JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

BrainOK: Brain Stroke Prediction using Machine Learning

Mrs. Neha Saxena
Department of Computer Engineering
Universal College of Engineering, Vasai, India
nehasaxena031@gmail.com

Mr. Deep Singh Bhamra
Department of Computer Engineering
Universal College of Engineering, Vasai, India
deepsbhamra@gmail.com

Mr. Arvind Choudhary
Department of Computer Engineering
Universal College of Engineering, Vasai, India
choudharyarvind182@gmail.com

Mr. Preet Maru
Department of Computer Engineering
Universal College of Engineering, Vasai, India
preetmaru49@gmail.com

Abstract **- A stroke, also known as a cerebrovascular accident or CVA is when part of the brain loses its blood supply and the part of the body that the blood-deprived brain cells control stops working. This loss of blood supply can be ischemic because of lack of blood flow, or haemorrhagic because of bleeding into brain tissue. A stroke is a medical emergency because strokes can lead to death or permanent disability. There are opportunities to treat ischemic strokes but that treatment needs to be started in the first few hours after the signs of a stroke begin. The patient, family, or bystanders should activate emergency medical services immediately should a stroke be suspected. A transient ischemic attack (TIA or mini-stroke) describes an ischemic stroke that is short-lived where the symptoms resolve spontaneously. This situation also requires emergency assessment to try to minimize the risk of a future stroke. By definition, a stroke would be classified as a TIA if all symptoms resolved within 24 hours. According to the World Health Organization (WHO) stroke is the 2nd leading cause of death globally, responsible to approximately 11% of total deaths . For survival prediction, our ML model uses dataset to predict whether a patient is likely to get stroke based on the input parameters like gender, age, various diseases, and smoking status. Unlike most of the datasets, our dataset focuses on attributes that would have a major risk factors of a Brain Stroke.**

Keywords - Machine learning, Brain Stroke. Ischemic Stroke, transient ischemic attack.

I. INTRODUCTION

Machine Learning (ML) delivers an accurate and quick prediction outcome and it has become a powerful tool in health settings, offering personalized clinical care for stroke patients. An application of ML and Deep Learning in health care is growing however, some research areas do not catch enough attention for scientific investigation though there is real need of research. Therefore, the aim of this work is to use ML algorithms like Logistic regression, SVM, KNN, Decision Tress and Random Forest to determine and predict the risk of Brain Strokes. A total of 39 studies were identified from the results of ScienceDirect web scientific database on ML for brain stroke from the year 2007 to 2019[2]. Support Vector Machine (SVM) is obtained as optimal models in 10 studies for stroke problems. Besides, maximum studies are found in stroke diagnosis although number for stroke treatment is least thus, it identifies a research gap for further investigation. Similarly, CT images are a frequently used dataset in stroke. Finally, SVM and Random Forests are efficient techniques used under each category [2]. The present study showcases the contribution of various ML approaches applied to brain stroke.

II. LITERATURE SURVEY

In the research conducted by Manisha Sirsat, Eduardo Ferme, Joana Camara, the main aim of the research was to classify state-of-arts on ML techniques for brain stroke into 4 categories based on their functionalities or similarity, and then review studies of each category systematically. The study further discusses the outcomes and accuracies obtained by using different Machine Learning models using text and image-based datasets.

In this study, the authors discussed many stroke related problems from the state-of-art. The reviewed studies were grouped in several categories based on their similarities. The study notes that it is difficult to compare studies as they employed different performance metrics for different tasks, considering different datasets, techniques, and tuning

parameters. Hence, it only mentions the research areas which were targeted in more than one study and the studies which report highest classification accuracy in each section [1].

Harish Kamal, Victor Lopez, Sunil A. Sheth, in their study discus how Machine Learning (ML) through pattern recognition algorithms is currently becoming an essential aid for the diagnosis, treatment, and prediction of complications and patient outcomes in several neurological diseases. The evaluation and treatment of Acute Ischemic Stroke (AIS) have experienced a significant advancement over the past few years, increasingly requiring the use of neuroimaging for decision-making. This study offers an insight into the recent developments and applications of ML in neuroimaging focusing on acute ischemic stroke. The implementations of machine learning are numerous, from early identification of imaging diagnostic findings, estimating time of onset, lesion segmentation, and fate of salvageable tissue, to the analysis of cerebral edema, and predicting complications and patient outcomes after treatment.

The paper finally concludes by discussing how Machine learning applications are expanding in the medical field for diagnostic and therapeutic purposes, and the rapidly expanding and increasingly neuro-imaging reliant field of AIS is proving to be fertile ground. There is a particular need for ML solutions in this field, which is faced with the challenge of increasingly complex data, with limited human expert resources. Future directions in ML for AIS may require collaborative approaches across multiple institutions to build a robust dataset for efficient training of ML networks [2].

In the research conducted by Chuloh Kim, Vivienne Zhu, Jihad Obeid and Leslie Lenert, they have assessed performance of natural language processing (NLP) and machine learning (ML) algorithms for classification of brain MRI radiology reports into acute ischemic stroke (AIS) and non-AIS phenotypes. The method followed included All brain MRI reports from a single academic institution over a two-year period were randomly divided into 2 groups for ML: training (70%) and testing (30%). Using “quanteda” NLP package, all text data were parsed into tokens to create the data frequency matrix. Ten-fold cross-validation was applied for bias correction of the training set. Labelling for AIS was performed manually, identifying clinical notes. They applied binary logistic regression, naïve Bayesian classification, single decision tree, and support vector machine for the binary classifiers, and we assessed performance of the algorithms by F1-measure. They also assessed how n-grams or term frequency-inverse document frequency weighting affected the performance of the algorithms.
The paper concluded with the understanding how supervised ML based NLP algorithms are useful for automatic classification of brain MRI reports for identification of AIS patients. Single decision tree was the best classifier to identify brain MRI reports with AIS [3].

In the research conducted by R. Punitha Lakshmi et al [4], they put forward their work on SVM Classifier Based On Otsu Thresholding For Ischemic Stroke Detection. The dataset used in order to train the algorithms/models were a set of 32 different types of brain MRI images which were in JPEG format. Both the classifiers i.e. the Random Forest Classifier and SVM Classifier were trained with the help of these images but with different procedure. All the MRI Images were first transformed using the wavelet transformation and the segmentation of those images were carried out by Otsu Thresholding. The images are obtained from Open Access Series of Imaging Studies (OASIS) which makes the MRI data sets of the brain which is available for the research purpose. Noise reduction of these images were done in pre-processing so as to get accurate results. After that, the data was then fed as input to the SVM Classifier. Thus, the maximum accuracy was given by SVM Classifier being 88% and Random Forest Classifier being at 81%.
The paper concluded with the understanding of how maximum accurate segmentation of brain and brain lesions is achieved with the help of SVM Classifier based on Otsu Thresholding and the dataset with scattering lesion tissues can also help to improve further accuracy rates of this Classification [4].

In the research conducted by Jaehak Yu et al, an implementation of system for semantic analysis of early detection of stroke and also the recurrence of stroke in Koreans over the age of 65 years based on the National Institute of Health (NIH) Stroke Scale was done by the researchers. The research was made possible with the help of data which was collected from the emergency medical center of the Chungnam National University Hospital consisting of 287 stroke patients out of which 16 patients, which had no stroke symptoms were excluded. Final NIHSS Data consisted of 227 patients, excluding the 60 patients whose data included missing values or outlier values among the NIHSS questionnaires. Patient subjects were the elderly over 65 years old, and consisted of 117 men and 110 women. The Machine Learning Algorithm which was used was C4.5 Decision Tree Algorithm. The researchers found out that it is the most advanced algorithm and its function of classification and prediction is already proven. The proposed system in this experiment classifies and predicts stroke severity score into four classes using representative classification and prediction models of machine learning and data mining methodology. To measure the experiment accuracy of the proposed system, the recall and precision are used as the measurements. The experiment resulted in faster and more accurate predictions of stroke severity and efficient system operation with the help of various Machine Learning algorithm used and C4.5 decision tree and Random Forest classified and predicted the performance with high accuracy.
The paper concluded with the understanding of how efficient use of Machine Learning Methodologies and a proper dataset to build a model to predict Brain Stroke and also assess the severity of symptoms to predict results with high precision can be implemented to build a system providing an alarm service to visit a medical centre or hospital in real-time [5].

Gangavarapu Sailasya, Gorli L. Aruna Kumari, in their study discuss how Brain Stroke, which is the fouth leading cause of death in India, can be predicted with the help of trained Machine Learning Models so as to minimize risk of death due to Brain

Stroke. For the purpose of prediction of Brain Stroke, the dataset was first acquired from Kaggle having 5110 rows and 12 columns and had attributes such as 'id', 'gender', 'age', 'hypertension', 'heart_disease','ever_married', 'work_type', 'Residence_type', 'avg_glucose_level', 'BMI', 'smoking_status' and 'stroke'. The dataset was highly imbalanced in terms of 'stroke' attribute due to '0'(No Stroke) outweighing '1'(Stroke) by 4861 values being of No stroke class. Therefore, for better accuracy, data pre-processing was performed to balance the data. Data Pre-processing is required before model building to remove the unwanted noise and outliers from the dataset, resulting in a deviation from proper training. Label Encoding was also performed for encoding the string value in dataset to numerals.

The classification algorithms that were used for training purpose were Logistic Regression, Decision Tree, Random Forest, K-nearest neighbors, Support Vector Machine and Naïve Bayes classifier. Out of all the algorithms chosen, Naïve Bayes Classification performs best with an accuracy of 82%. Even in recall and precision, Naïve Bayes has performed better in comparison with other algorithms.

The paper concluded with the understanding of how prediction of brain stroke can be made possible with the help of Machine Learning. Efficient use of trained models will also help to improve the accuracy and precision of prediction and will provide better insight to the user of the application to take further actions [6].

III. PROPOSED SYSTEM

The proposed system acts as a prediction support machine and will prove as an aid for the user with diagnosis. The algorithms used to predict the output have potential in obtaining a much better accuracy then the existing system. In proposed system, the practical use of various collected data has turned out to be less time consuming.

Advantages:

1. High performance and accuracy rate.
2. Data and information collected for prediction is easily available to the users.

System provides users with precaution that can be taken to reduce risk factor.

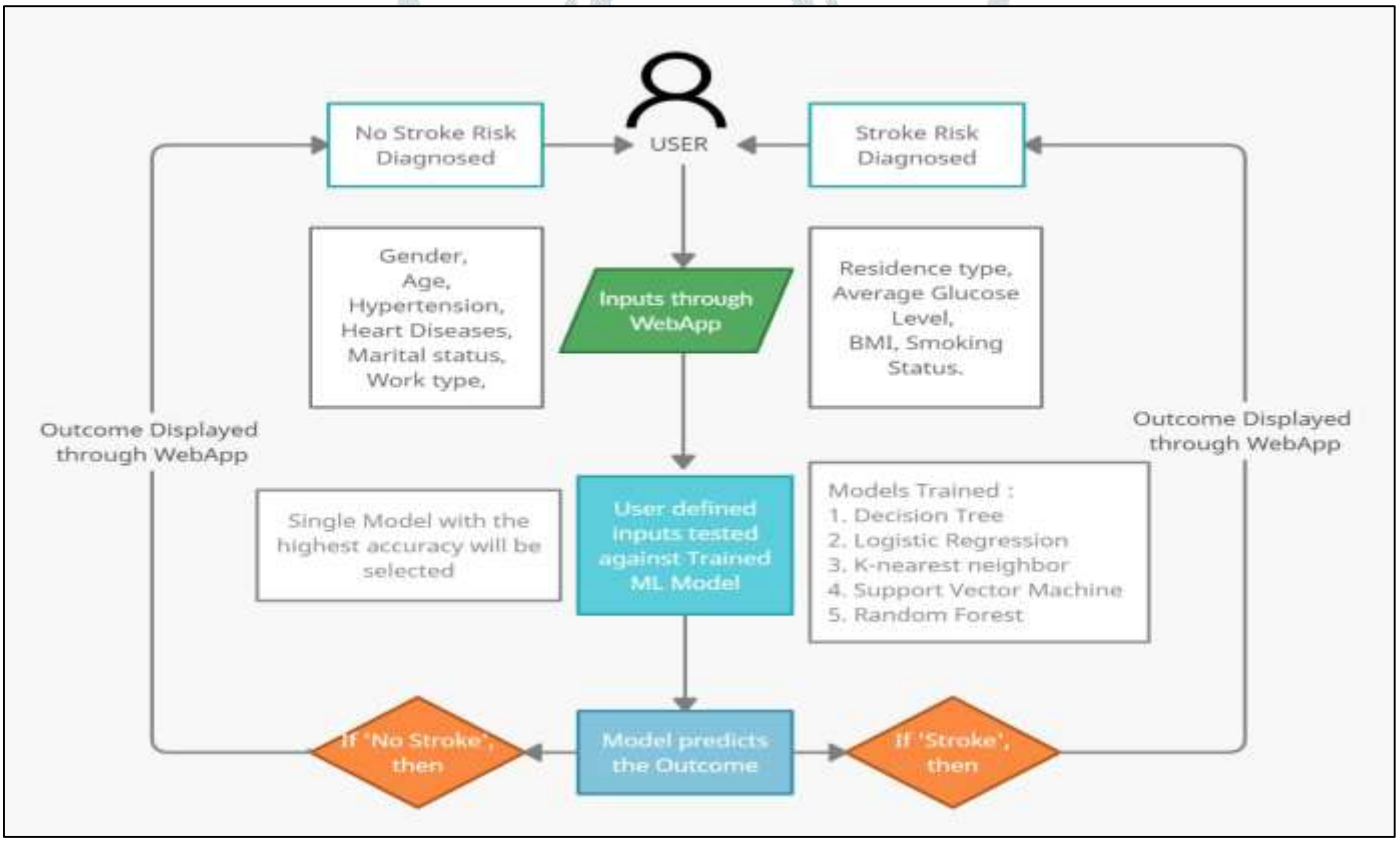

Figure 1 - System Architecture

Detailed Description of System Architecture shown in Fig. 1:

• USER: The person using our Web Application will be the user who wants to know whether they have a risk of having Brain or not.

• Inputs through WebApp: The user will be asked about some details regarding their gender, age, hypertension, heart diseases, marital status, work type, residence type, average glucose level, BMI and smoking status. All these details are necessary for the prediction of stroke possibility for that individual.

• User defined inputs tested against ML Model: Total of 5 Machine Learning Algorithms were trained so that the algorithm that yields best accuracy score will be considered as

the Trained ML Model that will help to predict stroke possibility against new data from the user side. Machine Learning Algorithms such as Decision Tree, Logistic Regression, K-Nearest Neighbor, Support Vector Machine and Random Forest.

• Model predicts the Outcome: The possibility of the user having stroke will be determined with the help of the Trained ML Model and if the user has risk of having brain stroke, depending on the accuracy of the model, it will predict the output for it and the same goes for no stroke.

• Stroke Risk Diagnosed: Through our Web Application, the user will get to know about the outcome of its input data. In the case for "Stroke" as an outcome, it will be displayed as "Stroke Risk Diagnosed

• No Stroke Risk Diagnosed: Through our Web Application, the user will get to know about the outcome of its input data. The outcome for "No Stroke" will be displayed as "No Stroke Risk Diagnosed".

The explanation of working of our Web Application is simplified with the help of modules that helps to predict the stroke risk of its user

The modules are:

A. Taking Inputs from the user through our web application.
The first step for our web application will be to take some basic input from the user so as to process that data with the trained data.

B. Processing the input data against training data.

As explained in the later of the previous module, the data that has been collected from the user will then be processed against the trained data and getting accurate results at the end of it.

C. Getting the test results.

The final step will be to provide accurate and precise results to the user using our web application so that they can take necessary steps depending upon the results that they have obtained.

The system has been implemented using 5 different Machine Learning Algorithms to obtain the best possible outcome and accuracy. The Machine learning model has been developed using Logistic Regression, Support Vector Machine (SVM), K Nearest Neighbour (KNN), Decision Tree and finally Random Forest algorithms.com

- Front End:

HTML (Hyper Text Markup Language), CSS (Cascading Style Sheets) and Bootstrap.

- Framework:

Flask: Python API to build web-applications.

- Runtime Environment:

Google Colaboratory: Colaboratory, or "Colab" for short, is a product from Google Research. Colab allows anybody to write and execute arbitrary python code through the browser, and is especially well suited to machine learning, data analysis and education. More technically, Colab is a hosted Jupyter notebook service that requires no setup to use, while providing access free of charge to computing resources including GPUs.

- Dataset

The Brain Stroke Prediction Dataset comprises of a total of 5110 rows data of data with 11 columns and had attributes such as 'id', 'gender', 'age', 'hypertension', heart_disease','ever_married', 'work_type', 'Residence_type', 'avg_glucose_level', 'BMI', 'smoking_status' and 'stroke'

- Libraries

Pandas, Numpy, Seaborn, Matplotlib, Sklearn/Scikit-learn, Pickle, Joblib.

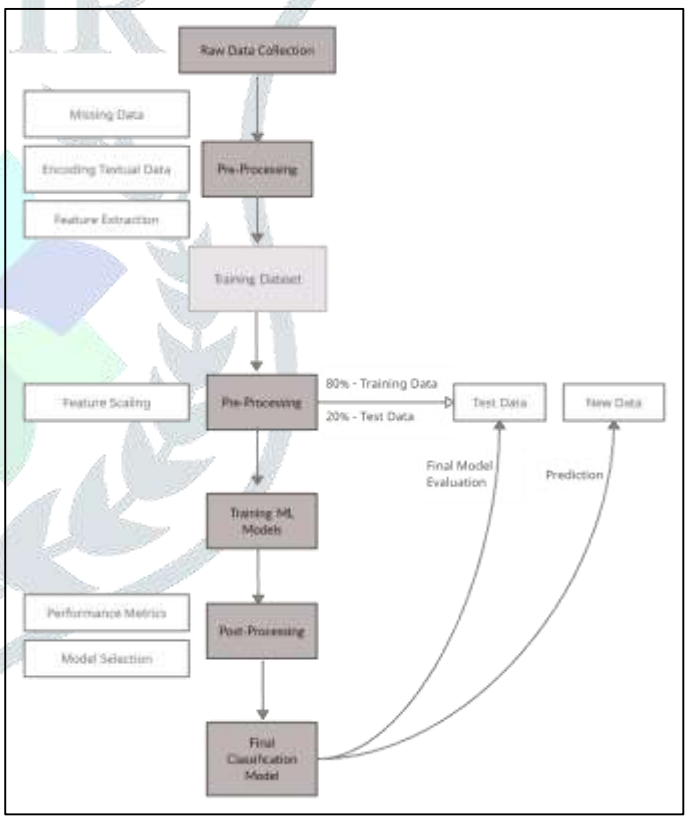

Figure 2 – Workflow

1) Clean the missing values in both training and testing data.
2) Applying Label Encoder to convert objects into integer.
3) Splitting the data in training and testing data.
4) Training ML Models:
   a) Decision Tree
   b) Logistic Regression
   c) KNN (K-nearest neighbor)
   d) SVM (Support Vector Machine)
   e) Random Forest

5) Calculating accuracy score for each model.
6) Model Selection with highest accuracy score
7) Create a GUI and extract that model into GUI module
8) Enter the new data for which stroke has to be predicted
9) Result: -Predicted data with respect to selected model.

After implementing the methodology, modules, algorithms and codes, the system will finally yield expected outcome. The homepage will help user to input the necessary data required for stroke prediction linked and the GUI part is made quite flexible for common people. Knowledge of the desired output helps to reach the destination effectively.

As given in the Implementation, the system has been build using 5 different ML algorithms to get the best possible accuracy using our dataset.

The following Results have been generated:

- The lowest accuracy is given by Logistic Regression Algorithm i.e. (76.96%)
- The highest accuracy is given by Random Forest Algorithm i.e. (98.56%)

Below is the Table of Evaluation parameters vs Models result (This was obtained from best dataset split for training and testing i.e. 70% for training and 30% for testing):

Evaluation parameters / Models	Accuracy Score	Recall Score	Precision Score	F1 Score
Decision Tree	97.39	100.00	94.95	97.41
Logistic Regression	76.96	82.22	73.87	77.82
KNN	91.64	100.00	85.86	92.16
SVM	82.00	88.84	77.73	82.92
Random Forest	**98.56**	**100.00**	**97.54**	**98.56**

Table 1 – Evaluation Parameters vs Models Result

So, we then imported the random forest trained model for testing against user-defined data and it is shown in Figure 3.

Figure 3 – Exporting Trained Model

The Comparison between all the Trained Models accuracy score is shown in the Figure 4 with the help of Matplotlib library of Python.

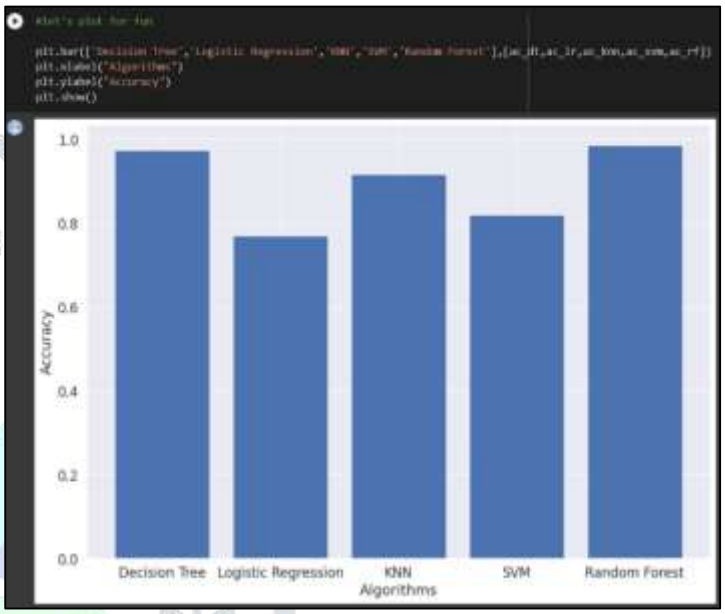

Figure. 4 – Trained Models Accuracy Score comparison

The proposed system helped us to analyse the best possible way to take inputs from the user in the GUI implementation part of our project shown in the Figure 5 and with the data that is provided to the GUI for stroke risk prediction, the Random Forest model, which was trained with the dataset, was used and the new data provided by the user was tested against the trained model.

Figure 6 - For Stroke

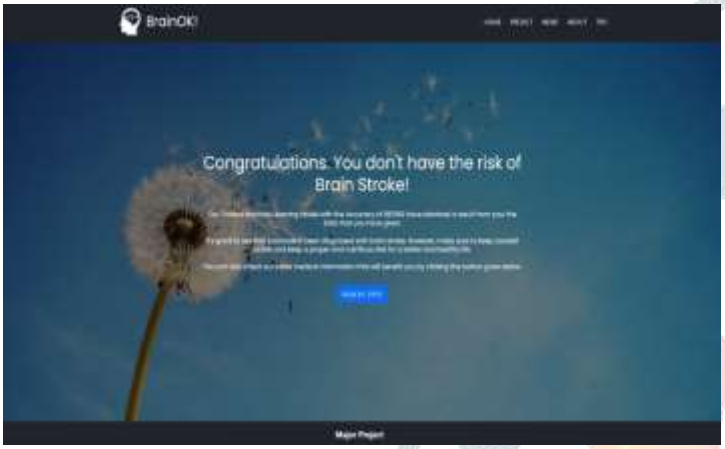

Figure 7 - For No Stroke

IV. Conclusion

After the literature survey, we came to know various pros and cons of different research papers and thus, proposed a system that helps to predict brain strokes in a cost effective and efficient way by taking few inputs from the user side and predicting accurate results with the help of trained Machine Learning algorithms. Thus, the Brain Stroke Prediction system has been implemented using the given 5 Machine Learning algorithm given a highest accuracy of 98.56%. The system is therefore designed providing simple yet efficient User Interface design with an empathetic approach towards their users and patients. The system has a potential for future scope which

Figure 5 – GUI

could lead to better results a better user experience. This will help the user to save their valuable time and will help them to take appropriate measures based on the results provided.

The future scope for the implemented system can be:

1. Increasing the accuracy of the model.
2. Additional information about brain stroke can be explained.
3. Allowing users to visualize their results based on their inputs.

V. References

[1] Manisha Sirsat, Eduardo Ferme, Joana Camara, “Machine Learning for Brain Stroke: A Review,” Journal of stroke and cerebrovascular diseases: the official journal of National Stroke Association (JSTROKECEREBROVASDIS), 2020.

[2] Harish Kamal, Victor Lopez, Sunil A. Sheth, “Machine Learning in Acute Ischemic Stroke Neuroimaging,” Frontiers in Neurology (FNEUR), 2018.

[3] Chuloh Kim, Vivienne Zhu, Jihad Obeid and Leslie Lenert, “Natural language processing and machine learning algorithm to identify brain MRI reports with acute ischemic stroke,” Public Library of Science One (PONE), 2019.

[4] R. P. Lakshmi, M. S. Babu and V. Vijayalakshmi, "Voxel based lesion segmentation through SVM classifier for effective brain stroke detection,” International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), 2017.

[5] J. Yu et al., "Semantic Analysis of NIH Stroke Scale using Machine Learning Techniques," International Conference on Platform Technology and Service (PlatCon), 2019,

[6] Gangavarapu Sailasya and Gorli L Aruna Kumari, “Analyzing the Performance of Stroke Prediction using ML Classification Algorithms,” International Journal of Advanced Computer Science and Applications (IJACSA), 2021.

[7] "Stroke Prediction Dataset". Kaggle.Com, 2021, https://www.kaggle.com/fedesoriano/stroke-prediction-dataset. Accessed 6 Oct 2021.