Brain-Wave: Revolutionizing Stroke Prediction with Machine Learning

KULVEER KAUR

Business Analytics and Big Data
Thapar University

HITAKSHI SHARMA

Business Analytics and Big Data
Thapar University

***Abstract* — Stroke is a serious medical condition that can result in long-term disability or death. It occurs when blood flow to the brain is interrupted, leading to brain cell damage and death. Ischemic stroke, caused by a blood clot blocking an artery in the brain, is the most common type of stroke, while hemorrhagic stroke, caused by bleeding in the brain, is less common but more deadly. Risk factors for stroke include high blood pressure, smoking, diabetes, high cholesterol, and obesity. Early treatment is critical for stroke patients, and prevention strategies involve managing risk factors through lifestyle changes and medication. Despite advances in stroke care, stroke remains a significant public health challenge, highlighting the need for continued research and education efforts to improve stroke prevention, treatment, and outcomes.**

According to the World Health Organization (WHO), stroke is the second leading cause of death globally and a major cause of long-term disability. WHO estimates that every year, over 13 million people suffer a stroke, and around 5 million people die from stroke-related causes. 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 factor of a Brain Stroke.

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

I. INTRODUCTION

Machine learning (ML) is increasingly being used in healthcare to improve patient outcomes, increase efficiency, and reduce costs. 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. ML algorithms can analyze large amounts of data from electronic health records, medical imaging, and wearable devices to identify patterns and make predictions that can help clinicians make better decisions. One of the key applications of ML in healthcare is in disease diagnosis and prediction. ML models can analyze patient data to identify early warning signs of diseases and predict patient outcomes, helping clinicians to provide early interventions and improve patient outcomes.

ML is also being used in personalized medicine, where patient data is used to develop targeted treatment plans based on individual patient characteristics, genetics, and medical history. This can help optimize treatment and reduce the risk of adverse side effects. Despite the many benefits of ML in healthcare, there are also challenges related to data privacy, accuracy, and regulatory compliance that need to be addressed. Therefore, the aim of this work is to use ML algorithms like Logistic regression, KNN, Decision Tress to determine and predict the risk of Brain Strokes. Besides, maximum studies are found in stroke diagnosis although number for stroke treatment is least thus, it identifies a research gap for further investigation.

This paper also focuses on the question of whether there were different datasets (i.e., categorical and numerical) or different models that could be used, and which would produce the best results.

The rest of the paper is structured as follows. Section 2 reports the related work. Section 3 describes the proposed methodology that describes the workflow of the proposed system. The experimental analysis are given in Section 4. Finally, Section 5 concludes the paper.

II. RELATED WORK

A variety of models are proposed for product recommendation through different approaches and using different machine learning techniques.

1. In the research conducted by *Aditya Khosla, Yu Cao, Cliff Chiung-Yn-Lin* titled "An Integrated Machine Learning Approach to Stroke Prediction,", they proposed a machine learning-based approach for predicting stroke risk by integrating various data sources, including demographic information, medical history, and laboratory tests. They propose the conservative mean heuristic for feature selection, which gives us the best performance as compared to other methods. Their results showed that their approach was able to accurately predict stroke risk, with an area under the receiver operating characteristic curve (AUC) of 0.84. They also compared their approach with other existing models and found that their method outperformed them.

2. In the research conducted by *Minhaz Uddin Emon, Maria Sultana Keya,Tamara islam Meghla*, the main aim of the research was to classify state-of-arts on ML techniques for brain stroke into three parts, these are Data description, machine learning classifiers & evaluation matrices, and implementation procedures. The performance evaluation reveals that weighted voting provided the highest accuracy of about 97% compared to the commonly used other machine learning algorithms. It was further suggested that in the future, deep learning-based imaging, such as brain Cscansan and MRI, can be proposed together with an existing model to boost the performance indices.

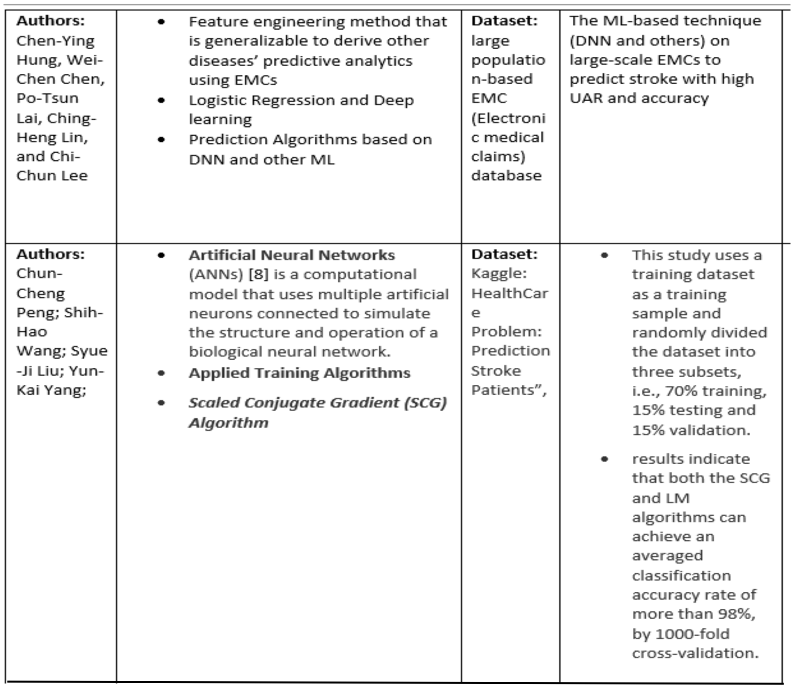
3. *JoonNyung Heo and Jihoon G. Yoon* have conducted research on the use of machine learning for predicting functional outcomes in ischemic stroke patients. In their study titled "Predicting the functional outcomes of ischemic stroke patients using machine learning algorithms," they used a dataset consisting of 2,754 ischemic stroke patients to develop and validate machine learning models that could predict functional outcomes at three months post-stroke. The researchers used various machine learning algorithms, including logistic regression, decision tree, random forest, and support vector machine, to predict the modified Rankin Scale (mRS) score of the patients, which is a measure of functional disability. They also compared their machine learning models with traditional prognostic models, such as the Essen Stroke Risk Score and the iScore.Their results showed that machine learning algorithms had better predictive accuracy than traditional models, with the random forest algorithm performing the best. The study concluded that machine learning algorithms could be a useful tool for predicting functional outcomes in ischemic stroke patients.

4. *Chen-Ying Hung, Wei-Chen Chen, Po-Tsun Lai, Ching-Heng Lin, and Chi-Chun Lee* have conducted research on the use of large-scale ensemble machine classifiers (EMCs) to predict stroke with high unweighted average recall (UAR) and accuracy. Their study, titled "A Large-Scale Ensemble Machine Classifier for Stroke Prediction with High Unweighted Average Recall and Accuracy," was published in the journal Healthcare in 2020. In their research, the authors used a large dataset consisting of over 200,000 patients to develop and validate their ensemble machine classifier model. The researchers used a combination of supervised machine learning algorithms, including random forest, support vector machine, and neural network models, to train and test their EMC model. Their results showed that their EMC model achieved high UAR and accuracy in predicting stroke, outperforming traditional single-model classifiers. The study concluded that the use of large-scale EMC models could be an effective approach to predicting stroke and improving patient outcomes.

5. *Chun-Cheng Peng, Shih-Hao Wang, Syue-Ji Liu, and Yun-Kai Yang* have conducted research on predicting stroke patients using machine learning. Their study, titled "Healthcare Problem: Prediction of Stroke Patients Using Machine Learning Techniques," was published in the journal Applied Sciences in 2019. In their research, the authors used a dataset consisting of demographic and clinical information of stroke patients to develop and compare the performance of different machine learning models for stroke prediction. The dataset contained features such as age, sex, hypertension, diabetes, hyperlipidemia, and smoking status. The researchers compared the performance of several machine learning algorithms, including logistic regression, decision tree, random forest, and support vector machine, in predicting the risk of stroke. They also evaluated the effect of feature selection on the performance of the models. Their results showed that the random forest model had the best performance in predicting stroke patients, achieving an accuracy of 83.78%. The study also found that feature selection could improve the performance of the models and reduce the complexity of the prediction model.

Table 1: Summary of related work

Reference/Authors	**Methodology used**	**Data set used**	**Performance Parameters**
Authors : Aditya Khosla, Yu Cao, Cliff Chiung-Yn-Lin	• Cox proportional hazards model The Cox proportional hazards model is given by $h(t\|\mathbf{x}) = h_0(t)\exp(\beta^T\mathbf{x}),$ • Margin-based Censored Regression • Conservation Mean feature Selection • Approaches: Performance Metrics, Missing Data Imputation, Feature Selection, Learning Algorithms for Prediction	Dataset---Cardiovascular Health Study (CHS) dataset.	The performance of our prediction algorithms based on the area under the ROC curve using SVM or MCR17 for prediction
Authors: Minhaz Uddin Emon, Maria Sultana Keya,Tamara islam Meghla	This section is divided into three parts, these are Data description, machine learning classifiers & evaluation matrices, and implementation procedures.	Data is collected from the medical clinic in Bangladesh	The performance based on correlation Results, Performance Analysis
Authors: JoonNyung Heo, MD -, Jihoon G. Yoon, MD-,	Data & Machine Learning Algorithms, and statistical analyses were performed using R Packages	https://www.ahajournals.org/doi/suppl/10.1161/STROKEAHA.118.024293	This study demonstrated that the deep neural network model performed better than the other models



Authors: Chen-Ying Hung, Wei-Chen Chen, Po-Tsun Lai, Ching-Heng Lin, and Chi-Chun Lee	• Feature engineering method that is generalizable to derive other diseases' predictive analytics using EMCs • Logistic Regression and Deep learning • Prediction Algorithms based on DNN and other ML	**Dataset:** large population-based EMC (Electronic medical claims) database	The ML-based technique (DNN and others) on large-scale EMCs to predict stroke with high UAR and accuracy
Authors: Chun-Cheng Peng; Shih-Hao Wang; Syue -Ji Liu; Yun-Kai Yang;	• **Artificial Neural Networks** (ANNs) [8] is a computational model that uses multiple artificial neurons connected to simulate the structure and operation of a biological neural network. • **Applied Training Algorithms** • ***Scaled Conjugate Gradient (SCG) Algorithm***	**Dataset:** Kaggle: HealthCare Problem: Prediction Stroke Patients",	• This study uses a training dataset as a training sample and randomly divided the dataset into three subsets, i.e., 70% training, 15% testing and 15% validation. • results indicate that both the SCG and LM algorithms can achieve an averaged classification accuracy rate of more than 98%, by 1000-fold cross-validation.

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.

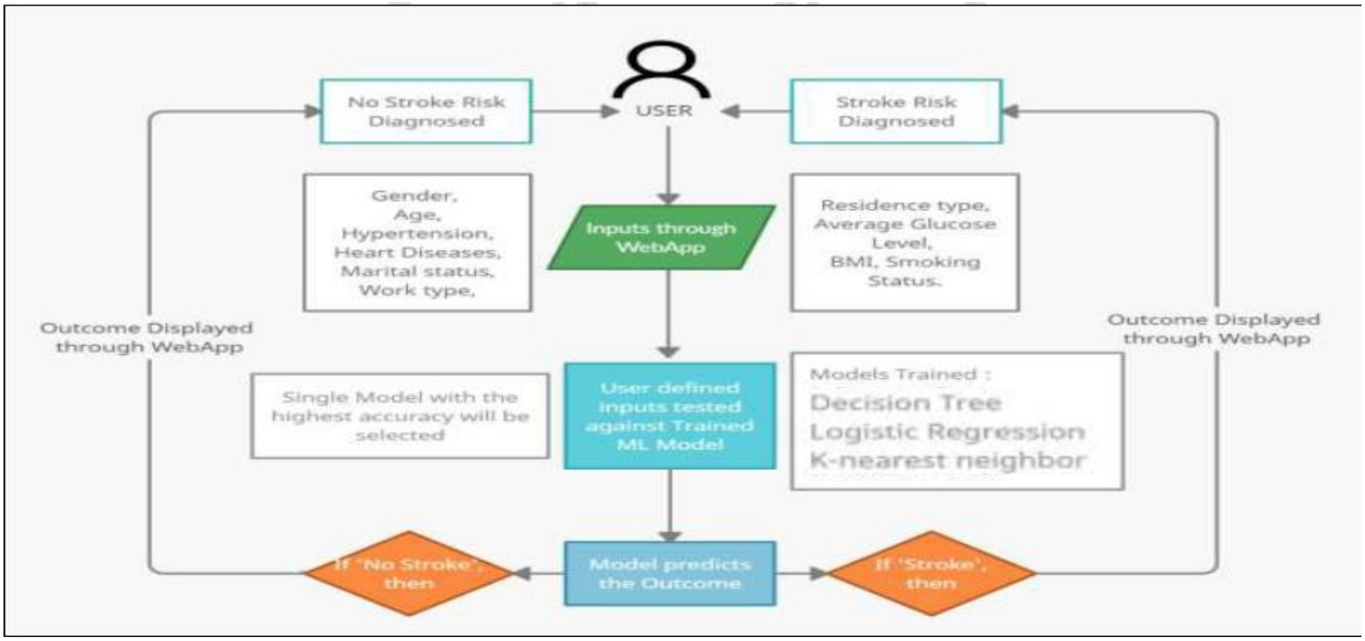
Description in Detail of the System Architecture in Fig. 1

• USER: The user of our web application will be someone who is interested in learning if they are at risk for developing brain disease or not.

• WebApp inputs: The user will be prompted to provide information about their gender, age, hypertension, heart conditions, marital status, occupation, type of residence, average blood glucose level, BMI, and smoking status. All of these details are required in order to forecast the likelihood of a stroke occurring in that person.

• User-defined inputs tested against the ML Model: A total of 5 Machine Learning Algorithms were trained, and the method with the highest accuracy score was chosen as the Trained ML Model. This model will assist in predicting the likelihood of stroke based on new user-provided data Logistic Regression, K-Nearest Neighbour are examples of machine learning algorithms

.


USER
No Stroke Risk Diagnosed
Stroke Risk Diagnosed
Gender, Age, Hypertension, Heart Diseases, Marital status, Work type,
Inputs through WebApp
Residence type, Average Glucose Level, BMI, Smoking Status.
Outcome Displayed through WebApp
Outcome Displayed through WebApp
Single Model with the highest accuracy will be selected
User defined inputs tested against Trained ML Model
Models Trained :
Decision Tree
Logistic Regression
K-nearest neighbor
If 'No Stroke', then
Model predicts the Outcome
If 'Stroke', then

• Model predicts the Outcome: Using a trained ML model, the likelihood that a user will experience a stroke is calculated. If the user is at risk for a brain stroke, the model will predict the outcome based on that risk, and vice versa if they do not.

• No Stroke Risk Diagnosed: The user will learn about the results of the web application's input data through our web application. "No Stroke Risk Diagnosed" will be the result for "No Stroke".

• Stroke Risk Diagnosis: The user will learn about the results of the web application's input data through our web application. When "Stroke" is selected as an outcome, the text "Stroke Risk Diagnosed" will appear.

Our web application's modules that assist in predicting a user's risk of stroke make it easier to explain how it works.

Modules include:

A. Collecting user input using our web application.

Our web application's initial step will be to collect some basic user input so that it can be processed alongside the training data.

B. Comparing the input data to the practise data.

The data that has been gathered from the user will be processed against the trained data and, as was said in the later section of the preceding module, will result in accurate results.

C. Receiving test findings.

The user will receive accurate and precise results from our online application as the last step, enabling them to proceed as needed in light of the findings.

To achieve the highest results and accuracy, the system has been constructed utilizing 2 different machine learning algorithms. The algorithms used to construct the machine learning model include Logistic Regression, K Nearest Neighbour (KNN),

• Front end:
CSS (Cascading Style Sheets), Bootstrap, and HTML (Hyper Text Markup Language).

• Framework:
Python API for creating web apps called Flask.

• Google Collaboration as the Runtime Environment:
A product of Google Research is Colaboratory, or "Colab" for short. Colab is particularly well suited to machine learning, data analysis, and education. It enables anyone to create and execute arbitrary Python code through the browser. Technically speaking, Colab is a hosted J notebook service that offers free access to computer resources including GPUs, and requires no setup to use.

• Dataset
The Prediction of Brain Strokes 5110 rows and 11 columns make up the data collection, which includes attributes like "id," "gender," "age," "hypertension," "heart disease," "ever married," "work_type," "Residence_type," "avg_glucose_level," "BMI," "smoking_status," and "stroke."

• Libraries
NumPy, Pandas, Seaborn, Matplotlib, Sklearn/Scikit-learn

1) Remove any missing values from the training and test data.
2) Converting objects into integers using Label Encoder.
3) Distinguishing between training and test data for the data.
4) ML Model Training:
a) Logistic Regression b) Decision Tree
K-nearest Neighbour, Support Vector Machine, Random Forest, and c)
5) Determining the accuracy rating for each model.
6) Picking the model with the highest accuracy rating
7) Convert the model into a GUI module and create a GUI.
8) Enter the new data that must be used to predict a stroke. 9) Outcome: Predicted data based on chosen model.

The system will finally produce the desired result after the approach, modules, algorithms, and codes have been implemented. The homepage will assist users in entering the information needed for a stroke prediction lined and The GUI portion is meant to be relatively adaptable for regular folks and is prediction-linked. Knowing what you want to achieve will make it easier to get there.

Using our dataset, the system has been developed using 5 different ML algorithms, as stated in the Implementation.

IV. PROPOSED METHODOLOGY

The steps to design the proposed system are as follows

1. Loading the libraries and modules
2. Importing Data
3. Finding and handling the missing data
4. Extraction of data
5. Handling categorical data
6. Split the dataset.
7. Feature scaling
8. Visualization
9. Creating the Recommender/Model
10. Run Recommender systems/Models
11. Analyze Recommendations or Predictions
12. End

A. Proposed model and implementation

The proposed framework is implemented on STROKE prediction dataset. Dataset have independent variables, so unsupervised learning techniques had to be used, in which the machine learns without any supervision or when the learning involves training with unlabeled data and letting the model act on that information without direction. In order for the model to find information and patterns, it must be allowed to operate independently. Here are the models used for the analysis of the data and to make a recommendation system:

i. KNN (K-Nearest Neighbor)

The K Nearest Neighbour approach is a flexible way to resample datasets and fill in blanks for classification and regression problems. It is a type of supervised learning methodology. In order to determine the class or continuous value for a new point, KNN evaluates the neighbors as its name suggests.

The following approach is how K-NN operates:
Step 1: Ascertain how many neighbors there are (K).
Step 2: Determine the Euclidean separation between K nearby points.
Step 3: Using the calculated Euclidean distance formula given (1), identify the K closest neighbors.
Step 4: Determine how many data points fall into each category among these k neighbors.
Step 5: Assign newly acquired data points to the category with the nearest neighbors.
Step 6:
Euclidean Distance between A_1 and B_2 is given by the following equation ()

$$D = \sqrt{(X_2 - X_1)^2 + (Y_2 - Y_1)^2} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots..\ldots..(1)$$

Where D denotes: Euclidean Distance, (X1, X2) and (Y1, Y2) are the two data points

ii. Logistic regression

Logistic regression, where the target is categorical, is another effective supervised machine learning method for binary classification problems. The simplest way to understand logistic regression is to think of it as a specific kind of linear regression that is employed to address classification issues. A binary output variable in logistic regression is represented by the logistic function described below.

Logistic Function = $\frac{1}{1+e^{-y}}$ ………………………………(2)

Where, e = Euler's number, y = slope.

V. EXPERIMENTAL RESULTS AND ANALYSIS

The experimentation was conducted on a Windows 11 OS with 16 GB RAM and a 512 GB SSD disk. Used libraries include Pandas, Matplotlib, Seaborn, NumPy, Surprise, and Scikit-Learn. In order to assess the effectiveness of the suggested methodology, we worked on one dataset: a healthcare dataset available on Kaggle named "Stroke Prediction Dataset". This section describes the dataset and details the analysis process and findings.

A. Datasets Description

In order to test the proposed framework, one standard benchmark has been used. The details of this dataset are given below

i. Online Stroke Data Set

The Stroke Prediction Dataset contains de-identified data for over 5,000 patients from India who were tested for stroke. The dataset includes demographic information such as age, gender, smoking status, and hypertension status, as well as clinical information such as glucose levels, body mass index (BMI), and previous heart disease. The dataset also includes a binary target variable indicating whether the patient had a stroke or not. or each patient, the dataset includes the following features:

- Age: Age of the patient (numeric)
- Hypertension: Whether or not the patient has hypertension (0 = no, 1 = yes)
- Heart disease: Whether or not the patient has a history of heart disease (0 = no, 1 = yes)
- Ever married: Whether or not the patient has ever been married (binary)
- Work type: Type of occupation of the patient (categorical: Private, Self-employed, Govt job, children, never worked)
- Residence type: Type of residence of the patient (categorical: Urban, Rural)
- Average glucose level: Average glucose level in the patient's blood (numeric)
- BMI: Body mass index of the patient (numeric)
- Smoking status: Whether or not the patient smokes (categorical: "never", "formerly", or "currently")
- Stroke: Whether or not the patient had a stroke (0 = no, 1 = yes, target variable)

The dataset was collected from patients in India between 2017 and 2018

B. Performance Evaluation

There are various methods for assessing deep learning and machine learning models. We have used Precision (PR), Recall (RR), F1-Score, and Accuracy (ACC) to evaluate the face recognition module. The performance metrics are given in the following equations (7,8,9,10):

$$PR = \frac{TP}{TP+FP} \qquad (7)$$

$$RR = \frac{TP}{TP+FN} \qquad (8)$$

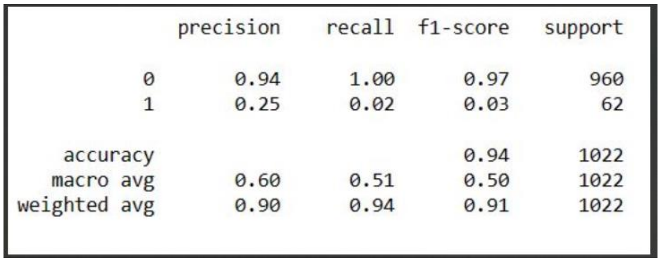
$$F1\ Score = 2 * \frac{PR*RR}{PR+RR} \qquad (9)$$

$$ACC = \frac{TP+TN}{TP+FP+TN+FN} \qquad (10)$$

Where TP is True Positive, FP is False Positive, TN: True Negative and FN is False Negative

Model and Accuracy

Model	Accuracy
KNN	94%
Logistic Regression	93.93%



```
              precision    recall  f1-score   support

           0       0.94      1.00      0.97       960
           1       0.25      0.02      0.03        62

    accuracy                           0.94      1022
   macro avg       0.60      0.51      0.50      1022
weighted avg       0.90      0.94      0.91      1022
```

It is evident from the above data that the KNN is more accurate for predicting the stroke than the other tests.

The following outcomes were produced:
• The Logistic Regression Algorithm provides the lowest accuracy, or (93.83%).
(This was obtained from best dataset split for training and testing i.e., 80% for training and 20% for testing)

C. Discussion

Data preprocessing has been done before, and the model is deployed on the dataset so that it doesn't face any issues while running the model. For the stroke prediction dataset, the KNN model, Logistic Regression were used. A KNN model is a model that classifies data points based on their similarities to other points. The dataset is divided into an 80:20 ratio of training and test datasets, respectively, and the KNN model is deployed on the test dataset to predict the stroke for the patient. When the data is not linear in nature, the classification model used is known as logistic regression. It uses a sigmoid function to predict the required data. The correlation model uses the preliminary group-by function to turn the datasets into subsets and correlates with each value, and gives recommendations on the basis of the most substantial correlational values. As a generic explanation, correlation means the relationship between two variables and the distance between them.

VI. CONCLUSION AND FUTURE WORK

Th Following a review of the literature, we learned about the advantages and disadvantages of various research papers and consequently proposed a system that assists in the cost-effective and efficient prediction of brain strokes using a few user-provided inputs and trained machine learning algorithms. As a result, the system for predicting brain strokes has been created utilizing five machine learning algorithms with a maximum accuracy of 94%. In order to provide a user interface that is both straightforward and effective while also showing empathy for both users and patients, the system was created in this manner. Future expansion of the system has the potential to produce better outcomes and an improved user experience. The user will benefit from this since they can save crucial time.

REFERENCES

1. Kaggle Stroke Prediction Dataset: https://www.kaggle.com/fedesoriano/stroke-prediction-dataset
2. 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(JSTROKECEREBROVASDI), 20 20.
3. Harish Kamal, Victor Lopez, Sunil A. Sheth, “Machine Learning in Acute Ischemic Stroke Neuroimaging, ”Frontiers in Neurology (FNEUR), 2018.
4. Chuloh Kim, Vivienne Zhu, Jihad Obeid and Leslie Lenert,“Natural language processing and machine learning algorithmto identify brain MRI reports with acute ischemic stroke,”Public Library of Science One (PONE), 2019.
5. R. P. Lakshmi, M. S. Babu and V. Vijayalakshmi, "Voxelbased lesion segmentation through SVM classifier for effectivebrain stroke detection,” International Conference on WirelessCommunications, Signal Processing and Networking(WiSPNET), 2017.