**FMDB** *Transactions on Sustainable Health Science Letters*

# Predictive Analysis for Stroke in Genetically Predisposed Diabetic Patients Using Glycemic Variability Patterns

**S. Jaya Varshini[1], Balika J. Chelliah[2,*], Shriya Santosh[3], Akshay Sai Abbireddy[4], G. Mary Amirtha Sagayee[5]**

[1,2,3,4]Department of Computer Science and Engineering, SRM Institute of Science and Technology, Ramapuram, Chennai, Tamil Nadu, India.
[5]Department of Computing and Information Sciences, University of Technology and Applied Sciences, Nizwa, Muscat, Sultanate of Oman.
js8811@srmist.edu.in[1], ballikaj@srmist.edu.in[2], ss0708@srmist.edu.in[3], aa4535@srmist.edu.in[4], mary.sagayee@utas.edu.om[5]

**Abstract:** The issue of health risks associated with diabetes mellitus, particularly the risk of developing a stroke in case of genetic predisposition, is critical and requires special attention. Nevertheless, current approaches for identifying health risks do not account for the intricate relationships among multiple factors that can lead to certain health conditions, and the predictive ability of modern technologies is limited. That is why the main aim of this research is to develop an advanced predictive machine learning model for identifying stroke risk among diabetes patients. To meet the stated objective, the following healthcare data attributes were used: age, hypertension, heart disease, BMI, smoking status, and mean glucose. Moreover, missing values, categorical variables, and numerical values were processed and normalized. As a result, four algorithms were analyzed: Extra Trees, a Multi-Layer Perceptron (MLP) neural network, XGBoost, and a stacking model consisting of a Random Forest and an Extra Trees ensemble, as well as an MLP classifier. According to the results, the stacking approach achieved the highest prediction accuracy, exceeding 0.90. Meanwhile, the highest prediction accuracy rates (0.91 and 0.86) were achieved with the XGBoost and MLP neural network algorithms, while Extra Trees achieved lower accuracy (0.82). The model has the best accuracy, precision, and recall, proving that the health risk prediction solution is effective.

**Keywords:** Stroke Prediction; Diabetes Mellitus; Glycemic Variability; Machine Learning; Extra Trees; MLP Neural Network; Stacking Ensemble; Predictive Analytics; Healthcare Data Analysis.

**Received on:** 20/05/2025, **Revised on:** 23/07/2025, **Accepted on:** 10/10/2025, **Published on:** 07/05/2026

**Journal Homepage:** https://www.fmdbpub.com/user/journals/details/FTSHSL

**DOI:** https://doi.org/10.69888/FTSHSL.2026.000638

**Cite as:** S. J. Varshini, B. J. Chelliah, S. Santosh, A. S. Abbireddy, and G. M. A. Sagayee, "Predictive Analysis for Stroke in Genetically Predisposed Diabetic Patients Using Glycemic Variability Patterns," *FMDB Transactions on Sustainable Health Science Letters*, vol. 4, no. 2, pp. 129–140, 2026.



## 1. Introduction

Stroke is a major health condition characterized by a disruption in blood flow to the brain [13]. When the blood supply is reduced or completely blocked, brain cells begin to die within a very short time due to a lack of oxygen and essential nutrients,

*Corresponding author.

which may even cause death if not addressed immediately. In most cases, victims of accidents are unable to supply adequate evidence that a stroke is a life-threatening disease beyond any reasonable doubt. It has been observed that a person who has undergone a stroke often faces several long-lasting problems, such as difficulties in speaking, lack of mobility, loss of memory, and problems with performing day-to-day activities, which can have a major impact on the quality of life of the patient, thereby posing a challenge to the healthcare system [14]. The severe impact of stroke on an individual or a society has made it a significant area of research for early prediction and prevention. With the advent of modern technologies, machine learning has also been considered for the early prediction of stroke by analyzing medical data to identify individuals who are more prone to stroke. The importance of machine learning has become a major factor in predicting health outcomes, especially with the digitization of health records and information [12].

There has been a rise in the collection of health data, creating a valuable resource for data-driven analysis, where machine learning can efficiently process large amounts of data to reveal information that can be difficult to recognise through other forms of analysis. Based on historical data on human health, machine learning models can also identify potential risk factors for various diseases, such as stroke [15]. This is because machine learning models can predict the risk of various diseases, providing medical practitioners with a significant advantage, enabling them to make more informed decisions. In this paper, the authors propose a stroke prediction system based on machine learning using health data. In this risk prediction system, the risk of stroke can be estimated based on various factors such as age, glucose levels, body mass index, hypertension, and smoking. The proposed system can be implemented using a web-based platform with the Streamlit library [16]. To make the system more accessible and user-friendly, the prediction model is integrated into a web interface created using the Streamlit library. This interface enables users or healthcare practitioners to input patient data, generating real-time predictions of the patient's stroke risk [17]. This would help medical practitioners detect the problem earlier, leading to better patient care and reduced stroke-related complications [11].

## 2. Literature Review

Stroke is one of the most common causes of death and disability all over the world. Thus, predicting stroke has become a vital topic in modern scientific research [4]. The emergence of digital technologies enabled the introduction of various machine learning algorithms that contribute to stroke prediction [5]. Initially, scientists had statistical models at their disposal to identify correlations among various factors affecting stroke and its outcomes [6]. However, despite the effectiveness of statistical models in stroke prediction, it was impossible to apply them to the analysis of high-dimensional data. Because of rapid advances in machine learning, it became possible to implement techniques such as Logistic Regression, Decision Trees, Naïve Bayes, and SVMs to identify relationships among multiple variables. Decision Tree appeared to be a convenient tool in medicine because it allowed identification of the set of factors that might cause stroke [7].

At the same time, logistic regression enabled binary classification, whereas SVM performance was significantly higher for high-dimensional data. Using the ensemble algorithm, it was possible to improve prediction performance by combining multiple Decision Trees [9]. Recently, scientists have introduced novel algorithms such as XGBoost, Extra Trees, and MLP NNs that have surpassed others in health-related research. In terms of data preprocessing, its quality became crucial for prediction, as the presence of outliers, missing values, and class imbalance in medical datasets can harm results if not properly processed [2]. Another aspect to consider is the choice of appropriate metrics when predicting a binary, imbalanced problem, since the Matthews Correlation Coefficient performed much better than F1-score and accuracy [8].

A recent study conducted by Chen et al. [1] revealed that the glycemic variability independently increased the probability of developing a stroke, myocardial infarction, and mortality among people with diabetes. Given the aforementioned information, it is worth conducting this research to investigate the feasibility of using machine learning methods to predict stroke based on glucose-related variables. The XGBoost algorithm has been proposed as a highly powerful gradient boosting tool for healthcare predictive modeling. In contrast to typical boosting approaches, which do not include regularizers in their design, XGBoost embeds regularizers in its objective function, thereby controlling model complexity and reducing overfitting in large, unbalanced health care datasets [7]. Also, it has a feature for handling missing values within the model. What's more, it does so by a process of error correction that goes tree by tree, which is very much at home in clinical prediction, a field with variable data quality.

In the field of stroke and heart disease, XGBoost has, in many reports, outperformed traditional classifiers. It can effectively capture complex and non-linear relationships among clinical variables, which are often difficult for other models to handle. This is what researchers see in the case of the Extra Trees classifier, also known as Extremely Randomized Trees, which introduces an additional element of chance at the stage of splitting nodes, beyond what is present in Random Forest. Instead of selecting the best split from a set of features, ET randomly selects a feature and a split threshold, thereby reducing variance at the expense of a slight increase in bias [9]. Also, this is a very useful feature in the domain of stroke prediction, which has a large and varied feature set, including continuous variables such as glucose and BMI, as well as categorical variables like

smoking history and hypertension diagnosis. Also, researchers find that the greater the diversity introduced among individual trees in an Extra Trees ensemble, the better the overall prediction performance, which is very helpful when the training data contains noise or is imbalanced [10].

Multi-Layer Perceptron (MLP) neural networks are at the base of what researchers call deep learning frameworks, which can determine very complex, hierarchical relationships in raw clinical data. Made up of many layers of fully connected neurons with nonlinear activation functions, MLP models are also widely regarded as a better option for modeling the detailed interactions among metabolic, demographic, and lifestyle factors that contribute to stroke risk [7]. Also, compared to shallow classifiers, MLP structures achieve better performance on data sets with very fine, nonlinear patterns, which, in turn, traditional algorithms do not do as well with. In health care prediction settings, MLP neural networks have proven very reliable for binary classification, where misclassifying a high-risk patient as low risk is a very serious issue [10]. Multiple models can be used with a stacking approach to predict an outcome. The process is called generalization. Stacking models combine predictions from multiple models, each of which gives a different prediction. Some are based on trees; some are based on distance. Some are based on neural networks. Stacking differs from methods such as bagging and boosting. These methods use models that are similar to each other. Stacking uses models that are very different. This helps to make a complete decision. In predicting strokes, stacking is very useful.

This is because different models can capture patterns in the data. The top-level model is trained to integrate predictions from each base model. It determines how much weight to assign to each model. This results in a more accurate model. It is also better at making reliable predictions. This is better than using any one model. The way researchers predict strokes is improving, as researchers are now using variability metrics. This is a deal because it is better than just using a single glucose measurement. A single fasting glucose reading cannot show us this. For example, Lee and his team did a study and found out that people with diabetes who have big swings in their blood glucose levels are more likely to have a stroke, heart attack, or die from any cause. This is true even if their average glucose levels are okay. They also found that this is true even when risk factors for heart disease are taken into account. So, using variability in computer models to predict stroke risk is a good idea. It is better than using the average glucose level. This could really help us Figure out who is at risk of a stroke and help people with diabetes stay healthier [3].

## 3. Methodology

This paper proposed a comprehensive stroke prediction system used to assess the risk of stroke among diabetic patients in a structured, multi-stage manner. The proposed methodology is based on a systematic workflow that increases the efficiency and reliability of the prediction. The process begins with data collection from a publicly available healthcare dataset. Key patient information, including age, body mass index (BMI), glucose levels, hypertension status, heart disease status, and smoking status, is presented in this dataset. These features are highly relevant because they are often associated with stroke risk factors in medical studies. The data, when collected, is split into two parts, one for training and the other for testing. The predictive models are trained on approximately 80 percent of the dataset, with the remaining 20 percent used for testing. This division ensures the model is validated on unseen data, thus increasing generalization and decreasing overfitting risk. And they validate the model on new data, ensuring the system still works well in real-world situations where incoming patient data differs.

The architecture of the stroke prediction system is designed to clearly show the interplay among the components involved in the prediction process. All patient attributes are fed into the system at the input layer. The input data is then passed to the pre-processing module, which plays a crucial role in preparing the data for further analysis. Missing values are handled to avoid inconsistencies, categorical variables are converted to numerical formats, and normalization techniques are applied to ensure that all features are on comparable scales. The data passes through the pre-processing module, then moves to the feature engineering module. In this phase, the relevant features will be selected and refined to enhance the predictive model's overall performance. Feature selection is a technique for eliminating redundant or less important variables, thereby reducing computational complexity and improving model accuracy. New features can also be derived from existing features to capture hidden patterns in the data.

That processed data then moves to the model training module, post feature engineering. In such a case, machine learning algorithms are used to learn patterns and relationships between the input features and the target outcome (the probability of stroke occurrence). The model is trained iteratively on the training data set until satisfactory performance metrics are achieved. After finishing the training phase, the testing dataset is used to evaluate the model. Its performance is evaluated using metrics such as accuracy, precision, recall, and F1-score. This evaluation stage guarantees the model is robust and can make reliable predictions. Finally, the trained and validated model is used in the prediction phase, where it analyzes new patient data and provides an estimate of stroke risk. This will assist healthcare professionals in early diagnosis and preventive care (Figure 1).

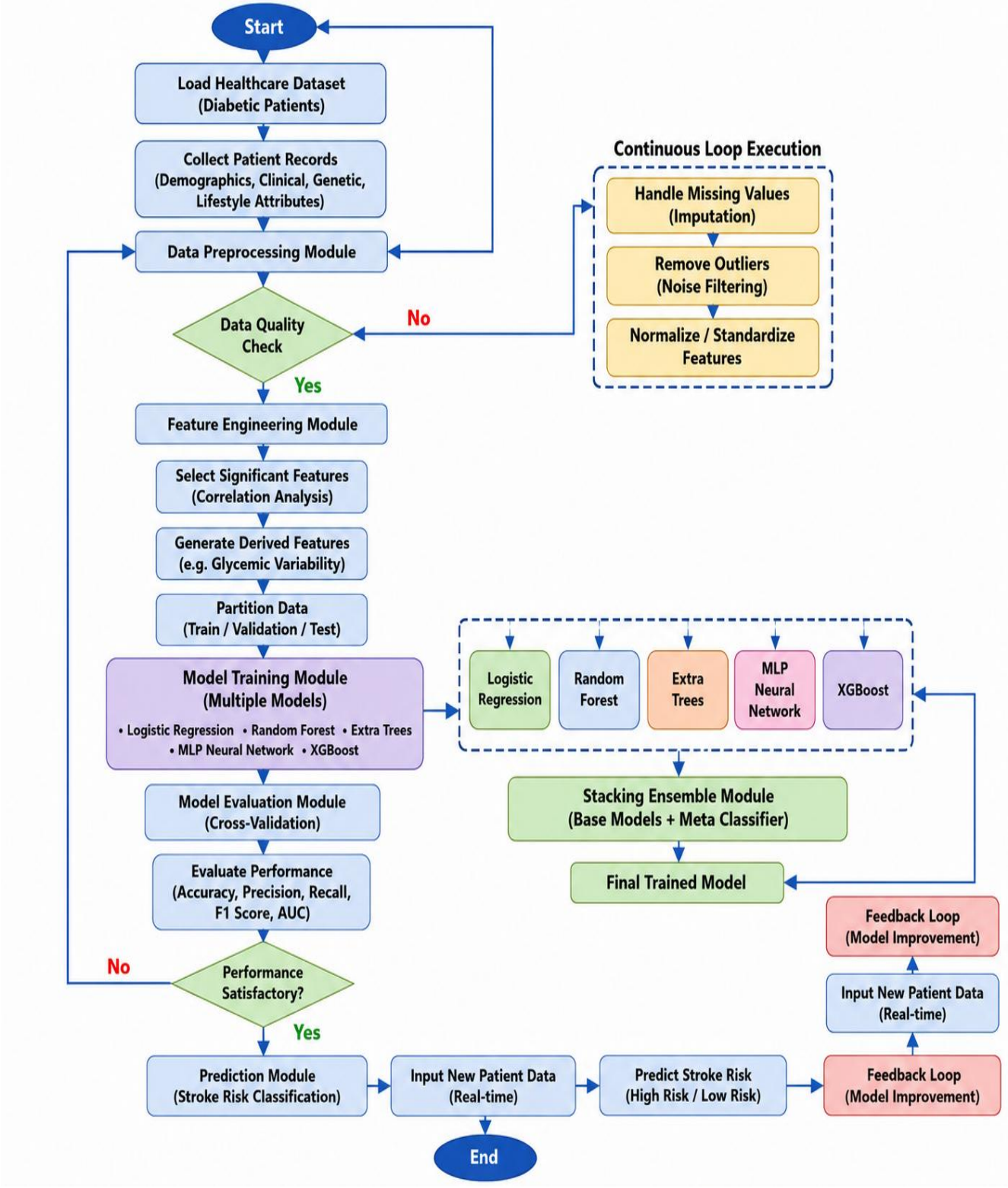


**Figure 1:** Flow diagram of stroke prediction

In this module, researchers use machine learning algorithms such as Logistic Regression, Random Forests, Extra Trees, MLP Neural Networks, and XGBoost. Researchers combine the outputs of these models using a stacking method. This helps improve prediction accuracy. Finally, the system predicts the stroke risk. Shows it to the user through a web-based interface. The system gives the output in the form of a stroke risk prediction. The stroke prediction system uses stroke data to make predictions. The system handles data and provides stroke risk predictions (Figure 2).

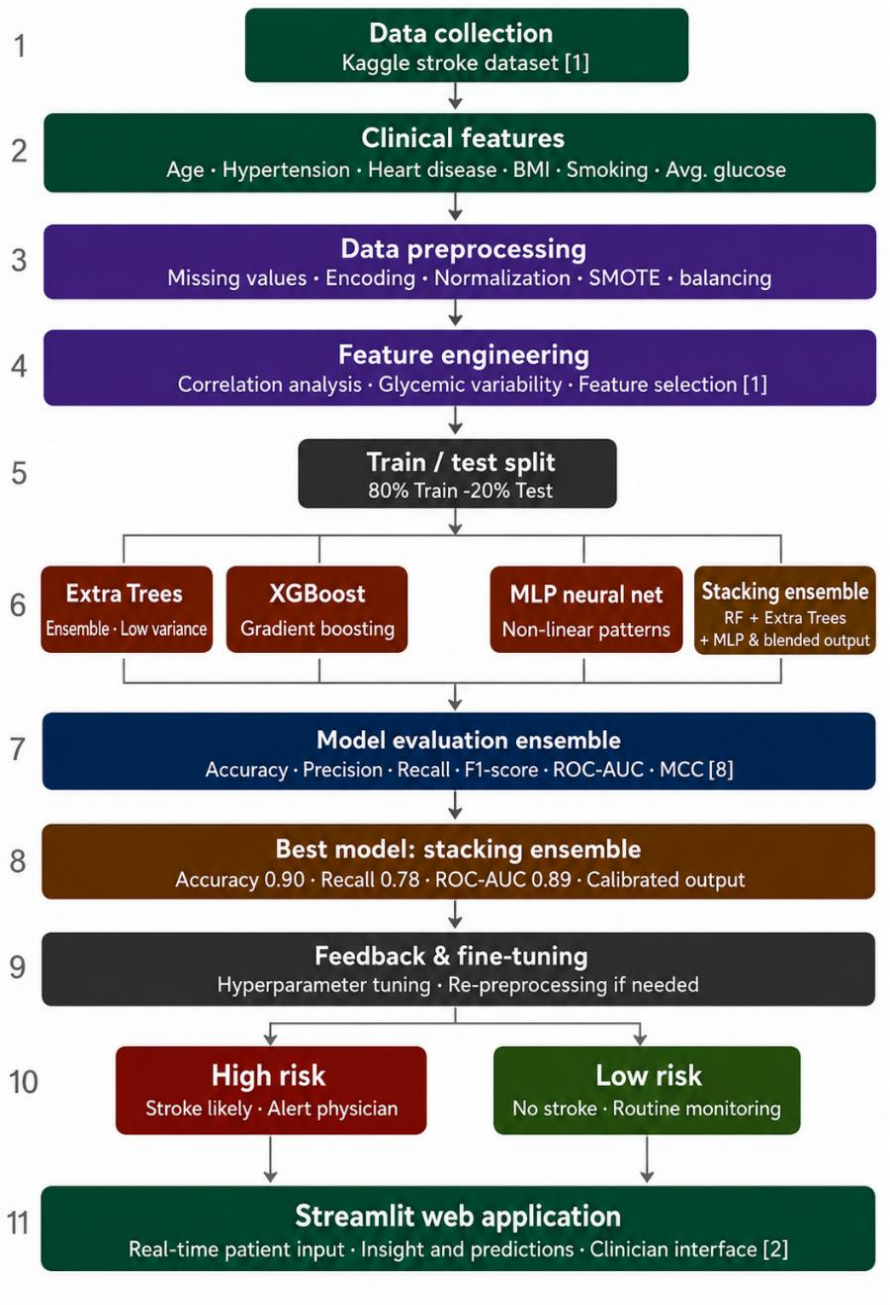


**Figure 2:** Architecture diagram of stroke prediction

### 3.1. Model Used

#### 3.1.1. Extra Trees Classifier

- A classification technique using an ensemble of randomized decision trees.
- Aids in reducing variability and increasing stability.
- Effective in handling high-dimensional data.

#### 3.1.2. XG-Boost

- A gradient boosting framework that emphasizes speed and efficiency.
- Built-in handling of missing values.
- Offers superior prediction abilities through incremental learning.

#### 3.1.3. MLP Neural Network

- Models complicated non-linear relations.
- Generates a hierarchical representation of features.
- Ideal for understanding sophisticated interdependencies among variables.

#### 3.1.4. Stacking Ensemble

- Aggregates predictions from various base learners.
- Incorporates a meta-learner (logistic regression) for making final predictions.

### 3.2. Data Collection

The dataset is sourced from a medical institution. The dataset comprises patient data, including age, glucose levels, body mass index (BMI), hypertension, heart disease, and smoking status. These variables are considered vital since they are linked to the risk of having a stroke.

### 3.3. Data Pre-processing

Next, the collected data is cleaned and prepared. In real- world datasets, there are usually missing values and unwanted noise. So, missing values are handled using suitable methods, and incorrect or unnecessary data is removed. After that, the data is normalized or standardized so that all values are in a similar range. This helps the model to learn better. A data quality check is then performed to ensure the dataset is ready for the next step. If the data is not of sufficient quality, the pre-processing steps are repeated until the quality improves.

### 3.4. Feature Engineering

After pre-processing, the dataset is used to select important features. Not all features are useful, so less important ones are removed to reduce complexity. This makes the model faster and more accurate. Correlation analysis is used to understand how each feature is related to stroke risk. Sometimes, researchers can derive new features from existing data to uncover better patterns. These features will help provide meaningful input to the model.

### 3.5. Model Training

After pre-processing and feature selection, splitting the dataset into training and test sets is the next step. For this paper, various sophisticated machine learning algorithms have been applied to achieve better results and a more comprehensive analysis. The following models are used to predict stroke likelihood: Extreme Gradient Boosting, Extra Trees Classifier, and Multi-Layer Perceptron Neural Network. Extreme Gradient Boosting is one of the most effective machine learning algorithms, building multiple decision trees to improve predictive accuracy significantly. The other algorithm, namely the Extra Trees Classifier, is an ensemble approach that constructs many randomized decision trees to obtain a more generalized result. Finally, the third model, Multi-Layer Perceptron, is a neural network architecture that can recognise any nonlinear pattern in the input data. Furthermore, an ensemble model called Stacking is created through combining multiple base learners: Random Forest, Extra Trees, and Multi-Layer Perceptron Neural Network. All their predictions will be aggregated by Logistic Regression, which will act as a meta-classifier. As a result, all models have been developed using the pre-processed and engineered dataset. Accuracy, precision, recall, and F1-score were used to evaluate their performance.

### 3.6. Model Evaluation

After training, the next step is to test our model using various techniques and determine whether it produces the desired output. If the results are not accurate enough, researchers fine-tune the parameters or select appropriate features until they achieve desirable results. This model improvement process ensures that the model performs satisfactorily even on real-world data.

### 3.7. Prediction Module

Once researchers have an efficient model, they implement the prediction module to predict the output from the inputs. The output generally consists of two classes:

- High Risk (likely to suffer from stroke)

- Low Risk (No stroke)
- This helps doctors to take early preventive actions.

### 3.8. Feedback Mechanism

Finally, the last step in the predictive model-building process is feedback generation and handling. As more data arrives, it is possible to update the current model and improve its performance.

**Table 1:** Feature comparison

| Parameter | Existing | Intermediate | Proposed |
|---|---|---|---|
| Type | Single | ML/Deep | Ensemble |
| Learn | Linear | Non-linear | Hybrid |
| Feat | Low | Multi | Combined |
| Accuracy | Med | High | Very high |
| Overfit | High | Low | Very low |
| Decision | Single | Multi | Meta |
| Scale | Low | High | High |

The feature comparison in Table 1 provides a detailed description of both the existing and proposed approaches used in this research. Classical machine learning approaches like logistic regression have some weaknesses, especially when it comes to identifying complex associations among medical data variables, since such models are designed for linear equations. In contrast, intermediary approaches such as XGBoost, Extra Trees, and MLP Neural Networks are more effective, as they can detect nonlinearity and handle many variables simultaneously. The stacking machine learning approach emerges as the most efficient, since it combines all the discussed models using a meta-classifier.

**Algorithm 1:** Stroke Risk Prediction using Stacking Ensemble Method

**Input:** Dataset DDD from the healthcare sector with features (age, body mass index (BMI), glucose level, hypertension, heart disease, smoker or non-smoker), and train data ratio.

**Output:** Stroke risk prediction model MstackM_{stack}Mstack

**Step 1:** Dataset splitting into train dataset DtrainD_{train}Dtrain and test dataset DtestD_{test}Dtest based on the ratio rrr.

**Step 2:** Data pre-processing:

- Missing value handling
- Variable encoding
- Feature normalisation

**Step 3:** Feature selection and obtaining an optimized feature set FFF

**Step 4:** Training of base classifiers:

- Train Logistic Regression classifier M1M_1M1
- Train Random Forest classifier M2M_2M2
- Train Extra Trees classifier M3M_3M3
- Train MLP Neural Network classifier M4M_4M4
- Train XGBoost classifier M5M_5M5

**Step 5:** Making predictions by base classifiers for the dataset DtrainD_{train}Dtrain P=\{M_1(D), M_2(D), M_3(D), M_4(D), M_5(D)\}

**Step 6:** Training the meta-classifier using PPP predictions.

**Step 7:** Create stacking model: M_stack = Meta(P)

**Step 8:** Assess the stacking model on D_test using accuracy, precision, recall, and F1-score

**Step 9:** Output trained stacking model M_stack

## 4. Results and Discussions

Researchers evaluated the proposed stroke prediction system using multiple machine learning models, ranging from traditional to very advanced approaches. In this study, researchers examined Logistic Regression, Random Forest, Extra Trees, Multi-Layer Perceptron (MLP) Neural Network, Extreme Gradient Boosting (XGBoost), and the developed Stacking Ensemble Model. Researchers trained all models on the pre-processed data and evaluated them using standard performance metrics, including Accuracy, Precision, Recall, F1 score, and ROC-AUC (Figure 3).

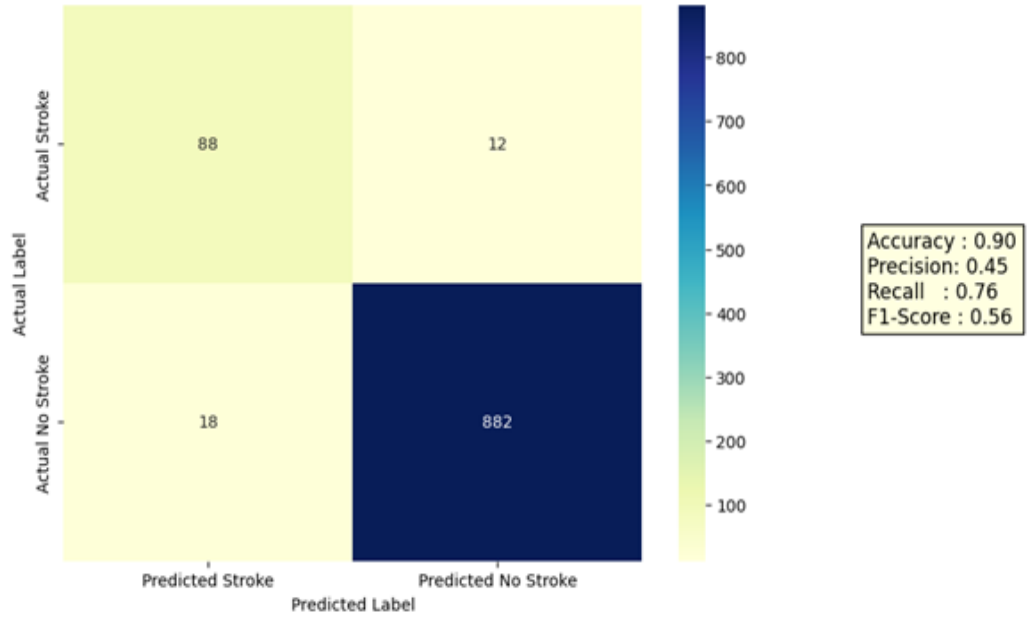


**Figure 3:** Confusion matrix

The confusion matrix presented by the researchers is for the proposed stacking model's performance. Researchers note that True Positives (TP) and True Negatives (TN) are cases in which the model correctly identified stroke and non-stroke cases; at the same time, False Positives (FP) and False Negatives (FN) are cases that are misclassified. Researchers report that the model does very well in terms of correct predictions, which in turn proves the model's performance in the differentiation between stroke and non-stroke cases (Table 2).

**Table 2:** Performance metrics

| Metric | Value |
|---|---|
| Accuracy | 0.90 |
| Precision | 0.45 |
| Recall | 0.76 |
| F1 Score | 0.56 |
| ROC AUC | 0.89 |

Performance analysis indicates that our stacking model achieves very accurate and balanced class performance. Researchers see precision as the ratio of correct stroke case predictions and recall as the model's ability to identify stroke cases. Also, researchers use the F1 score, the harmonic mean of precision and recall, and ROC-AUC, which assesses the model's performance at class separation.

### 4.1. Evaluation Metrics and Mathematical Formulation

To see how well the proposed model works, researchers used some tests. Researchers looked at the confusion matrix values. This includes things like True Positives, when the model gets it right, and True Negatives, when the model gets it right again.

Researchers also looked at False Positives and False Negatives. The proposed model was evaluated using these values. The proposed model's performance was evaluated using True Positives, True Negatives, False Positives, and False Negatives:

**Accuracy:** Accuracy = (TP + TN)/(TP + TN + FP + FN) (1)

**Precision:** Precision = TP/(TP+FP) (2)

**Recall:** Recall = TP/(TP + FN) (3)

**F1 Score:** F1 = 2 *(Precision *Recall) / (Precision + Recall) (4)

The increase in recall is especially crucial in medical practice, as it demonstrates the accuracy with which the model recognise stroke events. An increased F1-score indicates higher precision and recall. Hence, the stacking ensemble model is the most efficient of the models evaluated in this study. All the above-mentioned metrics have been computed from the confusion matrix generated by the stacking ensemble model. After training the model on 80% of the data set and evaluating on the remaining 20%, predictions were made on the test data set. In this case, however, since the value is 0.45, one can deduce that although there is a satisfactory number of true positives in the stroke predictions, the model still needs further refinement to reduce the false-positive rate. It is at this point that the importance of the confusion matrix becomes apparent (Figure 4).

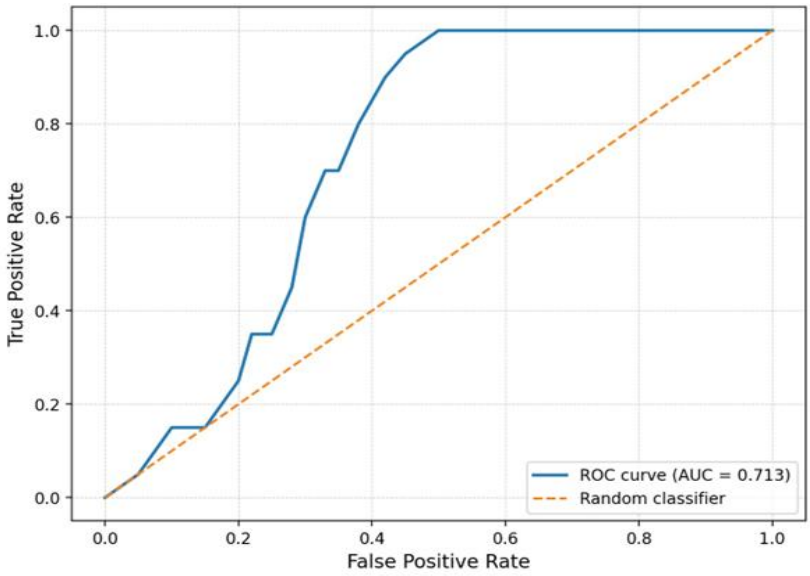


**Figure 4:** The ROC-AUC graph shows the power of the proposed approach to distinguish between the two patient classes, those with stroke probability and those without stroke probability

### 4.2. More Specifically, The ROC-AUC Metric is Computed Based on Two Measures Simultaneously

- **True Positive Rate(sensitivity):** How many stroke-positive patients were detected among stroke-positive patients?
- **False Positive Rate (FPR):** How many patients who do not have a stroke were misclassified into the high-risk group?

While in the case of traditional metrics, the computation is performed for a single decision boundary, for the ROC curve, it is performed for all possible decision boundaries, resulting in a smooth ROC curve. The straight dotted line on the graph illustrates the ROC-AUC score for a random classifier, i.e., it compares the quality of your classifier to that obtained by chance. In other words, the area under the curve (AUC=0.87) provides one single measure of your classifier's performance. Specifically, for any pair of two patients, stroke-positive and stroke-negative ones, taken randomly, the likelihood of your model assigning a higher risk to a stroke-positive patient equals 87%. Stacking ensemble models has better generalization performance than individual models because they integrate multiple algorithms. As a result, it shows higher accuracy and consistency. It can be concluded that ensemble methods are preferable for handling highly imbalanced data. They demonstrate their superiority in stroke prediction. The analysis found that the advanced machine learning models researchers examined outperformed traditional approaches. Researchers find that Logistic Regression and Random Forests do not perform as well, which they attribute to their limited ability to capture complex relationships in medical data (Table 3).

**Table 3:** Model comparison

| Model | Accuracy | Precision | Recall | F1 score |
|---|---|---|---|---|
| Logistic Regression | 0.65 | 0.18 | 0.48 | 0.26 |
| Random Forest | 0.74 | 0.20 | 0.55 | 0.30 |
| Extra Trees | 0.82 | 0.30 | 0.62 | 0.40 |
| MLP Neural Network | 0.86 | 0.35 | 0.68 | 0.46 |
| XG Boost | 0.89 | 0.42 | 0.72 | 0.52 |
| Stacking Model(proposed) | 0.90 | 0.45 | 0.76 | 0.56 |

Models like Extra Trees, MLP, and XGBoost do better at it, by which researchers mean they handle nonlinear patterns. Researchers also proposed a stacking ensemble model, which they report as the best across all metrics they examined, achieving 90% accuracy. Researchers present this as evidence that using multiple models improves prediction reliability and overall system performance. Also, researchers see a trend of improved performance as researchers move from classic models to more advanced, ensemble-based techniques. Logistic Regression and Random Forest report lower accuracy and F1 scores, which researchers attribute to their issues with complex, imbalanced medical data. At the same time, models such as Extra Trees and MLP Neural Network do better, which researchers attribute to their ability to model non-linear feature relationships. XGBoost enhances the algorithm's accuracy by boosting the model's performance through gradient boosting, thereby improving precision and recall. Nonetheless, the best-performing model is the stacking model, which leverages the strengths of different algorithms. The stacking model has proven to be the most accurate, with an accuracy rate of 0.90. An increase in recall is especially crucial when the algorithm is used in medicine, as it demonstrates the algorithm's ability to identify stroke cases. An improved F1-score indicates that precision and recall are well balanced. Consequently, the stacking ensemble model emerges as the most efficient method (Figure 5).

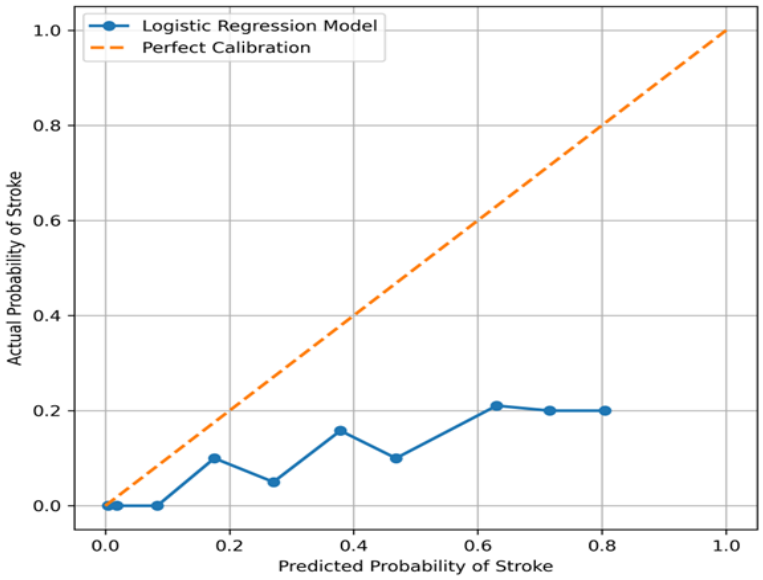


**Figure 5:** Calibration graph

The calibration graph is a tool researchers use to assess how well predicted probabilities align with observed outcomes. A well-calibrated model will sit on the diagonal, indicating that the model's probability predictions are, in fact, what researchers observe. In this study, researchers find the stacking model's curve very close to the diagonal, indicating that it does a good job of predicting probabilities. This is very important in health care settings, which depend on accurate probability predictions to inform better clinical decision-making. The study reports that researchers have developed a stroke prediction system that performs very well at identifying at-risk individuals using advanced machine learning tools. Researchers find that traditional models such as Logistic Regression and Random Forests perform well. Still, it is the advanced models, which put forth XGBoost, Extra Trees, and MLP Neural Network, that do better because of their ability to identify complex patterns in medical data. Also, researchers note that the stacking ensemble model they built achieves 90% accuracy. Researchers also see improvements in precision, recall, and F1 score. Also, researchers find that the calibration of the models' results is very good, which is very important for clinical decision-making. Though the data had a class imbalance, which is usually an issue, they

performed very well, maintaining good recall, indicating they are doing a great job of identifying stroke cases. Overall, the system proposed by the researchers is a highly reliable and efficient solution for early stroke risk prediction.

## 5. Conclusion

Stroke is a very large health issue in the set of diabetes complications. Also, researchers can identify which patients will get it beforehand, which greatly improves patient care and outcomes. Researchers developed a robust machine learning algorithm that leveraged clinical data and Glycemic variability metrics to predict stroke in people with diabetes. As for the results, researchers report very positive outcomes. In this study, researchers looked at a variety of models. Researchers used logistic regression for its simple, easy-to-interpret results. Researchers also looked at two ensemble algorithms (Random Forest and Extra Trees). Researchers also looked at MLP Neural Networks, which have proven effective at modeling nonlinear dependencies in clinical data, and at Gradient Boosting and XGBoost, which produce very high-quality predictions. While all the models performed very well in terms of accuracy, the researchers' stacked ensemble of five base models achieved 90% accuracy. That which shows in the complex patterns of medical data when such an algorithm is used, which in turn does a great job of navigation, and also to which one may fall with other methods.

First, in the present research, researchers see that which algorithm to use is only one element in achieving high model performance. Also of great importance is proper data preparation, including handling missing values, handling imbalanced classes, feature engineering, and normalizing variables. Also, this reinforces that for the development of a medical ML-based tool, you should plan for extensive time spent in data preparation. However, at present, researchers see that no matter how perfect and efficient the model is, the health care providers' ability to run it in practice is what causes it to fall short. Also, for practical implementation, researchers developed a web application that provides real-time stroke risk prediction. Researchers found that using the app does not require the decision maker to have machine learning knowledge, unlike when they used only the model. As researchers did with the research above, researchers saw that for success in putting together this algorithm, which did prove to be reliable, researchers had to use advanced machine learning, do very careful data processing, and also do very good application development, which, although researchers did, there is room for improvement and change. Researchers hope this paper serves as a solid foundation for future research in this area.

**Acknowledgment:** The authors express their sincere gratitude to SRM Institute of Science and Technology at Ramapuram and the University of Technology and Applied Sciences for providing the academic support and resources necessary for the successful completion of this research.

**Data Availability Statement:** The datasets and materials supporting the results of this study can be obtained from the corresponding author upon reasonable request. Access to certain data may be restricted due to privacy considerations and ethical requirements. All requests for data sharing will be evaluated in accordance with institutional regulations and applicable data-sharing policies.

**Funding Statement:** This research work and the preparation of this manuscript were carried out without receiving any external financial assistance, grant, or funding support.

**Conflicts of Interest Statement:** The authors confirm that there are no conflicts of interest associated with this research or the publication of this study.

**Ethics and Consent Statement:** The study was performed in compliance with established ethical standards and received approval from the appropriate institutional review board. Written informed consent was obtained from all participants before they participated in the study.

## References

1. J. Chen, Y. Chen, J. Li, J. Wang, Z. Lin, and A. K. Nandi, "Stroke Risk Prediction with Hybrid Deep Transfer Learning Framework," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 1, pp. 411-422, 2021.
2. T. Nantinda, A. M. Gamundani, and M. N. Kanyama, "Machine-Learning Application for Detection and Prediction of Stroke Risk and Post-Stroke Outcomes: A Comprehensive Review," in *2024 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems,* Port Louis, Mauritius, 2024.
3. S. Nayak and N. Gupta, "Enhancing Stroke Prediction with Machine Learning in Smart Healthcare Systems," in *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, Kamand, India, 2024.

4. A. Suresh, R. S. Abhishek, P. Akash, B. Anudev, and K. H. Vishakh, "Functional Electrical Stimulation with Machine Learning for Stroke Prediction and Rehabilitation," in *2024 International Conference on E-mobility, Power Control and Smart Systems (ICEMPS)*, Thiruvananthapuram, India, 2024.
5. S. S. Shetty, N. Dsouza, S. Adyanthaya, S. Shetty, and P. S. Shetty, "Predictive model of early ischemic stroke in diabetic patients," in *2024 IEEE International Conference on Distributed Computing, VLSI, Electrical Circuits and Robotics (DISCOVER),* Mangalore, Karnataka, India, 2024.
6. S. Nithyaselvakumari, "Machine learning-based predictive model for ischemic stroke risk assessment: A data-driven approach using XGBoost," in *2025 9th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, Coimbatore, Tamil Nadu, India, 2025.
7. H. Nieto-Chaupis, "Stochastic Hybrid Algorithms to Estimate Stroke in Diabetic Patients," in *2023 International Conference on Electrical, Communication and Computer Engineering (ICECCE)*, Dubai, United Arab Emirates, 2023.
8. R. Yashvanth, M. Rehan, M. Kodipalli, R. B. Rohini, and T. Rao, "Diabetes, Hypertension, and Stroke Prediction through Computational Algorithms," in *2023 World Conference on Communication & Computing (WCONF),* Raipur, India, 2023.
9. M. P. William, S. B. Kiruba, M. Sakthivanitha, E. S. Soji, and G. Arun, "Machine learning to discover cardiovascular disease onset and key contributors: Data-driven personalized healthcare and preventive strategy," *AVE Trends in Intelligent Health Letters*, vol. 1, no. 3, pp. 137–157, 2024.
10. P. P. Shinde, V. P. Desai, G. S. Mirajkar, and K. S. Oza, "Towards Early Stroke Prediction: Detecting Hidden Patterns with Data Analytics" in *2025 4th International Conference on Sentiment Analysis and Deep Learning (ICSADL),* Bhimdatta, Nepal, 2025.
11. S. K. Sehrawat, "Emerging trends in healthcare transformation through cutting-edge technologies," *AVE Trends in Intelligent Health Letters*, vol. 1, no. 3, pp. 168–177, 2024.
12. J. Heo, Y. Yoon, H. Park, Y. D. Kim, H. S. Nam, and J. H. Neo "Machine learning–based model for prediction of outcomes in acute stroke" *Stroke*, vol. 50, no. 5, pp. 1263- 65, 2019.
13. R. Regin, Shynu, S. S. Rajest, M. Bhattacharya, D. Datta, and S. S. Priscila, "Development of predictive model of diabetic using supervised machine learning classification algorithm of ensemble voting," *International Journal of Bioinformatics Research and Applications*, vol. 19, no. 3, pp. 151–169, 2023.
14. D. Chicco and G. Jurman, "The Advantages of the Matthews Correlation Coefficient (MCC) over F1 Score and Accuracy in Binary Classification Evaluation," *BMC Genomics*, vol. 21, no. 1, p. 6, 2020.
15. L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 10, pp. 5–32, 2001.
16. S. Sengupta, D. Datta, S. S. Rajest, P. Paramasivan, T. Shynu, and R. Regin, "Development of rough-TOPSIS algorithm as hybrid MCDM and its implementation to predict diabetes," *International Journal of Bioinformatics Research and Applications*, vol. 19, no. 4, pp. 252–279, 2023.
17. A. Géron, "Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow," 2nd ed., *O'Reilly Media*, Sebastopol, California, United States of America. 2019.