset used comprises twelve features. The column id is firstly ignored because it has no bearing on how the model is constructed. Following that, any null values detected in the dataset are filled in. In this case, the null values in the BMI column are filled using the mean of the data column.

*2) Relationship Matrix*

In the heatmap shown above, we can observe that there is no multicollinearity and that some of the features with the highest correlation to stroke include age and glucose level.

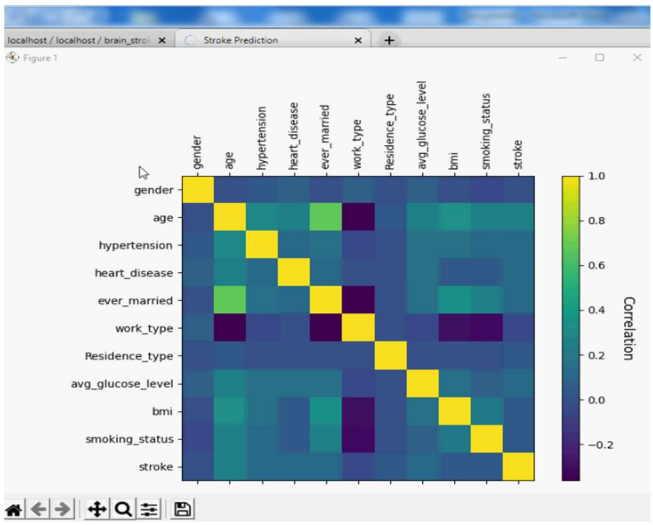



Figure 9. Results of Correlation

In the heatmap shown above, we can observe that there is no multicollinearity and that some of the features with the highest correlation to stroke include age and glucose level.

*3) Chi-Square test for the best features*

We can observe from the above table that the top 3 factors having the most influence on the output "Stroke" are age, average blood glucose level, and hypertension. Chi-Square Test is performed to find out this result.

*C. Assessment Matrix*

An instrument for assessing the effectiveness of machine learning classification algorithms is a confusion matrix. The effectiveness of each model developed has been evaluated using the confusion matrix. The confusion matrix shows how frequently our models estimate erroneously and how frequently they forecast accurately. False positives and false negatives have been attributed to values that were incorrectly anticipated, whilst genuine positives and true negatives have been assigned to values that were correctly anticipated. After grouping all predicted values in the matrix, the accuracy, precision-recall trade-off, and AUC of the model were used to evaluate its performance.

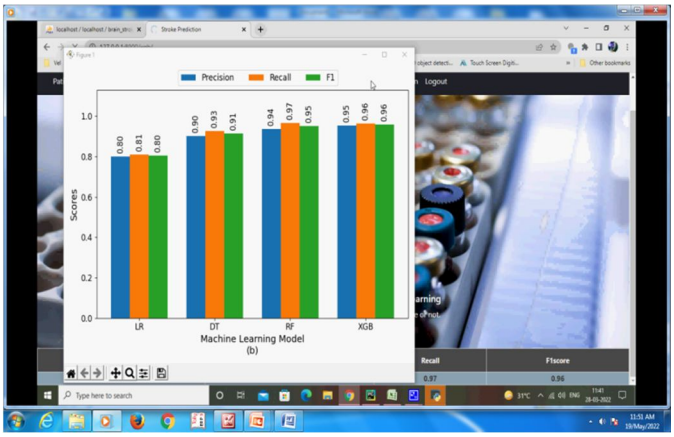
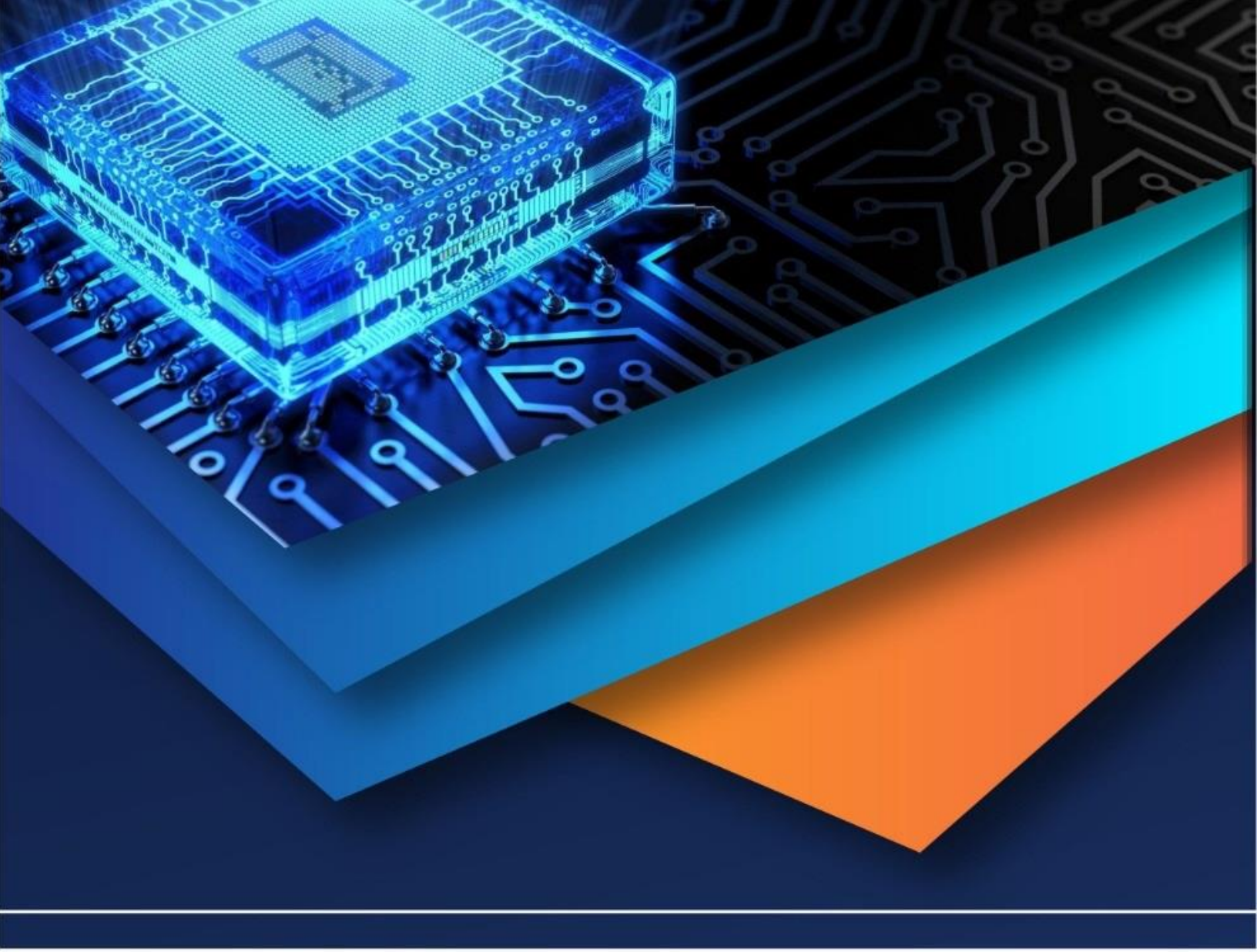



Fig10. Algorithm Comparison

The study highlights the usefulness of categorization techniques for structured entities, such as patient case sheets, in categorizing strokes according to specified characteristics (symptoms) and circumstances. Based on classification approaches, this study forecasts the type of stroke that a patient would experience. According to this study, stroke is more common in males than in women and in people between the ages of 40 and 60. Those who experienced an ischemic stroke outnumbered those who experienced a hemorrhagic stroke. The impact of the patient's modifiable and non-modifiable risk factors, as well as the specific symptoms of each patient, are factors in determining the type of stroke. In these hard times, it is crucial to be aware of and recognize the dangers associated with brain stroke. The model uses commonplace daily factors that are known to all individuals to estimate the likelihood of a brain stroke. Due to this, the initiative is both very relevant and necessary for society. The idea was designed to be implemented on a web platform in order to reach as many people as feasible. Someone who may have a stroke risk can be saved by receiving an early warning.

## REFERENCES

[1] Virani SS, Wong ND, Woo D, Turner MB, Soliman EZ, Sorlie PD, Sotoodehnia N, Turan TN, Go AS, Lloyd-, Nichol G, Paynter NP (2012) heart disease and stroke statistics—2012 update: a report, executive summary. 125(1):188–197
[2] In circulation https://www.strokejournal.org/article/S1052-3057(19)30523-3/fulltext#seccesectitle0006
[3] Hansen AT, Hvas AM, Pahus SH (2016) Testing for thrombophilia in young people who have had an ischemic stroke. Stroke 137:108–112
[4] Wijdicks EF, Lanzino G, Rabinstein AA, Dupont SA (2010) A review of aneurysmal subarachnoid hemorrhage for practicing neurologists. 30(5):45-54 Semin Neurol
[5] Stroke disease classification using machine learning methods Govindarajan Priya Premaladha Jayaraman1, Ravichandran Kattur Soundarapandian2, Amir H. Gandomi3, Rizwan Patan4, Ramachandran Manikandan2, Amir H. Gandomi3
[6] Jae-woo Lee, Hyun sun Lim, Dong-Wook Kim, Soon-ae Shin, Jink won Kim, Bora Yoo, and Myung-Hee Cho are the authors of "Computer Techniques and Programs in Biomedicine".
[7] "Stroke prediction using artificial intelligence" by Prakash Choudhary and M. Sheetal Singh. (IEEE - 2017)
[8] Deep learning algorithms for key finding detection in head CT scans: retrospective research, by Rohit Ghosh, Swetha Tanamala, Mustafa Biviji, Norbert G Campeau, and Vasantha Kumar Venugopal.
[9] "Predicting the outcome of stroke thrombolysis using machine learning from CT brain" Paul Bentley, Jeban Ganesalingam, Anoma Lalani, Carlton Jones, Kate Mahady, Sarah Epton, Paul Rinn, Pankaj Sharma, Omid Halse, Amrish Mehta, and Daniel Rueckert Stroke Risk Profile from the Framingham Study, Probability of Stroke.
[10] William B. Kannel, MD; Albert J. Belanger, MA; Ralph B. D'Agostino, Ph.D.; and Philip A. Wolf, M.D.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY