IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Stroke Prediction Using SHAP And SMOTE Techniques

Fowza Firoz Kasamji[1], Arshiya A.F.A Khan[2], Firdous F.A Khan[3], R. S. Deshpande[4]
[1] Student, [2] Head of Department,[3]Professor, [4]Dean
[1] School of Computational Sciences
[1] JSPM University, Wagholi, Pune, India

Abstract: Stroke is a leading threat to human life and lasting disability worldwide thus needing immediate access to accurate early prediction systems. Research explores stroke brain prediction through machine learning using publicly accessible medical data that contains information about patient demographics and lifestyle along with clinical data. The medical dataset showcases an extreme class imbalance where stroke cases made up only 5% of recorded whole data. SMOTE was used to balance the dataset therefore improving stability during model training. The proposed feature preprocessing methodology along with categorical encoding with robust scaling through RobustScaler and employ SHAP (SHapley Additive exPlanations) analysis for explainable feature selection. Through a GridSearchCV process the XGBoost classifier reached 75.77% accuracy along with 0.76 ROC-AUC score. The SHAP identification process discloses work type together with BMI and glucose levels and smoking history as important prediction variables. The implementation of this model provided accurate performance measurements of both classes through solid visual diagrams from confusion matrix and ROC curve outputs. The proposed framework offers an efficient and interpretable stroke risk prediction solution which allow healthcare practitioners to detect early high-risk subjects.

Keywords - Brain Stroke Prediction, Machine Learning, XGBoost, SMOTE, SHAP.

I. INTRODUCTION

A stroke is a serious medical emergency that can result in death, long-term disability, or irreversible neurological damage when the blood supply to the brain is cut off or interrupted [1]. Since it is one of the main causes of death and morbidity worldwide, early detection and prevention are essential. Conventional stroke risk assessment techniques frequently depend on clinical judgement and simplistic statistical analyses, which might not adequately account for the intricate relationships between different risk factors. Machine learning (ML) has become a potent tool in recent years for creating predictive models that use structured medical data to identify stroke risk. However, issues like class imbalance, overfitting, a lack of model transparency, and limited generalisation are common problems with current methods. In this study, we use a publicly accessible dataset of clinical, behavioural, and demographic characteristics to present a reliable and interpretable machine learning framework for brain stroke prediction. Only 5% of the total records in the original dataset were stroke cases, indicating a significant imbalance [2]. In order to solve this problem and achieve class balance, synthetic samples for the minority class were created using the Synthetic Minority Over-sampling Technique (SMOTE) [3]. Categorical encoding, feature scaling with RobustScaler to lessen the effect of outliers, and type normalisation to guarantee compatibility with machine learning algorithms were all examples of data preprocessing. An XGBoost classifier was selected Because of its robust performance on tabular data and capacity to manage intricate feature interactions [4]. In order to enhance model interpretability and aid in determining the most significant features, SHAP (SHapley Additive exPlanations) analysis was also included.In the past, medical settings have shown success with Explainable AI (XAI) systems such as SHAP, LIME, and Eli5, especially when used to assess

stroke patients using EEG recordings [5]. GridSearchCV was used to optimise the model's hyperparameters, and metrics like accuracy, ROC-AUC, precision, recall, F1-score, confusion matrix, and ROC curve visualisation were used to assess the model's performance. With a ROC-AUC score of 0.76 and an accuracy of 75.77%, the suggested model showed balanced and trustworthy predictive power. This study demonstrates the potential of integrating Explainable Artificial Intelligence (XAI) tools with cutting-edge machine learning techniques to facilitate early stroke detection and aid in proactive healthcare decision-making.

II. LITERATURE REVIEW

2.1 HYBRID AND ENSEMBLE MODELS FOR STROKE PREDICTION

In stroke prediction, ensemble learning has emerged as a crucial technique that allows models to take advantage of the advantages of several algorithms for increased robustness and accuracy. By merging structured feature learning and deep pattern recognition, the study by [6] presented a potent hybrid ensemble that combined Deep Neural Networks (DNN) with XGBoost, attaining 96.76% accuracy. Their method's applicability in clinical settings was limited, though, by its lack of discussion of fairness, explainability, and how to deal with missing values. The study conducted by [1] a comparative analysis of several machine learning models, such as SVM, XGBoost, and ANN, and discovered that Random Forest performed the best, achieving 99% accuracy. Despite being extremely accurate, the authors admitted that because interpretability was not included, the model lacked transparency and generalisability to actual healthcare settings.

The research by [7] created a thorough ensemble model that combined a variety of classifiers, such as boosting (XGBoost, LightGBM, CatBoost), bagging (Random Forest), and base (SVM, Naive Bayes) learners. Their approach comprised two stages of model training, both with and without hyperparameter adjustment. Using a Max Voting technique, the final ensemble had a high classification accuracy of 95.76 %. The study did not assess the model's interpretability or use explainability techniques, which limited the model's use in clinical settings despite its high accuracy. Our study suggests a single, interpretable XGBoost model that is improved with SHAP analysis and trained on a balanced dataset using SMOTE, which is more transparent and useful for real-world healthcare situations than their intricate stacking approach.

The study suggested [8] a Dense Stacking Ensemble (DSE) architecture that combines several advanced models into a single stroke prediction framework. SMOTE and multiple imputation approaches were used in the study to handle missing values and data imbalance. The authors improved and merged the top-performing classifiers into a meta-classifier after baseline evaluations, which produced a balanced dataset with over 96 %. accuracy and a 98.92 % AUC. The DSE model's intricacy and absence of integrated explainability tools restrict its transparency and interpretability in clinical applications, despite the method's thoroughness and impressive outcomes. On the other hand, our research expands upon a simplified, interpretable XGBoost pipeline enhanced with SHAP analysis and SMOTE balance, providing a workable, comprehensible approach better suited for actual healthcare implementation.

By combining XGBoost with principal component analysis (PCA) and SHAP, the research by [9] created a stroke risk prediction model that demonstrated high AUC values and up to 98% accuracy. Their work increased transparency, but it was more difficult to directly interpret clinical features because it relied on dimensionality reduction and synthetic datasets. A hybrid model developed by Ferdib-Al-Islam and Ghosh that combined SMOTE with an ensemble of KNN and XGBoost

[3] enhanced the classification of stroke risk instead lacked clearness for clinical decision-making. In a similar vein, the research by [2] used Random Forests and ExtraTrees in conjunction with SMOTE to balance the dataset in their ensemble learning pipeline, which produced an accuracy of 98.24%. Nevertheless, the study did not investigate the ensemble's interpretability, so the reasoning behind the predictions is unclear.

Together, these studies show that although ensemble and hybrid models can have a high predictive power, they frequently lack external validation, transparency, and deployment viability. Our work, on the other hand, suggests a simplified yet understandable method that makes use of a single optimised XGBoost model. While SMOTE addresses class imbalance without adding complexity to the pipeline, the integration of SHAP values improves model transparency. The suggested approach is intended for practical, clinicianfriendly implementation in actual healthcare settings by fusing interpretability and performance.

2.2 Deep Learning and Imaging-Based Stroke Prediction

Promising outcomes in predicting strokes and prognosis have been demonstrated by deep learning techniques, especially when combined with medical imaging. Using magnetic resonance imaging (MRI)

scans, [10] created an ensemble deep learning model that produced ROC-AUC scores higher than 0.95. Despite the model's high diagnostic accuracy, its lack of interpretability mechanisms and real-time evaluation limits its clinical adoption. [4] also used an image-based method, extracting features from brain MRI with 98.7% accuracy by combining MobileNet V3 with LightGBM and CatBoost. Their model did not incorporate any explainable AI techniques that are necessary for radiologist trust, and it lacked transparency despite its high performance.

In order to predict the prognoses of acute ischaemic stroke,[11] presented the Optimised Ensemble of Deep Learning (OEDL) framework, which integrated radiomics and clinical features. Although the ensemble performed better than singlemodality models, it lacked external validation and an analysis of the impact of individual features, both of which are essential for transferring research into clinical settings. In a similar vein, an ensemble of deep classifiers trained on dynamic radiomics features (DRF) taken from perfusion-weighted MRI (PWI) data was proposed by [12], and it achieved perfect recall in stroke detection with AUC values as high as 0.976. However, the model's practical implementation was restricted by its intricacy and inability to be explained. Ye et al. also pointed out that while employing union features increased prediction accuracy, it did not provide information about the distinct contributions of each feature.

These studies demonstrate the great predictive ability of deep learning when paired with radiological inputs, but they additionally highlight three frequent drawbacks: Explainability issues, the requirement for external validation, and the viability of real-time deployment. In contrast, our study employs a computationally lightweight XGBoost model, incorporates SHAP for interpretability, and focusses on structured health data rather than complex imaging inputs, making it more deployable in a variety of clinical settings.

2.3 Explainable AI Techniques for Interpretable Stroke Risk Modeling

Transparency and interpretability have become crucial components for clinical acceptance as predictive models are increasingly incorporated into healthcare. Hypertension and previous transient ischaemic attacks were found to be the most significant factors when [1] applied SHAP (Shapley Additive Explanations) to an XGBoost model trained on stroke cohort data. The study did not concentrate on improving predictive performance or external validation, despite offering insightful information about feature contributions. By combining SMOTE-Tomek sampling with a neural network, [13] addressed class imbalance and reported an 85.8% stroke prediction accuracy. However, the study's clinical utility was limited due to its lack of external validation and strong explainability tools.

To improve prediction and interpretation, a novel pipeline that combines XGBoost, PCA, and SHAP was suggested by [9]. The model was tested on two datasets and obtained AUC values close to 0.99; however, its practical adoption was impeded by its dependence on synthetic data and dimensionality reduction. Using SHAP values, [14] presented an explainable stroke prediction pipeline and showed that the most important predictors were age, BMI, and glucose levels. However, the study lacked validation across a variety of patient cohorts and had high false-negative rates.

Using LIME and SHAP, [15] created an interpretable ensemble model that performed well (94%) and provided case specific explanations. The model was not, however, tested in real-time settings or across various healthcare systems in this study. Last but not least, [16] improved model transparency but provided limited generalisability by integrating explainability using permutation importance to interpret neural network decisions on a small dataset.

On the other hand, our study uses SHAP for both feature selection and post-hoc interpretation, integrating it directly into the modelling process. This enables accurate and interpretable real-time stroke risk prediction when paired with a balanced dataset and an optimised XGBoost model, which makes it more appropriate for clinician use in decision-support systems.

III. PROPOSED METHODOLOGY / APPROACH

The entire machine learning pipeline used to predict brain strokes is described in this section, including data preprocessing, feature engineering, model development, hyperparameter optimization, and evaluation. The following technical subsections comprise the methodology's structure.:

This flowchart in **Fig. 1,** highlights the absolute ML workflow which includes dataset acquisition and preprocessing to SMOTE balancing, XGBoost model training, SHAP-based feature selection, hyperparameter tuning, and final evaluation.It provides a condensed, comprehensible illustration of the suggested methodology.

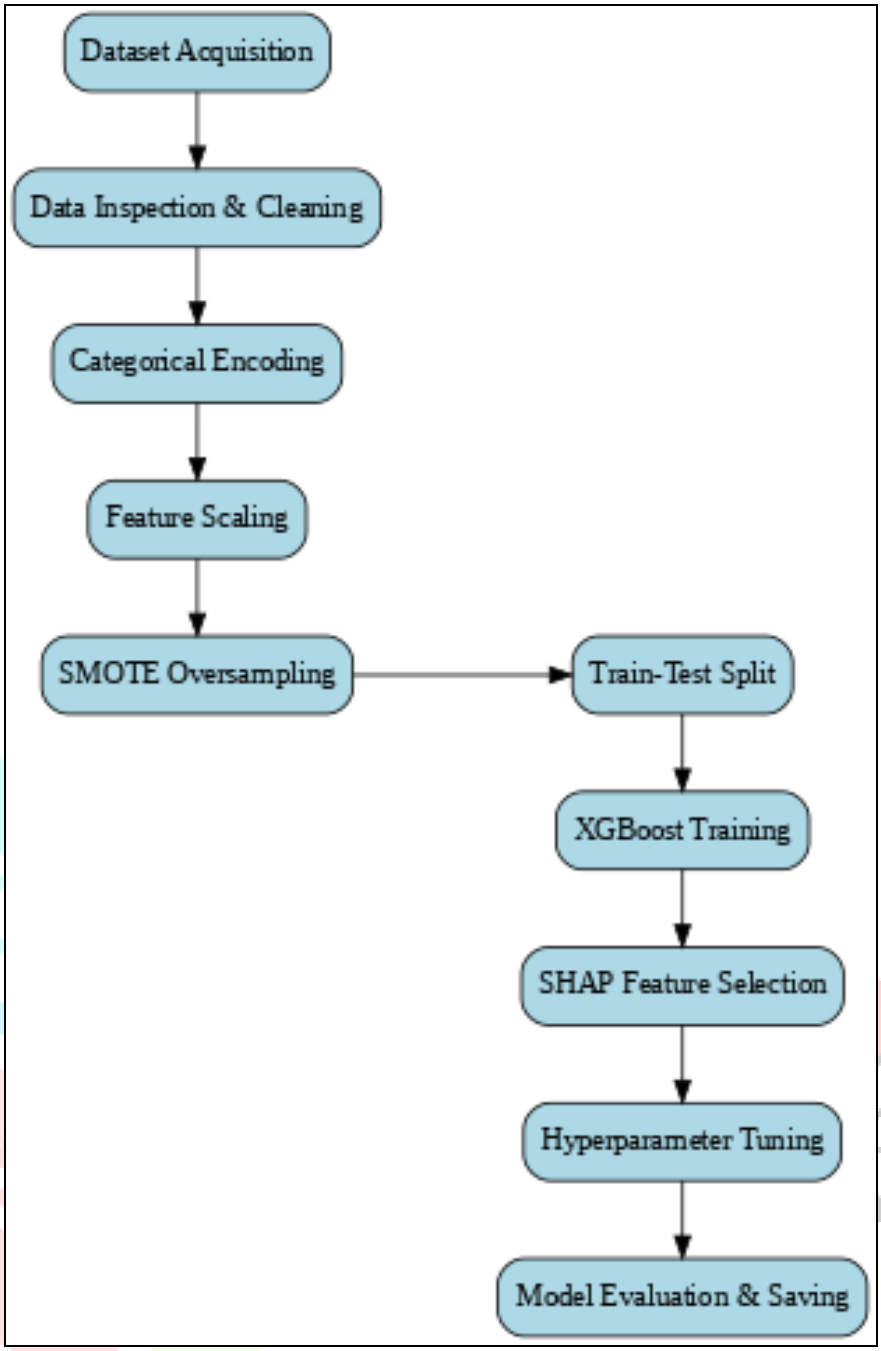

Fig. 1: Flowchart Of The Proposed Brain Stroke Prediction Pipeline

3.1 Dataset Description

A publicly accessible medical dataset Kaggle (https://www.kaggle.com/datasets/jillanisofttech/brain-strokedataset), provided 4,981 patient records for this investigation. The data set includes the most significant health and demographic characteristics that can be used to predict strokes. The data collection comprises real-world patient data and has been widely used in prior stroke-related research, making it perfect for machine learning algorithm training. The model's predictive power for strokes was enhanced by using SMOTE to address the class imbalance and ensure that both stroke and non-stroke occurrences were adequately represented.

3.1.1 HeatMap Distribution

The linear links between the dataset's numerical properties were revealed by the correlation heatmap, which is shown in **Fig. 2,** Age showed the greatest positive connection with stroke (0.25) among the associated variables, indicating that stroke is more common in elderly people. Additionally, there was a moderate connection (0.37) between age and body mass index (BMI), suggesting that BMI may gradually rise with age. The significance of other risk variables as possible contributors to stroke risk was further supported by the slight positive correlations they showed with stroke (both about 0.13). These risk factors included heart disease and hypertension. On the other hand, characteristics such as BMI and average glucose level had a weak connection with stroke (0.13 and 0.06, respectively), suggesting that although

they might still have clinical significance, their direct linear relationship with the incidence of stroke is limited. These results demonstrate that even if some variables have a poor correlation with stroke, they could still be significant when combined using non-linear models like XGBoost.

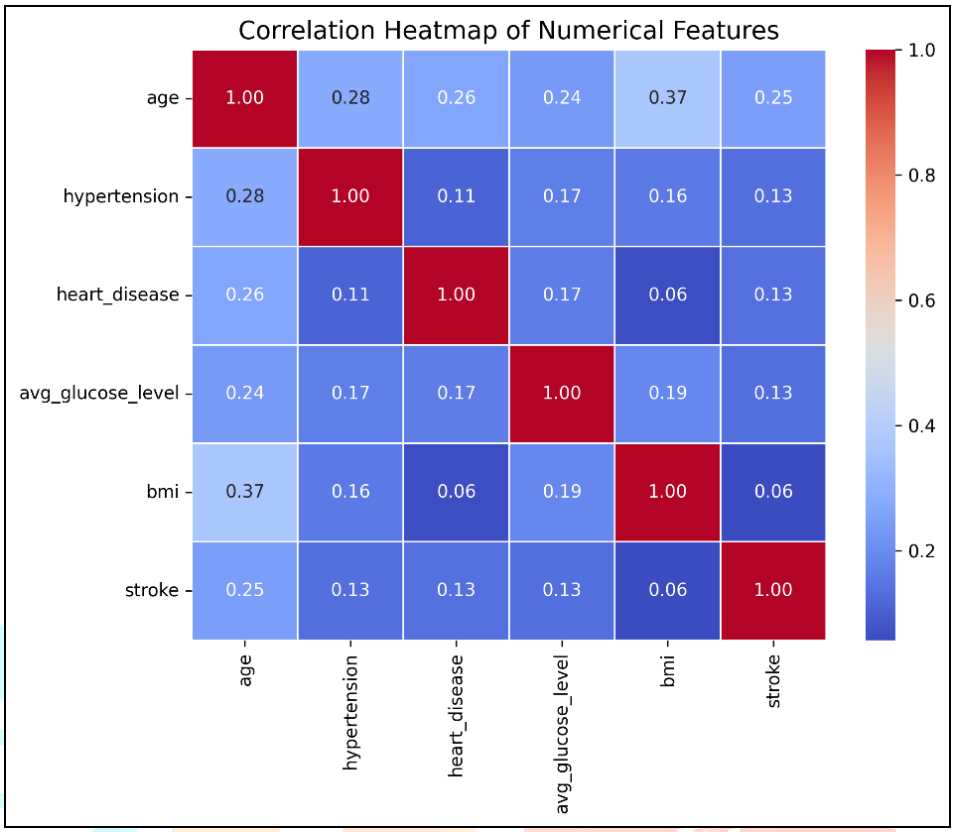

Fig. 2: Heatmap Distribution

3.1.2 Box Plot Distribution

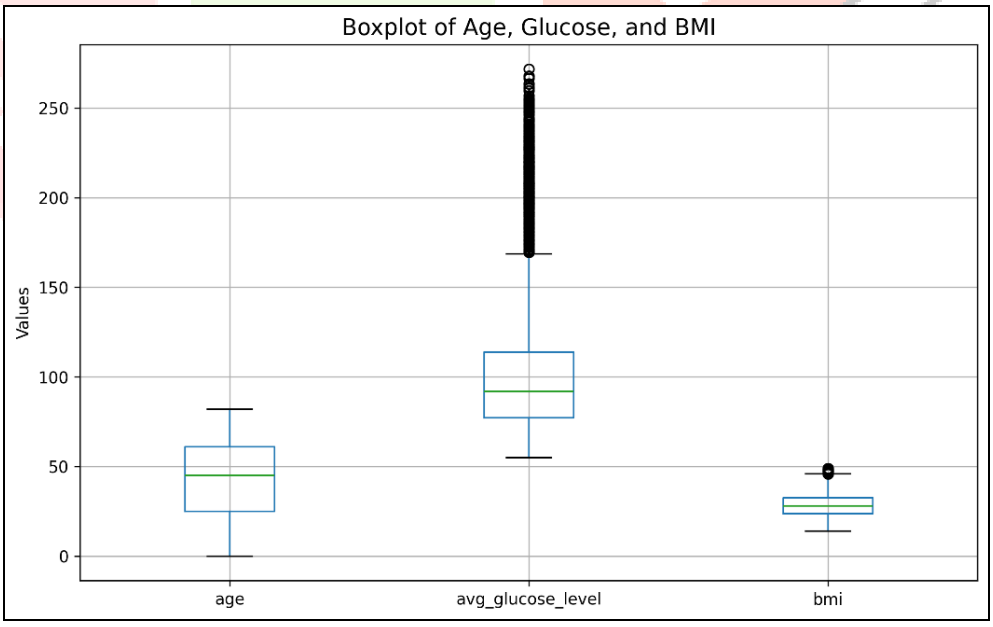

Fig. 3: Box Plot Distribution

Age, BMI, and average glucose level are three significant numerical parameters whose distribution and spread are compared in the boxplot displayed in **Fig. 3.** With a median age of 40 and a range of infancy to almost 80 years, the age distribution seems to be somewhat distributed. Some people have abnormally high glucose levels, which may indicate that they are diabetic or high-risk individuals. The average glucose level shows a larger variance and is severely skewed with many extreme outliers. Although there are a few modest outliers in the BMI values, which indicate rare cases of underweight or obesity, the distribution is generally compact, with a median of about 28 to 30. These observations draw attention to the existence of outliers, particularly in glucose readings, which, if ignored, may skew model performance. In order to

manage this skewness and preserve model stability, robust scaling strategies were used during preprocessing.

There are considerable disparities between stroke and nonstroke populations, as shown by the age distribution plot in **Fig. 4.** The red density curve indicates that older persons, especially those 60 and older, account for the majority of stroke cases. Non-stroke patients, on the other hand, are more evenly distributed among all age categories, with a noticeable concentration in the 30- to 60-year-old age range.

For non-stroke cases, the density curve shows a comparatively constant count until it progressively drops after age 60. Age is a significant risk factor for stroke, as seen by the sharp increase in stroke incidence with advancing years. These findings highlight the importance of age-focused screening and prevention, especially for older people.

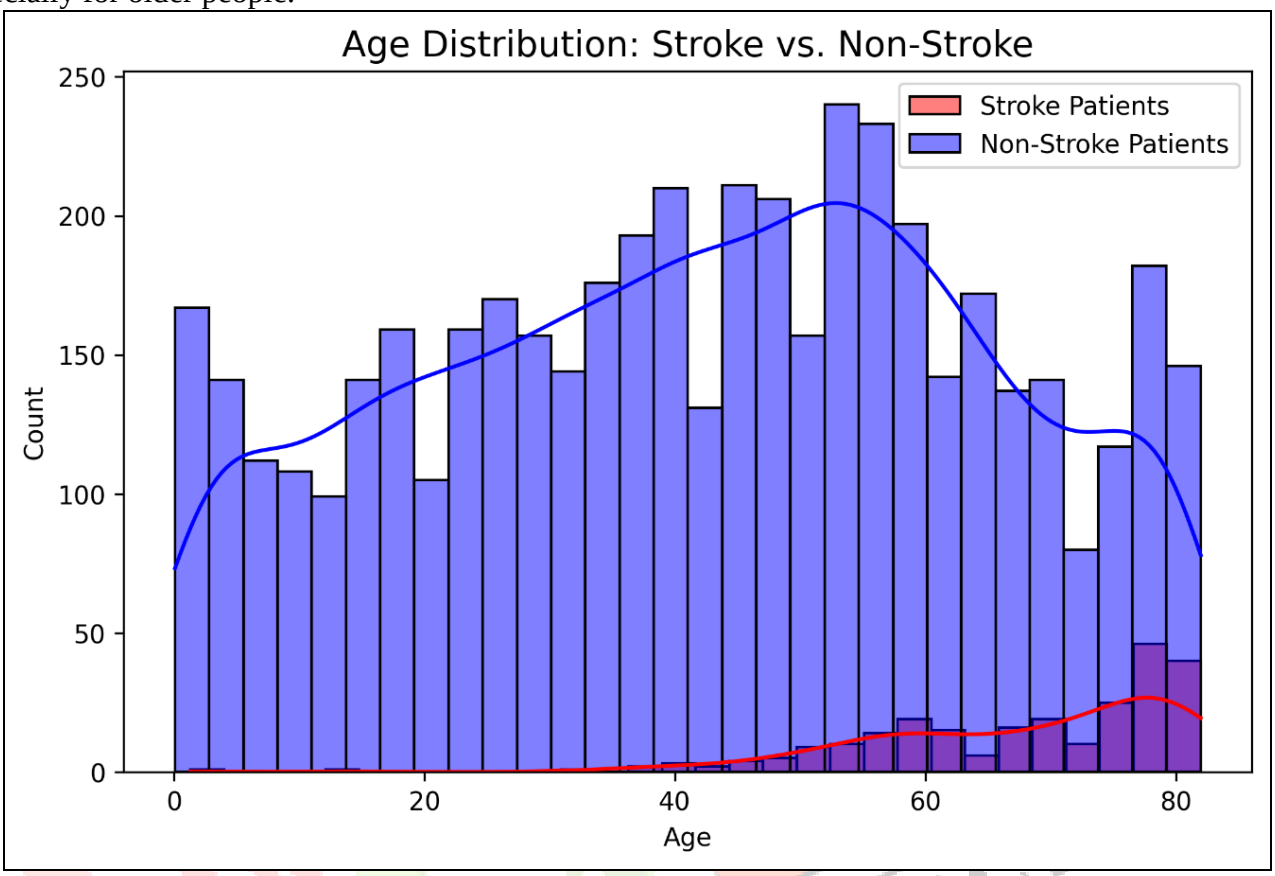

Fig. 4: Age Distribution (Stroke vs Non-Stroke)

3.2. Data Preprocessing

The first analysis of the dataset confirmed that no values were missing, allowing for a streamlined preprocessing pipeline. While categorical variables like gender and work type were one-hot encoded, binary variables like ever married and residence _type were label-encoded. The mapping of the ordinal feature smoking _status to integer values represented the change from non-smoker to current smoker.The RobustScaler was used to scale numerical features including age, avg glucose level, and bmi in order to lessen the influence of outliers. Additionally, Boolean columns created after encoding were deliberately transformed to integer types to ensure compatibility with machine learning methods.

The dataset's significant class imbalance, with strokepositive cases accounting for only 5% of all records, was a huge issue. This was reduced using the Synthetic Minority Over-sampling Technique (SMOTE). SMOTE generates synthetic samples of the minority class by interpolating between instances that already exist. The dataset was balanced, with 4,733 stroke (1) and non-stroke (0) cases, respectively. This prevented bias towards the majority class and enabled the classifier to learn from both classes effectively. This approach is consistent with previous research, such as the UCO method, which combined various sampling techniques in an effort to enhance classifier performance in unbalanced medical datasets [17].

3.3 Feature Engineering

The relevance of characteristics and their role in predictions were interpreted using SHAP (SHapley Additive exPlanations) Technique analysis.Features like work _type children and residence type were eliminated from the feature set to improve model performance and generalisation after it was determined that they had little effect based on SHAP values.

3.4 Machine Learning Models

XGBoost (Extreme Gradient Boosting) Algorithm was selected as the main model, Because of its ability to adapt when dealing with imbalanced and structured datasets. It is ideal for tasks involving the prediction of medical data, as it is an ensemble approach based on decision trees and allows for regularisation. The pre-processed training data was used to train the model, and a different test set was used to assess it.Additionally, XGBoost's performance and scalability have been confirmed on a number of public health datasets, including the Framingham Heart Study and NHANES [18], confirming its applicability for predictive modeling in medical settings.

3.5 Hyperparameter Tuning

XGBoost model performance was maximized by using GridSearchCV with a specified parameter grid. To systematically modify parameters such as max _depth, learning rate, n estimators, colsample bytree, and subsample, 3-fold crossvalidation was employed. The model with the highest ROCAUC score was selected as the best one during crossvalidation.

3.6 Evaluation Metrics

Performance criteria such as accuracy, precision, recall, F1score, and ROC-AUC score were employed to evaluate the final model. A confusion matrix was used to visualize correct and incorrect predictions, and the ROC curve provided details about the model's capacity for class discrimination. With a ROC-AUC score of 0.76 and an accuracy of 75.77%, the optimized model showed outstanding predictive power.

IV. RESULTS AND DISCUSSION

4.1 Model Performance Comparison

The final XGBoost model demonstrated good predictive power,with GridSearchCV optimisation. The dataset was balanced using SMOTE, and then 80% of the data was used to train the model and 20% for testing. With a ROC-AUC score of 0.76 and balanced precision, recall, and F1-score values for both stroke and non-stroke classes, it attained an accuracy of 75.77%. These outcomes demonstrate how well the model generalises to new data and successfully learnt to differentiate patterns between the two groups.

The performance of the XGBoost model can be reviewed in **Table** 1. Having an accuracy of 75.77 %, the model correctly diagnosed most stroke and non-stroke cases. The ROC-AUC score of 0.76 indicates that there is good distinguishing between the two classes. The precision, or the number of instances properly predicted, was 0.77 for stroke and 0.75 for non-stroke. Balanced sensitivity was shown by recall (the number of real cases found) of 0.74 for stroke and 0.76 for non-stroke. The F1-score of 0.75 for both classes indicates that the model kept a fair balance between false positives and false negatives.

TABLE 1: Performance Metrics Of the Final XGBoost Model

Metric	Value
Accuracy	75.77%
ROC-AUC	0.76
Precision	0.75 (Class 0), 0.77 (Class 1)
Recall	0.76 (Class 0), 0.74 (Class 1)
F1-score	0.75 (both classes)

4.2 Confusion Matrix

The model's ability to distinguish between stroke and non-stroke patients is shown in the confusion matrix. It offers a thorough analysis of the XGBoost model's right and wrong classifications on the test dataset.

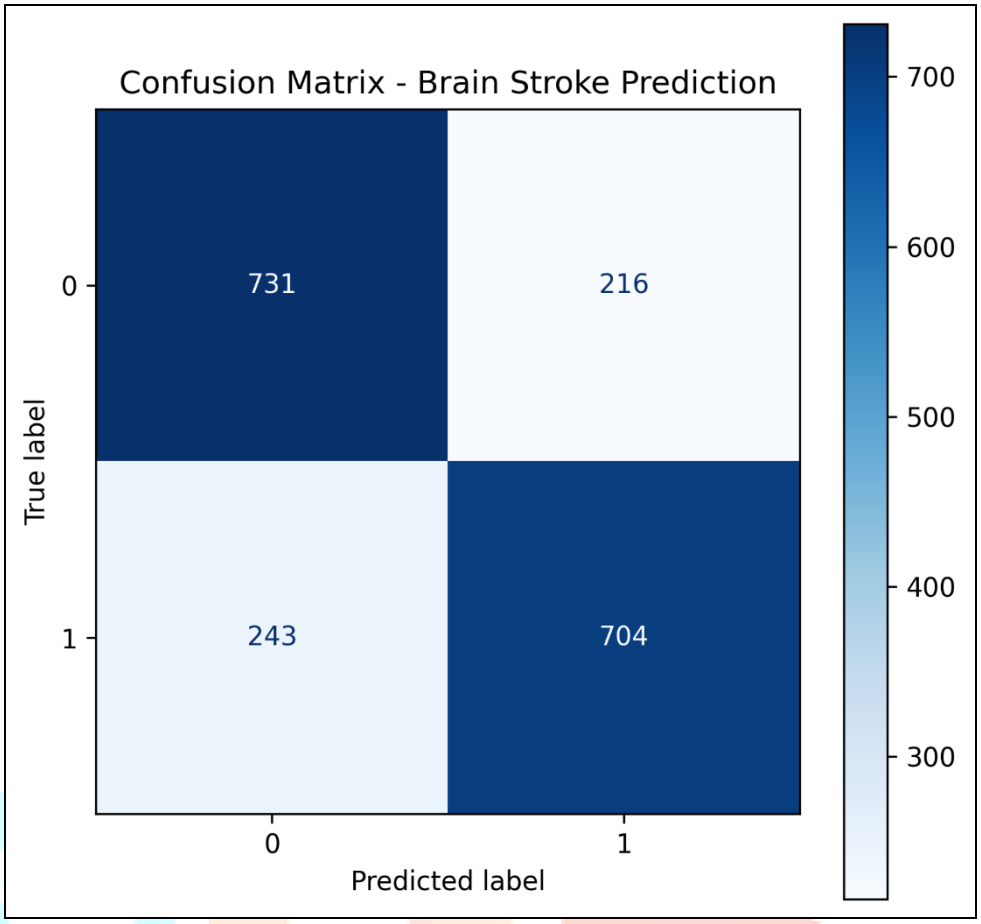

Fig. 5: Confusion Matrix For Brain Stroke Prediction

In **Fig. 5,** the performance of the XGBoost model in differentiating between stroke and non-stroke instances is shown. Whereas off-diagonal cells display incorrect classifications, diagonal cells reveal correctly classified occurrences. In particular, 731 non-stroke (class 0) and 704 stroke (class 1) instances were accurately predicted by the model. It did, however, also result in 216 false positives and 243 false negatives, underscoring the significance of minimizing errors in medical diagnosis scenarios.

The model's classification of stroke and non-stroke instances was made clear using the confusion matrix. Out of 1,894 test samples, the model correctly predicted 704 stroke cases (true positives) and 731 non-stroke cases (true negatives). However, it misidentified 216 non-stroke cases as stroke (false positives) and 243 stroke patients as non-stroke (false negatives). The prevalence of false negatives highlights a limitation, even though the overall accuracy is encouraging. This is especially true in medical applications where it can be critical to overlook a genuine stroke case.

The distribution of expected versus actual labels for the stroke classification work is provided in Fig. 5. A total of 1,435 cases (sum of true positives and true negatives) were accurately identified by the model. But it also misclassified 459 cases, including 243 false negatives, which missed real stroke cases. This raises a serious clinical problem since stroke patients who go undiagnosed might not get prompt treatment. Thus, the confusion matrix offers crucial information about the model's predictive capabilities' advantages and disadvantages, particularly in high-stakes medical situations.

4.3 Feature Importance / SHAP Analysis

To make the concept more transparent, a study called SHAP (SHapley Additive exPlanations) was conducted. With the use of SHAP values, the most significant attributes impacting model predictions were identified. The primary affecting factors were Marital Status (Ever Married), Body Mass Index (BMI), Average Glucose Level, Smoking Status, and Work Type (Private, Self-employed). This method made it easier to understand how specific attributes influence model decisions, which improved the system's interpretability in a clinical setting [19].

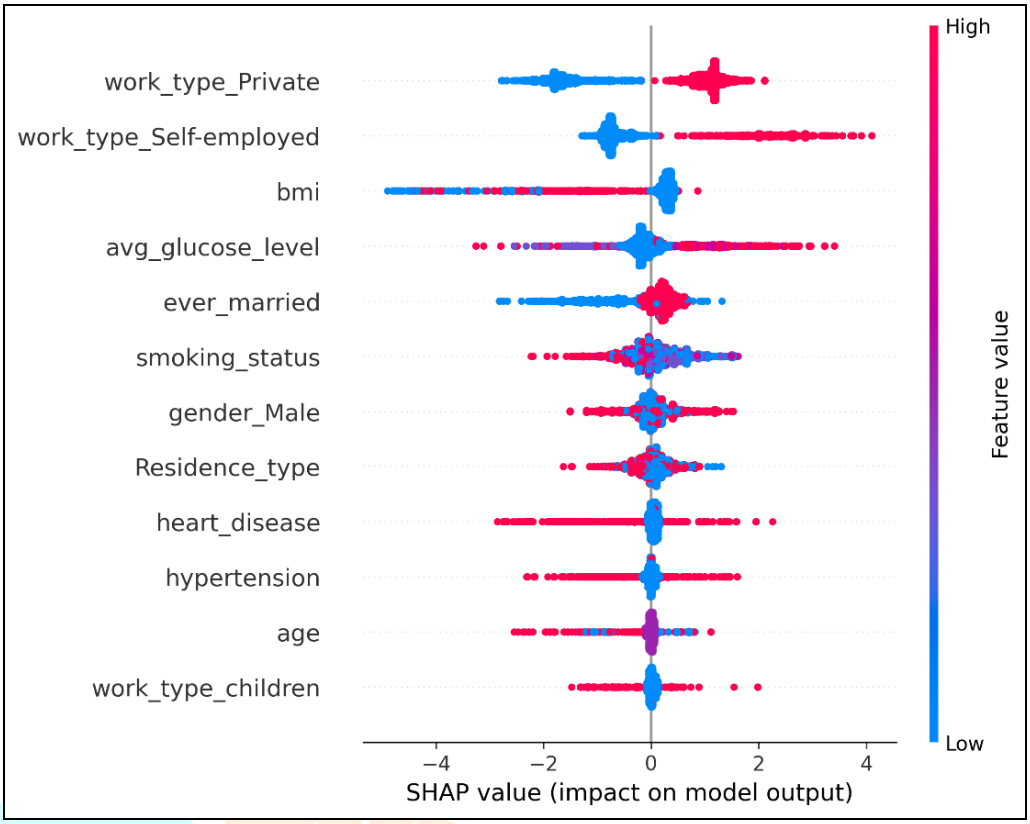

Fig. 6: Shap Value Analysis

Figure. 6, SHAP Summary Plot for Stroke Prediction Using XGBoost. The plot determines the ranking position of features according to their overall influence on model predictions. The color scale depicts high features with red and low features with blue.

According to the SHAP summary graphic, a higher risk of stroke was linked to higher BMI and glucose levels.

Additionally, categorical characteristics like smoking status and work type were crucial. In order to simplify the model and enhance generalisation, low-impact elements such as work type children and residence type were eliminated based on SHAP values.

4.4 Justification for Model Choice

When paired with SMOTE, XGBoost's shown efficiency in managing structured tabular data, robustness to multicollinearity, and inherent capacity to manage class imbalance led to its selection. In this study, interpretability and performance on unbalanced data were given priority using a single optimised model with explainability (SHAP), even if baseline models like SVM and MLP are frequently employed in medical prediction tasks. In subsequent research, XGBoost might be compared against other classifiers using identical preprocessing and sampling setups.

4.5 Strengths and Limitations of the Proposed Approach

The suggested approach is appropriate for clinical decisionsupport systems since it successfully strikes a compromise between interpretability and predictive accuracy. Combining XGBoost and SHAP results in a model that not only accurately classifies stroke risk but also clearly explains feature contributions, increasing the model's usability and credibility for medical experts. SMOTE improves the detection of minority (stroke-positive) instances by addressing class imbalance.The model is also lightweight and simple to use using joblib, which makes it easier to integrate into practical applications. But there are drawbacks to the framework. It may not be as generalizable to various populations or clinical contexts because it was learned on a single structured dataset. Additionally, the model generates a few false negatives, which might be problematic in applications related to medicine. Additionally, it doesn't use multimodal data that could improve prediction accuracy, including MRI, CT scan images, or wearable device signals. Future studies ought to investigate combining various data sources and confirming the model on larger, longer-term datasets.

4.6 ROC Curve Analysis

With the Receiver Operating Characteristic (ROC) curve, the model's classification ability is visually evaluated over threshold settings. The graph, which plots the True Positive Rate (Sensitivity) against the False Positive Rate, shows the trade-off between precisely identifying stroke cases and avoiding false alarms. The optimized XGBoost model's Area Under the Curve (AUC) of 0.76 showed a strong

discriminative ability to distinguish between stroke and non-stroke instances. This result supports the model's reliability in clinical settings where sensitivity and specificity are critical for an accurate diagnosis.

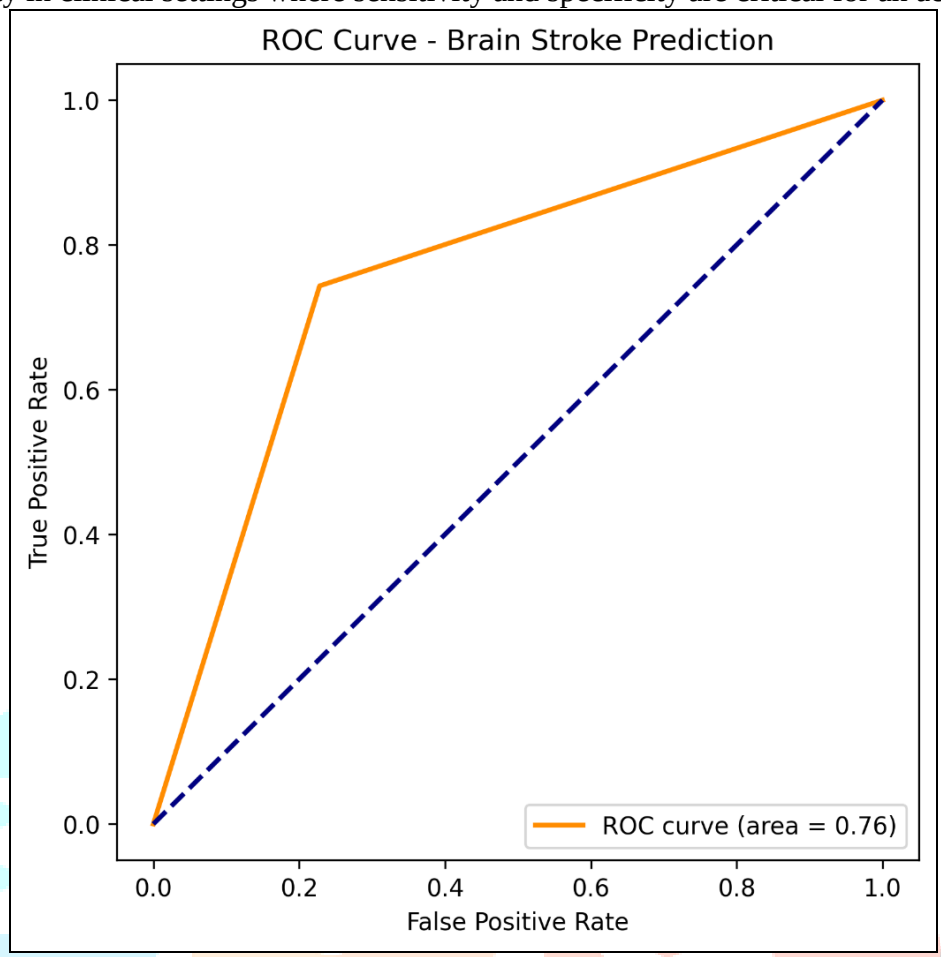

Fig. 7: Roc Curve for XGBoost Stroke Prediction Model

The curve in **Fig. 7** shows how well the model distinguishes stroke from non-stroke diagnoses. The 0.76 AUC value signifies strong class discrimination which results in dependable two-class categorization.

V. CONCLUSION AND FUTURE WORK

Using structured medical data, this study offered a reliable and understandable machine learning framework for early brain stroke prediction. The strategy used SMOTE to solve the class imbalance issue, thorough preprocessing and feature engineering methods, and the XGBoost classifier because of its capacity to handle both structured and unbalanced data. To improve model interpretability, SHAP analysis was incorporated, which aided in identifying important characteristics including work type, average glucose level, smoking status, and BMI. GridSearchCV was used to further refine the model, which ultimately yielded a ROC-AUC score of 0.76 and an accuracy of 75.77%. These findings suggest that the suggested model has potential as a decision-support tool in healthcare settings and is both accurate and clinically relevant.

This study's main contribution is the integration of explainability with predictive performance. In order to build confidence in medical applications, the framework goes beyond a black-box model by integrating SHAP values, which provides insights into feature contributions. Furthermore, using SMOTE to balance the dataset greatly enhanced the model's capacity to identify minority-class (stroke) cases, hence bolstering its usefulness in risk-based patient screening.

Despite showing encouraging outcomes, the current study has many drawbacks. One structured dataset was used for both training and validation, which would limit the model's applicability to a variety of clinical settings and populations. False negative results continue to be an issue, especially in medical applications where failing to diagnose a stroke can have major consequences. This constraint can be overcome in future research by using strategies like cost-sensitive learning, threshold adjustment, or more robust ensemble approaches. Furthermore, adding multimodal data sources to the model, such as wearable sensor readings, electronic health records (EHRs), and MRI and CT scan pictures, could greatly improve the predictive ability of the system. The model would be able to acquire regional and anatomical data by integrating imaging datasets, facilitating a more thorough and precise evaluation of stroke risk. Furthermore, two crucial stages toward actual application will be deploying the method in real-time clinical settings and verifying it on longitudinal datasets.

REFERENCES

[1] J. Li, Y. Luo, M. Dong, Y. Liang, X. Zhao, Y. Zhang, and Z. Ge, "Treebased risk factor identification and stroke level prediction in stroke cohort study," *BioMed Research International*, vol. 2023, no. 1, p. 7352191, 2023. R. Wijaya, F. Saeed, P. Samimi, A. M. Albarrak, and S. N. Qasem, "An ensemble machine learning and data mining approach to enhance stroke prediction," *Bioengineering*, vol. 11, no. 7, p. 672, 2024.

[2] R. Wijaya, F. Saeed, P. Samimi, A. M. Albarrak, and S. N. Qasem, "An ensemble machine learning and data mining approach to enhance stroke prediction," *Bioengineering*, vol. 11, no. 7, p. 672, 2024.

[3] F. Islam and M. Ghosh, "An enhanced stroke prediction scheme using smote and machine learning techniques," in *International conference on computing communication and networking technologies (ICCCNT), Kharagpur, India*, 2021.

[4] S. Rahman, M. Hasan, and A. K. Sarkar, "Prediction of brain stroke using machine learning algorithms and deep neural network techniques," *European Journal of Electrical Engineering and Computer Science*, vol. 7, no. 1, pp. 23–30, 2023.

[5] M. S. Islam, I. Hussain, M. M. Rahman, S. J. Park, and M. A. Hossain, "Explainable artificial intelligence model for stroke prediction using eeg signal," *Sensors*, vol. 22, no. 24, p. 9859, 2022.

[6] R. Agrawal, A. Ahire, D. Mehta, P. Hemnani, and S. Hamdare, "Optimizing stroke risk prediction using xgboost and deep neural networks." *International Journal of Advanced Computer Science & Applications*, vol. 15, no. 11, 2024.

[7] P. Premisha, S. Prasanth, M. Kanagarathnam, and K. Banujan, "An ensemble machine learning approach for stroke prediction," in *2022 International Research Conference on Smart Computing and Systems Engineering (SCSE)*, vol. 5. IEEE, 2022, pp. 165–170.

[8] A. Hassan, S. Gulzar Ahmad, E. Ullah Munir, I. Ali Khan, and N. Ramzan, "Predictive modelling and identification of key risk factors for stroke using machine learning," *Scientific Reports*, vol. 14, no. 1, p. 11498, 2024.

[9] L. Mochurad, V. Babii, Y. Boliubash, and Y. Mochurad, "Improving stroke risk prediction by integrating xgboost, optimized principal component analysis, and explainable artificial intelligence," *BMC Medical Informatics and Decision Making*, vol. 25, p. 63, 2025.

[10] A. W. Abulfaraj, A. K. Dutta, and A. R. W. Sait, "Ensemble learningbased brain stroke prediction model using magnetic resonance imaging," *Journal of Disability Research*, vol. 3, no. 5, p. 20240061, 2024.

[11] W. Ye, X. Chen, P. Li, Y. Tao, Z. Wang, C. Gao, J. Cheng, F. Li, D. Yi, Z. Wei *et al.*, "Oedl: an optimized ensemble deep learning method for the prediction of acute ischemic stroke prognoses using union features," *Frontiers in Neurology*, vol. 14, p. 1158555, 2023.

[12] M. M. Yassin, J. Lu, A. Zaman, H. Yang, A. Cao, X. Zeng, H. Hassan, T. Han, X. Miao, Y. Shi *et al.*, "Advancing ischemic stroke diagnosis and clinical outcome prediction using improved ensemble techniques in dsc-pwi radiomics," *Scientific Reports*, vol. 14, no. 1, p. 27580, 2024.

[13] C. Rana, N. Chitre, B. Poyekar, and P. Bide, "Stroke prediction using smote-tomek and neural network," in *2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*. IEEE, 2021, pp. 1–5.

[14] C. Kokkotis, G. Giarmatzis, E. Giannakou, S. Moustakidis, T. Tsatalas, D. Tsiptsios, K. Vadikolias, and N. Aggelousis, "An explainable machine learning pipeline for stroke prediction on imbalanced data," *Diagnostics*, vol. 12, no. 10, p. 2392, 2022.

[15] P. Srinivasu, U. Sirisha, K. Sandeep, S. Praveen, L. Maguluri, and T. Bikku, "An interpretable approach with explainable ai for heart stroke prediction. diagnostics 2024, 14, 128."

[16] A. Rehman, T. Alam, M. Mujahid, F. S. Alamri, B. Al Ghofaily, and T. Saba, "Rdet stacking classifier: a novel machine learning based approach for stroke prediction using imbalance data," *PeerJ Computer Science*, vol. 9, p. e1684, 2023.

[17] M. Wang, X. Yao, and Y. Chen, "An imbalanced-data processing algorithm for the prediction of heart attack in stroke patients," *IEEE Access*, vol. 9, pp. 25394–25404, 2021.

[18] N. S. Rajliwall, R. Davey, and G. Chetty, "Cardiovascular risk prediction based on xgboost," in *2018 5th Asia-Pacific World Congress on Computer Science and Engineering (APWC on CSE)*. IEEE, 2018, pp. 246–252.

[19] X. Zheng, F. Wang, J. Zhang, X. Cui, F. Jiang, N. Chen, J. Zhou, J. Chen, S. Lin, and J. Zou, "Using machine learning to predict atrial fibrillation diagnosed after ischemic stroke," *International Journal of Cardiology*, vol. 347, pp. 21–27, 2022.

Dataset Link:
Brain Stroke Prediction Dataset:
https://www.kaggle.com/datasets/jillanisofttech/brain-strokedataset
Viewed on:02/03/2025