International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 04 | April - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Brain Stroke Prediction Using Machine Learning Algorithm

Mr. Pradip Bhakare[1], Anushka Burade[2], Kiran Mali[3], Tapasya Gawai[4], Mayuri Hingane[5]

[1]Asst.Professor, Department of ENTC, MGICOET Shegaon, Maharashtra.
[2,3,4,5] Students, Department of ENTC, MGICOET Shegaon, Maharashtra.

***

Abstract - Brain stroke is a serious medical illness that may result in severe neurological damage or even death if left undetected and untreated. Predicting the risk of stroke at an early stage can substantially improve clinical decision-making and prevention care. The proposed work involves using machine learning for predicting the risk of brain stroke based on patient health data. The used dataset includes related features like age, Hypertension, heart disease, smoking status, BMI, glucose level, and work type was employed. Different machine learning algorithms such as Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN) were tried and evaluated against performance measures such as accuracy, precision, recall, and F1-score. The highest prediction accuracy was found in the random Forest model, implying that it could be used to accurately classify based on this problem. This research demonstrates the potential of machine learning to enhance stroke diagnosis and assist Healthcare decision – making.

Key Words: Brain Stroke, Stroke Prediction, Random Forest, Health Data, Classification, Early Detection, Health Care Analytics.

I. INTRODUCTION

Brain stroke is a serious medical condition that results when the blood supply to a region of the brain is cut off or diminished, depriving brain tissue of oxygen and nutrients. This can cause damage or death to brain cells within minutes, leading to long-term disability or death if not treated immediately. Strokes are generally categorized into two forms: ischemic strokes, resulting from blockages, and hemorrhagic strokes, resulting from bleeding within the brain.

Stroke is ranked among the primary causes of death and permanent disability in the world, according to world health statistics.

A brain stroke results when a section of the brain is deprived of its blood supply or has its blood supply diminished, thereby not allowing the brain tissue to receive oxygen and Nutrients, which may result in to brain damage or death. Strokes usually result from a blocked artery (ischemic stroke) or the breaking of a blood vessel (hemorrhagic stroke).

In a few instances, there is a temporary blockage in the form of a transient ischemic attack (TIA), commonly known as a “mini-stroke.” Sudden numbness or weakness in the Face, arm, or leg (especially on one side), difficulty speaking or understanding speech, Vision problems, dizziness, and severe headache. Stroke is a medical emergency that needs immediate treatment to reduce brain damage and enhance chances of recovery. Early diagnosis and treatment can very much minimize the possibilities of long-term disability or death.

PROBLEM STATEMENT

Brain stroke is a severe medical disorder that may cause long-term disability or fatality if it is not detected and treated in its early stage. The project focuses on designing a machine learning model with the help of a Random Forest Classifier to make an estimate of how likely a person is to get a brain stroke. The system will examine the data of a patient his or her age, hypertension, heart disease, BMI, level of glucose, and other relative features to ascertain individuals at maximum risk. Early prediction will help the system support doctors in their decisions and thereby save lives earlier.

OBJECTIVE

- To develop a predictive model for brain stroke based on the Random Forest Classifier algorithm.
- To predict patient health parameters (e.g., age, hypertension, heart disease, glucose level, BMI).

- To categorize people based on their risk of stroke.
- To help healthcare professionals with early detection and decision-making.
- To enhance patient outcomes through timely prediction and intervention.

II. LITERATURE REVIEW

This study employs a variety of physiological parameters and machine learning algorithms, including Logistic Regression (LR), Decision Tree (DT) Classification, Random Forest (RF) Classification, and Voting Classifier, to train four models for accurate prediction. Random Forest was the top performing algorithm for the task with an accuracy of about 96 percent. The dataset employed in method development was the open-access Stroke Prediction dataset. [1]

In this paper, we propose a novel stroke detection algorithm to predict the occurrence of stroke. We employed a dataset with must detail of crucial parameters which are responsible for the brain stroke such as: Age: Body Mass Index (BMI): Gender: Heart Disease: Smoking Status etc. to generate a prediction model. The performance of the models was compared using various metrics like accuracy, F1 score, recall, precision. Our results indicated that the Support Vector Machine algorithm performed better than other models, with an accuracy of 99.5%, precision of 99.9%, recall of 99.1%, F1-score of 99.5% and specificity of 99%.[2]

The primary objective of this research is to forecast the likelihood of a brain stroke occurring at an early stage using machine learning algorithms. Different classification models such as, K-Nearest Neighbor (K-NN), Logistic Regression (LR), Random Forest (RF), were effectively applied in this study for classification research. The performance of the method is tested by Precision, Recall, and F1 Score evaluation metrics. [3]

The major aim of the project is to enhance the accuracy of stroke detection, which encompass real-time deployment and interpretability problems with logistic Regression. Our stroke prediction algorithm performance was tested by confusion Metrics-including accuracy, precision, recall and F1-score. At the conclusion of research, Random Forest model provided an accuracy of 98.5% against the current model logistic Regression that has 86% accuracy. [4]

We investigate into the clinical, biochemical and neuroimaging factors related with the outcome of stroke patients to generate a predictive model utilizing machine learning techniques for prediction of mortality and morbidity 3-months post admission. Finally, machine learning algorithms RF could be effectively utilized in stroke patients for long-term outcome prediction of mortality and morbidity. [5]

The aim of this paper is to forecast the stroke vulnerability status of patients using a random forest (RF) machine learning model. The model was developed based on Python computer programming language from healthcare dataset stroke data collected from the Kaggle machine learning dataset repository. The dataset was well cleaned and the clean dataset was utilized to train the random forest machine learning model for effective forecasting of stroke the results obtained from the random forest model were evaluated using a confusion matrix and it was found that random forest is a very good choice of algorithm for predicting stroke Vulnerability as evidenced in its prediction accuracy of 93%. [6]

This study explores these deep learning models' efficiencies in contrast to other Approaches that rely on stacking techniques. High accuracy will be obtained by the trained models based on CNN algorithms, VGG-16, and ResNet, which achieve precision, recall, specificity, and an F1-score. [7]

In other work, they have trained the models for correct prediction with machine learning approaches such as Logistic Regression Algorithm, Decision Tree Classification algorithm, Random Forest Classification Algorithm, K-Nearest Neighbors Algorithm, Support Vector Machine Algorithm, and Naïve Bayes Classification algorithm. The algorithm that did this job with the best accuracy was Support Vector Machine, with an accuracy of nearly 94.6%. In this suggested analysis, the process of predicting stroke has been improved by utilizing Random Forest Classification and XGBoost Algorithm techniques. Random Forest Classification and XGBoost Algorithm, with an accuracy of around 98%, was the algorithm that did these tasks best. [8]

This paper delves into the creation of an effective framework for the prediction of chronic stroke prevalence with the use of CNN, ResNet, and the VGG-16 deep learning algorithms. These have enhanced decision support leading to the early identification of stoke because it can analyze big image dataset. Measures such as reliability, F1-score, recall, precision, and specificity are constructed using the Sigmoid and ReLU activation functions. This research investigates these deep learning models' efficiencies compared to other methods based on stacking techniques. The trained models will achieve high accuracy with the algorithms of CNN, VGG-16, and ResNet, which achieve precision, recall, specificity, and an F1 score. [9]

III. METHODOLOGY

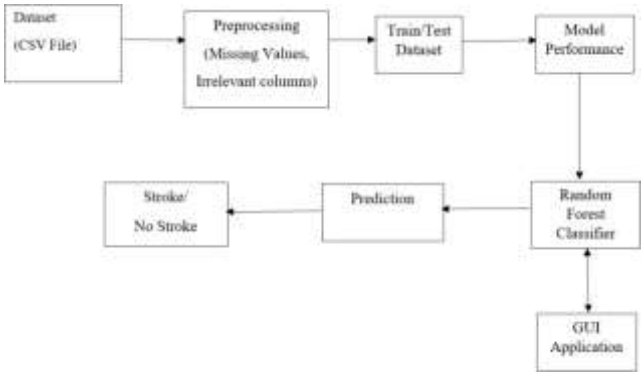

Fig.3.1: Block Diagram.

Data Collection:-Kaggle or publicly available medical data on brain stroke prediction. It contains 5,110 records, where each row corresponds to a single patient.

Table. 3.1: Patient Health Dataset.

ATTRIBUTE	DESCRIPTION
Id	Unique Identifier
Gender	"Male", "Female", "Other"
Age	Age of the Patient
Hypertension	0 if the patient doesn't have hypertension, 1 if the patient has hypertension
Heart Diseases	0 if the patient doesn't have heart diseases, 1 if the patient has heart diseases
Ever Married	"No" or "Yes"
Work Type	"Children", "Govt_job", "Never_worked", "Private" or "Self-employed"
Residence Type	"Rural" or "Urban"
Average Glucose Level	Average glucose level in blood
BMI	Body Mass Index
Smoking Status	"Formerly smoked", "Never smoked", "Smokes" or "Unknown"
Stroke	1 if the patient had a stroke or 0 if not

Data Preprocessing:-Handling Missing Values: Fill or drop rows containing missing bmi or smoking_status. Encoding Categorical Variables: Use one-hot encoding or label encoding for non-numeric fields.

Feature Selection:-The Feature importance plot shows how much each feature contributes to predictions. The more important features have higher bars. From Result created a Feature importance plot, look at the bars. The higher the importance value of the feature, the more it contributes to the predictions of the model. This can be used to find out which factors are important for predicting stroke. The most important feature is age (0.16) while the most insignificant feature is bmi (0.09).

Division of the Dataset:-• Training Set: 80%, Used to train the machine learning model. The model Learns patterns and relationships from this data. • Testing Set: 20%, Used for testing the model's performance on unseen data, assisting in verifying how well it generalizes.

Model Selection-Random Forest classifier:-Random Forest Classifier was chosen for this brain stroke prediction model because it is strong, highly accurate, and can efficiently handle both numerical and categorical variables. It's an ensemble learning method where multiple decision trees are constructed during training and the mode of the class is returned for classification problems. This method over fitting and enhances generalization to a single decision tree.

A Graphical User Interface (GUI):-The GUI was implemented with Tkinter, which is a built-in Python library for developing desktop applications. It offers input fields for users to input patient information like age, gender, hypertension status, average glucose level, BMI, and smoking status. After the user provides the necessary input and presses the "Predict" button, the GUI executes the data and provides the outcome of the prediction either "Stroke" or "No Stroke" on the basis of the trained Random Forest Classifier model.

Table. 3.2: Software Requirements for Brain Stroke Prediction System.

COMPONENT	DESCRIPTION
Programming Language	Python 3.8+
Libraries Used	Pandas, Numpy, Scikit-learn, tkinter
Development Tools	Visual Studio Code/ Google Colab
Operating System	Window 10/11
Machine Learning Model	Random Forest Classifier(Supervised Learning)
GUI Toolkit	Tkinter(Standard Python Library for GUI)
Dataset Format	CSV(Comma-Separated Values)

IV. RESULT ANALYSIS

This study predicts the stroke for a patient based on classification methodologies. This study brings out the effectiveness of the classification methods for structured entities like patient case sheets to detect the stroke based on the parameters (symptoms) and factors.

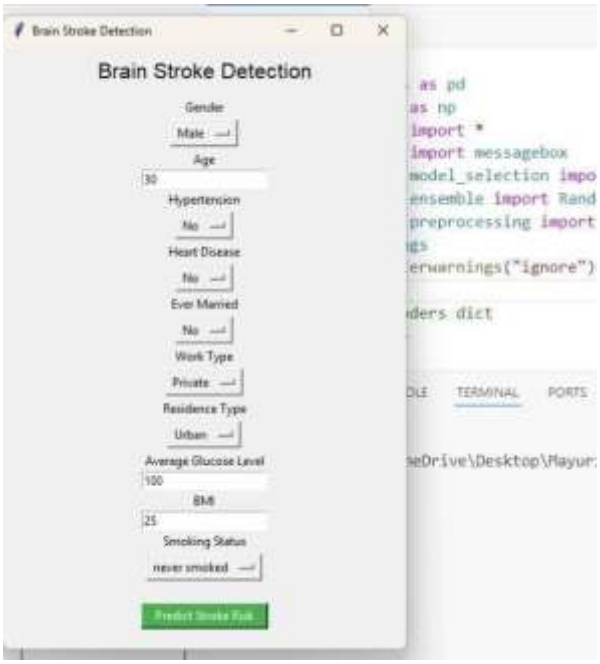

Fig.4.1: Input Values for Stroke Prediction. (This are the Input Values Given by the User)

By the classification algorithm the data given by the user is classified that user had a chance to get Stroke then it shows YES else it shows No.

4.1. Having No Stroke

Fig. 4.2: Representation of Output for Having No Stoke.

The "Brain Stroke Detection" application presents a simplistic form-based interface to predict the stroke risk. The users have to enter personal and health related information such as age, hypertension, history of heart disease, marital status, type of work, type of residence, average glucose level, BMI, and smoking. Once the data is entered by clicking on the "Predict Stroke Risk" button, the system calculates the data displays the result of the prediction as a pop-up message. Here, despite certain risk factors such as heart disease and smoking, the model had diagnosed the user with no risk of stroke, providing a congratulatory message. The backend is supported by Python libraries such as Pandas, NumPy, and a Random Forest Classifier from scikit-learn, and the interface is most probably developed using Tkinter.

4.2. Having Stroke

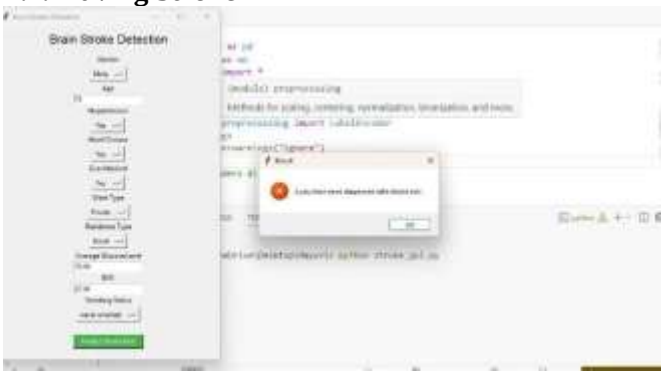

Fig. 4.3: Representation of Output for Having Stoke.

The "Brain Stroke Detection" app offers a straightforward and easy-to-use interface for stroke risk prediction based on user input. The form captures information like gender, age, hypertension, heart disease, marital status, work type, residence type, average glucose level, BMI, and smoking status. Once the information is filled in and the "Predict Stroke Risk" button is clicked, the system processes the data and shows the result in a pop-up window. In this case, the user was diagnosed with a possible stroke risk. The app is built using Python, using libraries such as Pandas, NumPy, and scikit-learn for data preprocessing, and the GUI is probably developed using Tkinter.

V. CONCLUSION

In conclusion, the application of the Random Forest machine learning algorithm to predict brain stroke has shown encouraging outcomes in terms of accuracy, stability, and interpretability. Random Forest as an ensemble learning technique takes advantage of the collective decision making of a set of decision trees to minimize the risk of over fitting. In the case of stroke prediction, variables like age, hypertension, heart disease, smoking status, and glucose level are important factors in determining the risk of stroke. The project assists us in predicting the patients who are diagnosed with brain stroke by preprocessing the dataset and using Random Forest model to achieve an accuracy of an average of 96.99%.

FUTURE SCOPE

Future work can investigate the application of deep learning methods such as Convolutional Neural Networks (CNNs) for brain scan or MRI image analysis, which can provide more accurate identification of stroke-prone regions. The integration of wearable health devices and IoT (Internet of Things) can enable real-time monitoring and early warning systems for high-risk patients. Furthermore, implementing the model into a mobile or web- based system with an easy-to-use GUI can help healthcare professionals and patients make quicker, data-informed decisions.

REFERENCES

[1] Sathya Sundaram. M,Pavithra. K,Poojasree. V (June 2022),"STROKE PREDICTION USING MACHINE LEARNING", International Advanced Research Journal in Science, Engineering and Technology Vol 9 Issue 6, https://www.researchgate.net/publication/362990184.

[2] Shehzada Mushtaq, Kamaljit Singh Saini, Saimul Bashir (June),"Machine Learning for Brain StrokePrediction", 2023 InternationalConference, https://ieeexplore.ieee.org/document/10151148.

[3] .K. D. Mohana Sundaram, G. Haritha,A. Abhilash(Volume 8 ~ Issue 4 (2022)),"Detection of Brain Stroke Using Machine Learning Algorithm", www.questjournals.org.

[4] Jarapala Parvathi, Saikiran Ellambotla(VOL. 12 NO. 4 (2024)),"Machine Learning based Brain Stroke Prediction using Light Gradient Boosting Machine Algorithm. https://ijisae.org/index.php/IJISAE/article/view/6734.

[5] Nnanna Chidera Egegamuka, Nnanna Ekedebe,April 2024International Journal of Innovative Science and Research Technology 9(4):2783- 2795,"Development of Random Forest Model for Stroke Prediction", https://www.researchgate.net/publication/380606767.DOI:10.38124/ijisrt/IJI SRT24APR2566.

[6] Carlos Fernandez-Lozano,Pablo Hervella,Ramón Iglesias-Rey ,Article Number: 10071 (12 May 2021),"Random forest-based prediction of stroke Outcome", https://www.nature.com/articles/s41598-021-89434-7.

[7] Moses Adah Agana,Obono I. Ofem,Ofem Ajah Ofem,2024 – Volume 6 [Issue 1],"Stroke Prediction with Random Forest Machine Learning Model", https://www.jofscience.com/index.php/ARJOCS/article/view/111.

[8] Nnanna, Chidera Egegamuka,Nnanna, Ekedebe Ajoku, Kingsley Kelechi,Volume 9, Issue(4, April – 2024),"Development of Random Forest Model for Stroke Prediction", https://doi.org/10.38 124/ijisrt /IJISRT24APR2566.

[9] Kamarajugadda Raviteja, J.Mythri,Bhuvaneshwari Jolad,(13 April 2025),"Brain Stroke Detection Using Deep Learning Approaches", https://link.springer.com/chapter/10.1007/978-981-96-3358-6_42.

[10] Kallam, B. & Shaik, A. (2022). Brain stroke prediction using supervised machine [11] Kallam, B. & Shaik, A. (2022). Brain stroke prediction using supervised machine learning. International Journal of Creative Research Thoughts, 10(6):a371-a374. www.ijcrt.org .

[11] Gupta, P., & Verma, A. (2024). Multi-Disease Prediction System with Voice-based Input Using Streamlit and Machine Learning. AI in Healthcare, 14(1), 33–45

[12] Rebouc¸as ES, Marques RCP, Braga AM, Oliveira SAF, deAlbuquerque VHC, Filho PPR (2018) New level set approach based on Parzen estimation for stroke segmentation in skull CT images. Soft Comput. https://doi.org/10.1007/s00500-018-3491-4.

[13] A. Sudha, P. Gayathri, and N. Jaisankar, "Effective analysis and predictive model of stroke disease using classification methods, "International Journal of Computer Applications, vol. 43, no. 14, pp. 26– 31, 2012.

[14] Yadav, A., & Rao, D. V. (2023). Fake News Detection Using Naive Bayes Classifier: A Comparative Study. Journal of Management http://jmss.a2zjournals.com/index.php/mss/article/view/22.

[15] Saadi, S. M., & Al-Jawher, W. A. M. (2023). Image Fake News Prediction Based on Random Forest Gradient-boostingMethods.JournalResearch.https://www.iasj.net/iasj/download/01257e0b0f4e1.