Brain Stroke Prediction Using Machine Learning

DR. SHANTHI D L[1], ASHWINI R L[2], B D N S THANMAI[3], BHAVYATH M[4], THUMBURU RAVICHANDRAKANTH[5]

[1] *Assistant Professor, Student, Department of Information Science Engineering, BMS Institute of Technology and Management, Bengaluru, INDIA*

[2, 3, 4, 5] *Student, Department of Information Science Engineering, BMS Institute of Technology and Management, Bengaluru, INDIA*

Abstract*— Strokes are serious medical emergencies that require immediate attention. They can lead to significant impairments in various bodily functions and cognitive abilities. Understanding the importance of maintaining healthy blood flow to the brain is crucial for preventing strokes and minimizing their impact. It's also essential to recognize the signs of a stroke, such as sudden weakness or numbness in the face, arm, or leg, especially on one side of the body, sudden confusion, trouble speaking or understanding speech, sudden trouble seeing in one or both eyes, sudden trouble walking, dizziness, loss of balance, or coordination, and sudden severe headache with no known cause. Seeking prompt medical care if you or someone you know experiences these symptoms can make a significant difference in the outcome. This project centers on creating a web application for Stroke Prediction utilizing Machine Learning. A stroke, a critical medical event, happens when blood flow to the brain is interrupted, resulting in potentially fatal outcomes. Numerous elements influence the risk of stroke, such as high blood pressure, diabetes, and lifestyle preferences. To refine stroke prediction, we utilize machine learning algorithms, including Logistic Regression, K-Nearest Neighbors (KNN), Extratree Classifier, Gaussian Naive Bayes, Decision Tree, and Random Forest Classifier. These algorithms scrutinize user input data, taking into account a variety of risk factors. The experimental result shows that the Extratree Classifier achieves highest accuracy of 84%. We have developed the Web Application using flask framework to demonstrate the brain stroke prediction using Machine Learning.

Index Terms – Machine Learning, Classifier, Accuracy, Brain Stroke, Prediction.

I. INTRODUCTION

Strokes are indeed a critical medical condition that impact the brain similarly to how heart attacks affect the heart. Understanding the causes, risk factors, and prevention measures is essential for reducing the incidence of strokes and improving outcomes for those who do suffer from them. Several factors increase the risk of having a stroke. These can be divided into lifestyle risk factors and medical risk factors such as High blood pressure , Smoking , Poor diet Physical inactivity and obesity. The different body parts and how they function are the foundation of human life. A hazardous condition that ends human lives is stroke. After the age of 65, this condition is frequently discovered. Heart attacks influence the working of the heart, and strokes affect the brain similarly. One of these two conditions—a blood supply restriction to the brain or the rupture and bleeding of brain blood vessels—is what causes strokes. If there is a rupture or a blockage, blood and oxygen cannot reach the brain's tissues. In both industrialized and developing nations, it is currently the fifth greatest cause of death. A stroke victim's chances of making a full recovery are improved the earlier they receive medical care. Any stroke victim needs to see a doctor right away. Otherwise, it will result in death, permanent disability, and brain damage. Patients can develop stoke for a variety of reasons. Diet, inactivity, alcohol, tobacco, personal history, medical history, and complications are the main causes of stroke, according to National Heart, Lung, and Blood Institute. Predicting brain strokes is crucial for early detection and prevention of potentially life-threatening events. Brain strokes, also known as cerebrovascular accidents, occur when blood flow to the brain is interrupted, either by a blockage (ischemic stroke) or bleeding (hemorrhagic stroke). Predictive models for brain stroke aim to identify individuals at high risk based on various factors, including medical history, lifestyle choices, and physiological indicators. These models leverage advanced techniques such as machine learning algorithms and statistical analysis to analyze large datasets and identify patterns that precede stroke occurrence. By leveraging these predictive models,

healthcare professionals can proactively intervene, applying preventive strategies and lifestyle changes to lessen stroke risk factors in individuals identified as at risk. Successful stroke prediction not only preserves lives but also decreases the long-term effects and strain on healthcare systems by averting severe strokes and their related complications. Brain Strokes, often dubbed “brain attacks,” come in two distinct types: ischemic and hemorrhagic.

Ischemic Stroke: Imagine a tiny traffic jam inside your brain’s highways. Ischemic strokes occur when a blood clot or plaque build-up blocks the flow of oxygen-rich blood within an artery. This blockade can be triggered by conditions like sclerosis, clotting disorders, heart defects, or microvascular diseases.

Hemorrhagic Stroke: Now, picture a catastrophic explosion in a narrow alley. Hemorrhagic strokes happen when a blood vessel bursts open leaking blood inside the brain. High blood pressure, brain aneurysms, and brain tumours are common instigators of this explosive event.

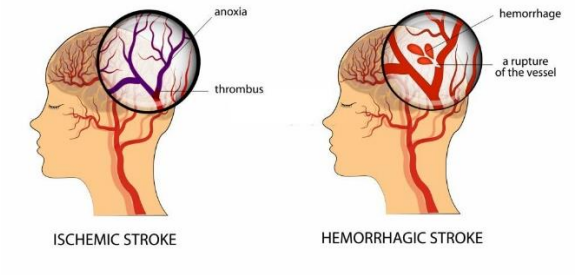


Figure 1: Ischemic And Hemorrhagic Stroke

A brain stroke prediction project might encompass a range of activities including gathering data, selecting key features, building models using techniques such as machine learning, as well as testing and deploying these models. It entails evaluating elements such as a person's medical background, lifestyle behaviors, and potentially genetic factors to assess the risk of stroke. Moreover, the inclusion of real-time monitoring tools or wearable technology could improve the precision of predictions.

Our goal is to use machine learning techniques on extensive existing datasets to accurately predict individual risk of stroke. To collect datasets of brain stroke from hospitals, dataset repositories Design and develop an algorithm for early detection of brain stroke to take precautions. Developing a web application for early stroke detection, prediction and providing remedies to the above solution.

II. RELATED WORK

Senjuti Rahman, Mehedi Hasan et al. [1] has predicted the early detection of the numerous stroke warning symptoms can lessen the stroke's severity. The main objective of this study is to forecast the possibility of a brain stroke occurring at an early stage using deep learning and machine learning techniques. To gauge the effectiveness of the algorithm, a reliable dataset for stroke prediction was taken from the Kaggle website. Several classification models, including Extreme Gradient Boosting (XGBoost), Ada Boost, Light Gradient Boosting Machine, Random Forest, Decision Tree, Logistic Regression, K Neighbors, SVM - Linear Kernel, Naive Bayes, and deep neural networks (3-layer and 4-layer ANN) were successfully used in this study for classification tasks. The Random Forest classifier has 99% classification accuracy, which was the highest (among the machine learning classifiers). The three layer deep neural network (4-Layer ANN) has produced a higher accuracy of 92.39% than the three-layer ANN method utilizing the selected features as input. The research's findings showed that machine learning techniques outperformed deep neural networks. Jeena R S et al. [2] explained the Early diagnosis of stroke is essential for timely prevention and treatment. Investigation shows that measures extracted from various risk parameters carry valuable information for the prediction of stroke. This research work investigates the various physiological parameters that are used as risk factors for the prediction of stroke. Data was collected from International Stroke Trial database and was successfully trained and tested using Support Vector Machine (SVM). In this work, we have implemented SVM with different kernel functions and found that linear kernel gave an accuracy of 90 %. research reports predictive analytical techniques for stroke diseases using deep learning model applied on heart disease dataset. The atrial fibrillation symptoms in heart patients are a major risk factor of stroke and share common variables to predict stroke. The outcomes of this research are more accurate than medical scoring systems currently in use for warning heart patients if they are likely to develop stroke

research reports predictive analytical techniques for stroke diseases using deep learning model applied on heart disease dataset. The atrial fibrillation symptoms in heart patients are a major risk factor of stroke and share common variables to predict stroke. The outcomes of this research are more accurate than medical scoring systems currently in use for warning heart patients if they are likely to develop stroke Pattanapong Chantamit-o-pas et al. [3] reports predictive analytical techniques for stroke diseases using deep learning model applied on heart disease dataset. The atrial fibrillation symptoms in heart patients are a major risk factor of stroke and share common variables to predict stroke. The outcomes of this research are more accurate than medical scoring systems currently in use for warning heart patients if they are likely to develop stroke. Kunder Akash et al.[4] The majority of strokes are classified as ischemic embolic and Hemorrhagic. An ischemic embolic stroke happens when a blood clot forms away from the patient brain usually in the patient heart and travels through the patient bloodstream to lodge in narrower brain arteries. Hemorrhagic stroke is considered another type of brain stroke as it happens when an artery in the brain leaks blood or ruptures. Stroke is the second leading cause of death worldwide and one of the most life- threatening diseases for persons above 65 years. It injures the brain like “heart attack” which injures the heart. Once a stroke disease occurs, it is not only cost huge medical care and permanent disability but can eventually lead to death. Every 4 minutes someone dies of stroke, but up to 80% of stroke can be prevented if we can identify or predict the occurrence of stroke in its early stage.

III. RESEARCH METHODOLOGY

Stroke is the world's second leading cause of death and poses a significant health challenge to individuals and national healthcare systems alike. Key modifiable risk factors for stroke include hypertension, cardiac conditions, diabetes, glucose metabolism issues, atrial fibrillation, and lifestyle choices. Our project aims to harness machine learning techniques to analyze large existing datasets for effective stroke prediction based on these modifiable risk factors. We plan to develop an application that offers personalized risk assessments and sends lifestyle modification recommendations to educate users about managing their stroke risk factors, thereby enhancing their understanding of the relevant health concepts.

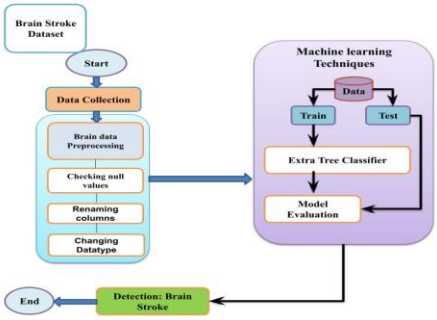


Figure 2: Proposed System of Brain Stroke

Dataset: The attribute information encompasses a range of data types, such as electronic health records (EHRs), medical imaging (MRI, CT scans), lab results, genetic data, lifestyle factors, and data from wearable devices. It's important to set up secure systems for storing, transmitting, and integrating these data sources, while also adhering to privacy laws and data security protocols.

- gender: "Male", "Female" or "Other"
- age: age of the patient
- hypertension: 0 if the patient doesn't have hypertension, 1 if the patient has hypertension
- heart_disease: 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease
- ever_married: "No" or "Yes"
- work_type: "children", "Govt_jov", "Never_worked", "Private" or "Self-employed"
- avg_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.

Dat Preprocessing: Data preprocessing is a vital stage in the machine learning process, where raw data is cleaned and transformed to a format conducive for training machine learning models. This step enhances data quality, addresses problems like missing values and outliers, and ensures compatibility with the algorithms to be applied.

Feature Extraction: Identify significant features pertinent to stroke prediction from the gathered data sources. Engage in feature engineering to derive

valuable insights and generate new features that encapsulate relevant patterns and correlations. Employ both domain expertise and data-driven methodologies to cherry-pick a subset of features that substantially enhance the predictive capability of the model.

Machine Learning Models: Select suitable machine learning algorithms like logistic regression, random forests, support vector machines, or deep learning models for stroke prediction. Utilize labeled data to train these models, integrating both conventional statistical methods and advanced machine learning techniques. In a machine learning context, an architecture diagram illustrates the system's high-level structure, depicting its components, data flow, and interactions. It offers a visual representation of data processing, model training, and prediction processes within the system. Optimize hyperparameters and model architecture to improve predictive accuracy and the ability to generalize.

Decision Tree

A decision tree in machine learning is a versatile, interpretable algorithm used for predictive modelling. It structures decisions based on input data, making it suitable for both classification and regression tasks. This article delves into the components, terminologies, construction, and advantages of decision trees, exploring their applications and learning algorithms. A decision tree is a type of supervised learning algorithm that is commonly used in machine learning to model and predict outcomes based on input data. It is a tree-like structure where each internal node tests on attribute, each branch corresponds to attribute value and each leaf node represents the final decision or prediction. The decision tree algorithm falls under the category of supervised learning. They can be used to solve both regression and classification problems.

Logistic Regression

Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables. Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1. Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems. In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1). The curve from the logistic function indicates the likelihood of something such as whether the cells are cancerous or not, a mouse is obese or not based on its weight, etc. Logistic Regression is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets.

K-Nearest Neighbors (KNN)

The K-Nearest Neighbors (KNN) algorithm is a supervised machine learning method employed to tackle classification and regression problems. Evelyn Fix and Joseph Hodges developed this algorithm in 1951, which was subsequently expanded by Thomas Cover. The article explores the fundamentals, workings, and implementation of the KNN algorithm. KNN is one of the most basic yet essential classification algorithms in machine learning. It belongs to the supervised learning domain and finds intense application in pattern recognition, data mining, and intrusion detection.bIt is widely disposable in real-life scenarios since it is non-parametric, meaning it does not make any underlying assumptions about the distribution of data (as opposed to other algorithms such as GMM, which assume a Gaussian distribution of the given data). We are given some prior data (also called training data), which classifies coordinates into groups identified by an attribute.

ExtraTrees Classifier

ExtraTreesClassifier is an ensemble learning method fundamentally based on decision trees. ExtraTreesClassifier, like RandomForest, randomizes certain decisions and subsets of data to minimize over-learning from the data and overfitting. Extra Trees is like Random Forest, in that it builds multiple trees and splits nodes using random subsets of features, but with two key differences: it does not bootstrap observations (meaning it samples without replacement), and nodes are split on random splits, not best splits.

Validation: Validate the trained models using the validation dataset to assess their generalization ability and robustness. Conduct external validation using independent datasets or cross-validation techniques to further validate model performance.vEvaluate the clinical utility of the predictive models, considering factors such as interpretability, ease of integration into clinical workflows, and potential impact on patient outcomes. The accuracy score is a common metric used to evaluate the performance of classification models in machine learning. It represents the proportion of correctly classified instances out of the total instances in the dataset. In other words, it measures how accurately the model predicts the class labels of the data.

Model Deployment: Deploy the trained models into clinical workflows or decision support systems for real-time risk assessment. Integrate the predictive models with electronic health record systems or other healthcare IT infrastructure to enable seamless data exchange and interoperability. Provide user-friendly interfaces or dashboards for healthcare providers to interact with the predictive models and interpret the results effectively.

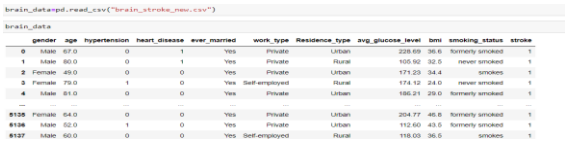
Model Deployment: The web application is developed using flask framework and application is predicting the brain stroke model, identifying individuals at high risk of stroke allows for early intervention and preventive measures. Early detection of stroke risk enables healthcare providers to implement timely interventions, potentially reducing the severity of strokes and improving patient outcomes. By identifying high-risk individuals and providing appropriate care, the project can contribute to better long-term health outcomes for patients.

IV. EXPERIMENTAL RESULTS

The implementation can have classified into different modules of project and are listed as:

- Data Collection
- Data Preprocessing
- Feature Selection / Extraction
- Model Training And Testing
- Machine Learning Model Deployment
- Performance Evaluation
- Comparison Study of the Model
- Exploratory Data Analysis (EDA)

Once you've chosen your tools, you can load the dataset into your programming environment. For example, if you're using Python and your dataset is in CSV format, you can use Pandas to read the CSV file into a Data Frame.



	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	Male	67.0	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
1	Male	80.0	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
2	Female	49.0	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
3	Female	79.0	1	0	Yes	Self-employed	Rural	174.12	24.0	never smoked	1
4	Male	81.0	0	0	Yes	Private	Urban	186.21	29.0	formerly smoked	1
...	...	...	...	...	...	...	...	...	...	...	...
5105	Female	64.0	0	0	Yes	Private	Urban	204.77	46.8	formerly smoked	1
5106	Male	52.0	1	0	Yes	Private	Urban	112.60	43.6	formerly smoked	1
5107	Male	60.0	0	0	Yes	Self-employed	Rural	118.03	36.5	smokes	1

Figure 2: Dataset Description

Comparing models is a critical step in the machine learning workflow, helping you select the best-performing model for your particular task. In this project ExtraTree classifier gives more accuracy.

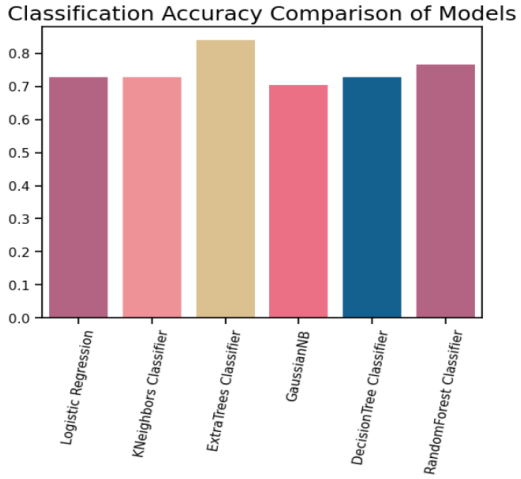
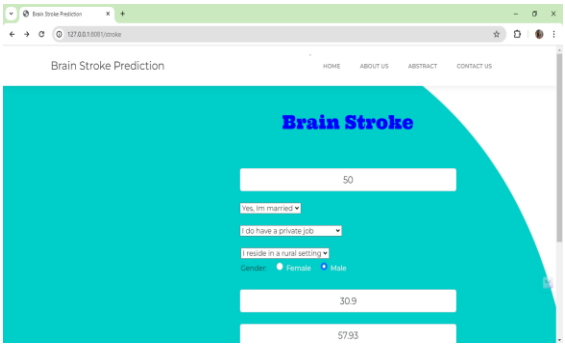


Figure 3: Accuracy Comparison of the Model

The web application implementation as follows .

Figure 4: Landing page of the Application



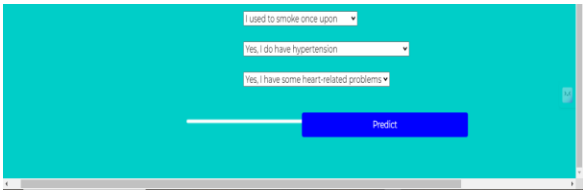


Figure 5: Input the Parameters for Stroke Prediction

Table 4.1: Classification Report of the Classifiers

	No Brain Stroke				Brain Stroke				
Algorithm	Precision	Recall	F1-Score	Support	Precision	Recall	F1-Score	Support	Accuracy
Logistic Regression	0.74	0.70	0.72	40	0.72	0.76	0.74	41	72.84
K-Neighbor Classifier	0.78	0.62	0.69	40	0.69	0.83	0.76	41	76.54
Extratree Classifier	0.86	0.80	0.83	40	0.82	0.88	0.85	41	83.95
GaussianNB	0.69	0.72	0.71	040	0.72	0.68	0.70	41	70.37
Decision Tree Classifier	0.74	0.70	0.72	40	0.72	0.76	0.74	41	72.84

Table 4.1 displayed the report of the classifiers , also shows the main results that includes Precision , Recall , F1 Scroe and Support . As per the above table The Extra Tree Classifier is giving the highest accuracy.

REFERENCES

[1] Senjuti Rahman, Mehedi Hasan, and Ajay Krishno Sarkar, " Prediction of Brain Stroke using Machine Learning Algorithms and Deep Neural Network Techniques", EJECE, European Journal of Electrical Engineering and Computer Science ISSN: 2736-5751, Vol 7, Issue 1, January2023.

[2] Jeena R S and Suksh Kumar, " Stroke prediction using SVM", International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT), 2016.

[3] Pattanapong Chantamit-o-pas and Madhu Lata Goyal, " Prediction of Stroke Using Deep Learning Model", International Conference on Neural Information Processing, 2017.

[4] Kunder Akash Mahesh , Shashank H N , Srikanth S , Thejas A M , " Prediction of Stroke using Machine Learning", Researchgate, June 2020.

[5] M. J. Ferdous and R. Shahriyar, "A Comparative Analysis for Stroke Risk Prediction Using Machine Learning Algorithms and Convolutional Neural Network Model," International Conference on Electrical, Computer and Communication Engineering (ECCE), Chittagong, Bangladesh, 2023, pp. 1-6, doi: 10.1109/ECCE57851.2023.10101567, 2023.

[6] B. R. Gaidhani, R. R.Rajamenakshi, and S. Sonavane, "Brain Stroke Detection Using Convolutional Neural Network and Deep Learning Models,", 2nd International Conference on Intelligent Communication and Computational Techniques (ICCT), Jaipur,

India, 2019, pp. 242-249, doi: 10.1109/ICCT46177.2019.8969052, 2019.

[7] Gupta, Saumya & Raheja, Supriya, "Stroke Prediction using Machine Learning Methods". 553-558. 10.1109/Confluence52989.2022.9734197, 2022.

[8] T. I. Shoily, T. Islam, S. Jannat, S. A. Tanna, T. M. Alif and R. R. Ema, "Detection of Stroke Disease using Machine Learning Algorithms," 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Kanpur, India, 2019, pp. 1-6, doi: 10.1109/ICCCNT45670.2019.8944689, 2019.

[9] S. Kumar Satapathy, H. K. Kondaveeti and D. Parmar, "An Effective Framework for Predicting Stroke Prediction using Machine Learning Technique," Fifth International Conference on Electrical, Computer and Communication Technologies (ICECCT), Erode, India, 2023, pp. 01-08, doi: 10.1109/ICECCT56650.2023.10179766, 2023.

[10] Gary H, Gibbons L. National Heart, Lung and Blood Institute. 2022 [updated 2022 March 24]. Available from: https://www.nhlbi.nih.gov/health/stroke.