© November 2025 | IJIRT | Volume 12 Issue 6 | ISSN: 2349-6002

Automated Stroke Prediction Using Machine Learning: Exploratory Study with a Web Application for Early Intervention

Khatal S.S[1], Gholap. P.S[2], Vaishnavi Mane[3], Aarti Jadhav[4], Sanika Naykodi[5]
[1,2,3,4,5]*Student, Department of Computer Engineering, Sharadchandra Pawar College of Engineering and Technology, Junnar, Pune, Maharashtra, India*

***Abstract*—Stroke is a serious medical condition that happens when the blood supply to the brain is blocked or stopped. This causes damage to the brain and leads to problems with how the body and mind work. Stroke is a major health issue around the world and has serious effects on both people's health and the economy. To tackle this, scientists are creating computer programs that can predict stroke before it happens. These programs help doctors take action early, which can save lives. As people get older, more people are at risk of having a stroke, so having good prediction tools is more important than ever.**

In this study, the performance of a new machine learning method was tested against six other common types of classifiers. The goal was to see how well the method could predict stroke and how well it worked in different situations. To understand how these complex computer models make decisions, the study also looked at two methods for explaining model behaviour: SMOTE. These are well-known and trusted ways to make sense of how machine learning models work, especially in healthcare performed better than simpler ones The best model achieved almost 91% accuracy, while the other models had accuracy between 83% and 91%. The new system, which includes both overall and detailed explanations, helps make complex models easier to understand and use. This can improve how stroke is diagnosed and treated.

***Index Terms*—Stroke prediction, machine learning Algorithms, SMOTE, HTML, FLASK.**

I. INTRODUCTION

Stroke is a serious health problem that happens when blood flow to the brain stops, causing damage to brain tissue and possibly long term disability or even death. Around the world, stroke as one of the main reasons people east it puts a lot of pressure on healthcare systems, it's very important to detect stroke early and act quickly to reduce brain damage and help patients recover better. But traditional ways of diagnosing it stroke like using imaging tests and checking symptoms, often take a lot of time and need expert doctors, which can slow down treatment.

New developments in machine learning and predictive tools offer a better way to find stroke early These tools look at patient information like medical records, age, Lifestyle, and test results to spot people who might be at risk for a stroke. This helps doctors decide faster and take prevent or treat the problem Predictive analytic in stroke many types of data, such as a person's health history, vital signs, daily habits, and lab results Machine learning models lite Logistic Regression, Random Forest, and Boost can find patient this data. Using multiple models together, called ensemble methods, helps make mam accurate predictions. These systems can help reduce mistakes in diagnosis, make better use of Hospital remark, and save more lives.

II. PROBLEM STATEMENT

Early diagnosis of stroke is still a big challenge in medicine because the first signs are often very small and not clearly pointing to a stroke. These signs can be missed or misunderstood, which can make it harder to get help quickly. The usual ways to diagnose a stroke, like using brain scans or checking how a patient acts, take a lot of time and need doctors with special training. This can lead to different results depending on who is doing the test, especially in places where there are not enough doctors or medical facilities. Also, the best time to act in a stroke situation is very short usually within the first few hours after symptoms start.

If help is delayed, it can cause serious brain damage, long-term problems, or even death. Because of this, there is a big need for a better way to find strokes early.

To solve these problems, a new system using machine learning is being suggested. This system would look at patient information like past medical history, vital signs, and test results to find patterns that might mean a stroke is happening or about to happen. By using smart computer programs that can work quickly with real-time data, this system could help doctors diagnose strokes more accurately, cut down on the need for expert judgment, and get patients the care they need faster. This would help save lives and make healthcare systems work better.

III. RELATED WORK

A. Literature Survey

A study in BMC Bio informatics used a stacked model that combined Random Forest, Gradient Boosting, and SVM, and it achieved more than 90% accuracy in predicting strokes.

Research in IEEE Access showed that XGBoost worked better than other classifiers when dealing with uneven stroke data, especially when paired with SMOTE for balancing the data. Methods like PCA and Recursive Feature Elimination (RFE) help improve model results by cutting down on extra information and simplifying the data. Explainable Al tools like SHAP and LIME are becoming more common to help understand model decisions, which makes machine learning more trusted in medical settings.

B. Existing Systems

Both commercial and research systems use advanced image processing and machine learning to find, separate, and prioritize patients with acute stroke using CT and MRI scans. Companies like Rapid Al, Viz.AI, and Brainbox offer complete tools that create perfusion maps, automatically estimate ASPECTS and penumbra-core areas, and send real-time alerts to doctors. These tools have shown benefits in several studies, but they don't always agree with expert doctors across different hospitals.

In academic research, most work uses CNN or U-Net types of models and takes part in competitions like ISLES or Deep Isles. These models do a good job at segmentation, but many haven't been tested widely across many hospitals. Common issues across these systems include differences in data, changes in how images are labelled, strict rules from regulators, and not enough long-term results showing if they really help patients. Solving these problems is still a big focus in research, which is why this study is being done.

IV. PROPOSED SYSTEM

A. Data Collection

Data is gathered from clinical datasets that include: age, gender, whether someone has hypertension, heart disease, marital status, type of work, type of residence, average glucose level, BMI, and smoking status.

B. Data Preprocessing

Missing values are filled in using the mean of the available data. Categorical data is converted into numerical format using one-hot encoding Features are scaled using a method called standardization.

C. Machine Learning Algorithms

1.Logistic Regression (LR)

It predicts the chance of someone having a stroke. A special function called the sigmoid function is used on a combination of input features. The formula is: P(Y-1/X)-1/(1+(808101+ βηχαλ.

2. Random Forest (RF)

It uses many decision trees together. The final result is decided by the majority vote from all the trees. It works well with complex, nonlinear relationships in the data.

3.XGBoost (XGB)

It's a boosting method that builds models in a step-by-step way. It keeps improving by focusing on the errors from previous predictions. It is known for performing very well with data that has clear structures.

B. System Architecture

- 1-High level overview.
- The system is divided into three main parts: Data this part deals with collecting, storing, and labelling data.

- ML/Backend- this part handles preprocessing data, creating features, training models, keeping track of models, and providing an At for using models**.**
- Application/ frontend- This part includes a dashboard for doctors and users, as well as reports and alerts.
- Additional important areas, keeping data secure and private, keeping logs and monitoring the system having a process for updating models, and checking for explainable and fairness Component breakdown and responsibilities Data ingestion.
- Sources data comes from electronic health records (EHR), patient surveys, lab results, metadata item medical images (not the images themselves unless working on image-based models), data from medical devices (blood pressure and heart rate), and CSV files uploaded for research purposes Validation data is checked against a set of rules, missing values are handled, and personal information is removed to protect privacy.

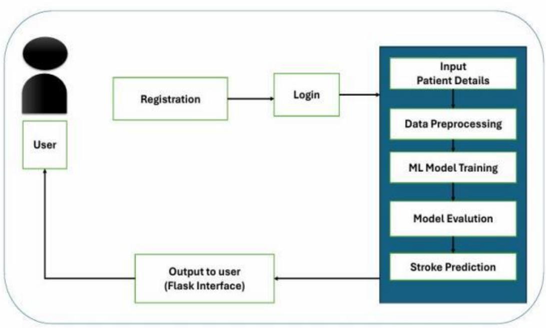

Fig. 1 System Diagram [27]

C. Module Descriptions

1. Data Collection Module

This module gathers all the important information that can influence the chance of someone having a brain stroke. The information includes details like a person's age, gender, body mass index, blood sugar levels, smoking habits, high blood pressure, and heart disease. This data can come from hospitals, medical files, wearable gadgets, or public medical data sources like Kaggle.

2. Data Preprocessing Module

This module gets ready the raw information for the machine learning model. It deals with missing data, repeated entries, and information that doesn't match. It also changes text-based information, like gender or smoking status, into numbers and makes sure numbers like blood sugar and BMI are in a standard format.

3. Feature Selection Module

This module finds the most useful information that helps in predicting a stroke. Choosing the right features makes the model work better and keeps it simpler.

4. Model Training Module

This is the main part where different machine learning methods are used to teach the model using the cleaned data. Common methods include Logistic Regression, Random Forest, and XGBoost, which help in guessing whether someone might have a stroke.

5. Model Evaluation Module

This module checks how well the model works by testing it with new data. It helps find out if the model is accurate and can make good predictions for people it has not seen before.

6. Prediction Module

This module lets users enter patient information and get a prediction about stroke risk. It uses the trained model to determine if a person is likely to have a stroke and gives a probability score showing how likely that is.

7. Visualization and Dashboard Module

This module shows the results of predictions and health data in an easy-to-read format using a dashboard. It uses charts and graphs to help users and medical professionals understand the results clearly

8. Report Generation Module

This module creates a detailed health report after the prediction The report includes the patient's input data, the prediction result, the probability score, and advice for staying healthy.

9. User Interface Module

The user interface module connects the user with the prediction system. It lets users enter patient details, send them for prediction, and see the results through a web or desktop application.

V. METHODOLOGY

Machine learning is becoming more popular in medical diagnostics because it can efficiently analyse large amounts of medical data, including pictures of skin lesions. This is especially useful for identifying types of skin cancer. When it comes to predicting strokes, the main goals of using machine learning are to improve the accuracy of diagnosis and make the classification process more efficient. Different machine learning models are used to build an automated system for predicting strokes, and these models are evaluated using measures like accuracy, recall, and F1 score to find the best one for the job. In this study, the method for automatically predicting strokes involves making a "Yes" or "No" prediction

The process of building the model follows a five-step approach, as shown in Figure 2. First, they get a dataset of electronic health records. Then, they preprocess the data by re-scaling and normalizing it. Next, they extract features from the data. After that, they build a classifier using the extracted features. Finally, they use SHAP and LIME methods to understand how the model makes its decisions. This improved strategy aims to make stroke predictions more accurate and help doctors make better treatment decisions

A. Technology Stack

1. Programming Language

- Python

Python is the main language used because it is easy to learn, has a lot of tools and libraries. and works well for machine learning and data science tasks.

2. Development Environment

- Jupyter Notebook/VS Code/PyCharm

These tools are used for looking at data, building models, and fixing any problems in the code.

3.Machine Learning Libraries

These libraries help with preparing data, making models, and checking how well they work.

- NumPy- Works with numbers and arrays.
- Pandas -Helps organize and clean data.
- Scikit-learn (sklearn) -Has basic machine learning tools and ways to check model permanence.
- XGBoost/LightGBM -Better tools for making accurate predictions.
- Matplotlib & Seaborn - Create charts and graphs to show data.
- Joblib/Pickle - Save and load models for use later.

3. SHAP/LIME

Help explain how models make decisions Data Storage and Management

4. CSV/Excel/Database MySQL or SQLite)

These are used to store the original data, cleaned data, and results from the models.

5. Flask

These is used to create the back-end part of the website, which connects the user interface with the machine learning model.

6. Frontend Technologies

HTML, CSS, Bootstrap These are used to build the website so users can interact with it and see the results.

B. Implementation and Training

1) Overview/Pipeline

- Load dataset explore and clean.
- Preprocess data (fill in missing values, convert categories into numbers, scale values).
- Create and choose features.
- Deal with unbalanced classes (use SMOTE or adjust class weights).
- Train several models (Logistic Regression, Random Forest, XGBoost, Ligh GBM.
- Tune model settings (use Random Search, Grid Search, or Optina).
- Check model performance (confusion matrix, precision, recall, FI score, ROC-AUC, PR-AUC).
- Explain model decisions (using SHAP) Save and export the final model and pipeline.

VI. EVALUATION AND EXPECTED OUTCOMES

1 The evaluation part of the Brain Stroke

Prediction system checks how well the machine learning model works and how reliable it is, The model is trained using a dataset with information about patients, including things like age, gender, high blood pressure, heart disease, body mass index, blood sugar levels, and whether they smoke. This evaluation happens in the following steps:

1. Data Splitting

The data is split into two parts:
-Training Set (80%) -This is used to teach the model.
-Testing Set (20%) This is used to see how well the model can predict new data and how well it works in general.
2. Performance Metrics
To get accurate and useful results, several measures are used:
-Accuracy-This shows how often the model's predictions are correct
-Precision-This tells us how many of the cases the model predicted as strokes were actually strokes.
-Recall (Sensitivity)-This shows how well the model finds real stroke cases.
-Confusion Matrix-This is a table that helps us understand how the model classifies stroke and non-stroke cases
3. Model Comparison
Several machine learning methods are tested to see which own works best
-Logistic Regression
-Random Forest Classifier
-Decision Tree
-XGBOOST
This helps us find the model that gives the best results in terms of accuracy and how easy it is to understand.

2 Expected Outcomes
The Brain Stroke Prediction System is designed to achieve the following outcomes.
1. Accurate Stroke Risk Prediction
The model will be able to determine if a patient is at risk of having a stroke with a high level of accuracy, typically between 80% and 95%, depending on the quality of the data and how it is prepared.
2. Early Detection and Prevention
The system helps find patients who are at high risk of stroke early on, so healthcare professionals can take action to prevent strokes from happening.
3. Feature Importance Insights
The model will identify the most important factors that increase the risk of stroke, such as high blood pressure, high blood sugar, body mass index, and smoking.
4. User-Friendly Dashboard
There will be a web-based interface that shows the prediction results, the probability of a stroke, and personalized health tips for the user.
5. Support for Medical Professionals
The model can help doctors make better decisions by providing clear, data-based information to support diagnosis and patient care.

VII. CONCLUSION AND FUTURE ENHANCEMENTS

The Brain Stroke Prediction System using Machine Learning shows how using data can help find stroke risks early and stop them before they happen. By looking at patient health details like age, high blood pressure, heart problems, blood sugar level s, body mass index, and lifestyle choices, the system can tell how likely someone is to have a stroke.
Machine learning methods like Logistic Regression, Random Forest, and Support Vector Machine (SVM) were used to group people into two groups: those who had a stroke and those who didn't. Among these, models that use groups of techniques, such as Random Forest or XGBoost, worked better in terms of accuracy, reliability, and ease of understanding compared to older models.
Testing the model with accuracy, precision, recall, F1-score, and ROC-AUC showed that the system works well and consistently with real patient data.
Plus, the system has charts and an easy-to-use interface that make it simple for doctors and patients to understand the results and make informed decisions.

REFERENCES

[1] A. Alloubani, A. Saleh, and I. Abdelhafiz, “Hypertension and diabetes mellitus as a predictive risk factors for stroke,” Diabetes Metabolic Syndrome, Clin. Res. Rev., vol. 12, no. 4, pp. 577–584, Jul. 2018.
[2] A. K. Boehme, C. Esenwa, and M. S. V. Elkind, “Stroke risk factors, genetics, and prevention,” Circ. Res., vol. 120, no. 3, pp. 472–495, Feb. 2018.
[3] I. Mosley, M. Nicol, G. Donnan, I. Patrick, and H. Dewey, “Stroke symptoms and the decision to call for an ambulance,” Stroke, vol. 38, no. 2, pp. 361–366, Feb. 2007.
[4] J. Lecouturier, M. J. Murtagh, R. G. Thomson, G. A. Ford, M. White, M. Eccles, and H. Rodgers, “Response to symptoms of stroke in

the UK: A systematic review," BMC Health Services Res., vol. 10, no. 1, pp. 1–9, Dec. 2010.

[5] L. Gibson and W. Whiteley, "The differential diagnosis of suspected stroke: A systematic review," J. Roy. College Physicians Edinburgh, vol. 43, no. 2, pp. 114–118, Jun. 2013.

[6] A. R. Alhawaimil, "Segmentation of Brain Strokes image," International Journal of Advanced Research in Computer and Communication Engineering, vol. 4, no. 9, pp.375-378,2015.

[7] S. Keerthana and K. Sathiyakumari, "Brain Strokes Prediction using fuzzy C-means Clustering" International Journal of Computer Applications, vol. 154, no. 4, pp. 26-30, 2016.

[8] I Kesavamurthy, Subha Rani and N Malmurugan, "Early Diagnosis of Acute Brain Infarct Using Gabor Filter Technique for Computed Lomography Images", Biomedical Soft Computing and Human Sciences, Vo1.l4, NO. I, pp. ll-16 (2009).

[9] Tslem Rekik, Stephanie All assonniere, Irevor K. Carpenter, Joanna M. Wardlaw, Medical image analysis methods in MRlCI-imaged acute-sub acute ischemic stroke lesion: Segmentation, prediction and insights into dynamic evolution simulation models. A critical appraisal, Neuroimage: Clinicall (2012), pp 164- 178.

[10] Fukhaylang,DouglasK.S.Ng,DanieIH.K.Chow, An image feature approach for computer-aided detection of ischemic stroke, 2011, Computers in Biology and Medicine 41, pp 529- 536.