Brain Stroke Prediction Using Machine Learning Models

1st Emil Bluemax
CSE
PES UNIVERSITY
Bangalore, India
emil.bluemax@gmail.com

2nd J.P.DANIEL CHRISTOPHER
CSE
PES UNIVERSITY
Bangalore City, India
danielchristopher513@gmail.com

3rd Aditya Rajendra Khot
CSE
PES UNIVERSITY
Bngalore City, India
adityakhot55@gmail.com

Abstract—
Stroke or a cerebral vascular accident is the sudden death of brain cells due to inadequate blood flow and oxygen resulting from a blood clot occluding an artery in the brain or a blood vessel rupturing. When either of these things happens, brain cells begin to die and brain damage occurs. About two million brain cells die every minute during stroke with loss of abilities controlled by that area of the brain which include speech, movement and memory. A high number of dead brain cells are associated with increased risk of permanent brain damage, disability or death. Stroke is divided into two broad categories that define its pathophysiology: Ischemic stroke, Hemorrhagic stroke.Death from stroke is as a result of co-morbidities and/or complications. Complications of stroke may arise at different time periods This research work proposes an early prediction of stroke diseases by using different machine learning approaches with the occurrence of hypertension, body mass index level, heart disease, average glucose level, smoking status, previous stroke and age. Using these high features attributes, Some classifiers have been trained, they are Logistics Regression, Decision Tree Classifier, AdaBoost Classifier, K-Neighbors Classifier, and XGBoost Classifier for predicting the stroke The proposed study has an accuracy of:
A Decision Tree Classifier (DTC):93.72%, XGBoost (XGB):96.45%,Naïve Bayes Algorithm:78.76%,Random Forest classifier:97.22%

Index Terms—**Stroke, Machine learning, Classification, Data pre-processing, Confusion matrix**

I. Introduction

Stroke is a disease that affects the arteries leading to and within the brain. A stroke occurs when a blood vessel that carries oxygen and nutrients to the brain is either blocked by a clot or ruptures. According to the WHO, stroke is the 2nd leading cause of death worldwide.

Globally, 3% of the population are affected by sub-arachnoid hemorrhage, 10% with intracerebral hemorrhage, and the majority of 87% with ischemic stroke. 80% of the time these strokes can be prevented, so putting in place proper education on the signs of stroke is very important. The existing research is limited in predicting risk factors pertained to various types of strokes.

Early detection of stroke is a crucial step for efficient treatment and ML can be of great value in this process. To be able to do that, Machine Learning (ML) is an ultimate technology which can help health professionals make clinical decisions and predictions. During the past few decades, several studies were conducted on the improvement of stroke diagnosis using ML in terms of accuracy and speed. The existing research is limited in predicting whether a stroke will occur or not.

When a stroke occurs,identify the affected patient's condition and begin therapy within a half hour or one hour is very important. Otherwise, saving the patient's life becomes tough. It can be difficult to predict stroke based on risk factors since risk variables are complicated and inconsistent . In this experiment, wecompare some machine learning methods which may be predicted stroke early based on physical characteristics.

Machine Learning techniques including Random Forest, KNN , XGBoost , Catboost and Naive Bayes have been used for prediction.Our work also determines the importance of the characteristics available and determined by the dataset.Our contribution can help predict early signs and prevention of this deadly disease.

II. RELATD WORK

Aditya Khosla et al. presented a feature extraction method which find strong features using their proposed method: conservative mean. They compared their work Cox Proportional Hazards model with different machine learning process to find stroke possibility using well known Cardiovascular-Health-Study (CHS) dataset.

Govindarajan et al. conducted a study to categorize stroke disorder using a text mining combination and a machine learning classifier and collected data for 507 patients. For their analysis, they used various machine learning approaches for training purposes using ANN, and the SGD algorithm gave them the best value, which was 95%.

Jaehak, et al. discussed an AI model for prediction of stroke illness based on the real-time diagnostic procedure. They proposed a stroke prediction process which can identify stroke victimization period biological-signals using artificial intelligence system.

In 2020, Yu et al. conducted a study to detect stroke using bio-signals in real-time with Artificial Intelligence where both Long Short-Term Memory (LSTM) and Random Forest algorithms were used. Here they collected the EMG bio-signals in real-time and important attributes were determined and the prediction models were trained and utilized to predict stroke. Random Forest gave a correctness of90.38% whereas LSTM gave 98.958% correctness [14].

Monteiro et al. performed a study to get a functional outcome prediction of ischemic stroke using machine learning. In their research, they apply this technique to a patient who was passing three months after admission. They got the AUC value above 90%.

III. Research Methods

A. Data Dictionary

dataset consist of 5110 people's information and now all the attributes are described:
age: This attribute means a person's age. It's numerical data.
gender: This attribute means a person's gender. It's categorical data.
hypertension: This attribute means that this person is hypertensive or not. It's numerical data.
work-type: This attribute represents the person work scenario. It's categorical data.
residence-type: This attribute represents the person living scenario. It's categorical data.
heart-disease: This attribute means whether this person has a heart disease person or not. It's numerical data.
avg_glucose_1eve1: This attribute means what was the level of a person's glucose condition. It's numerical data.
bmi: This attribute means body mass index of a person. It's numerical data.
ever-married: This attribute represents a person's married status. It's categorical data.
smoking-Status: This attribute means a person's smoking condition. It's categorical data.
stroke: This attribute means a person previously had a stroke or not. It's numerical data.
In this all attribute stroke is the decision class and rest of the attribute is response class.

B. Exploratory Data Analysis

- The data pre-processing step deals with the missing values in the dataset and converts the categorical values converting them into nominal numeric values

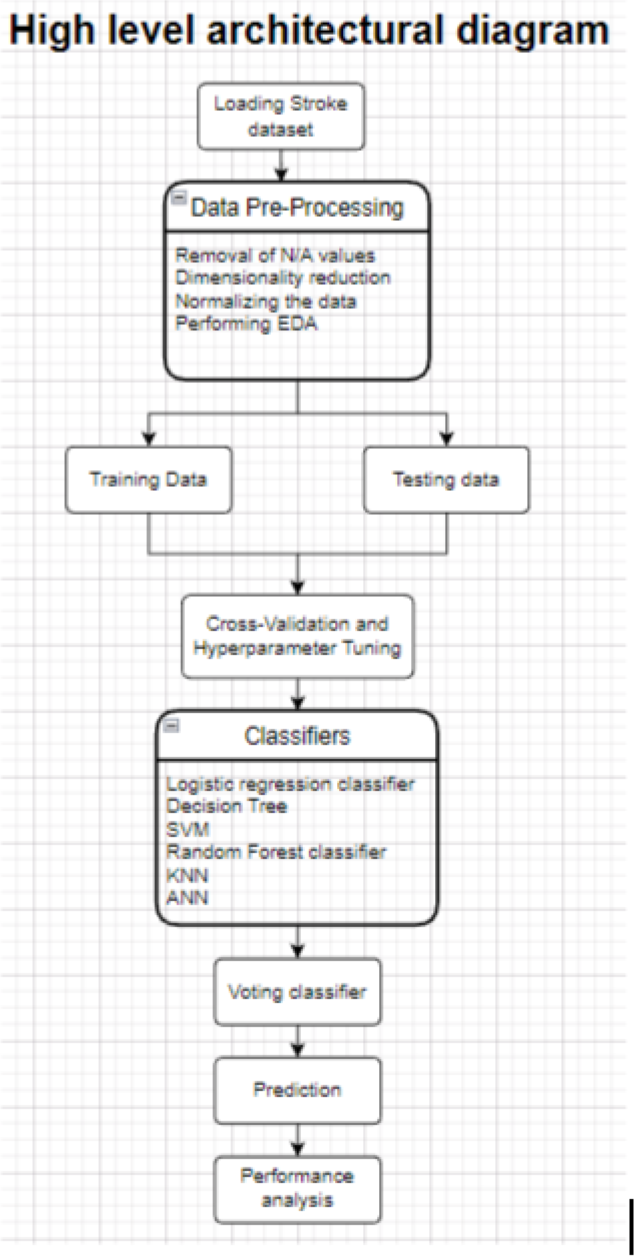


Fig. 1. Architectural diagram

- we observe that there is very low correlation among the attributes, the highest correlation observed was between age and bmi with a value of 0.32 all other correlation value's were less than 0.3
- This is followed by data analysis which helps us better comprehend the relationship amongst the variables. It also indicated the need to oversample the data, due to the huge imbalance between minority class and majority class.
- These NaN values had been substituted by the mean/average value of the bmi which was 28.893237.
- Label encoding technique was applied using the LabelEn-

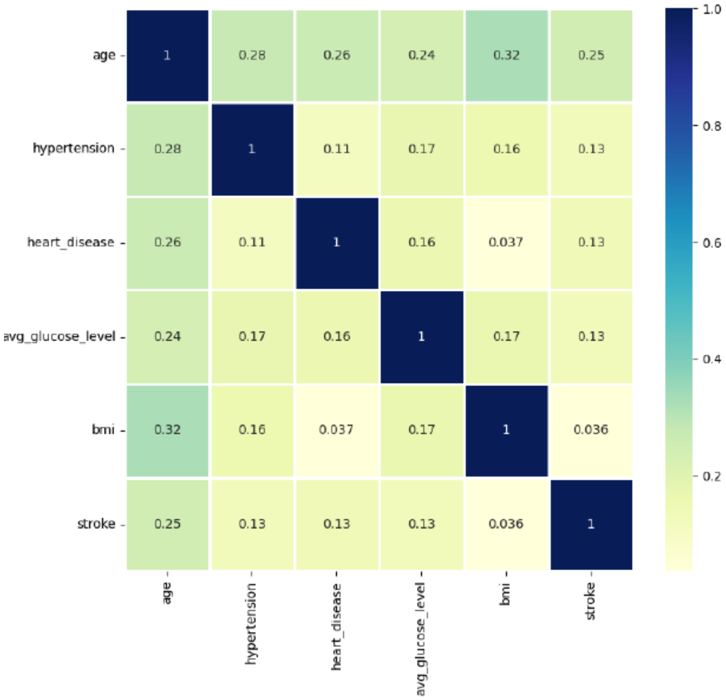


Fig. 2. Correlation Matrix

coder() method found in sklearn's preprocessing library, where all the categorical values were replaced using numbers starting from zero till n-1, where n refers to the number of classes in the variable.

- During EDA, it turned out to be that the dataset used in this paper had only 249 entries of people who suffered stroke and 4861 people didn't have a stroke, making the dataset highly imbalanced with only about 4.8% of the total entries of minority class stroke. If machine learning algorithms are applied on such dataset it would have resulted in poor performance on minority class whose performance is the most important
- In order to overcome this problem of imbalanced classes, Synthetic Minority Oversampling Technique (SMOTE) is used.
- his technique works by selecting an instance of minority class at random and then finds its k nearest neighbors. Then a randomly selected neighbor is chosen, and a synthetic instance is created at a randomly selected point between the two examples in feature space.

C. *Machine Learning Classifier*

In our paper, we have used different machine learning classifiers. The classifiers are namely Gaussian Naive Bayes, Logistic Regression, Decision Tree Classifier, K-Nearest Neighbours, AdaBoost Classifier, XGBoost Classifier, and Random Forest Classifier.

All these classifiers are well known, and we can compare our results with other similar research work. We use 80% data of our dataset to train our algorithm, and the remaining 20% of the data is considered to assess the trained model. For ML model validation, we use k-fold cross validation process. In the k-fold cross validation technique, total dataset is employed to training and testing the classification process. The dataset is splited into k parts, that known as fold. In training procedure, this process uses k-1 folds to train ML model and one-fold is employed to test model. This process is repeated k times, and every fold can be considered as test data-set. The ultimate outcome is the common for the testing group performance using all the used folds. Advantage of this technique, that the all samples within the data-set are used for train and test, which removes the high variance. Confusion matrices are used to evaluate the model's performance by calculating accuracy, recall, precision, f-1 score, false-positive rate, false-negative rate. Analyzing these values, we find out the best model to predict stroke.

D. *Naïve Bayes Algorithm*

Naive Bayes Algorithm is a classification based on the statistical probability that calculates a set of probabilities by adding up the frequency and combination of values from a given dataset. This algorithm uses the Bayes theorem and assumes all attributes are independent or not interdependent given by the value of the class variable Therefore, the above Naive Bayes method is adjusted as follows: To explain the Naive Bayes method, please note that the classification process requires several instructions to determine what class is suitable for the sample being analyzed. Therefore, the above Naive Bayes method is adjusted as follows:

$$P(C \mid x) = \frac{P(x \mid C)P(C)}{P(x)}$$

Fig. 3. Naïve Bayes eq-1

$$P(C \mid F1 \ldots . Fn) = \frac{P(C)P(F1 \ldots Fn \mid C)}{P(F1 \ldots Fn)}$$

Fig. 4. Naïve Bayes eq-2

E. *Decision Tree Algorithm*

The C4.5 algorithm is a development of the ID3 Algorithm. Both of these algorithms were created by a researcher in the field of artificial intelligence named J. Rose Quinan in the late 1970s.

In general, the steps of the C4.5 algorithm in making a decision tree are:

- Choose an attribute as the root.
- A harrow for each value
- Divide cases into branches
- Repeat the process on each branch until all cases in the branch have the same class.

To select an attribute as the root, the highest gain value is based on the existing attributes. Gain can be calculated using the formula as shown in equation 1 below.

$$Entropy(s) = \sum\nolimits_{i=1}^{n} P_i * \log_2 Pi$$

Fig. 5. Decision tree eq-1

$$Gain(S, A) = Entropy\ (S) - \sum\nolimits_{i=1}^{n} \frac{|S_i|}{|s|} *\ Entropy\ (S_i)$$

Fig. 6. decision tree eq-2

Meanwhile, to calculate the entropy value can be used with the following equation

F. Random Forest Algorithm

A Random forest consists of a combination of decision trees in which each tree rests on a random vector value sampled independently and with the same distribution for all trees in the forest. Classification in random forest consists of a collection of classifications h(xk),k=1,2,... where k is an independently distributed random vector and each tree assigns a voting unit to the most popular class in input . The following explanations are given for the sake of identify the accuracy of the Random Forest. 1. **Random Forests Converge** Centralizing random forests by determining the margin function can determine more accurate results. If the classification ensemble is h1(x),h2(x),...,hk(x) with a random training set from the random vector distribution Y,X. The margin can be determined by the following equation:

$$\begin{aligned} mg(X, Y) &= av_k I\left(h_k(X) = Y\right) \\ &- \max av_k I\left(h_k(X) = j\right). \end{aligned}$$

Fig. 7. Random forest eq-1

The indicator function is I(). The margin function is used to measure how far the average number of votes in Y,X for a class exceeds the average vote for other classes. The larger the margin obtained, the more accurate the classification results.

2. **Strength and Correlation** The upper bound on the random forest can be derived for generalization error by

$$PE^* \leq \bar{P}\left(1 - s^2\right)/s^2$$

Fig. 8. Random forest eq-2

G. XGboost Algorithm

XGBoost is a supervised learning algorithm based on ensemble trees. It aims at optimising a cost objective function composed of a loss function (d) and a regularization term ():

where yiˆ is the predictive value, n the number of instances in the training set, K is the number of trees to be generated and fk is a tree from the ensemble trees. The regularization term is defined as:

$$\Omega(\theta) = \underbrace{\sum_{i=1}^{n} d\left(y_i, \widehat{y_i}\right)}_{Loss} + \underbrace{\sum_{k=1}^{K} \beta\left(f_k\right)}_{regularization},$$

Fig. 9. XGBoost eq-1

$$\beta(f_t) = \gamma T + \frac{1}{2}\left[\alpha \sum_{j=1}^{T} |c_j| + \lambda \sum_{j=1}^{T} c_j^2\right],$$

Fig. 10. XGBoost eq-2

where is the minimum split loss reduction, is a regularization term on the weight and c is the weight associated to each leaf. Let ft(xi)=cq(xi), where q is in [1,T], where T is the number of leafs. A greedy approach is performed to select the split that increases the most the gain. The detailed procedures and equation's derivations are given in Appendix A. Table III outlines the ten XGBoost key parameters, ranges and default values of each parameter.

H. Classification Metrics

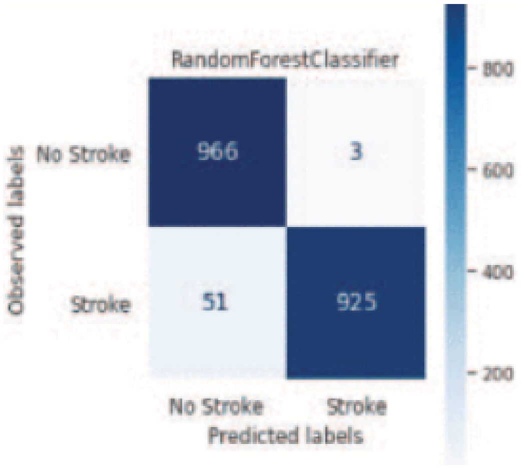


Fig. 11. Random forest confusion matrix

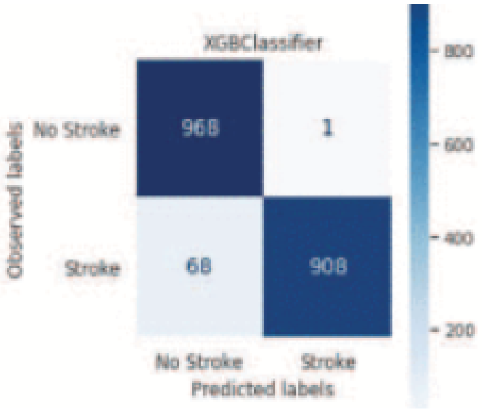


Fig. 12. XG boost confusion matrix

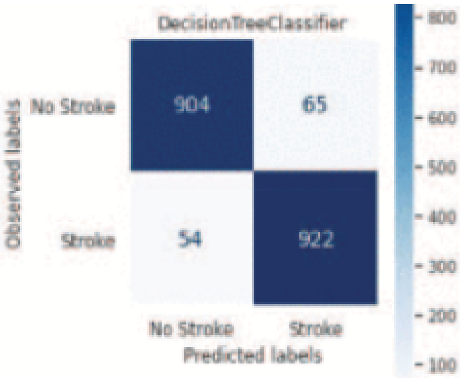


Fig. 13. Decision tree confusion matrix

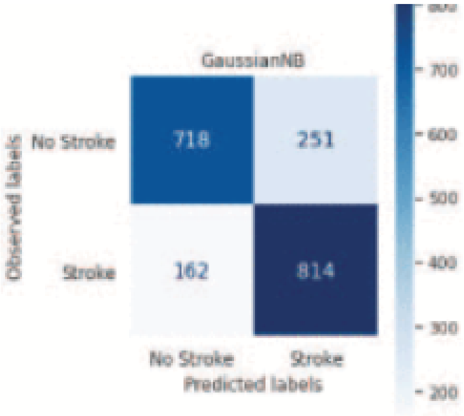
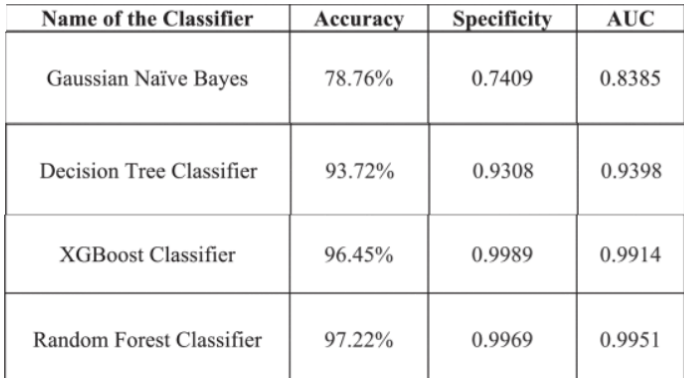


Fig. 14. Naive bayes confusion matrix

IV. Results and Conclusion

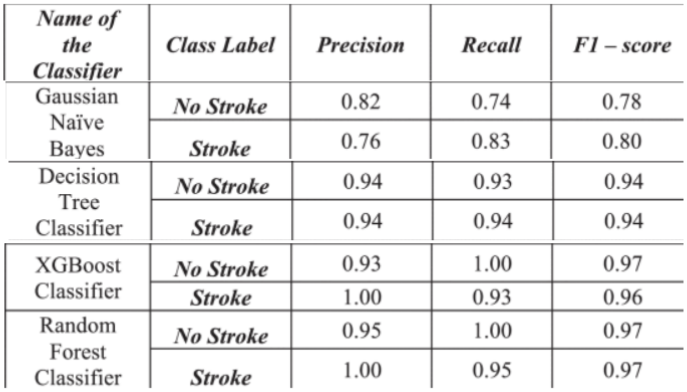
The highest accuracy value was obtained by Random Forest with a result of 97.22%, in the test to predict stroke disease using 12 variables with a total data of 5,109 data. Random Forest has the advantage in classifying data because it works for data that has incomplete attributes, and is good for handling large sample data.

Further Deep Learning Models can also Be used in order to predict the risk of stroke more effectively without performing clinical test.



Name of the Classifier	Accuracy	Specificity	AUC
Gaussian Naïve Bayes	78.76%	0.7409	0.8385
Decision Tree Classifier	93.72%	0.9308	0.9398
XGBoost Classifier	96.45%	0.9989	0.9914
Random Forest Classifier	97.22%	0.9969	0.9951

Fig. 15. Accuracy speceficity AUC



Name of the Classifier	*Class Label*	*Precision*	*Recall*	*F1 – score*
Gaussian Naïve Bayes	*No Stroke*	0.82	0.74	0.78
	Stroke	0.76	0.83	0.80
Decision Tree Classifier	*No Stroke*	0.94	0.93	0.94
	Stroke	0.94	0.94	0.94
XGBoost Classifier	*No Stroke*	0.93	1.00	0.97
	Stroke	1.00	0.93	0.96
Random Forest Classifier	*No Stroke*	0.95	1.00	0.97
	Stroke	1.00	0.95	0.97

Fig. 16. Class Label, Precision, Recall, F1-Score

We can try using other classifying models like voting classifier, Since the proportion of positive and negative brain stroke cases are highly imbalanced some transformations and boosting needs to be done to represent equal representation

Propose a framework that uses brain Magnetic Resonance Imaging (MRI) with deep learning to improve the state of the art. It exploits advancements in deep learning to improve brain stroke prediction performance further.

An application or smart system suggested to be made in predicting the diagnosis of stroke, by adding other algorithms in machine learning to create better and more accurate models. It is recommended also to add attributes to the dataset such as recent strenuous activity, and occupations to strengthen the prediction.

References

[1] S. Gupta and S. Raheja, ”Stroke Prediction using Machine Learning Methods,” 2022 12th International Conference on Cloud Computing, Data Science Engineering (Confluence), 2022, pp. 553-558, doi: 10.1109/Confluence52989.2022.9734197.

[2] N. S. Adi, R. Farhany, R. Ghina and H. Napitupulu, ”Stroke Risk Prediction Model Using Machine Learning,” 2021 International Conference on Artificial Intelligence and Big Data Analytics, 2021, pp. 56-60, doi: 10.1109/ICAIBDA53487.2021.9689740.

[3] M. U. Emon, M. S. Keya, T. I. Meghla, M. M. Rahman, M. S. A. Mamun and M. S. Kaiser, ”Performance Analysis of Machine Learning Approaches in Stroke Prediction,” 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2020, pp. 1464-1469, doi: 10.1109/ICECA49313.2020.9297525.

[4] R. Islam, S. Debnath and T. I. Palash, ”Predictive Analysis for Risk of Stroke Using Machine Learning Techniques,” 2021 International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2), 2021, pp. 1-4, doi: 10.1109/IC4ME253898.2021.9768524.

[5] A. Devaki and C. V. G. Rao, ”An Ensemble Framework for Improving Brain Stroke Prediction Performance,” 2022 First International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT), 2022, pp. 1-7, doi: 10.1109/ICEEICT53079.2022.9768579.

[6] V. Krishna, J. Sasi Kiran, P. Prasada Rao, G. Charles Babu and G. John Babu, ”Early Detection of Brain Stroke using Machine Learning Techniques,” 2021 2nd International Conference on Smart Electronics and Communication (ICOSEC), 2021, pp. 1489-1495, doi: 10.1109/ICOSEC51865.2021.9591840.

[7] M. Sheetal singh, Prakash choudhary, " Stroke Prediction using Artificial Intelligence ", 8th Annual Industrial Automation and Electromechanical Engineering conference(IEMECON) 2017 DOI: 10.1109/IEMECON.2017.8079581.

[8] Tasfia Ismail Shoily, Tajul Islam, , Sumaiya Jannat, Sharmin Akter Tanna,Taslima Mostafa Alif, Romana Rahman Ema. " Detection of Stroke disease using Machine Learning Algorithms " 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT) DOI: 10.1109/ICCCNT45670.2019.8944689

[9] V. J. Jayalaxmi, V geetha, M. Ijaz, " Analysis and Prediction of Stroke using Machine Learning Algorithms " 2021 International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA) — 978-1-6654-2829-3/21/$31.00 ©2021 IEEE — DOI: 10.1109/ICAECA52838.2021.9675545

[10] I. L. Cherif and A. Kortebi, "On using eXtreme Gradient Boosting (XGBoost) Machine Learning algorithm for Home Network Traffic Classification," 2019 Wireless Days (WD), 2019, pp. 1-6, doi: 10.1109/WD.2019.8734193.