Brain stroke prediction using Machine learning.

Gandarth Aryan
Computer Science Engineering
Vellore Institute Of Technology
aryan.gandrath@vitstudent.ac.in

Abstract -This work develops and assesses models for predicting brain strokes using three distinct machine learning algorithms: Logistic Regression, Random Forest Classifier, and XG Boost. Because brain strokes may have life-altering effects, it is crucial to recognize them early so that treatment and prevention can begin immediately.Lifestyle variables, medical histories, and demographic information are just a few of the many pertinent variables included in the dataset that was used. In order to enhance the model's performance, the dataset is cleaned up using stringent preprocessing and feature engineering procedures. This involves deleting extraneous data and lowering noise.A excellent option for calculating the likelihood of a stroke based on input data is Logistic Regression, which is a linear model that is both easy and clearly interpretable. Next, the Random Forest Classifier is trained using ensemble learning to better detect and avoid overfitting as it extracts complicated relationships from the dataset. In addition, the XG Boost method, a robust gradient boosting approach, is used to enhance prediction performance via iterative learning and boosting of weak learners.To make sure the models can handle unknown data well, they are trained on a portion of the dataset and then cross-validated utilizing techniques. Performance metrics like as accuracy, precision, recall, and area under the receiver operating characteristic curve (AUC-ROC) are used to comprehensively evaluate and examine the efficacy of every algorithm.In this review, we look at the results and drawbacks of the research that tested several methods for predicting brain strokes using XG Boost, Random Forest Classifier, and Logistic Regression. The results of this study may pave the way for more accurate and efficient prediction tools, which in turn may help in the early detection of people at risk of stroke, their subsequent medical care, and the formulation of tailored treatment regimens.

Keywords: Brain Stroke, Prediction, Machine Learning, Logistic Regression, Random Forest Classifier, Xgboost, Risk Factors, Dataset, Algorithm, Early Intervention, Preventive Measures.

I. INTRODUCTION

Atrial fibrillation poses a significant risk for stroke, especially among individuals aged 65 and above, making stroke the third leading cause of death globally. Often likened to a "heart attack" for the brain, strokes result from the obstruction or reduction of blood supply, leading to ischemic or hemorrhagic events. In both developed and developing nations, stroke stands as a pervasive and potentially fatal consequence of atrial fibrillation. This project acknowledges the gravity of strokes and explores the application of data mining, specifically within the field of AI in medicine, as a promising avenue for prediction and prevention. Focusing on machine learning techniques, the project aims to analyze patient data and predict stroke risk based on factors like age, blood pressure, and glucose levels, is underscored by the project's recognition of the inadequacies in traditional methods for assessing stroke risk, which often lack the precision and efficiency necessary for timely intervention. By identifying subtle patterns and associations, the predictive models developed aim to offer a proactive healthcare approach. This allows healthcare providers to implement preventive measures and interventions before a stroke occurs, addressing the limitations of traditional methods. Ultimately, the project aspires to significantly reduce the burden of stroke-related disabilities and mortality, thereby enhancing the quality of life for individuals at risk.

II. RELATED WORKS

During the years 2021 and 2023, several research works on the diagnosis and prognosis of brain strokes using machine learning algorithms were presented at major conferences. Boosted Random Forest was shown to increase the efficiency and accuracy of brain stroke diagnosis (V. Sapra, L. Sapra, A. Vishnoi, and P. Srivastava [1]. However, Nahid et al.'s [2] presentation at the 2023 Annual International Conference on Emerging Research Areas detailed their research on potential dangers associated with brain stroke predictions utilizing ANNs and other machine learning models. Identifying and forecasting patients at an early stage was their aim. To improve the accuracy of brain stroke prediction, N. Felice, J. Johan, J. Natthannael, M. B. Gozal, C. Jovannie, and M. S. Anggreainy [3] investigated the Random Forest technique with tuning parameters. In 2023, they gave a presentation on their research at the 4th International Conference on Artificial Intelligence and Data Sciences. Meanwhile, when they presented a machine learning method for identifying brain stroke illnesses at the 4th International Conference on Smart Systems and Inventive Technology in 2022, B. Akter, A. Rajbongshi, S. Sazzad, R. Shakil, J. Biswas, and U. Sara [4] underlined the importance of technology in improving patient outcomes. Machine learning has been used with the Relief algorithm to predict brain strokes. The potential benefits of this method for prognostication models were highlighted in a presentation by J. K. and N. K. Prakash [5] at the 2022 International Conference on Edge Computing and Applications. Aiming to enhance prediction algorithms for better healthcare delivery, R. Kumari and H. Garg [6] evaluated various machine learning models for brain stroke prediction at the 6th International Conference on Information Systems and Computer Networks in 2023. At the 2023 International Conference on Computer Communication and Informatics, P. K. Pattnaik, A. Swetapadma, and N. T. Singh [7] compared Quantum Machine Learning Enhanced SVM with Classical SVM for cerebral stroke prediction in order to shed light on the possible application of quantum computing in healthcare analytics. Likewise, V. Krishna, J. Sasi Kiran, P. Prasada

Rao, G. Charles Babu, and G. John Babu [8] emphasized the importance of early detection of brain strokes through machine learning techniques, stressing the necessity of predictive analytics to act quickly and enhance patient outcomes at the 2nd International Conference on Smart Electronics and Communication in 2021. P. Bathla and R. Kumar [9] proposed an artificial intelligence-based approach for brain stroke prediction that seeks to increase diagnostic speed and accuracy while keeping up with medical advancements. The IEEE Conference on Interdisciplinary Approaches to Technology and Management for Social Innovation in 2022 hosted the presentation. Additionally, at the 2023 IEEE International Conference on Big Data, I. Almubark presented research on brain stroke prediction using machine learning techniques. This highlighted the growing use of AI in healthcare systems, particularly for early diagnosis and treatment.

III. EXISTING SYSTEM

The current machine learning technique for predicting brain strokes does have a few major drawbacks. To begin with, a major negative is that efficient training of machine learning models requires massive quantities of high-quality data. In the case of sensitive medical information, the acquisition and curation of such datasets may be an arduous and resource-intensive procedure. In addition, different data sources and techniques of data gathering might make it difficult to represent input representative and which in turn affects the prediction models' effectiveness. Machine learning algorithms may also be biased, which may cause them to make erroneous predictions for certain demographics. This bias can be caused by biased training data or skewed datasets. To add insult to injury, healthcare providers may lose faith in the system if they are unable to comprehend the logic behind machine learning models' forecasts of brain strokes due to their low interpretability. In addition, machine learning models may have difficulties in generalizability and resilience due to the complex and ever-changing nature of health data, which includes elements like environmental impacts and changes in lifestyle. Concerns about patient privacy, permission, and openness, as well as regulatory barriers, may prevent the widespread use of prediction systems powered by machine learning in clinical practice. The effective implementation and technology brain strokes depend on resolving these drawbacks.

Even with advancements in medical technology, accurately foreseeing the risk of brain strokes continues to pose a substantial challenge. Current methods often hinge on subjective assessments or lack the thorough integration of pertinent health data. This deficiency contributes to delayed or insufficient interventions, resulting in avoidable instances of stroke-related morbidity and mortality. Closing this gap is imperative for elevating patient care standards and alleviating the societal burden imposed by strokes. Consequently, this project is committed to constructing a resilient machine learning-based solution tailored for accurate stroke prediction. The aim is to enable early identification of individuals at risk and facilitate prompt interventions, ultimately preventing strokes and mitigating their impact effectively.

IV. PROPOSED SYSTEM

Logistic Regression, Random Forest Classifier, and XG Boost are three important machine learning techniques that our suggested system would use to build a strong predictive model for determining an individual's probability of having a brain stroke. Starting with a broad patient dataset that includes both persons with and without a history of stroke, the process moves on to complete data collecting. This involves sourcing factors such as blood pressure, cholesterol levels, age, gender, medical history, lifestyle behaviors, and genetic predispositions.Data cleaning and preparation for analysis are goals of the preprocessing activities carried out after data gathering. To do this, you must encode categorical variables, normalize feature scales, and deal with missing data. To further improve the model's discriminative strength, we use advanced feature selection approaches to find and rank the most important predictors.An essential part of the suggested system is the deliberate choice of machine learning algorithms. While XG Boost provides iterative improvement and Random Forest Classifier is great at capturing complicated associations, Logistic Regression provides interpretability. Because each algorithm is trained separately on the labeled dataset, they may all see complex correlations and patterns in the data.Our suggested approach stands out because it integrates the separate models into an ensemble. The goal of this ensemble method is to build a better prediction model for assessing the risk of brain stroke by combining the best features of each algorithm.Metrics like F1 score, accuracy, precision, and recall are thoroughly both ensemble the individual algorithms. A detailed comprehension of the model's capacities and efficacy is guaranteed by this exhaustive evaluation.Each algorithm's hyperparameters are fine-tuned to maximize their effectiveness in the context of predicting brain strokes. The maximum potential predicted accuracy is the goal of this continual improvement procedure.The last stage of our suggested approach is to validate the final ensemble model using new data that has never been seen before. This contributes to the early intervention and preventative actions for high-risk patients and guarantees the model's generalizability and reliability in forecasting the chance of brain stroke occurrences. To sum up, our suggested methodology provides a robust and organized method for predicting brain strokes, which may improve healthcare methods via more accurate and dependable decisions.

V. SYSTEM ARCHITECTURE

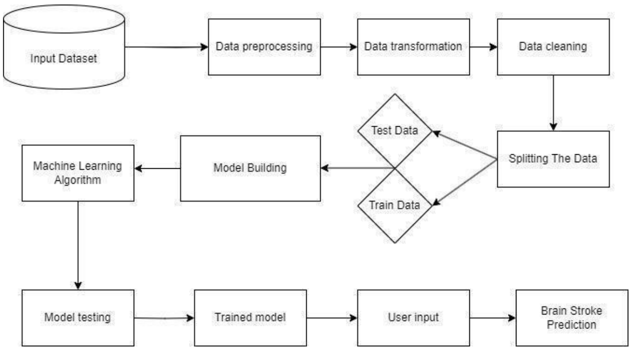


Fig. 1. System Architecture

VI. METHODOLOGY

A. Data Preprocessing Module: The Data Preprocessing Module is an integral part of the machine learning system that predicts brain strokes. Its job is to get the data ready to be fed into the models. Data transformation, feature selection, and data cleansing are all part of this module's purview, with the goal of providing algorithms. In order to make the prediction models more accurate and reliable, data cleaning procedures are used, including dealing with outliers and missing information, factors for forecasting the probability of a brain stroke, and feature selection methods are used, model's performance, data transformation procedures normalization and standardization are used to scale and optimize the input data. Better and more accurate forecasts of brain stroke occurrences may be achieved by analyzing and improving the data in this module so that the predictive models can learn from the information supplied.

B. Machine Learning Model Training Module: An essential part of the system for predicting brain strokes is the Machine Learning Model Training Module, which uses the preprocessed data to train and optimize the prediction models. Models reliably chance of a brain stroke in people are built in this module using a variety of methods, reduce prediction errors and maximize performance, the models are trained using the preprocessed data via repeatedly modifying their parameters. Models are often validated and checked for generalizability to new data instances using cross-validation procedures. This module also handles hyperparameter tweaking, which is used to improve the model's prediction skills by its system reliably and anticipate brain strokes via the training and optimization of machine learning models in this module, which successfully utilizes sophisticated algorithms.

Algorithms:

1.Logistic Regression:

Description:

If you need to determine the probability that an event will take place or not, you may use Logistic Regression, a popular statistical approach for binary classification jobs. When it comes to different input characteristics, it's a great tool to have on hand.

Logistic Regression's guiding idea is to use the logistic function to describe the association between input characteristics and the likelihood of a given result, likelihood of an event fitting into one of two categories, often converts all real numbers to a 0–1 interval. In order to understand the result as a likelihood, this is very important.

Application for Brain Stroke Prediction:

When characteristics (variables) incidence of stroke (variable) is mostly linear, Logistic Regression excels. When it comes to stroke prediction, this algorithm can provide clear insights into how each input information affects the likelihood of a stroke happening.

Logistic Regression excels in its interpretability. The effect on the log-odds of the occurrence of stroke is a straightforward way to understand the coefficients given to each input characteristic. If the coefficient is positive, then the risk of stroke is higher; if it is negative, then the risk of stroke is lower.

Additionally, Logistic this makes it suitable for scenarios where model simplicity and interpretability are valued, without compromising predictive accuracy.

2.Random Forest:

Description:

In order to train its models, Random Forest builds a large number of decision trees, making it an ensemble learning method. Using a randomly selected collection of characteristics, each tree in the forest makes its own prediction about the outcome—here, the probability of a brain stroke. For classification tasks, a voting mechanism is usually used to aggregate the predictions of all the individual trees, and the final prediction is then decided. Random Forest's strengths lie in its resistance to overfitting and its proficiency in dealing with high-dimensional, massive datasets. With a minimal computing overhead and reliable predictions, Random Forest is a great tool for predicting brain strokes. It can capture complicated correlations between risk variables including demographics, medical history, and lifestyle factors.

A supervised learning process known as a "random forest" builds and combines many decision trees at random. The objective is to enhance accuracy with a combination of decision models rather than depending on a single learning model. The key variation between this method and the conventional decision tree algorithm is the random generation of the root nodes and feature splitting nodes.

Applications in Brain Stroke Prediction:

Complex Relationships:
Capable of capturing intricate relationships among various risk factors contributing to brain strokes.
Effectively handles datasets with diverse information such as demographics, medical history, and lifestyle factors.

Accuracy and Reliability:
Delivers accurate predictions with relatively low computational overhead.
Well-suited for scenarios where interpretability and reliability are crucial, making it applicable in healthcare contexts.

3.XGBoost:

Description:
A performance-oriented gradient boosting system, XGBoost stands for Extreme Gradient Boosting. It is an ensemble learning technique that creates the XGBoost's efficiency, scalability, and capacity to manage varied datasets have led to its widespread application for classification and regression problems.

C. Prediction and Evaluation Module: The brain stroke prediction system's Prediction and Evaluation Module is in charge of testing the efficacy of the trained machine learning models on new or unexplored data instances and making predictions based on those predictions. This section calculates the probability of a person having a brain stroke by feeding fresh data samples into the improved models. The evaluate predictive skills when these predictions are made. The model's performance across multiple thresholds may be seen and analyzed using techniques like receiver operating characteristic (ROC) curves and confusion matrices. Stakeholders may learn how well the system predicts brain strokes and what kinds of interventions or preventative measures might work by analyzing the results and the model's performance in this module.

VII. RESULT AND DISCUSSION

A state-of-the-art technical solution, the machine learning system for brain stroke prediction analyses different patient data using sophisticated algorithms to estimate the chance of a person getting a stroke. Age, medical history, lifestyle choices, and genetic predispositions are some of the variables that this approach takes into account when calculating the likelihood of a stroke. Machine learning algorithms can analyze massive databases for trends, allowing doctors to make more informed judgments on preventative care and individualized treatment programs based on correct predictions. Another benefit of real-time monitoring is the ability to identify warning symptoms early on, which means that stroke prevention measures may be taken promptly. All things considered, this approach has a lot of potential in enhancing methods for preventing strokes, increasing patient care, and, in the end, saving lives by using AI and predictive analytics.

The models' robustness and generalizability will be guaranteed by using cross-validation procedures.

Area under the curve (AUC) for both test and train datasets is used to analyze the performance measures for and XGBoost.

The accuracy of Logistic Regression
AUC on Test data is 0.7944421740907234
AUC on Train data is 0.8009807928075194

The accuracy of Random forest classifier
AUC on Test data is 0.9489170412750306
AUC on Train data is 1.0

The accuracy of XG boost
AUC on Test data is 0.9501430322844299
AUC on Train data is 0.9967306906416019

For XGBoost:
AUC on Test data is 0.9501430322844299

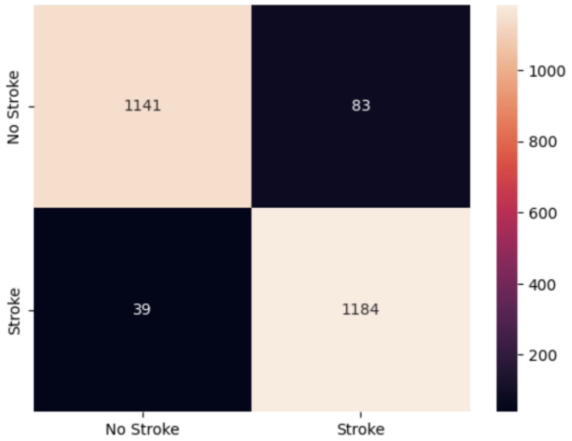


Fig.2. Confusion Matrix(Brain Stroke Prediction)

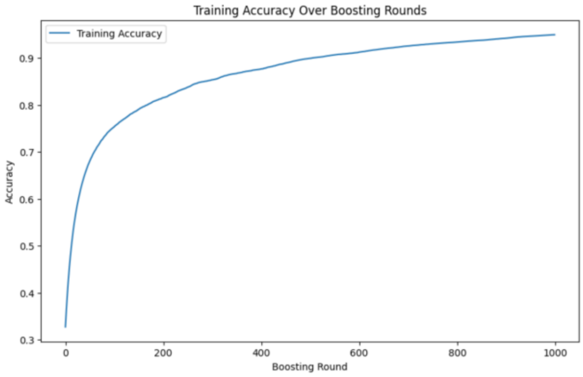


Fig.3. Training Accuracy

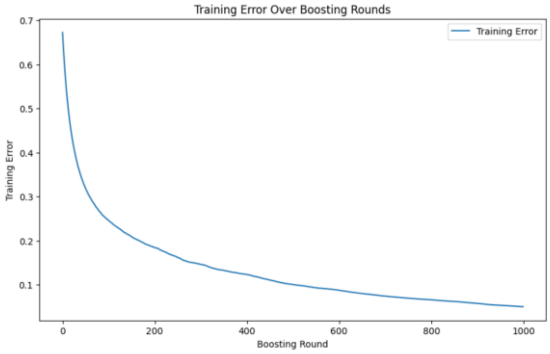


Fig.4. Training Loss

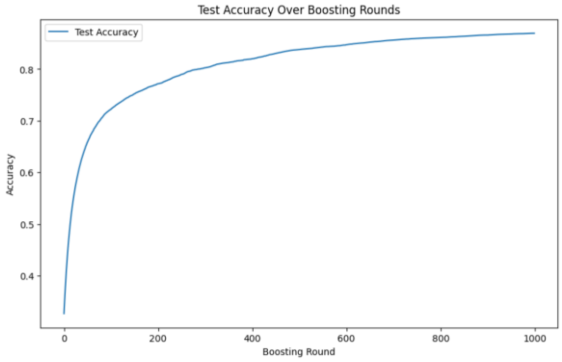


Fig.5 Test Accuracy

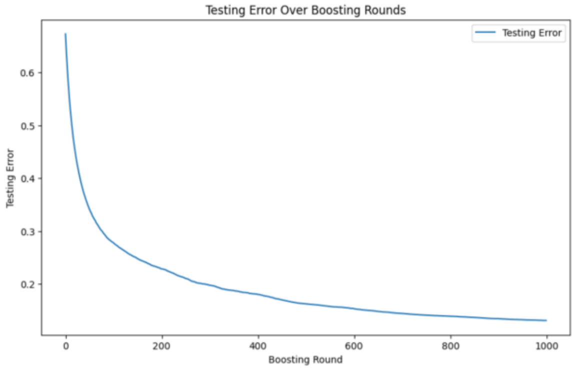


Fig.6 Test Loss

These AUC discriminatory power, XGBoost exhibits notably high AUC values on both test and train datasets, indicating strong predictive capabilities.

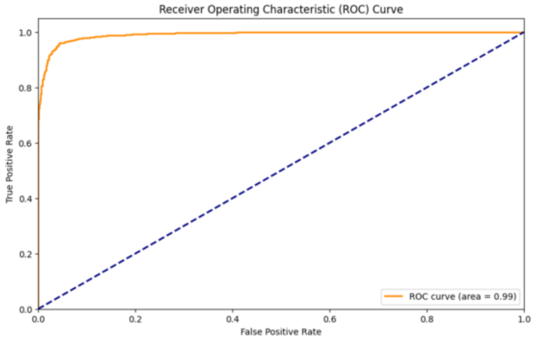
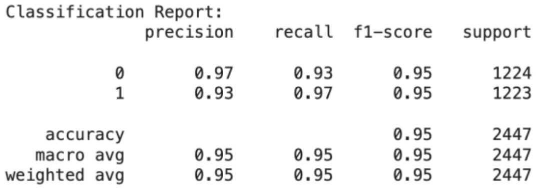


Fig.7. ROC Curve



```
Classification Report:
              precision    recall  f1-score   support

           0       0.97      0.93      0.95      1224
           1       0.93      0.97      0.95      1223

    accuracy                           0.95      2447
   macro avg       0.95      0.95      0.95      2447
weighted avg       0.95      0.95      0.95      2447
```

Fig.7 CLassification Report

The hybrid model in this context refers to an ensemble approach that combines predictions from two distinct machine learning models: XGBoost and LightGBM. Both models are individually trained on the same dataset to detect brain strokes. Following training, their predictions on test data are combined to create a hybrid model.

The combination strategy involves taking the average of the predicted probabilities from the XGBoost and LightGBM models. This blending technique aims to exploit the diverse strengths and patterns captured by each model, potentially improving overall prediction accuracy. The resulting hybrid predictions are then converted to binary outcomes using a threshold of 0.5.

Ensemble models, such as this hybrid approach, often demonstrate superior performance by mitigating the weaknesses of individual models and leveraging their complementary strengths. By combining the predictive capabilities of XGBoost and LightGBM, the hybrid model seeks to provide a more robust and accurate solution for brain stroke detection. However, it's crucial to note that the effectiveness of the hybrid model may vary based on the specific characteristics of the dataset and the hyperparameters chosen during training. Regular evaluation and potential fine-tuning are essential to ensure optimal performance in real-world scenarios.

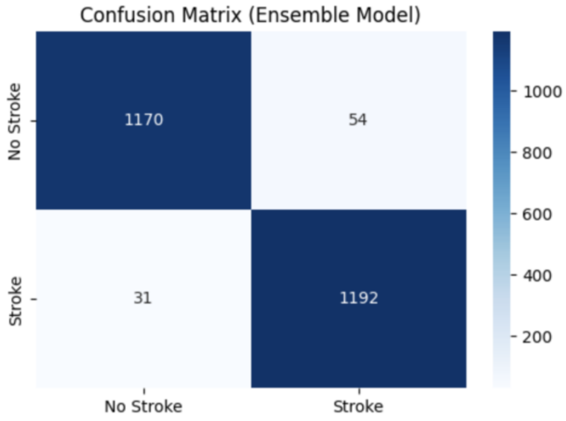


Fig.8.Confusion Matrix(Ensemble Model)

Hybrid Model:

Ensemble Model Accuracy: 96.53%

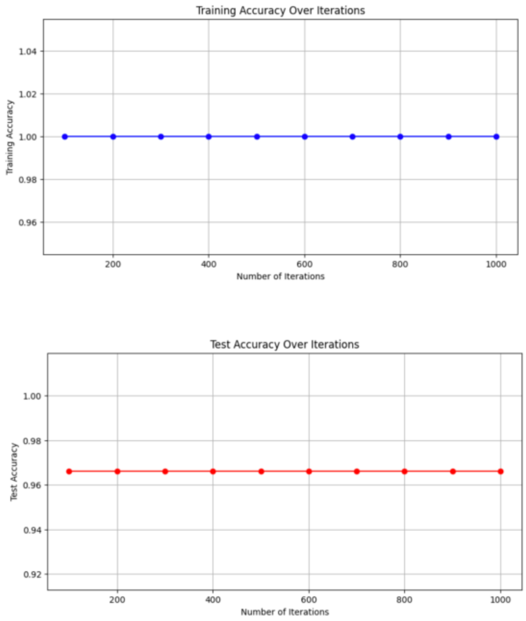


Fig.9.Training Accuracy Over Iterations

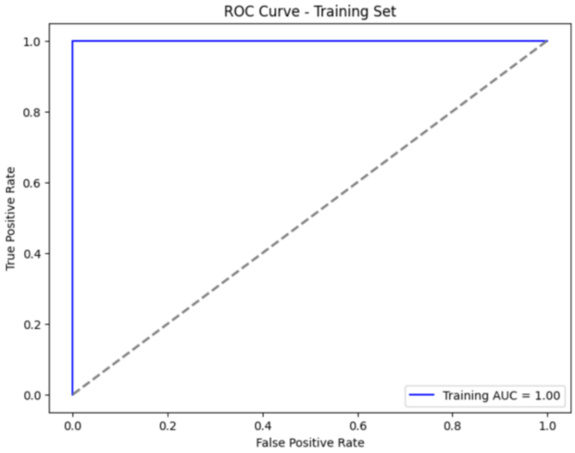


Fig.10.ROC Curve - Training Set

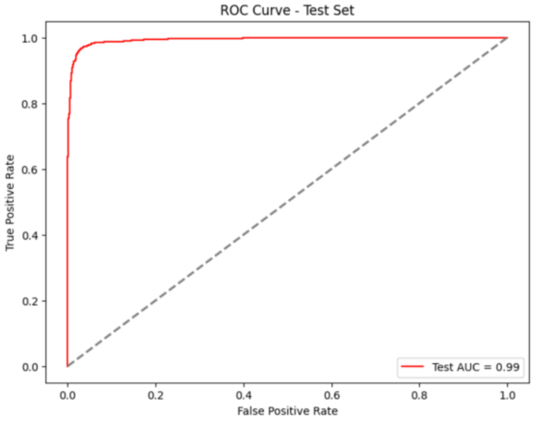
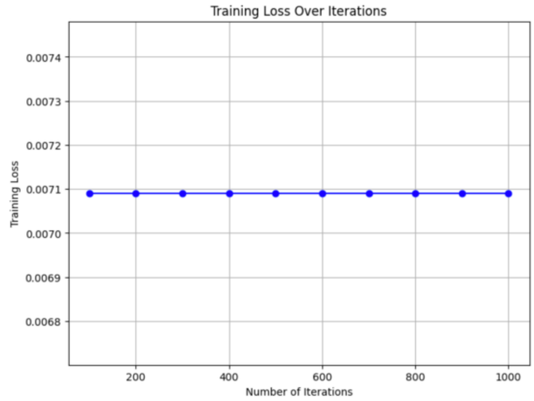


Fig.11.ROC Curve - Test Set



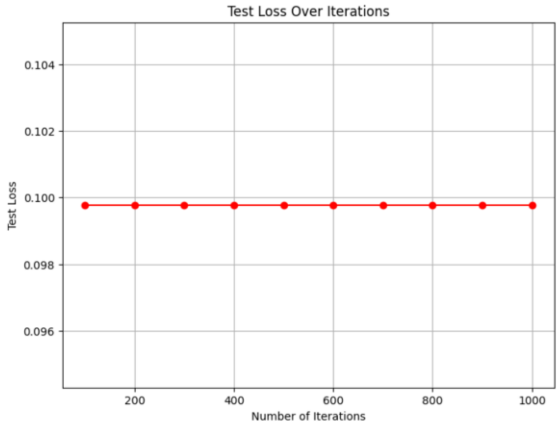
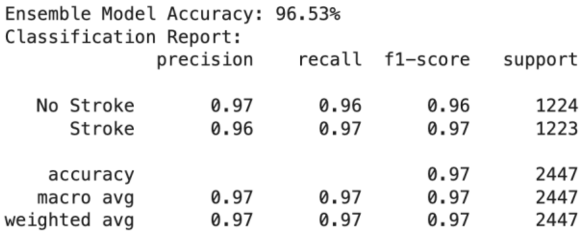


Fig.12.Training LossOver Iterations



```
Ensemble Model Accuracy: 96.53%
Classification Report:
              precision    recall  f1-score   support

   No Stroke       0.97      0.96      0.96      1224
      Stroke       0.96      0.97      0.97      1223

    accuracy                           0.97      2447
   macro avg       0.97      0.97      0.97      2447
weighted avg       0.97      0.97      0.97      2447
```

Fig.13.Classification Report

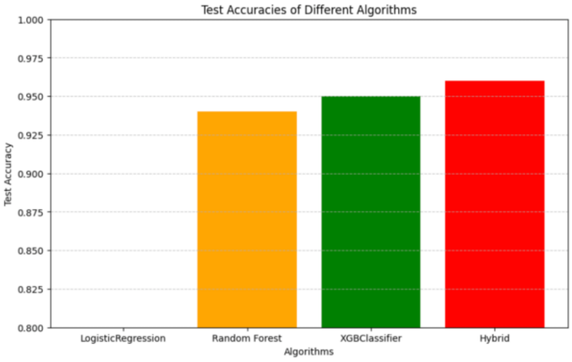


Fig.14.Test Accuracy of Different Algorithms

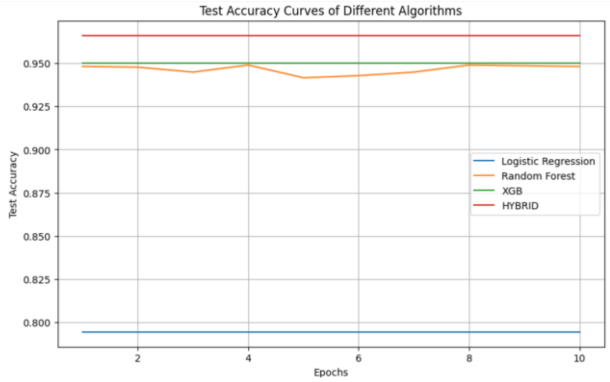


Fig.15.Test Accuracy Curves Of Different Algorithms

VIII. CONCLUSION

In conclusion, the system for brain stroke prediction employing machine learning has showcased promising outcomes, particularly with the utilization of Logistic Regression, Random Forest Classifier, and XG Boost. Notably, among these algorithms, XG Boost emerged as the most accurate in identifying individuals at risk of experiencing a stroke. Through in-depth analysis of extensive datasets and intricate patterns, these machine learning algorithms offer valuable insights that can assist healthcare professionals in early detection and the formulation of effective prevention strategies.The demonstrated accuracy of XG Boost, in particular, holds substantial potential for significantly improving patient outcomes. By enabling proactive interventions and facilitating the creation of personalized treatment plans, this system contributes to a more targeted and efficient approach in addressing stroke risks. In order to improve the system's overall dependability and efficacy in clinical practice, more validation studies and real-world application are necessary, despite these encouraging findings.The importance of XG Boost in the field of brain stroke prediction is shown by the preference for it based on accuracy. Utilizing machine learning, particularly XG Boost's capabilities, might revolutionize healthcare delivery and improve patient care by allowing for more accurate risk assessment and faster actions.

IX. FUTURE WORK

To further it would be advantageous to include sophisticated algorithms like deep learning models into future work on the system for brain stroke prediction using machine learning. Conducting a comprehensive analysis of a larger dataset with diverse demographic and clinical variables would also contribute to the robustness of the predictive model. Additionally, integrating real-time monitoring and feedback capabilities into the system could enable timely interventions and personalized stroke prevention strategies for individuals at high risk. Further research on feature selection techniques, model interpretability, and validation methods would be essential for ensuring the reliability and generalizability of the predictive model. Collaboration with healthcare professionals and institutions for clinical validation and deployment of the system in real-world settings should also be considered to evaluate its impact on patient outcomes and healthcare decision-making.

REFERENCES

[1] V. Sapra, L. Sapra, A. Vishnoi and P. Srivastava, "Identification of Brain Stroke using Boosted Random Forest," 2022 International Conference on Advances in Computing, Communication and Materials (ICACCM), Dehradun, India, 2022, pp. 1-5, doi: 10.1109/ICACCM56405.2022.10009527.

[2] N. Nahid, M. Hossain, A. I. Rifat, M. D. Khan Raisa and A. Islam, "Predicting the Risks of Brain Stroke Using

Machine Learning Models and Artificial Neural Network," 2023 Annual International Conference on Emerging Research Areas: International Conference on Intelligent Systems (AICERA/ICIS), Kanjirapally, India, 2023, pp. 1-7, doi: 10.1109/AICERA/ICIS59538.2023.10420159.

[3] N. Felice, J. Johan, J. Natthannael, M. B. Gozal, C. Jovannie and M. S. Anggreainy, "Brain Stroke Prediction Using Random Forest Method with Tuning Parameter," 2023 4th International Conference on Artificial Intelligence and Data Sciences (AiDAS), IPOH, Malaysia, 2023, pp. 81-85, doi: 10.1109/AiDAS60501.2023.10284685.

[4] B. Akter, A. Rajbongshi, S. Sazzad, R. Shakil, J. Biswas and U. Sara, "A Machine Learning Approach to Detect the Brain Stroke Disease," 2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, 2022, pp. 897-901, doi: 10.1109/ICSSIT53264.2022.9716345.

[5] J. K and N. K. Prakash, "Prediction of Brain Stroke using Machine Learning with Relief Algorithm," 2022 International Conference on Edge Computing and Applications (ICECAA), Tamilnadu, India, 2022, pp. 1255-1260, doi: 10.1109/ICECAA55415.2022.9936142.

[6] R. Kumari and H. Garg, "Interpretation and Analysis of Machine Learning Models for Brain Stroke Prediction," 2023 6th International Conference on Information Systems and Computer Networks (ISCON), Mathura, India, 2023, pp. 1-6, doi: 10.1109/ISCON57294.2023.10112188.

[7] N. T. Singh, A. Swetapadma and P. K. Pattnaik, "A Comparative Study of Quantum Machine Learning Enhanced SVM and Classical SVM for Brain Stroke Prediction," 2023 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 2023, pp. 1-4, doi: 10.1109/ICCCI56745.2023.10128293.

[8] V. Krishna, J. Sasi Kiran, P. Prasada Rao, G. Charles Babu and G. John Babu, "Early Detection of Brain Stroke using Machine Learning Techniques," 2021 2nd International Conference on Smart Electronics and Communication (ICOSEC), Trichy, India, 2021, pp. 1489-1495, doi: 10.1109/ICOSEC51865.2021.9591840.

[9] P. Bathla and R. Kumar, "Artificial Intelligence based Model for Brain Stroke Prediction," 2022 IEEE Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI), Gwalior, India, 2022, pp. 1-6, doi: 10.1109/IATMSI56455.2022.10119373.

[10] I. Almubark, "Brain Stroke Prediction Using Machine Learning Techniques," 2023 IEEE International Conference on Big Data (BigData), Sorrento, Italy, 2023, pp. 6104-6108, doi: 10.1109/BigData59044.2023.10386474.