ijamrsd.com
JAMRSD
ISSN: 2583-6404
Journal of Advanced Multidisciplinary Research Studies and Development
Peer Reviewed Referred Journal
Bi-monthly
Open Access

USING MACHINE LEARNING FOR PREDICTING STROKES AUTOMATICALLY: A RESEARCH STUDY THAT LOOKS AT EARLY WARNING SIGNS AND INCLUDES A WEBSITE TO HELP WITH EARLY INTERVENTION

Vaishnavi Mane[1], Aarti Jadhav[2], Sanika Naykodi[3]
[1,2,3] Student, Department of Computer Engineering, Sharadchandra Pawar College of Engineering and Technology, Junnar, Pune, Maharashtra, India.
Email: vaishnavimane8080@gmail.com[1], aartijadhav2572@gmail.com[2], sanunaykodi@gmail.com[3]

Abstract
Stroke is a serious brain condition that happens when there's a blockage or stoppage of blood flow to the brain. This causes damage to brain cells and can lead to loss of function. It is one of the main reasons people die or become disabled, and it puts a big strain on healthcare systems. Finding people at risk early is very important to prevent serious problems and help people live longer.This study looks at a new way to predict stroke using machine learning and compares it with six common types of classification methods. The goal is to check how well these models can predict stroke, how reliable they are, and how well they work in different medical situations. Since many health records don't have enough data on people who get strokes, a technique called SMOTE is used to make the data more balanced. This helps improve the accuracy and fairness of the models. To help doctors understand how the models work, explainable AI methods are also used. The results show that advanced techniques that combine multiple models work better than older ones. The best model reached an accuracy of about 91%, while other models performed between 83% and 91%. The new system is both accurate and easy to understand, making it useful for finding stroke risks early and helping doctors make better decisions.

Keywords: Stroke prediction, Python, Logistic Regression, machine learning Algorithms, SMOTE, HTML, FLASK.

► *Corresponding Author: Vaishnavi Mane*

I. Introduction

Stroke happens when the flow of blood to the brain is blocked or when a blood vessel in the brain breaks. This stops the brain from getting enough oxygen, which can cause lasting damage to brain cells. Stroke is one of the main reasons people die or become disabled around the world, and it puts a big strain on healthcare systems and the economy.

It is also called a brain attack or cerebrovascular accident (CVA). Brain stroke is one of the main reasons people die around the world [7].

In the past, doctors used imaging tests like CT and MRI scans along with patient exams to diagnose stroke.

These methods work, but they require time and skilled professionals, which might not be available everywhere, especially in places with fewer resources. Delaying a diagnosis can make it harder to treat stroke effectively.
New developments in machine learning offer a way to predict who is at risk of stroke before it happens.
Additionally, there is an increasing need for transparency and explainability in machine Learning models used in healthcare can help doctors understand the factors that increase a patient's risk of having a stroke. This information can assist in making better treatment decisions [13].
The World Stroke Organisation estimates that 13 million people worldwide have a stroke each year, leading to about 5.5 million deaths [13].
These tools use information like age, lifestyle, lab results, and medical history.
Models such as Logistic Regression, Random Forest, and Gradient Boosting can find patterns in health data. Using multiple models together, called ensemble learning, helps improve the accuracy of predictions.
According to global health data, stroke is the second biggest cause of death and a major cause of long-term disability.
That's why it's important to create strong predictive tools to help with early detection and prevention efforts.

II. Problem Statement

Despite medical advancements, early detection of stroke remains challenging.
Initial symptoms are often subtle and may be misinterpreted, delaying urgent care. Conventional diagnostic methods require advanced imaging and trained professionals, which may not be readily accessible in rural or underdeveloped regions.
Time plays a critical role in stroke management, as therapeutic interventions are most effective within a narrow treatment window.
Any delay can result in irreversible brain damage or fatal outcomes.
To address these limitations, this research proposes a machine learning-based predictive system capable of analyzing patient health indicators in real time.
By processing structured medical and lifestyle data, the system aims to identify high-risk individuals quickly and accurately. Such a solution can support clinicians, reduce dependency on manual assessments, and enhance decision-making efficiency in healthcare environments.

III. Dataset Description

The dataset used in this study consists of approximately 5,000 patient records containing demographic and clinical features such as age, gender, hypertension status, heart disease history, body mass index (BMI), average glucose level, smoking habits, and residential category.
The target variable is binary:
1 — Stroke Occurrence
0 — No Stroke
Since stroke cases represent a minority class within the dataset, class imbalance is addressed using SMOTE to synthetically generate minority samples.
This improves classifier sensitivity and reduces bias toward the majority class.

IV. Related Work

A. Literature Survey

Existing works in the literature have investigated various aspects of stroke prediction. Jeena et al. provides a study of various risk factors to understand the probability of stroke [12].
Previous studies demonstrate the effectiveness of ensemble learning in stroke prediction.
Research published in biomedical journals reports that stacked models combining Random Forest, Gradient Boosting, and Support Vector Machines achieve accuracy levels exceeding 90%.
To check how well the model is performing, we used Accuracy, Precision, Recall, and F1-score.
Accuracy is the percentage of correct predictions out of all predictions the model made [11].
Other studies highlight the effectiveness of XGBoost in handling imbalanced medical datasets, particularly when paired with SMOTE.
Feature reduction techniques such as Principal Component Analysis (PCA) and Recursive Feature Elimination (RFE) have also been shown to improve model performance.
Additionally, the use of Explainable AI (XAI) tools such as SHAP and LIME has gained importance in medical applications, enhancing transparency and clinical trust in automated systems.

B. Existing Systems

Commercial and research-based systems often rely on imaging-based analysis using deep learning techniques such as CNNs and U-Net architectures.
These systems primarily focus on CT and MRI image segmentation for acute stroke detection.
While imaging-based tools demonstrate strong segmentation capabilities, they require advanced infrastructure and are not always generalizable across institutions.
Furthermore, variability in data quality and regulatory constraints pose additional challenges.
The proposed system differs by emphasizing early structured-data-based risk prediction rather than image-based diagnosis, making it scalable and accessible.

V. Proposed System

A. Data Collection

Clinical datasets containing demographic and health-related parameters are used. Features include age, gender, hypertension, heart disease, marital status, occupation type, residence type, glucose level, BMI, and smoking status.

B. Data Preprocessing

Missing values are imputed using statistical techniques.
Categorical variables are encoded using one hot encoding.
Feature scaling is performed using standardization.
Class imbalance is corrected using SMOTE

C. Machine Learning Algorithms

1. Logistic Regression (LR)

A supervised classification algorithm that estimates the probability of a binary outcome using a sigmoid activation function. It provides interpretable coefficients and works efficiently on linearly separable data.

2. Random Forest (RF)

Random Forest is an ensemble learning method that constructs multiple decision trees during training. It reduces overfitting by averaging predictions from different trees. It performs well on complex datasets and can handle nonlinear relationships efficiently.

3. XGBoost (XGB)

A gradient boosting framework that builds trees sequentially by minimizing prediction errors. It incorporates regularization to prevent overfitting and is highly effective for structured datasets.

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.

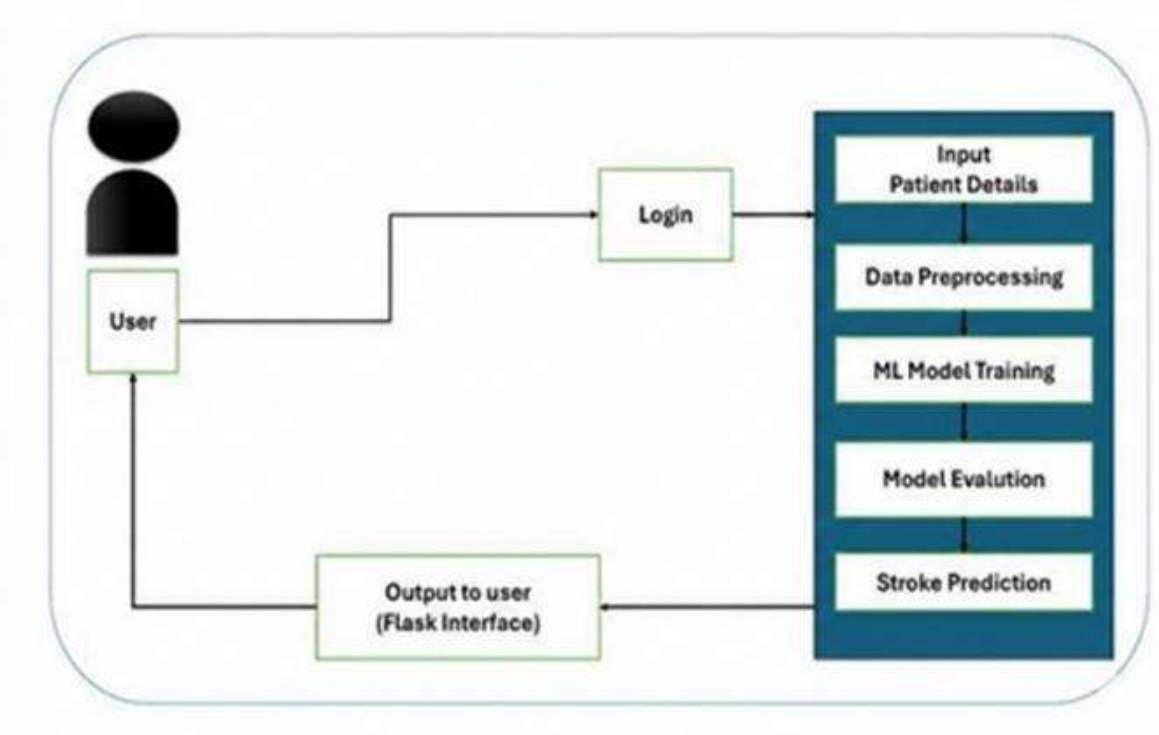

Fig. 1 System Diagram

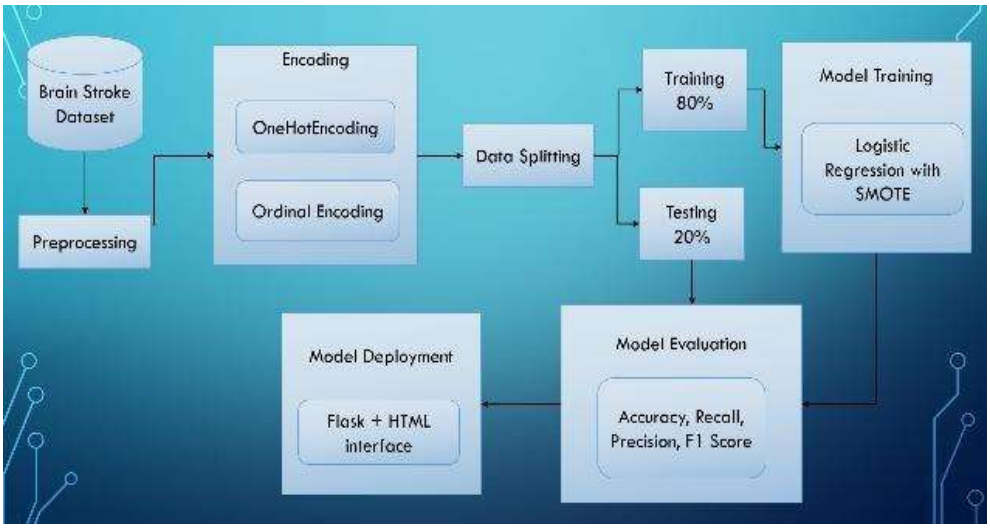

Fig.2 Project Workflow

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.

Accuracy measures overall correctness. Precision measures correctness of positive predictions.

Recall (Sensitivity) measures ability to detect actual stroke cases.

F1-score balances precision and recall. ROC-AUC evaluates model performance across different thresholds.

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.

VI. 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 development process follows a structured pipeline:

1. Data acquisition from healthcare records.
2. Data cleaning and normalization.
3. Feature extraction and selection.
4. Model training using multiple classifiers.
5. Hyperparameter tuning via Grid Search or Random Search.
6. Performance evaluation using metrics such as Accuracy, Precision, Recall, F1-score, and ROC-AUC.
7. Model interpretation using explainability techniques.
8. Deployment through a Flask-based web interface

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. CSV/Excel/
These are used to store the original data, cleaned data, and results from the models.
4. Flask
This is used to create the back-end part of the website, which connects the user interface with the machine learning model.
5. 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
- Train several models (Logistic Regression, Random Forest, XGBoost, Light GBM).
- Tune model settings (use Random

Search, Grid Search, or Optuna).
- Check model performance (confusion matrix, precision, recall, F1 score, ROC AUC, PR-AUC).
- Explain model decisions (using SHAP).
- Save and export the final model and pipeline.

VII. Evaluation and Expected Outcomes

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 one 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.

4. 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 healthcare professionals find patients who are at a high risk of having a stroke early, so they can take steps to stop it from happening.

3. Feature Importance Insights

The model will show which factors are most likely to 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 chance of a stroke, and personalized health advice for the user.

5. **Support for Medical Professionals**

The model can help doctors make better decisions by giving clear, data-based information to support diagnosis and patient care.

VIII. Final Result

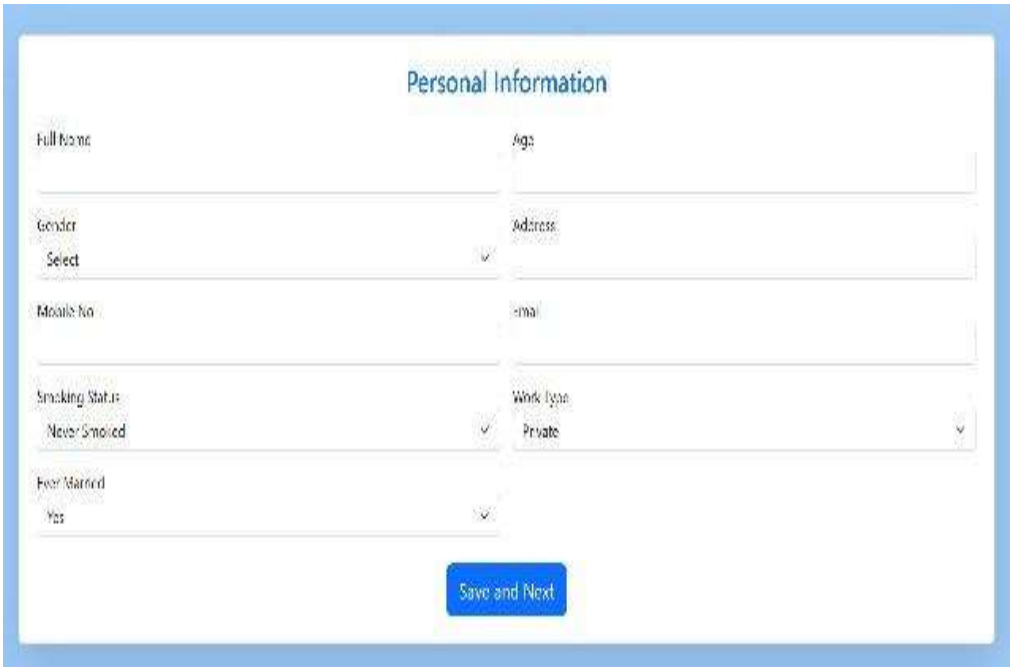

Fig. 3

Fig. 4

Fig. 5

Fig. 6

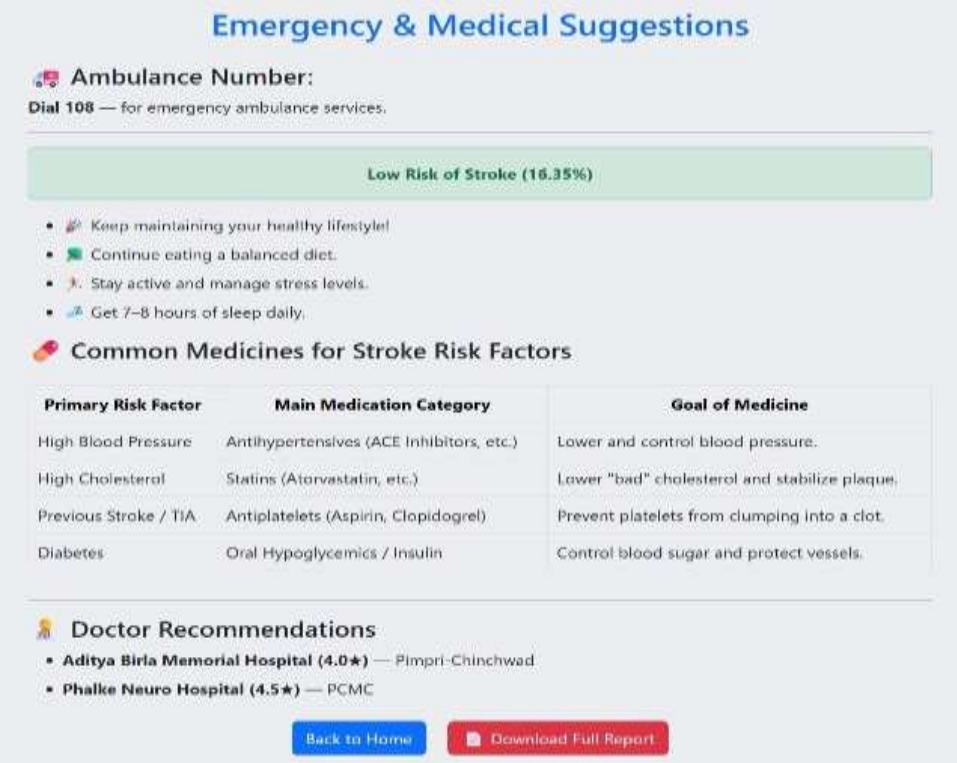

Fig. 7

IX. Conclusion and Future Enhancements

In this research, several supervised machine learning algorithms, including Logistic Regression, Random Forest, and XGBoost, were used and tested with real-world healthcare data.

Techniques like feature scaling, categorical encoding, and class balancing through SMOTE were applied, which greatly improved the models' reliability and fairness. Among the methods tested, the ensemble-based approaches worked best because they can better understand the complex relationships between different risk factors. The evaluation measures like Accuracy, Precision, Recall, F1-score, and ROC-AUC showed that the system has strong ability to make good predictions. Especially, Recall, which is about finding real stroke cases, is very important in medical diagnosis. It's more important to find real cases than just having a high overall accuracy. The system's performance is between 80% and 91%, which means it can help doctors find patients at high risk before serious symptoms appear.

This study shows a good system for predicting stroke using structured healthcare data.

It is scalable, easy to understand, and works efficiently. The system not only improves prediction results but also supports early intervention, prevention, and digital health. By reducing the need for slow manual checks, it can quickly assess risk and help improve patient care and use of healthcare resources.

Future Enhancement

Using a stacked cross-validation method is expensive, but it's done only once and offline.

Once the model is ready, it can make predictions very quickly, so it's ready to use almost immediately [13].

1. Using smart wearable health devices

The system can work with smart devices like fitness trackers and smartwatches.

These devices collect health data such as heart rate, activity levels, sleep quality, and other body information. By connecting these wearables to the system, health data can be checked in real time. This helps provide more accurate information and helps find possible health issues early.

2. Using cloud platforms

The system is designed to work on cloud platforms so it can support more users, be accessed from anywhere, and manage data well.

Using the cloud allows users and healthcare workers to access it through the internet. It also offers good storage, faster handling of large health data sets, and the ability to support many users at once without slowing down the system.

3. Using deep learning models

The system uses advanced deep learning models to improve disease prediction.

These models can analyze complex medical data, find hidden patterns, and learn from a lot of health information. This helps the system give better predictions and improves performance as more data is added over time.

4. Predicting multiple disease risks

The system is made to predict the risk of several health conditions based on patient data.

Instead of just looking at one illness, it checks for multiple diseases by looking at medical history, lifestyle choices, and body measurements. This overall approach helps users and healthcare professionals find possible health risks early and take steps to prevent them.

5. Connecting with hospital databases

The system can connect in real time with hospital databases to get and update patient medical records.

By linking directly with hospital systems, the platform can get important health data like past diagnoses, lab results, and treatment records. This connection improves prediction accuracy and ensures healthcare providers have the latest information for better 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 MRlCIimaged 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.C how, An image feature approach for computeraided detection of ischemic stroke, 2011 , Computers in Biology and Medicine 41 , pp 529- 536.
11. 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," in Proc. 10th Int. Conf. Computing, Communication and Networking Technologies (ICCCNT), 2019.
12. S. Dev, H. Wang, C. S. Nwosu, N. Jain, A. Davenport, and D. John, "A predictive analytics approach for stroke prediction using machine learning and neural networks," Healthcare Analytics, vol. 2, pp. 1–12, 2022.
13. K. Mridha, S. Ghimire, A. Aran, M. M. Uddin, and M. F. Mridha, "Automated Stroke Prediction Using Machine Learning: An Explainable and Exploratory Study With a Web Application for Early Intervention," IEEE Access, vol. 11, pp. 52288–52302, 2023.