Fusion of XGBOOST and Random Forest for Accurate Stroke Prediction

1st Md Saif Ali
Computer science and engineering
Gurunanak institutions technical campus
Ibrahimpatnam,Telangana
mohammadsaifali87@gmail.com

2nd Mamidipally Uday Kiran
Computer science and engineering
Gurunanak institutions technical campus
Ibrahimpatnam,Telangana
udaykiranmamidipally@gmail.com

3rd Lakkampelly Adithya
Computer science and engineering
Gurunanak institutions technical campus
Ibrahimpatnam,Telangana
adithya4446@gmail.com

4th Ms.Rajashree Sutrawe
Computer science and engineering
Gurunanak institutions technical campus
Ibrahimpatnam,Telangana
raj.sutrawe@gniindia.org

***Abstract*—Stroke is a life-threatening medical condition caused by disrupted blood flow to the brain, representing a major global health concern with significant health and economic consequences. Researchers are working to tackle this challenge by developing automated stroke prediction algorithms, which can enable timely interventions and potentially save lives. As the global population ages, the risk of stroke increases, making the need for accurate and reliable prediction systems more critical. In this study, we evaluate the performance of an advanced machine learning (ML) approach, focusing on XGBoost and a hybrid model combining XGBoost with Random Forest, by comparing it against six established classifiers. We assess the models based on their generalization ability and prediction accuracy. The results show that more complex models outperform simpler ones, with the best-performing model achieving an accuracy of 96%, while other models range from 84% to 96%. Additionally, the proposed framework integrates both global and local explainability techniques, providing a standardized method for interpreting complex models. This approach enhances the understanding of decision-making processes, which is essential for improving stroke care and treatment. Finally, we suggest expanding the model to a web-based platform for stroke detection, extending its potential impact on public health.**

Keywords– Model Fusion, Feature Selection, Predictive Modeling, Supervised Learning, Data Preprocessing

I. INTRODUCTION

The incidence of stroke has been increasing globally, and it is now considered one of the leading causes of death and dis-ability. Early intervention is crucial in preventing long-term disability and mortality associated with stroke. Traditional methods of predicting stroke risk, however, are often time-consuming and prone to errors. Additionally, there is a growing need for transparency and explainability in machine learning models in healthcare. The use of an interpretable machine learning model can provide clinicians with valuable insights into the factors that contribute to a patient's stroke risk, thereby aiding in treatment decisions.

The World Stroke Organization estimates that 13 million people worldwide experience a stroke each year, leading to 5.5 million fatalities. Stroke affects all aspects of a patient's life, including their family, social environment, and work, and is one of the top causes of mortality and disability in the world. A common misconception is that certain groups of people, such as the elderly or those with underlying illnesses, are the only ones who are affected by stroke.

Recently, machine learning algorithms have shown great promise in accurately predicting stroke risk based on various clinical risk factors. By leveraging these algorithms, clinicians can identify high-risk patients and intervene early, potentially reducing the number of stroke-related complications and improving patient outcomes.

II. EXISTING SYSTEM

A Convolutional Neural Network (CNN) is a type of deep learning model designed primarily for processing structured grid data, such as images. CNNs are especially well-suited for tasks like image classification, object detection, and even speech recognition. The core idea behind CNNs is their ability to automatically and hierarchically extract features from raw input data. convolutional layers, which apply filters to detect low-level features like edges or textures activation functions, typically ReLU, which introduce nonlinearity pooling layers, which reduce spatial dimensions to make the model computationally efficient; and fully connected layers, which interpret the extracted features to perform tasks like classification. By using these layers, CNNs can effectively detect increasingly complex features as the input data progresses through the layers of the network. One of the biggest advantages of CNNs is their ability to automatically learn relevant features from raw data, which eliminates the need for manual feature engineering, a common requirement in traditional machine learning algorithms. This ability to detect patterns in data also gives CNNs the advantage of translation invariance, meaning they can recognize objects even if their position in the image changes.

DRAWBACKS:

- CNNs are computationally expensive to train and deploy.
- Large Dataset Requirement.

• Memory and Storage Requirements

III. PROPOSED SYSTEM

The main component of our suggested method is making use of machine learning, more especially XGBoost and its combination with Random Forest. We seek to evaluate these methods' effectiveness by a thorough comparison study with six well-known classifiers. More complex models work better, as seen by the experimental findings, where the best-performing model achieves an astounding 96% accuracy, while other models continuously fall between 84 and 96% accuracy.

The fusion of XGBoost and Random Forest for stroke prediction offers several significant advantages. By combining the strengths of both models, the fused approach enhances overall prediction accuracy and reliability. Random Forest, known for reducing variance through bagging, and XGBoost, which effectively minimizes bias using boosting, together provide a balanced model that performs well on both training and unseen data. This balance leads to improved generalization and makes the system more robust against overfitting and noise—commo n challenges in medical datasets. Additionally, the hybrid model can capture complex patterns in patient data, enabling better identification of risk factors associated with stroke. As a result, the fusion model serves as a powerful decision-support tool in healthcare, offering more accurate and dependable predictions that can assist in early diagnosis and timely intervention for stroke prevention.

ADVANTAGES:

• Delivering accurate stroke prediction findings, which are essential for prompt intervention and care.

• Helping to identify important health and self-reported status factors that predict stroke and enabling focused therapy.

• Improving the capacity to predict stroke overall and to generalize results based on health and personal status information.

IV. RELATED WORK

Machine learning has gained significant attention in the medical field for predicting diseases, including stroke. Various studies have explored different algorithms to improve diagnostic accuracy and support clinical decision-making. Random Forest and XGBoost are two widely used ensemble learning techniques that have shown promising results in healthcare analytics. Random Forest has been applied in numerous stroke prediction studies due to its robustness and ability to handle high-dimensional data. Researchers have demonstrated its effectiveness in identifying critical features and managing imbalanced datasets common in stroke-related cases. On the other hand, XGBoost, a gradient boosting framework, has gained popularity for its superior performance in structured data problems. Its ability to learn from errors and optimize model parameters has led to improved accuracy in stroke and cardiovascular disease prediction tasks. This project builds on existing research by integrating both XGBoost and Random Forest in a fusion framework to achieve higher prediction accuracy and reliability. By leveraging the complementary strengths of both models, the proposed approach aims to offer a more accurate, generalized, and clinically useful tool for early stroke detection

V. METHODOLOGIES

MODULE NAME:

1) Dataset:

The preparation of the dataset for stroke prediction using XGBoost and XGBoost with Random Forest was the main emphasis of the first module. The dataset was retrieved from a CSV file called "stroke-dataset.csv" that contained a variety of personal information about the subjects, such as their age, gender, and lifestyle and health problems. The "stroke" column, which has binary labels (0 for no stroke, 1 for stroke), serves as the prediction target variable.

2) Importing Necessary Libraries:

The Python programming language was used for the analysis and modelling. Importing key libraries from sklearn.ensemble included pandas, numPy, matplotlib, scikit-learn, XGBoost, and RandomForestClassifier. These libraries make it easier to manipulate data, visualize it, develop machine learning models, and assess them.

3) Data Pre-Processing:

A pandas DataFrame containing 5110 entries was filled with the dataset. The "bmi" column of the dataset included missing values, which were later imputed using the k-nearest neighbours (KNN) technique after an initial examination of the dataset indicated their existence. StandardScaler was used to standardize numerical characteristics and encode categorical information. Based on age, BMI, and average glucose level categories, new category characteristics were developed.

4) Splitting the Dataset:

To resolve class imbalance, the dataset was divided into training and test sets using the Synthetic Minority Over-sampling Technique (SMOTE). Eighty percent of the SMOTE oversampled data was in the training set, and twenty percent was in the test set.

5) Model Selection:

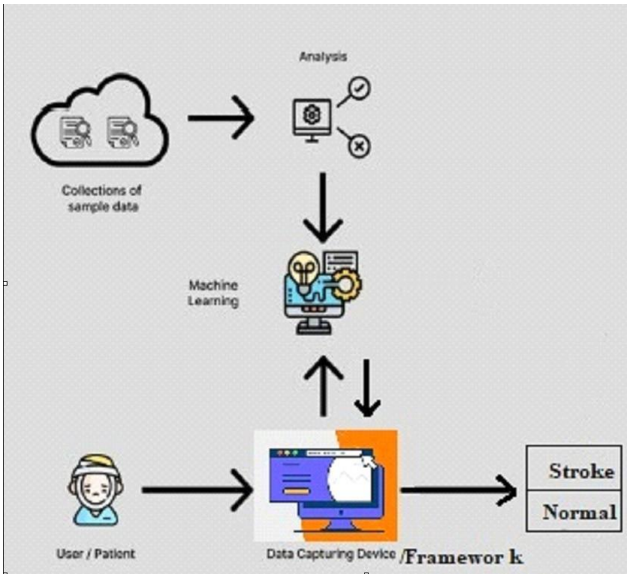
For stroke prediction, Random Forest (RF) and XGBoost (XGB) were two machine learning models that were put into practice. With an accuracy of 95%, the XGBoost model outperformed the Random Forest model, which came in at 81%. To assess the models' performance, metrics such ROC AUC score, F1-score, accuracy, and recall were computed. Furthermore, XGBoost with Random Forest (XGB_RF), a hybrid model, was created and produced with an accuracy of 84%. The oversampled training set of data was used to train each model.

6) Saving the Trained Models:

The pickle library was used to store the XGBoost and XGBoost with Random Forest models as ".pkl" files after they had been trained and assessed. Models that have been

saved can be later loaded and used in an environment that is ready for production.

VI. SYSTEM ARCHITECTURE



VII. CONCLUSION

In conclusion, our machine learning-based Stroke Prediction system functions as a useful backup to conventional diagnostic instruments in clinical settings for Computer-Aided Diagnosis. Although the AI models are quite good at predicting the future, the inclusion of Explainable Artificial Intelligence (XAI) approaches takes care of the important issue of decision-making transparency. Our method enables doctors to comprehend and have confidence in the system's outputs by offering interpretable explanations for predictions, such as emphasizing the influence of certain clinical parameters on confidence levels. This not only fosters a collaborative atmosphere where medical experts can critically review and improve the model's performance, but it also increases confidence in the decision support system. Furthermore, our diagnosis framework's perturbation-based explanation method shows promise for wider applications in a variety of medical domains, highlighting the potential contribution of explainability to the advancement of healthcare AI.

REFERENCES

[1] Learn About Stroke. Accessed: May 25, 2022. [Online]. Available: https://www.world-stroke.org/world-stroke-day-campaign/why-stroke-matters/learn-about-stroke

[2] T. Elloker and A. J. Rhoda, ‘‘The relationship between social support and participation in stroke: A systematic review,’’ Afr. J. Disability, vol. 7, pp. 1–9, Oct. 2018.

[3] M. Katan and A. Luft, ‘‘Global burden of stroke,’’ Seminar Neurol., vol. 38, no. 2, pp. 208–211, Apr. 2018.

[4] A. Bustamante, A. Penalba, C. Orset, L. Azurmendi, V. Llombart, A. Simats, E. Pecharroman, O. Ventura, M. Ribó, D. Vivien, J. C. Sanchez, and J. Montaner, ‘‘Blood biomarkers to differentiate ischemic and hemor-rhagic strokes,’’ Neurology, vol. 96, no. 15, pp. e1928–e1939, Apr. 2021.

[5] X. Xia, W. Yue, B. Chao, M. Li, L. Cao, L. Wang, Y. Shen, and X. Li, ‘‘Prevalence and risk factors of stroke in the elderly in northern China: Data from the national stroke screening survey,’’ J. Neurol., vol. 266, no. 6, pp. 1449–1458, Jun. 2019.

[6] A. Alloubani, A. Saleh, and I. Abdelhafiz, ‘‘Hypertension and diabetes mellitus as a predictive risk factors for stroke,’’ Diabetes Metabolic Syn-drome, Clin. Res. Rev., vol. 12, no. 4, pp. 577–584, Jul. 2018.

[7] A. K. Boehme, C. Esenwa, and M. S. V. Elkind, ‘‘Stroke risk factors, genet-ics, and prevention,’’ Circ. Res., vol. 120, no. 3, pp. 472–495, Feb. 2018.

[8] I. Mosley, M. Nicol, G. Donnan, I. Patrick, and H. Dewey, ‘‘Stroke symp-toms and the decision to call for an ambulance,’’ Stroke, vol. 38, no. 2, pp. 361–366, Feb. 2007.

[9] J. Lecouturier, M. J. Murtagh, R. G. Thomson, G. A. Ford, M. White, M. Eccles, and H. Rodgers, ‘‘Response to symptoms of stroke in the UK: A systematic review,’’ BMC Health Services Res., vol. 10, no. 1, pp. 1–9, Dec. 2010.

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