# Brain Stroke Prediction System

Niteen Dhutraj[1*], Bhushan Chaudhari[1], Divyani Patil[1], Shreya Patil[1], and Chetana Patil[1]

Dept. of Information Technology, SVKM's IOT, Dhule, Maharashtra, India
niteen125@gmail.com, chaudharibs@gmail.com, divyanipatil60@gmail.com, shreyapatil2855@gmail.com, chetanap292004@gmail.com

*Corresponding Author: Niteen Dhutraj (niteen125@gmail.com)

**Abstract.** Brain stroke is one of the major causes of death and disability, and this is mainly because of delayed diagnosis and restricted early opportunity evaluation. In conventional diagnosis, tools such as CT scans and MRI are used to detect stroke occurrence only after it has happened. Prediction models are also available, but only on limited information, which can be answered with a yes or no response. An explainable machine learning-based stroke prediction method is presented in this research. Various factors are considered by this system. To increase the accuracy of the proposed solution, fact-balancing tools like SMOTE and ADASYN are used with the Random Forest, Logistic Regression, and XGBoost algorithms. Medical trust is enhanced by explainable AI techniques like SHAP and LIME.

**Keywords:** Brain Stroke Prediction, Machine Learning, Artificial In-telligenc, Explainable AI (XAI), SHAP, LIME, Logistic Regression, Random Forest, XGBoost.

## 1 Introduction

Brain stroke is a crucial neurological disorder that occurs at the same time as the blood supply to part of the brain is interrupted, leading to rapid neuronal harm within minutes. Globally, stroke is diagnosed as the second most important cause of mortality and a high contributor to long-term disability [3]. Stroke incidence is increasing due to lifestyle changes and chronic conditions such as hypertension, diabetes, and cardiovascular diseases. A primary project in stroke control is that early signs and symptoms are frequently subtle, non-precise, or neglected, resulting in not on-time prognosis and bad medical check-ups [6]. Conventional diagnostic strategies, which include CT scans, MRI, and neurological examinations are correct but reactive in the nature, as they perceive stroke as the most effective after its incidence [7]. Existing prediction approaches maximumly depend upon periodic medical checkups constrained medical datasets, offering only binary consequences including "stroke" or "no stroke." These current systems do not incorporate continuous physiological changes, lifestyle behaviors and the everyday fitness variations that notably have an a effect on stroke risk. As a result there remains a significant hole in present day healthcare systems


M. Kaushik et al. (eds.), *Proceedings of the International Conference on Responsible, Risk-aware, and Regulated AI (RRRAI 2026)*, Advances in Intelligent Systems Research 210,
https://doi.org/10.2991/978-94-6239-723-1_46

wherein proactive, personalized, and continuous monitoring is required [9]. In real-world healthcare systems, non-stop tracking systems want to recall different challenges such as the security of statistics transmission and privacy preservation, and real-time techniqueing of wearable fitness facts. Therefore, the sensible utility of stroke prediction algorithms necessitates dependable healthcare gateways and backend infrastructures. Furthermore, backend infrastructure guide is also required for non-stop health information go with the flow within the implementation of wearable devices in healthcare monitoring in the real worldwide. The stable transmission and stroke chance prediction in due time are depending on various aspects.

## 2 Related Work

The way we predict the brain strokes has changed completely in current years moving away from old school methods toward the advanced AI and Machine Learning. As pointed out by Reddy Madhavi et al. (2022), there is a shift toward deep learning, specifically using CNNs to scan CT and MRI images for much quick results [4]. But it is not just about the images. Researchers at the moment are digging deep into the established clinical data. For example, Shan et al. (2023) demonstrate the accuracy by using Random Forest algorithm in the prediction of strokes [5]. At present, the majority of the researchers consider ensemble models to be the best way forward. In support of this statement, Sundaram et al. (2024), used combination of XGBoost and Random Forest algorithms [10]. This notion is transformed into the practice with the help of Devarakonda et al. (2023). The latter have conducted research on the various kinds of models including Naive Bayes and Decision Trees and proved that clinical data remains a backbone of good risk analysis [11]. This method is becoming popular because of combining different types of the clinical data like numbers and categories really help the models perform better on messy and real-world datasets [9, 12]. Of course, picking the right model is only half the story; the quality of your data matters just as much. Akinwumi et al. (2025) recently stressed that the certain medical factors are the real keys to saving lives through early action [13]. A major problem has been unbalanced data, but the use of SMOTE has helped researchers increase random forest accuracy to 96.94%. Even especially basic variables High blood pressure, BMI and age can be incredibly predictive; Srivastava et al. (2023) These were used to increase the logistic regression accuracy to 95.02% [6].Where sophisticated imaging cannot be used, doctors depend upon a patient's way of life. According to Deepthi et al., smoking and cardiac issues are the top prior indications [3]. Currently, the industry is moving towards hybrid models and "Explainable AI" in order to ensure that the final results are clear for the scientific community [8]. This entire approach follows the path taken by pioneers such as Islam et al. (2013) and Hossain et al. (2011) [3, 2].

# 3 Architecture and Methodology

This begins with data collection using historical data and devices, followed by preprocessing to address missing values, convert categorical variables (e.g., gender, smoking) to numerical variables, and scale numerical variables (e.g., age, BMI, glucose) to be of the same scale. In order to find out the class distribution, correlation, outliers, and data imbalance a frequent problem in medical data because of the small number of stroke patients Exploratory Data Analysis (EDA) comes next. The HC gateway, wearable technology, the Health Connect app, and the backend processing layers make up the brain stroke prediction system, as shown in Fig. 1.

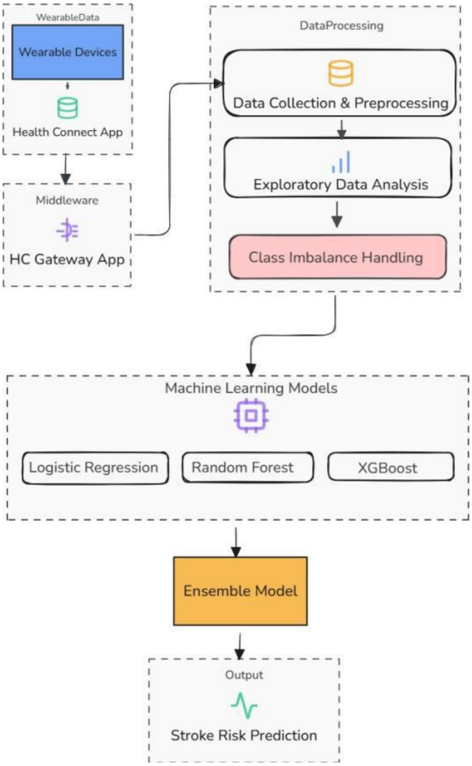


**Fig. 1.** Architecture of the Brain Stroke Prediction System

## 3.1 Wearable Devices (Data Collection Source)

The devices continuously monitor fitness parameters, heart rates, physical activity ranges, glucose levels, and blood flow patterns in a non-invasive manner.

As stroke risks arise due to sudden physiological changes, continuous monitoring is necessary. Wearable devices eliminate the need for accessing guide records, and correct real-time fitness signals are provided, which form a basis for early prediction of stroke risks.

### 3.2 Health Connect Application (Data Aggregation Layer)

The Health Connect Application functions as the first software layer between the wearable devices and the backend devices. It also stores users-unique health facts, ensuring that every users amassed records are nicely saved and prepared for similar processing.

### 3.3 HC Gateway Application (Middleware Layer)

The feature of the HC Gateway Application is that it performs as a middleware that connects the Health Connect App and the backend server. The gateway also facilitates secure communication through the API and efficient data transmission between the frontend and backend services. It enables the scalable processing of real-time health data streams generated from wearable devices. The backend processing offerings may be hosted on scalable cloud infrastructure for the high availability, dependable processing and green management of non-stop health statistics streams. The cloud-based totally microservices architecture can be used to increase an HC Gateway and offer scalability and toughness for its use in realistic eventualities. Orchestration systems which includes Kubernetes and containerization platforms which includes Docker may be used to distribute workload and offer device availability. The balancing strategies make certain that the statistics is being processed in case the structures are compromised or are experiencing intense loads.

### 3.4 Data Preprocessing

After data collection, the raw data is preprocessed to remove noise, handle missing values using methods like mean/median imputation, convert categorical features (like gender, smoking, and hypertension) into numerical features, and normalize numerical features (like age, blood pressure, BMI, and glucose levels). SMOTE and class weighting techniques are used to balance the dataset for the detection of high-risk stroke patients.

### 3.5 Model Training

Rather than relying on a single approach, we trained our system using a strategic blend of linear statistical modeling, decision-tree ensembles, and gradient-boosted frameworks. To maximize precision, our training strategy uses the linear baseline for simplicity alongside tree based ensembles that can decode complex health pattern. We also integrated a boosting layer specifically to iron out any remaining errors ,ensuring that the final stroke risk forecast is a both stable and highly reliable.

### 3.6 Ensemble Model Creation

Rather than using a single classifier, we developed an ensemble framework that merges the outputs of all three models into a unified prediction. By integrating these different perspectives, the ensemble strategy produces a much more stable and high-performing result that is remains same across diverse datasets.

### 3.7 Stroke Risk Prediction

The stroke threat prediction level is the major result obtained from the system. When new user information is provided, the trained ensemble system studying model examines more than one health indicator, i.e., age, BMI, glucose level, blood pressure, lifestyle, and physiological styles. According to the information, the version calculates the risk rating which represents the risk of stroke incidence. Instead of just showing raw scores, the system groups the final ratings into three clear risk levels are Low, Moderate, and High.

### 3.8 Result Display and Report Generation

Once the stroke risk prediction is made, the system show the percentage and the category of the predicted stroke risk, such as Low, Moderate, or High, in a user friendly interface. Display of charts and graphs is based on health trends, including the contributions. The process produces a structured report for assessing the risk of stroke enabling users to review the health parameters and share the prediction report with healthcare professionals for early awareness of the risks involved.

### 3.9 Stroke Risk Probability Calculation through Ensemble Learning Model

We implemented the hybrid ensemble methodology to generate the personalized stroke risk probabilities that are both precise and dependable. This process takes the comprehensive input feature vector consisting of the real time health metrics taken from the wearable devices via Google Health Connect and the HC Gateway. The final probability, which represents a user specific risk level, is determined through a systematic five-step calculation process.

**Feature Contribution and Weighting using Logistic Regression** Logistic Regression is the key probabilistic component modeling the linear relationship between the input public health factors and the phenomenon of strokes.

The stroke risk probability is calculated using a sigmoid function as follows:

$$P_{LR}(stroke) = \frac{1}{1 + e^{-(\beta_0 + \sum_{i=1}^{n} \beta_i x_i)}} \quad (1)$$

where $\beta_0$ is the bias (intercept) and $\beta_i$ represents the learned coefficients for each input feature $x_i$.

**Non-Linear Interaction Modeling using Random Forest** Random Forest is uses a group of decision trees to record intricate non-linear relationships. The average output of all trees is the stroke probability:

$$P_{RF}(stroke) = \frac{1}{K} \sum_{k=1}^{K} T_k(X) \tag{2}$$

where $T_k(x)$ represents a prediction of the $k^{th}$ decision tree. This increases robustness capturing feature interactions missed by a linear model.

**Optimized Risk Estimation using XGBoost** XGBoost improves stroke prediction by learning from the residual errors using gradient boosting. The update of the prediction in the iteration process, denoted as $t$, is defined as:

$$y_i^{(t)} = y_i^{(t-1)} + \eta f_t(x_i) \tag{3}$$

Where $\eta$ is the learning rate (shrinkage) and the $f_t(x_i)$ represents the newly added weak learner. This enhances accuracy and the sensitivity to minority stroke cases.

**Weighted Soft Voting Ensemble Strategy** The final stroke risk probability is obtained by integrating the output of Random Forest, XGBoost, and Logistic Regression using a weighted soft voting method:

where $w_1$, $w_2$, and $w_3$ are weights for each classifier, which are empirically set to satisfy:

$$w_1 + w_2 + w_3 = 1 \tag{4}$$

This ensemble reduces bias and improves the generalization by combining model strengths.

**Explainable AI Integration using SHAP** To achieve transparency, SHAP values are used in explaining the output during the post-hoc phase. The equation for predicting output is represented by:

$$P_{stroke} = \phi_0 + \sum_{i=1}^{n} \phi_i \tag{5}$$

where $\phi_0$ represents the base value (the average model output over the entire training set) and $\phi_i$ denotes the SHAP value for each feature $i$, quantifying it is specific contribution to the deviation from the base prediction. SHAP values help medical professionals see how each parameter affects stroke risk.

## 4 Experiments and Results

The Brain Stroke Prediction System using XGBoost, Random Forest, and Logistic Regression methods. The problem of class imbalance was addressed using the SMOTE method. Accuracy, Precision, Recall, F1-score, and AUC-ROC were used to evaluate the model.

### 4.1 Quantitative Performance Metrics

The overall performance results on the unseen test dataset are summarized in Table 1.

**Table 1.** Performance Comparison of Machine Learning Models

| Model | Accuracy (%) | Precision | Recall | F1-Score | AUC-ROC |
|---|---|---|---|---|---|
| Logistic Regression | 96.5 | 0.59 | 0.92 | 0.72 | 0.991 |
| Random Forest | 97.2 | 0.69 | 0.76 | 0.72 | 0.983 |
| XGBoost | 98.1 | 0.83 | 0.78 | 0.80 | 0.983 |

The results show the Accuracy Paradox, whereby while XGBoost had the highest accuracy at (98.1%), Logistic Regression had the highest recall at (92%) and the highest AUC at 0.991. Therefore, considering the importance of false negatives in stroke prediction, Logistic Regression is more appropriate for early warning. Furthermore, the system uses a pie chart to show the impacts of important health parameters such as age, blood pressure, BMI, average glucose levels, smoking, and physical activity. In the "Detailed Feature Impact" section, the system explains the impact of the parameters on the final prediction. In this case, the impact of the parameters is as follows: Age has a 25% impact, hypertension a 20% impact, BMI a 15% impact, average blood sugar a 12% impact, smoking status an 18% impact, and physical activity a -10% impact. With the help of the SHAP/LIME-based explainable AI approach, the system increases the level of transparency and helps the user to understand the stroke risk and the reasons behind the final prediction.

**Table 2.** Confusion Matrix for Random Forest Model

| Actual / Predicted | No Stroke | Stroke |
|---|---|---|
| **No Stroke** | 955 (TN) | 17 (FP) |
| **Stroke** | 12 (FN) | 38 (TP) |

Table 2 presents the confusion matrix illustrating how well the Random Forest version predicts strokes. The model correctly classifying 955 non stroke cases and 38 stroke cases, signifying a high level of general accuracy. However, it mistook 12 stroke sufferers as non-stroke, which is important in scientific diagnosis as a result of the potential for delaying treatment as a result of false negatives. Additionally, 17 non- stroke cases were misclassified as stroke, mainly as a result of false alarms.

**Table 3.** Confusion Matrix for XGBoost Model

| Actual / Predicted | No Stroke | Stroke |
|---|---|---|
| **No Stroke** | 964 (TN) | 8 (FP) |
| **Stroke** | 11 (FN) | 39 (TP) |

Table 3 demonstrates the confusion matrix of the XGBoost, which showed superior classification performance in all the metrics. The model produced the most straightforward 8 false positives and 11 false negatives while classifying 39 stroke cases and 964 non-stroke cases. In contrast to Random Forest, XGBoost improves the balance between sensitivity and precision and lowers the number of incorrect predictions

**Table 4.** Confusion Matrix for Logistic Regression Model

| Actual / Predicted | No Stroke | Stroke |
|---|---|---|
| **No Stroke** | 940 (TN) | 32 (FP) |
| **Stroke** | 4 (FN) | 46 (TP) |

Table 4 demonstrates the confusion matrix of logistic regression. It was able to accurately predict 46 stroke cases and 940 non-stroke cases, while having only 4 false negative predictions, which is the lowest number of false negative predictions of all models. Although it has 32 false positive predictions, it is appropriate for screening because it is more important to avoid missing stroke cases than to avoid false positive predictions

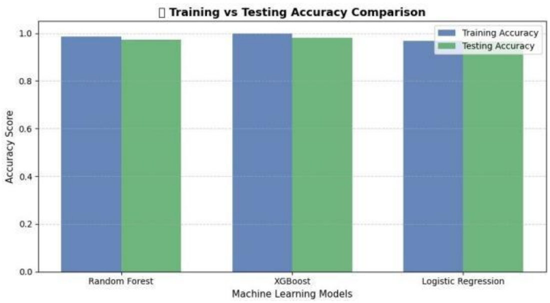


**Fig. 2.** Comparison of training and testing accuracy across Random Forest, XGBoost, and Logistic Regression models.

As illustrated in Fig. 2, the training and testing accuracy of the implemented models are compared using bar graph. XGBoost performed the very super checking out accuracy of 98.1%, accompanied by means of the use of Random Forest at 97.2% and Logistic Regression at 96.5%. The small distinction among education and trying out accuracy for all models shows minimum overfitting and accurate generalization to unseen statistics. This demonstrates that all models are solid and reliable.

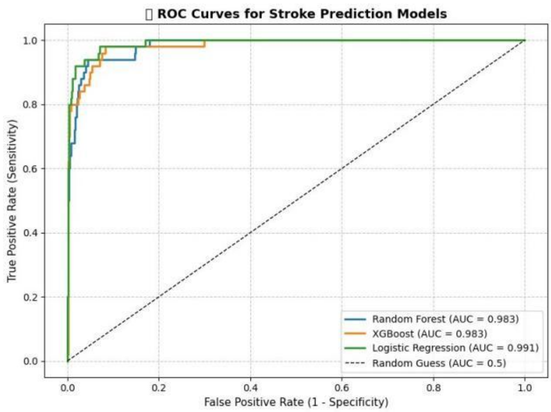


**Fig. 3.** The stroke prediction models that are being evaluated are described by Receiver Operating Characteristic (ROC) curves.

Fig. 3 shows how the three models' sensitivity and specificity are traded off. With a very good AUC value of 0.991, logistic regression demonstrated remarkable capacity to differentiate between stroke and nonstroke cases. Strong AUC values of 0.983 were found for both Random Forest and XGBoost. The model's overall quality improves as the curve gets closer to the top-left nook. Logistic regression is the most sensitive model in terms of performance.

## 5 Conclusion

The mortality rate due to stroke is significant, as well as the long-term disabilities that stroke causes. This further stresses the importance of efficient early risk assessment systems. This research presents a stroke prediction system that makes use of ensemble techniques like XG Boost, Random Forest, and Logistic Regression. The model's effectiveness would be improved by adding clinical, demographic, and lifestyle variables as well as by using techniques like SMOTE to address the issue of class imbalance. Additionally, applying Explainable AI techniques like SHAP would improve the model's effectiveness.

## References


1. Deepthi, T.N., Sharmila, Sk., Swarna, M., Gouthami, M., Akshaya, C.L.: Predic-tion of Brain Stroke in Human Beings Using Machine Learning. In: 2023 Second International Conference on Electronics and Renewable Systems (ICEARS), pp. 1314–1320. IEEE, Tamil Nadu, India (2023)
2. Hossain, A.M., Ahmed, N.U., Rahman, M., Islam, M.R., Sadhya, G., Fatema, K.: Analysis of Sociodemographic and Clinical Factors Associated with Hospital-ized Stroke Patients of Bangladesh. Faridpur Medical College Journal 6(1), 19–23 (2011)
3. Islam, M.N., Moniruzzaman, M., Khalil, M.I., Basri, R., Alam, M.K., Loo, K.W., Gan, S.H.: Burden of Stroke in Bangladesh. International Journal of Stroke 8(3), 211–213 (2013)
4. Reddy, K.M., Sakthivel, M., Kovuri, K., Avanija, J., Kaleru, S.: Brain Stroke Pre-diction Using Deep Learning: A CNN Approach. In: International Conference on Inventive Research in Computing Applications (ICIRCA), pp. 1–6. IEEE, India (2022)
5. Shan, X., Chen, Y., Qiao, Z.: A Study of Stroke Prevalence Prediction Based on Random Forest Algorithm. In: 2023 IEEE 3rd International Conference on Power, Electronics and Computer Applications (ICPECA), pp. 1–6. IEEE, Shenyang, China (2023)
6. Srivastav, S., Guleria, K., Sharma, S.: Machine Learning Models for Early Brain Stroke Prediction: A Performance Analogy. In: 2023 World Conference on Com-munication and Computing (WCONF), pp. 1–6. IEEE, India (2023)
7. De Bresser, J., Portegies, M.P., Leemans, A., Biessels, G.J., Kappelle, L.J., Viergever, M.A.: A Comparison of MR-Based Segmentation Methods for Mea-suring Brain Atrophy Progression. NeuroImage 54(2), 760–768 (2011)
8. Chaki, J., Wozniak, M.: Deep Learning and Artificial Intelligence in Action (2019–2023): A Review on Brain Stroke Detection, Diagnosis, and Intelligent Post-Stroke Rehabilitation Management. IEEE Access 12, 52161–52181 (2024)
9. Satapathy, S.K., Patel, A., Yadav, P., Thacker, Y., Vaniya, D., Parmar, D.: Machine Learning Approach for Estimation and Novel Design of Stroke Disease Predictions using Numerical and Categorical Features. In: 2023 International Conference for Advancement in Technology (ICONAT), pp. 1–6. IEEE, Goa, India (2023)
10. Sundaram, K., Lanitha, B., Kamaraj, K., Ramamoorthy, A.K.: Enhanced Brain Stroke Prediction: An Ensemble of Random Forest, Logistic Regression and XG-Boost. In: 2024 International Conference on Emerging Research in Computational Science (ICERCS), pp. 1–5. IEEE (2024)
11. Devarakonda, N., Sai, B.L.S., Pravalika, U., Naganjaneyulu, S.: Brain Stroke Pre-diction Using Machine Learning Techniques. In: 2023 Fifth International Confer-ence on Electrical, Computer and Communication Technologies (ICECCT), pp. 1–6. IEEE (2023)
12. Wang, B., Jiang, B., Liu, D., Zhu, R.: Early predictive accuracy of machine learning for hemorrhagic transformation in acute ischemic stroke: A systematic review and meta-analysis. Journal of Medical Internet Research 27, e71654 (2025)
13. Akinwumi, P.O., Ojo, S., Nathaniel, T.I., Wanliss, J., Karunwi, O., Sulaiman, M.: Evaluating machine learning models for stroke prediction based on clinical variables. Frontiers in Neurology 16 (2025)