Optimizing Stroke Risk Prediction Using XGBoost and Deep Neural Networks

Renuka Agrawal[1], Aaditya Ahire[2], Dimple Mehta[3], Preeti Hemnani[4], Safa Hamdare[5]
Department of Computer Science and Engineering, Symbiosis Institute of Technology,
Symbiosis International (Deemed University) Pune, India[1,2,3]
Department of Electronics and Telecommunication Engineering, SIES Graduate School of Technology, Mumbai, India[4]
Department of Computer Science, Nottingham Trent University, Nottingham, NG11 8NS, UK[5]

Abstract—**Predicting brain strokes is inherently complex due to the multifaceted nature of brain health. Recent advancements in machine learning (ML) and deep learning (DL) algorithms have shown promise in forecasting stroke occurrences to a certain extent. This research paper explores the predictive potential of ML and DL models by utilizing a comprehensive dataset encompassing diverse patient characteristics, including demographic factors, work culture, stress levels, lifestyle, and family history. Notably, this study incorporates 14 clinically significant attributes for prediction, surpassing the 10 attributes utilized by earlier researchers. To address existing limitations and enhance predictive accuracy, a novel ensemble model combining Deep Neural Networks (DNN) and Extreme Gradient Boosting (XGBoost) is proposed in this work. Also, a comparative analysis against individual DNN and XGBoost models, as well as Random Forest and Support Vector Machine (SVM) approaches are being done. The performance of the ensemble model is assessed using various metrics, including accuracy, precision, F1 score, and recall. The findings indicate that the DNN-XGBoost model exhibits superior predictive accuracy compared to standalone DNN and XGBoost models in identifying brain stroke occurrences.**

Keywords*—*DNN; XGBoost; stress level; stroke prediction

I. Introduction

Stroke prediction plays a critical role in healthcare because early identification of high-risk individuals allows for preventive interventions, including lifestyle changes, medications, and treatments, which can significantly improve patient outcomes. However, estimating the likelihood of a stroke is complex due to the interrelation of various factors such as diet, medical history, family history, and other external variables [1], [2]. Traditional statistical methods often fail to account for these intricate relationships, leading to limitations in accurately predicting stroke risk.

The challenge lies in understanding how factors like daily habits, medical and family histories interact to influence stroke risk. Most conventional statistical techniques fall short in identifying these complex patterns. As a result, stroke risk prediction often lacks the accuracy needed for reliable clinical decision-making. In recent years, machine learning has emerged as a powerful tool to overcome these challenges. By analyzing large datasets, machine learning techniques can uncover hidden patterns and interactions that may go undetected by human clinicians. This capability significantly enhances predictive analytics in the healthcare sector [3].Several machine learning models, including decision trees, random forests, and ensemble methods, have been employed in stroke prediction, each offering distinct advantages. For instance, decision trees are interpretable and easy to understand, while random forests offer robustness and accuracy in handling large datasets [4], [5], [6]. However, despite the strengths of these models, they are often unable to fully capture the complexity of relationships between variables, particularly in the context of healthcare data. Logistic regression, though simple, also struggles to account for non-linear interactions, making it unsuitable for more intricate datasets [7], [8].

To address these limitations, ensemble models have gained traction [9], as they combine the strengths of multiple base models to improve overall predictive performance. Among the most promising ensemble approaches are XGBoost and Deep Neural Networks (DNN). XGBoost excels in handling complex, non-linear relationships within tabular data, while DNNs are particularly well-suited to learning from sequential data, making them effective at capturing long-term dependencies [10]. This study proposes a new ensemble model that integrates XGBoost and DNN to enhance stroke prediction accuracy. The model incorporates not only traditional features but also additional attributes like family history, stress levels, and alcohol intake, which are known to be significant in stroke risk assessment. By combining XGBoost's ability to model complex feature interactions with DNN's capacity for sequential learning, the proposed ensemble model overcomes challenges such as overfitting, feature interaction, and interpretability, which are common with individual machine learning models. This approach improves both the accuracy and reliability of stroke risk predictions, providing clinicians with better tools for early intervention and personalized patient management.

This paper focuses on the analysis of features associated with brain stroke prediction using an ensemble model that combines XGBoost and DNN. The key contributions of this study can be summarized as follows:

- Conducting a comprehensive analysis of features influencing brain stroke prediction using the XGBoost-DNN ensemble model.
- Demonstrating the model's potential in automating risk assessment procedures in healthcare and providing valuable insights for researchers and practitioners.
- Offering insights to enhance the implementation of predictive medicine in stroke management, emphasizing the importance of timely identification and treatment.

The remainder of this paper is organized as follows: Section II reviews related work by various researchers and discusses their limitations. Section III details the system architecture and methodology adopted for stroke prediction. Section IV presents the results obtained from the ensemble model. Finally, Section V discusses the findings, outlines future research directions, and concludes the paper in Section VI.

II. Literature Review

Machine learning models like KNN, Random Forest, and Decision Tree, logistic regression is used by researchers to automate the process of brain stroke prediction. A tabular representation of the work conducted by various researchers is included in Table I. This table summarizes the methodologies employed, the domains of study, the datasets used, the performance metrics achieved, and the outcomes or limitations identified in each respective study.

T. Lumley et al. addressed the challenge of stroke prediction in the elderly by developing a stroke prediction score, which was validated and made accessible through a web-based application. Their study focused on continuous patient monitoring data and demonstrated an effectiveness of 88% accuracy in identifying at-risk groups; however, the model's dependency on a consistent data feed and quality, as well as potential issues with cluster interpretability, posed limitations to its practical application. Similarly, R.O. Ogundokun et al. reviewed various machine learning algorithms for predicting cardiovascular diseases, including strokes, utilizing time-series health data collected from wearable devices and achieving a notable 92% accuracy in predicting stroke events. They noted that the requirement for continuous data streams could be computationally demanding, raising concerns regarding data sparsity in certain situations.

In another study, S. Shareefunnisa et al. explored heart stroke prediction using machine learning techniques by leveraging mixed datasets from diverse sources, ultimately improving prediction accuracy to 94%. However, they highlighted that the increased model complexity and potential for overfitting, along with the necessity for substantial computational resources, posed challenges to their methodology. S. Dev et al. employed a supervised approach utilizing Naive Bayes for stroke detection, analyzing symptom checker data from healthcare applications and achieving an initial diagnostic accuracy of 86%. Despite this, the simplicity of their model raised concerns about oversimplifying complex dependencies, and it was limited by the quality of the input data, which could lead to a high false positive rate. In a similar context, M. S. Sheetal and P. Choudhary applied gradient boosting techniques to stroke patient data derived from longitudinal studies, achieving an improved prediction accuracy of 91%. Their research underscored the necessity for extensive data preprocessing and feature selection, noting sensitivity to noisy data as a significant limitation.

Expanding on this, S. Rahman et al. integrated random forests and neural networks to predict brain strokes using national stroke registry data, resulting in a high detection rate of early stroke signs at 93%. They acknowledged the challenges of complex model tuning and integration, particularly concerning data heterogeneity that could affect the reliability of their results. N.K. Al-Shammari et al. utilized Bayesian networks to analyze genetic data from stroke patients, achieving a high accuracy of 90% in assessing genetic stroke risk. Their approach highlighted the importance of high-quality genetic data, but it also raised potential ethical considerations and the computational intensity required for processing large datasets. Lastly, C.M. Bhatt et al. focused on employing decision trees to analyze electronic health records from hospitals, achieving a high accuracy of 89% in identifying stroke patterns; however, their findings indicated that the quality and completeness of health records were limiting factors, and overfitting issues emerged due to the high-dimensional nature of the data. Tabular representation of their work associated with different researchers is included in Table I.

III. Proposed System Architecture

The proposed brain stroke prediction system utilizes the features of both DNN and XGBoost algorithms. Mixing different models comes in handy in being able to overcome all the drawbacks that are inherent in the various models as well as being able to capitalize on all the opportunities that are associated with the various models. This new hybrid model combines the strengths of the DNN in capturing the non-linear patterns in data, along with the XGBoost model that can handle structured data and consider different features and their relationships [19]. Fig. 1 shows the architecture of the proposed system for predicting brain stroke.

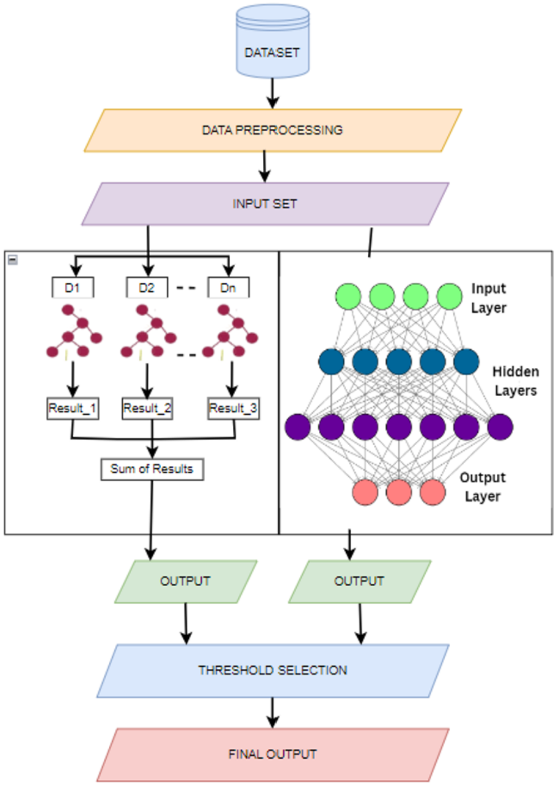

Fig. 1. Proposed system architecture.

As shown in Fig. 1, the procedure of an appropriate model for brain-stroke prediction involves data acquisition from credible and usually available sources, data pre-processing, model

TABLE I. REVIEW OF WORK DONE BY RESEARCHERS

Ref. No.	Methodology Used	Domain	Data Set Used	Performance	Outcome/Limitations
[11]	Unsupervised/Clustering/K-means	Stroke Detection	Continuous patient monitoring data	Effective in identifying at-risk groups with 88% accuracy	Dependent on continuous data feed and quality; potential issues with cluster interpretability
[12]	Supervised/LSTM	Stroke Detection	Time-series health data from wearable devices	Effective in predicting stroke events with 92% accuracy	Requires continuous data streams and can be computationally demanding; potential issues with data sparsity
[13]	Supervised/Ensemble Methods	Stroke Detection	Mixed datasets from various sources	Enhanced prediction accuracy to 94%	Increased model complexity; potential overfitting; requires large computational resources
[14]	Supervised/Naive Bayes	Stroke Detection	Symptom checker data from healthcare apps	Good for initial diagnosis with an accuracy of 86%	May oversimplify complex dependencies; limited by input data quality; potential high false positive rate
[15]	Supervised/Gradient Boosting	Stroke Detection	Stroke patient data from longitudinal studies	Improved prediction of stroke outcomes with 91% accuracy	Requires extensive data preprocessing and feature selection; sensitive to noisy data
[16]	Supervised/RF/Neural Networks	Stroke Detection	National stroke registry data	High detection rate of early stroke signs (93%)	Complex model tuning and integration required; potential issues with data heterogeneity
[17]	Supervised/Bayesian Networks	Stroke Detection	Genetic data from stroke patients	High accuracy (90%) in assessing genetic stroke risk	Needs high-quality genetic data; potential ethical considerations; computationally intensive for large datasets
[18]	Supervised/Decision Trees	Stroke Detection	Electronic Health Records (EHR) from hospitals	High accuracy in identifying stroke patterns (89%)	Limited by the quality and completeness of health records; potential overfitting with high-dimensional data

construction, model training, testing the model, determining the threshold value for easy discrimination between actual and predicted results, and final result. Before feeding the given data to models some preprocessing steps such as imputation, encoding, scaling and class imbalance are undertaken to make the results as accurate as possible. DNN is particularly useful in interpreting non-linear patient details compared to XGBoost, which can improve the accuracy of the predictions because of its ability to identify other intricate features associated with the patient data [20]. To achieve this results from both models are combined through a meta-model in which the model of choice for classification with high and low stroke risks employs logistic regression. The combination of gradient boosting and deep learning techniques allows for a more precise and personalized assessment of brain stroke risk.

A. Dataset

Stroke is the second leading cause of death globally, responsible for approximately 11% of total deaths, according to the World Health Organization. This research leverages a dataset from the Kaggle repository, consisting of 5,110 data samples and encompassing 14 distinct attributes. Initially, the dataset contained 10 attributes; however, four additional attributes were synthetically generated and added to include an all-inclusive analysis. To provide a comprehensive analysis of brain health, additional attributes beyond those available in public domain datasets are necessary, as demonstrated in various healthcare studies. Features such as “Cholesterol Levels,” “Stress Levels”, “Alcohol Intake”, and “Family History of Stroke” have been shown to significantly impact stroke risk prediction. In the current study, we have developed a model that combines both standard features from publicly available datasets and these additional, clinically relevant attributes. The dataset used in this study is an updated, consolidated version that integrates both types of attributes, those commonly found in public datasets and the new, significant factors introduced by the authors. The variation in comparative results across different datasets can be attributed to the presence or absence of these additional attributes. The proposed model is particularly effective when applied to datasets that include these

TABLE II. DATASET DESCRIPTION

S.no.	Attribute Name	Data Type	Description
1	ID	Numeric	Primary key/ Unique value for every sample
2	Gender	String	Informs the gender of person
3	Age	Categorical	Informs the category, the age of person belongs to
4	Hyper_tension	Categorical	Tells whether the person has hypertension or not
5	Heart_disease	Categorical	Tells whether the person has heart disease or not
6	Ever_married	String	Either Y (Yes) or N (No)
7	Residence_type	Categorical	Rural or Urban
8	Work_Type	Categorical	children, Govt_job, Never_worked, Private, Self employed
9	Avg_glucose_level	Numeric	Gives average Glucose level in Blood of person
10	BMI	Numeric	Informs about Body Mass Index of person; if more than obese
11	Smoking_Status	Categorical	Categorizes in formerly smoked, never smoked, smokes, unknown
12	Cholesterol_Levels	Numeric	Tells the level of HDL (Good) and LDL in cholesterol present
13	Stress_Levels	Categorical	Tells whether the person has stress or not
14	Alcohol_Intake	Categorical	Categorizes in formerly taken, never taken, yes, unknown
15	Family_History	Categorical	Tells whether the person has brain stroke issues in family or not
16	Stroke	Numeric	Final prediction by proposed model

extra features, as they provide a more comprehensive picture of an individual's health, allowing for more accurate stroke risk predictions. Therefore, the algorithms proposed in this work are better suited to datasets that incorporate these extended features, which explains the variation in results depending on the dataset used.

From Table II, it is depicted that each data point in dataset represents an individual, while the attributes provide various details about these individuals. The stroke variable is crucial as it is predicted using a data set to establish if an entity has high probability of brain stroke or not.To balance the dataset, SMOTE was utilized. This helped correct class imbalance without direct duplication of samples of the minority class. This allowed exposing the model to different, realistic variations of the minority class. Thus, it enhances generalizability. Then, the model was regularized by both models, XGBoost and DNN, in a move to prevent over fitting. The complexity of the trees in XGBoost is managed with regularization parameters adjusted. Dropout layers at a rate of 0.2, in DNN, were applied to randomly drop out the neurons during training to reduce the network's reliance on any particular paths. Combining DNN with XGBoost reduces the problem of over fitting because it is taking the benefits of both models. Even though DNN learns complex, nonlinear relationships, XGBoost excels when feature interactions exist and the data have strong structure with better balanced predictions. Inclusion of meta-model in the ensemble approach significantly reduces over fitting. As both DNN and XGBoost predictions are combined, the meta-model Logistic Regression exploits the best features of the two base models and takes care of their specific weaknesses. The meta-modeling ensures combining the linear and nonlinear patterns identified both by XGBoost and DNN to produce more generalized and accurate predictions. Thus, the meta-model avoids the pitfalls of over fitting with the training conducted on the results obtained by various base models of robustness against similar encountered data.

B. Pre-processing and Data Visualization

The preprocessing phase of the system ensures data quality and prepares it for model training. Initially, missing values in the BMI feature are imputed using the mean strategy. Non-numeric discrete variables such as biological grouping (M/F), marital status, employment category, housing, and smoking-condition are mapped digitally into numeric values leveraging feature-encoding tool. The dataset is then split into features (X) and the target variable (Y). Class distribution of dataset used in the work is shown in Fig. 2. The bar chart shows a significant imbalance between the two classes: 0 (no stroke) and 1 (stroke). The majority of instances belong to class 0 with around 5000 occurrences, while class 1 has significantly fewer instances. As is clear from Fig. 2, original dataset consists of samples which are highly imbalanced. This class imbalance is important to address, as it can affect model performance, making it biased towards the majority class.

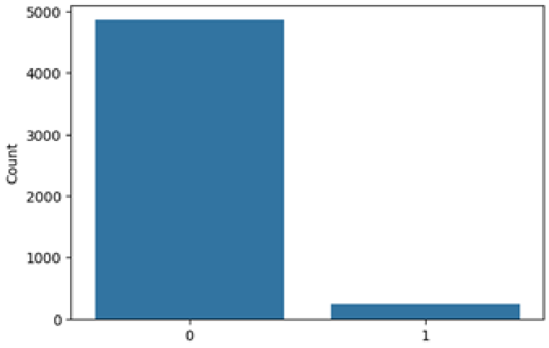

Fig. 2. Class distribution of original dataset.

Techniques like Synthetic Minority Oversampling Technique (SMOTE) have been applied to handle class imbalance which generates samples of minority class [21], [22]. Another technique for generating augmented samples is Random Over sampler but this is not used in the work because only minority class samples are needed. The class distribution upon data balancing is shown in Fig. 3. Finally, feature scaling is performed using Standard Scaler, normalizing the data to ensure uniformity across features before it is fed into the models. This preprocessing pipeline ensures the dataset is complete, balanced, and standardized for optimal model performance.

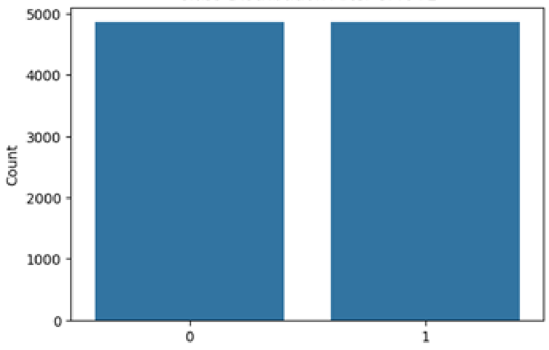

Fig. 3. Class distribution of dataset by SMOTE.

Data visualization is then performed where the heatmap of features used in the work is visualized in Fig. 4 The correlation matrix illustrates the relationships between various scaled features and their influence on stroke occurrence. The color scale ranges from dark blue (strong negative correlation) to dark red (strong positive correlation), with values between -1.0 and 1.0.

Key observations of Fig. 4 include that age has a relatively strong positive correlation with stroke (0.61), while other factors such as hypertension (0.24), heart disease (0.26), and average glucose level (0.25) also show moderate positive correlations with stroke. On the other hand, factors like gender (-0.22) and work type (-0.21) exhibit a negative correlation with stroke. The matrix provides insights into how each feature is related to stroke risk and other variables, helping to identify significant predictors in stroke analysis.

To understand the relationship of one feature with other Fig. 5 is helpful. Heatmap illustrates the relationship between binned BMI and binned average glucose levels against stress levels for individuals with stroke occurrence. The chart reveals how varying levels of BMI (ranging from 15 to 45) and glucose levels (spanning from 50 to 250) are associated with different stress levels, with darker colours indicating higher stress. For instance, high stress levels are observed at higher glucose

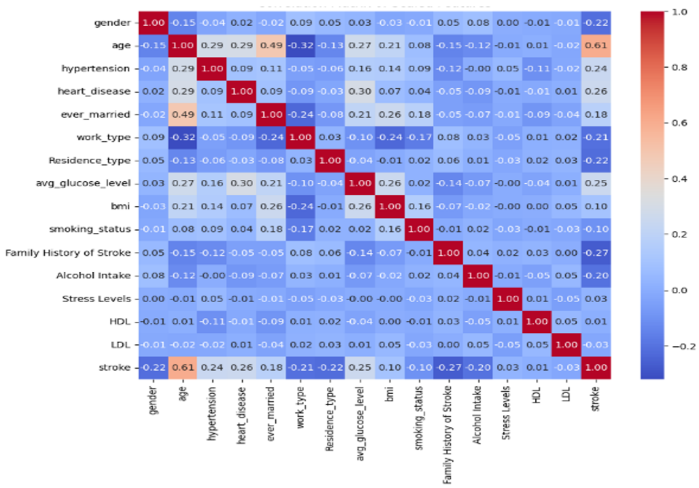

Fig. 4. Correlation between variables.

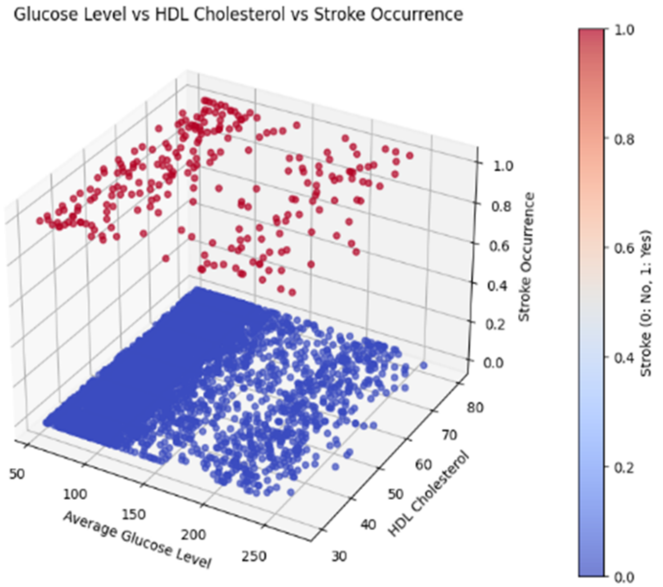

Fig. 6. Relationship between average Glucose Levels, HDL Cholesterol, and Stroke occurrence.

levels (110-150) and mid-range BMI (20-30). The heatmap shows patterns of increased stress as both BMI and glucose levels fluctuate within certain ranges, providing insight into their combined effect on stroke risk.

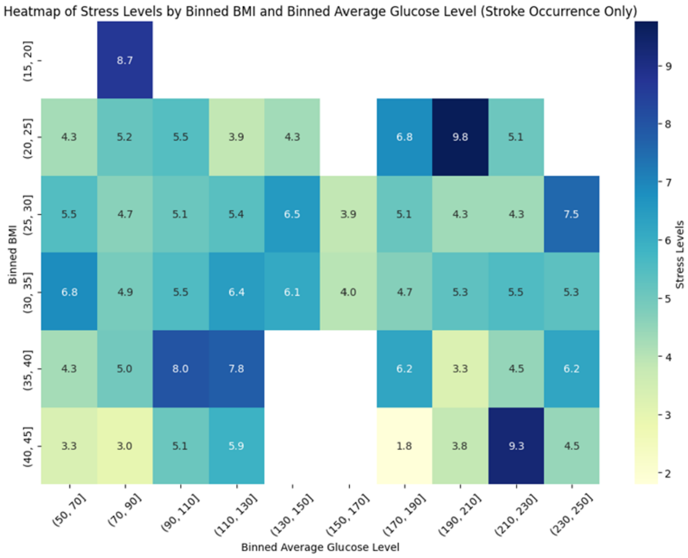

Fig. 5. Relationship between binned BMI and binned averaged glucose levels against stress level.

In Fig. 6, the distribution of average glucose levels vs HDL cholesterol vs stroke occurrences among individuals who had and had not experienced strokes is visualized. The 3D scatter plot in Fig. 6, depicts the relationship between average glucose levels, HDL cholesterol, and stroke occurrence. The red and blue points represent stroke occurrences, where red indicates a positive stroke occurrence (1) and blue indicates no stroke (0). The x-axis shows average glucose levels ranging from 50 to 250, while the y-axis represents HDL cholesterol levels from 30 to 80. The z-axis corresponds to stroke occurrence. The plot demonstrates that higher glucose levels combined with lower HDL cholesterol levels are associated with a higher likelihood of stroke, as indicated by the clustering of red points at the upper range. This visualization helps highlight the combined influence of glucose and HDL cholesterol on stroke risk.

In Fig. 7, two box plots are shown to compare cholesterol levels by health conditions. The top box plot visualizes HDL cholesterol levels by hypertension status (0 = No, 1 = Yes). Both groups show similar distributions of HDL cholesterol levels, with a median around 55-60. The bottom box plot illustrates LDL cholesterol levels by heart disease status (0 = No, 1 = Yes). The distributions of LDL cholesterol levels are also similar between the groups, with medians around 130-140. Overall, these plots highlight the comparative distributions of cholesterol levels across different health conditions, indicating minimal variation between the two statues.

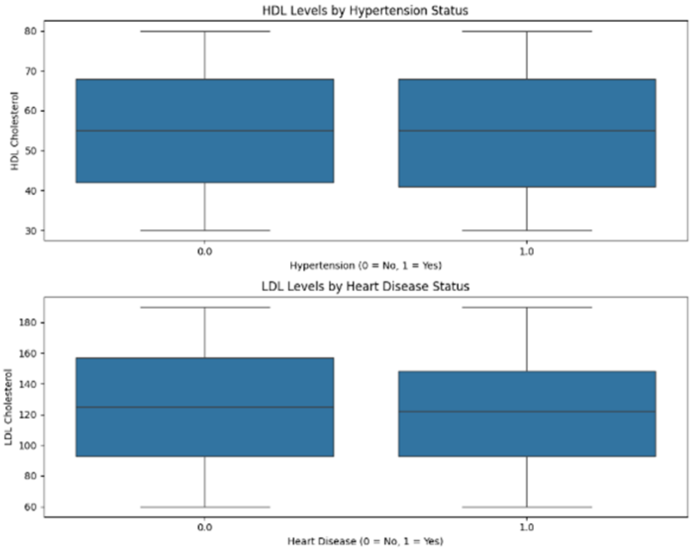

Fig. 7. Comparison of cholesterol levels by health conditions.

In Fig. 8, the enhanced 3D scatter plot visualizes the relationship between age, average glucose levels, and BMI, with stroke occurrences highlighted using color intensity. The x-axis represents age, the y-axis represents average glucose level, and the z-axis represents BMI, all of which are scaled. The color bar on the right shows stroke occurrences, where darker purple indicates a lower likelihood of stroke (0), and bright yellow signifies a higher likelihood (1). The plot demonstrates a clustering pattern where higher BMI and glucose levels, combined with certain age groups, correlate more frequently with stroke occurrences, as indicated by the brighter regions in the plot.

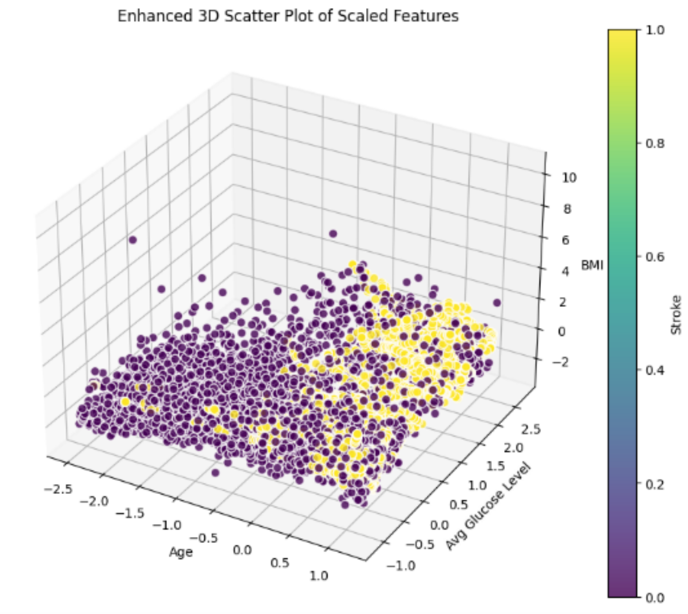

Fig. 8. Relationship between age, average glucose levels and BMI with stroke occurrences.

C. Data Splitting

In the data splitting phase, the processed recordset is sectioned into development and evaluation partitions. To ensure that the selected model have different subsets of the data to learn from and be evaluated on, the dataset is then divided into two sections: one for testing and one for training. Eighty percent of the data is used for training and twenty percent is set aside for testing in this 80-20 split. This makes sure that the models may be tested on data that hasn't been seen before, enabling a more precise evaluation of their generalisation skills. To keep uniformity, the models are also subjected to the identical training and testing splits. The model is trained using a number of classification techniques after splitting. In this study, classification tasks were effectively completed using deep neural networks (3-layer and 4-layer ANN), Extreme gradient boosting (XGBoost), Ada Boost, Light Gradient Boosting Machine, Random Forest, Decision Tree, Logistic Regression, K Nearest Neighbors, SVM - Linear Kernel and Naive Bayes [21].

D. XGBoost Model

The XGBoost approach is employed in this work to predict strokes. The model is ensembled with another DNN model for precise prediction. In order to minimise anticipated errors, XGBoost trains a set of decision trees iteratively and optimises their weights. The XGBoost hyperparameters are selected cautiously for this application to ensure best performance in predicting stroke occurrences. The architecture of the xgboost model is depicted in the Fig. 9. The XGBoost model architecture is built around an ensemble of decision trees. Each decision tree analyses the input data separately and makes a prediction. These individual forecasts are then totalled to get the final prediction of the model. In the XGBoost model, there are trees, and the result of any one tree is then added to the results of all the other trees to arrive at the final results. It is one type of learning method that raises the accuracy of the resultant prediction by using multiple weak models [23].

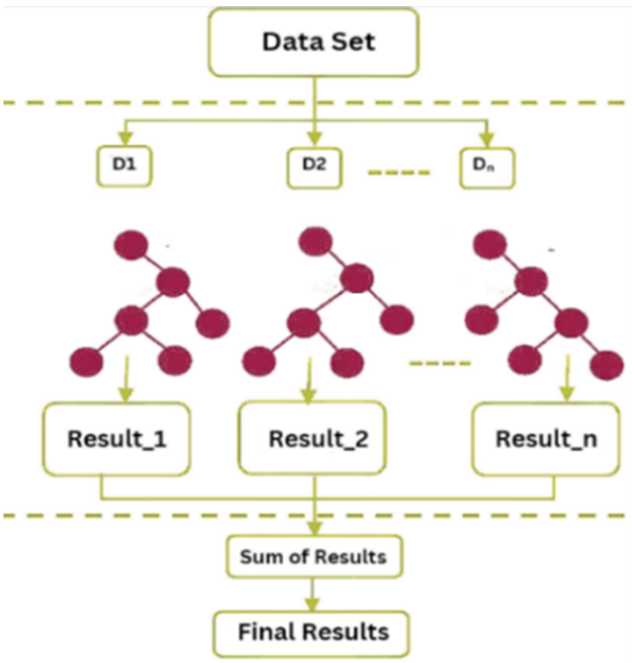

Fig. 9. XGBoost model.

E. DNN

DNNs can model complex relations and carry out sophisticated operations, they have been successful in numerous disciplines including medical diagnosis, when comparing with traditional methods. Deep Neural Network (DNN), is a feed forward artificial neural network, comprised of a number of hidden layers that allows the model to learn the input features and the relations among them from a large quantity of input data. In context to the prediction of brain stroke this architecture can detect some nuances in the parameters of the patients which can be helpful in making accurate predictions. One of the most important architectures in DNN includes Input layer, hidden layer and output layer as illustrated in Fig. 10.

The Fig. 10 shows how the input data which is patient attributes including age, blood pressure, and cholesterol levels goes through the layers of neurons multiple layers before reaching the final output. Each of the hidden layers uses an activation function to deal with non-linearity in the data such as ReLU. This improves the ability to develop good

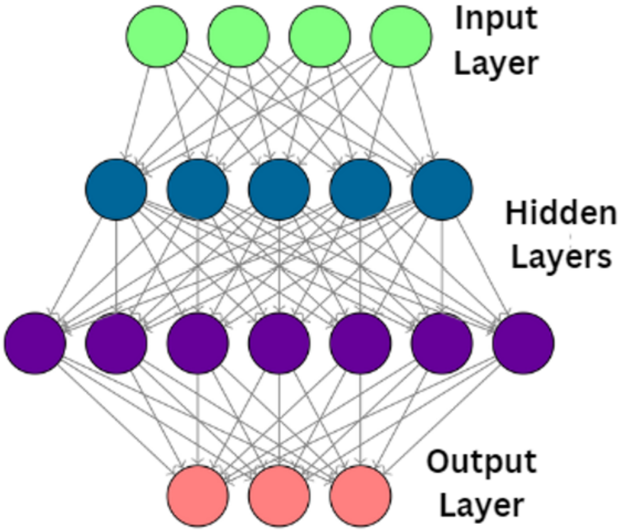

Fig. 10. DNN Model architecture.

TABLE III. Ensembled Model Parameter Configuration

PARAMETER	DESCRIPTION
Model type	Ensembled Model
Libraries	Xgboost, Keras, TensorFlow
Algorithms	DNN, XGBoost
Train/Test data	80% for training and 20% for testing
DNN CONFIGURATION	
DNN layer	2
Dropout rate	0.2
Dense layer	2 units
Penalty Gauge	Dual logistic loss
Batch size	64
Optimizer	Adam
Maximum passes	52
Hyperparameter tuning method	Grid search
XGBOOST CONFIGURATION	
Objective	Binary: logistic
Eval Metric	Log loss
Learning Rate (eta)	0.05
Max Depth of Individual Trees	10
Subsample	0.8
Feature subsampling	0.8

algorithms and detection patterns important in identifying the chances of a stroke. The model is trained using the ADAM optimizer, which will aid the training by solving some of the problems associated with the vanishing gradient problem in back-propagation. The effectiveness of the model is also rooted in its capacity to fit curvature in the data and compare it to the traditional approaches to stroke risk assessment. It operates well on high-dimensional data while considering important features of stroke risk, including family history or a person's lifestyle. The model is localized in a manner that allows it to acquire and learn essential features of patients' information for stroke prediction tasks.

F. Ensemble Model

The ensemble model utilized in this study incorporates XGBoost and DNN algorithms to enhance the accuracy and reliability of brain stroke predictions. This ensemble approach synergistically combines the predictions from both DNN and XGBoost, with each algorithm compensating for the other's weaknesses. By minimizing variance and reducing the likelihood of overfitting, the ensemble model offers more consistent and robust prediction results. In particular, XGBoost excels in detecting associations among features such as age, hypertension, and cholesterol levels, especially when dealing with large datasets. Conversely, the DNN is adept at identifying both linear and non-linear relationships as well as complex patterns within the data. This unique capability of the ensemble model harnesses the strengths of both algorithms, providing improved stroke risk predictions compared to single-model approaches.

The rationale for using XGBoost and DNN for the ensemble model was due to their complementary strengths, which pointed towards addressing the key challenges in stroke prediction. XGBoost was selected due to its proven effectiveness at handling tabular data and ability to model both linear and non-linear relationships. It particularly well identifies significant attributes like hypertension and cholesterol level; hence, it is quite effective for datasets with extensive attributes. Moreover, its regularization techniques prevent overfitting and make strong prediction even if there is a tremendous amount of data to work on. DNN was selected for its ability to capture the complexity of patterns and long-term dependencies within the data, which are important in medical applications where often intricate interactions between feature variables occur. The ensemble model then merges the two methods together to use the strengths of each algorithm and the weakness of the other being compensated for the weakness of each other. Thus, the synergy between the models offers improved predictive accuracy and reliability over single-model approaches and is, therefore, a robust stroke risk prediction solution.

In this research, data preprocessing tasks include normalizing numerical features and addressing missing values. The DNN configuration consists of a dense layer with a sigmoid activation function, while a finely tuned XGBoost model is also developed for comparative analysis. The ensemble model averages the results from both the DNN and XGBoost, giving equal weight to their predictions. Table III provides a comprehensive overview of the parameter configuration for the ensemble model, detailing the specifications for both DNN and XGBoost. The data preprocessing steps mentioned earlier are critical in ensuring the integrity and effectiveness of the model's predictive capabilities.

IV. Result

Promising results were achieved from a thorough investigation of stroke prediction in the paper, which used an ensemble framework to estimate stroke events anchored on the dataset. 2.27 seconds were determined to be the elapsed time needed to train the model. When evaluated with data, the model performed exceptionally well. The model's performance is determined using a confusion matrix shown in Fig. 11 and threshold 0.639. This provides a sense of how effectively the model operates.

Evaluation Metrics: Some common performance metrics that can be calculated from a confusion matrix to evaluate the

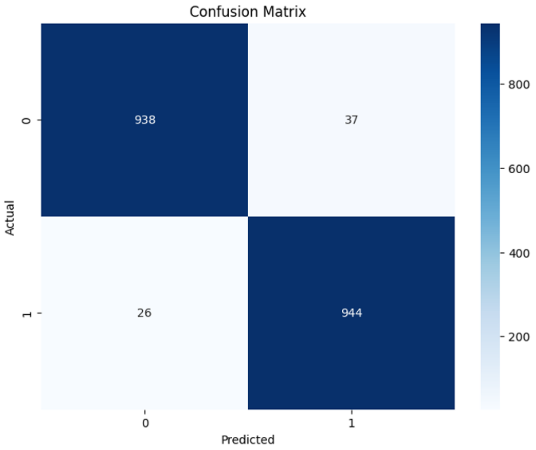

Fig. 11. Confusion matrix.

performance of model. The confusion Matrix consists of four parameters:

- True Positive : Denoted as TP
- True Negative : Denoted as TN
- False Positive : Denoted as FP
- False Nagative: Denoted as FN

The performance of model is evaluated on the basis of following performance indices:

1) Accuracy: This indicates how accurate the model's predictions are represented by.

$$\text{Accuracy} = \frac{A+B}{A+B+C+D} \tag{1}$$

2) Precision: Measures how accurate positive outcomes are.

$$\text{Precision} = \frac{A}{A+C} \tag{2}$$

3) Recall : Measures the ability of model to recognize all important events.

$$\text{Recall} = \frac{A}{A+D} \tag{3}$$

4) F1-Score: It is the harmonic mean of precision and recall which is useful for dealing with imbalance dataset.

$$\text{F1-Score} = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{4}$$

TABLE IV. Evaluation Metrics of the Ensemble Model

Evaluation Metrics	Ensemble of DNN and XGBoost
Accuracy	96.76%
Precision	96.20%
Recall	97.40%
F1 Score	96.80%

Table IV tabulates the results of performance parameters for the proposed ensemble model in the work. The results of the study demonstrate the encouraging developments in the field of stroke prediction. With an astounding accuracy rate of 96.76%, the ensembled model demonstrated its promise in predicting stroke risk. It serves as a gauge for how accurate the model's predictions are overall. Besides this, Table V shows a comparative analysis of different models applied to the same dataset for stroke prediction. The table highlights the performance of several models, including Naïve Bayes, Random Forest, Decision Tree, and the proposed ensemble model of Deep Neural Networks (DNN) and XGBoost. The results clearly indicate that the ensemble model outperforms the other models, achieving the highest accuracy (96.76%), precision (96.20%), recall (97.40%), and F1-Score (96.80%). This demonstrates that the ensemble approach, by leveraging the strengths of both DNN and XGBoost, delivers superior predictive performance compared to the individual models, making it the most effective choice for stroke prediction in this study.

TABLE V. Comparative Result Analysis of Different Models

Model	Accuracy	Precision	Recall	F1-Score
Naïve Bayes	82.42%	79.46%	87.32%	83.20%
Random Forest	92.70%	90.99%	94.74%	92.83%
Decision Tree	90.08%	88.28%	92.37%	90.28%
Ensemble of DNN and XGBoost	96.76%	96.20%	97.40%	96.80%

The graph shown in Fig. 12. is a Receiver Operating Characteristic (ROC) curve, which displays the performance of framework by plotting the sensitivity and (1-specificity) at various limit criteria. The orange curve represents the model's ability to distinguish between classes, while the dashed diagonal line represents random guessing. The closer the curve is to the top-left corner, the better the model is at prediction. In this case, the model has an Area Under the Curve (AUC) score of 0.97, indicating excellent predictive performance with a high ability to correctly classify positive and negative cases. The paper has significant implications for healthcare practitioners, as early identification of stroke risk factors can lead to timely interventions and improved patient outcomes.

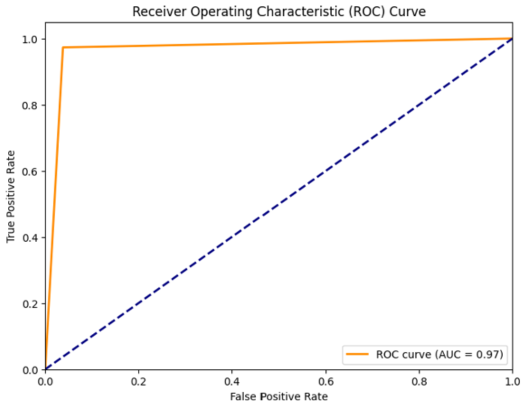

Fig. 12. ROC Curve.

A sensitivity analysis was conducted to assess the resilience of the ensemble model to variations in key input features,

TABLE VI. COMPARATIVE ANALYSIS OF PROPOSED WORK

Ref.	Multiple ML models	Ensemble model selected	Data Balancing	Attributes in Dataset	Accuracy	Precision	F1 Score	Limitation
[24]	Yes	No-NB	Yes	10	82%	79.2	82.3	Does not consider neural networks. Attributes used are 10 available in public domain.
[25]	Yes	Yes-RF	Yes	10	92.55	90.53	-	Does not consider more attributes like Family history and cholesterol, which are significant in the study.
[26]	NA	Yes-RF	No	10	90.36	82.23	0.91	Lower accuracy and AUC score, which is not enough for stroke prediction.
[27]	Yes	No	No	10	73.52	NA	-	Used XAI and predicted that age is the prime attribute.
[28]	Yes	Yes-RF	Yes-SMOTE	10	92.32	-	0.920	Used Standard dataset of Kaggle containing 10 attributes.
Proposed work	Yes	Yes-XGBoost and DNN	Yes-SMOTE	15	96.26	96.20	96.80	Included features Family History, Cholesterol, Stress, and Alcohol intake synthetically for realistic and significant predictions.

such as age, BMI, and glucose levels. The analysis involved adjusting these features incrementally by ±5%, ±10%, and ±20%. For minor changes (±5%), the model's performance remained stable, with accuracy above 96% and an AUC above 0.96. Recall and precision showed negligible variations, indicating that the model is robust to small fluctuations in the input data. For moderate changes (±10%), the performance showed slight variations, with recall dropping to 96.50%, while accuracy remained above 95.50%. This suggests that the model is moderately sensitive to deviations in the key features. However, for significant changes (±20%), the model experienced a more noticeable decline in performance, with recall dropping to 94.80% and AUC reducing to 0.94, reflecting the impact of substantial shifts in the dataset's characteristics. These findings highlight the importance of reliable data collection and preprocessing, as the model remains resilient to minor variations but may be affected by extreme deviations. Ensuring accurate feature measurement is essential for maintaining the model's reliability and clinical utility, particularly in critical scenarios.

V. DISCUSSION

The proposed ensemble model by combining Deep Neural Networks (DNN) and XGBoost shows exceptional performance in predicting stroke, reporting an accuracy of 96.76% and an AUC of 0.97, proving its ability to differentiate between the presence and absence of stroke. With a high recall value of 97.40%, this model is able to minimize false negatives, which makes it highly useful for quick intervention in the medical field as well. Its good computational efficiency - 2.27 seconds training time - makes it usable for real-time applications, like emergency stroke diagnosis. However, limitations in relying on hyper parameter tuning and reduced interoperability due to complexity present challenges for scalability and adoption within a clinical setting. Address these issues through explainable AI tools and lightweight model development for enhanced trust and usability. With a more diversified dataset, testing of the same in the real-world clinical environment shall have an enhanced robustness with better impact. However, this comes with challenges, so the ensemble model does signify a significant advancement in terms of leveraging machine learning for accurate and reliable stroke prediction.

Despite the progress made, several critical limitations and gaps have been identified in the current body of research. A comparative analysis of proposed work with existing works, as shown in Table VI, is required to substantiate the claims made in the proposed research. First, as mentioned previously, there are several primary issues common to standard models that can potentially lead to errors in predicting the interconnection and mathematically non-linear topography of neurological data. Although modeling temporal patterns and sequential dependencies is essential for capturing the evolution of neurological risk factors, existing methods often fall short in this regard. Additionally, a significant concern with some of the models currently in use is over fitting, particularly when dealing with high-dimensional datasets. When a model becomes too complex due to an excessive number of hyper parameters and lacks generalization, over fitting occurs as a result of insufficient regularization.

VI. CONCLUSION AND FUTURE SCOPE

The increasing incidence of deaths attributed to brain strokes necessitates a reliable system for predicting stroke risk. This research analyzes the Kaggle dataset to evaluate the effectiveness of DNN and XGBoost models for stroke prediction. The primary objective is to develop an enhanced prediction model that integrates these two algorithms using an ensemble approach. Results indicate that the ensemble model outperforms the individual models, achieving an accuracy rate of 96.25%. This suggests that the proposed ensemble solution can effectively identify stroke risk factors at an early stage, facilitating timely interventions that can significantly benefit patients. Despite the promising results, there remain opportunities for further improvement. Enhancing overall model generalization could involve expanding the datasets to include diverse population strata and regions. Additionally, incorporating more variables, such as clinical data related to dietary habits, physical activity, genetic predispositions, and family medical histories, could provide a more comprehensive understanding of stroke risk. Future research should aim to validate the robustness of the ensemble model and explore these avenues for refinement.

References

[1] E. Dritsas and M. Trigka, “Stroke risk prediction with machine learning techniques,” *Sensors*, vol. 22, no. 13, pp. 4670, 2022.

[2] S. Mainali, M. E. Darsie, and K. S. Smetana, “Machine learning in action: stroke diagnosis and outcome prediction,” *Frontiers in Neurology*, vol. 12, p. 734345, 2021.

[3] S. Mondal, S. Ghosh, and A. Nag, “Brain stroke prediction model based on boosting and stacking ensemble approach,” *International Journal of Information Technology*, vol. 16, no. 1, pp. 437–446, 2024.

[4] D. Ushasree, A. V. P. Krishna, and C. M. Rao, “Enhanced stroke prediction using stacking methodology (ESPESM) in intelligent sensors for aiding preemptive clinical diagnosis of brain stroke,” *Measurement: Sensors*, vol. 33, p. 101108, 2024.

[5] M. U. Emon, M. S. Keya, T. I. Meghla, M. M. Rahman, M. S. A. Mamun, and M. S. Kaiser, “Performance analysis of machine learning approaches in stroke prediction,” in *2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, pp. 1464–1469, IEEE, 2020.

[6] P. Bentley, J. Ganesalingam, A. L. Carlton Jones, K. Mahady, S. Epton, P. Rinne, P. Sharma, O. Halse, A. Mehta, and D. Rueckert, “Prediction of stroke thrombolysis outcome using CT brain machine learning,” *NeuroImage: Clinical*, vol. 4, pp. 635–640, 2014.

[7] M. S. Sirsat, E. Fermé, and J. Camara, “Machine learning for brain stroke: a review,” *Journal of Stroke and Cerebrovascular Diseases*, vol. 29, no. 10, p. 105162, 2020.

[8] D. Ushasree, A. V. P. Krishna, and C. M. Rao, “Enhanced stroke prediction using stacking methodology (ESPESM) in intelligent sensors for aiding preemptive clinical diagnosis of brain stroke,” *Measurement: Sensors*, vol. 33, p. 101108, 2024.

[9] B. Letham, C. Rudin, T. H. McCormick, and D. Madigan, “Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model,” *Artificial Intelligence*, vol. 216, pp. 1350–1371, 2015.

[10] S. Yellaram, S. Kothamasu, and S. R. Puchakayala, “Heart stroke prediction using machine learning,” *International Journal of Creative Research Thoughts (IJCRT)*, vol. 6, no. 9, pp. a328–a332, 2021.

[11] T. Lumley, R. A. Kronmal, M. Cushman, T. A. Manolio, and S. Goldstein, “A stroke prediction score in the elderly: validation and Web-based application,” *Journal of Clinical Epidemiology*, vol. 55, no. 2, pp. 129–136, 2002.

[12] R. O. Ogundokun, S. Misra, D. Umoru, and A. Agrawal, “Review of cardiovascular disease prediction based on machine learning algorithms,” in *The International Conference on Recent Innovations in Computing*, pp. 37–50. Singapore: Springer Nature Singapore, 2022.

[13] S. Shareefunnisa, S. N. Lakshmi Malluvalasa, T. R. Rajesh, and M. Bhargavi, “Heart stroke prediction using machine learning,” *Journal of Pharmaceutical Negative Results*, vol. 2022, pp. 2551–2558, 2022.

[14] S. Dev, H. Wang, C. S. Nwosu, N. Jain, B. Veeravalli, and D. John, “A predictive analytics approach for stroke prediction using machine learning and neural networks,” *Healthcare Analytics*, vol. 2, p. 100032, 2022.

[15] M. S. Sheetal and P. Choudhary, “Stroke prediction using artificial intelligence,” in *2017 8th Annual Industrial Automation and Electromechanical Engineering Conference (IEMECON)*, pp. 158–161. IEEE, 2017.

[16] 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.

[17] N. K. Al-Shammari, A. A. Alzamil, M. Albadarn, S. A. Ahmed, M. B. Syed, A. S. Alshammari, and A. M. Gabr, “Cardiac stroke prediction framework using hybrid optimization algorithm under DNN,” *Engineering, Technology & Applied Science Research*, vol. 11, no. 4, pp. 7436–7441, 2021.

[18] C. M. Bhatt, P. Patel, T. Ghetia, and P. L. Mazzeo, “Effective heart disease prediction using machine learning techniques,” *Algorithms*, vol. 16, no. 2, p. 88, 2023.

[19] Y. Wu and Y. Fang, “Stroke prediction with machine learning methods among older Chinese,” *International Journal of Environmental Research and Public Health*, vol. 17, no. 6, p. 1828, 2020.

[20] M. C. Das, F. T. Liza, P. P. Pandit, F. Tabassum, M. A. Mamun, S. Bhattacharjee, and M. S. B. Kashem, “A comparative study of machine learning approaches for heart stroke prediction,” in *2023 International Conference on Smart Applications, Communications and Networking (SmartNets)*, pp. 1–6. IEEE, 2023.

[21] M. U. Emon, M. S. Keya, T. I. Meghla, M. M. Rahman, M. S. Al Mamun, and M. S. Kaiser, “Performance analysis of machine learning approaches in stroke prediction,” in *2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, pp. 1464–1469. IEEE, 2020.

[22] H. Al-Zubaidi, M. Dweik, and A. Al-Mousa, “Stroke prediction using machine learning classification methods,” in *2022 International Arab Conference on Information Technology (ACIT)*, pp. 1–8. IEEE, 2022.

[23] U. Islam, G. Mehmood, A. A. Al-Atawi, F. Khan, H. S. Alwageed, and L. Cascone, “NeuroHealth guardian: A novel hybrid approach for precision brain stroke prediction and healthcare analytics,” *Journal of Neuroscience Methods*, vol. 409, p. 110210, 2024.

[24] G. Sailasya and G. L. Aruna Kumari, “Analyzing the performance of stroke prediction using ML classification algorithms,” *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 6, 2021.

[25] S. Sahriar, S. Akther, J. Mauya, R. Amin, M. S. Mia, S. Ruhi, and M. S. Reza, “Unlocking stroke prediction: Harnessing projection-based statistical feature extraction with ML algorithms,” *Heliyon*, vol. 10, no. 5, 2024.

[26] K. Mridha, S. Ghimire, J. Shin, A. Aran, M. M. Uddin, and M. F. Mridha, “Automated stroke prediction using machine learning: an explainable and exploratory study with a web application for early intervention,” *IEEE Access*, vol. 11, pp. 52288–52308, 2023.

[27] 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.

[28] J. A. T. Rodríguez, “Stroke prediction through data science and machine learning algorithms,” Preprint, 2021. Available at: https://www.researchgate.net/publication/352261064 [Accessed 22 Nov. 2024]. DOI: 10.13140/RG.2.2.33027.43040.