SHAP-Weighted Hybrid Ensemble for Interpretable Stroke Risk Prediction Using Tabular Clinical Data

Author: Abu Bakar Shabbir

School of Systems and Technology, University of Management and Technology, Lahore 54770, Pakistan

Email: abubakarshabbir64@gmail.com

Abstract

Stroke remains one of the main causes of death and long-term disability worldwide. Therefore, early and reliable prediction is necessary to allow timely intervention. To address this problem, a SHAP-weighted hybrid ensemble framework is proposed in this study. The framework integrates complementary machine learning models, gradient-boosted decision trees via XGBoost, bagged decision trees using Random Forest, and an attention-based deep tabular learner, TabNet, to deliver clinically interpretable and reliable stroke risk prediction from structured clinical data. The presented model leverages Shapley Additive Explanations (SHAP) to derive adaptive feature-driven ensemble weights, enabling complementary model fusion that balances strong discriminative performance with clinician-centric interpretability. Full-scale preprocessing of the clinical data with KNN imputation, categorical encoding, SMOTETomek resampling for class balance, etc., realized the data quality required for subsequent modeling. Benchmark experiments performed on the Brain Stroke dataset (a public dataset with n=4981) demonstrated that the hybrid ensemble is superior to each of the base learners, with the highest reported metrics of ROC-AUC, F1, and recall being 0.8099, 0.2165, and 0.42, respectively. Hence, the sensitivity and precision were balanced in this study. Moreover, the employment of SHAP as a forecasting method considerably helps in understanding the model's internal structure. Just to give you an example, as a result of the analysis, the model's output was found to be most impacted by age, average glucose level, and BMI, and these three not only go a step further but are also supported by clinical findings. The proposed framework moves the needle on the front of interpretable AI by reconciling the performance of black-box models with the transparency requirements of clinical decision-support systems. In other words, the SHAP-weighted ensemble is a stroke risk estimation tool that is open to the future, where AI is used in medical settings for the good of the patient and with minimized chances of unintended consequences.

Keywords: Stroke prediction, SHAP, Hybrid ensemble, Explainable AI, TabNet, XGBoost, Random Forest, Interpretable machine learning, Clinical decision support, Data imbalance correction

1. Introduction

Stroke remains a major cause of death and long-term disability worldwide, accounting for almost 12% of all deaths annually. Although there has been significant progress in medical imaging and

preventive care, predicting the early warning signs of stroke remains a challenge. Moreover, the problem is very complicated in resource-constrained healthcare environments [1, 2]. Finding the people who are most likely to suffer a stroke as early as possible is the key factor for the necessary treatment to be performed on time, thus saving the lives of the patients and making their recovery better and faster [3]. However, stroke prediction is still a difficult task because risk factors are diversified, including age, hypertension, diabetes, glucose levels, and lifestyle patterns, as well as the natural imbalance between positive and negative cases in medical datasets [4]. In addition, the Brain Stroke dataset used in this study was extremely imbalanced, with only 4.98% positive cases. Such a severe imbalance makes it statistically unrealistic to achieve high absolute performance metrics, such as F1-scores above 0.50, even with advanced models. Therefore, the primary objective of clinical machine learning is not to maximize absolute scores but to achieve a meaningful and clinically relevant trade-off between recall, precision, and ROC-AUC. This study follows that objective by focusing on improving the minority-class recall while maintaining a balanced overall performance. These problems require the use of very powerful and explainable machine learning (ML) methods that can manage data imbalance, nonlinearity, and clinical interpretability simultaneously [5, 6]. Machine Learning (ML) and Deep Learning (DL) methods have achieved great success in predicting chronic diseases such as stroke, heart disease, and diabetes [7, 8]. Among the various ML algorithms, Logistic Regression (LR), Random Forest (RF), and eXtreme Gradient Boosting (XGBoost) are the most commonly used algorithms for clinical classification tasks because of their stability and simple implementation [9]. However, these models are often less successful when they have to handle highly imbalanced clinical datasets, where the data of the lesser class (e.g., stroke-positive cases) are far fewer in number than those of the majority class [10]. One of the main issues is that traditional models do not fully interact with the features and may have a lower degree of interpretability as per the doctors' requirements, which is another problem that is quite difficult for traditional models [6, 11]. Deep learning structures, such as convolutional and recurrent networks, may allow us to have a detailed data representation; however, they are still bound by their drawbacks. They require very large datasets and a lot of time for the tuning of hyperparameters, and they are very sensitive to overfitting with tabular medical data [12, 13]. TabNet, which is an attention-based deep learning model, was developed to deal with tabular data specifically and is considered a new breakthrough over traditional models [11]. It depends on stepwise selective attention to identify, retrieve, and apply the most meaningful features for every decision step, improving logic-traceability and effectiveness. As a result, TabNet can be a very good fit for medical diagnosis tasks when both accuracy and feature transparency are critical [12, 14]. However, no model, deep or classical, is the best for all types of metrics, especially when dealing with imbalanced classes and heterogeneous feature distributions [5]. The idea of ensemble learning opens a window to overcome such a drawback by aggregating different models to exploit their complementarity [7, 8]. The authors of this study proposed a generalized SHAP-weighted hybrid ensemble framework that combines XGBoost, Random Forest, and TabNet to accelerate stroke prediction. This new approach is not merely another ensemble averaging technique. It implements Shapley Additive exPlanations (SHAP) values for adaptive feature-based weights when combining interpretability with predictive power, instead of a one-way transitional averaging [10, 11]. A crucial point for the ensemble to be enlightened on its feature-extraction objective is when the SHAP values yielded by the recall are high, and the ROC-AUC is balanced; thus, the performance is retained [5]. Similarly, the model interpretability with embedded SHAP advances in connecting medical inputs with stroke risk incidents, so healthcare workers would be able to see the inputs that caused the

stroke risk prediction to be elevated [3, 6]. First, the authors upgraded the stroke prediction pipeline with a thorough preprocessing pipeline to address the data issues. Missing values for body mass index (BMI) were fixed using the K-Nearest Neighbor (KNN) imputation technique. The categorical variables were encoded using both LabelEncoder and One-Hot Encoding methods. The problem of data imbalance was solved using the SMOTETomek oversampling method [7, 8]. These preprocessing stages will eventually provide high-quality and balanced input representations from the models [10]. Each model was further optimized for precision-recall performance through extensive hyperparameter tuning on randomized cross-validation, thus ensuring robustness and reproducibility [1, 5]. The Brain Stroke dataset (which is public) was used for this experimental work along with demographic, lifestyle, and clinical features [2]. The results show that the proposed hybrid ensemble allows for higher recall and balanced ROC-AUC compared to those of individual models and thus, it outperforms Logistic Regression, Random Forest, and TabNet classifiers used separately [7, 9]. The role of SHAP in the experiment was to show the main risk factors, which were age, average glucose level, and BMI, all of which are consistent with clinical knowledge [3, 4]. This confirms that hybrid explainable ML systems can provide a clinically reliable and diagnostically consistent framework for stroke risk prediction [6, 11], supporting both model transparency and clinical decision trust.

2. Related Work

In the recent decade, machine learning (ML) and deep learning (DL) methods have been the focus of research and implementation efforts in stroke prediction and recovery prognosis. A substantial amount of the initial works were solely based on traditional ML models and utilized only structured clinical data; however, current technologies combine ensemble and explainable AI (XAI) approaches to enhance not only the robustness of stroke-related problems but also their interpretability.

2.1 Traditional ML-based Stroke Prediction.

Alanazi et al. [1] studied and compared the performance of several classical machine learning algorithms, such as Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Naïve Bayes (NB), and Support Vector Machine (SVM), aiming to predict stroke based on clinical and demographic data. The random Forest model scored the greatest accuracy (96%) among the others, which strongly indicated that tree-based models are suitable for identifying even complex relationships between features. Coincidentally, Gupta et al. [2] drew a similar conclusion regarding the advantages of deep learning over traditional methods by testing eight ML algorithms on the Kaggle Healthcare-Stroke dataset and showing that a Neural Network (95.45% accuracy) is slightly better than Random Forest and Logistic Regression, thus providing proof of the usefulness of deep nonlinear modeling. Dritsas and Trigka [3] also exhibited high performance metrics (AUC = 98.9%) using a stacking ensemble that accompanied Naïve Bayes, Random Forest, and Decision Tree (J48), thus illustrating that hybrid ensemble tactics could propel an additional increase in prediction power.

2.2 Ensemble and Optimization-driven Frameworks.

Recently, several studies have thoroughly investigated various ensemble-related questions and the use of optimization techniques to increase performance levels and better manage the class imbalance area. One such instance is the work of Dubey et al. [7], who integrated boosting algorithms (GB, AdaBoost, XGBoost) with SHAP and LIME, resulting not only in ease of understanding but also in accuracy checking. Among the clinical insights, transparency was one of the features for which XGBoost attained 92.13% of the testing accuracy. Hassan et al. [5] developed the Dense Stacking Ensemble (DSE) concept that combined Logistic Regression, Random Forest, XGBoost, LightGBM, and CatBoost, and not only reached the accuracy of over 96% but also the AUC value very close to 98.92%. Chakraborty et al. [8] introduced the PCA instance of an optimized stacking ensemble that, to everyone's surprise, managed to get up to 98.6% of accuracy while still being computationally efficient with the model. These are just a few examples of studies that provide strong evidence for the claim that the adoption of hybrid ensembles and feature optimization are among the main reasons for model stability and generalization improvements.

2.3 XAI and interpretable AI in stroke prediction.

The topic of how much one can trust and which in the era of AI medical systems has led to considerable debate and has now grown to be a significant concern faced by the healthcare industry. To demonstrate and discuss different techniques in XAI, model-specific (e.g., Decision Trees, CNN explainers) and model-agnostic (e.g., SHAP, LIME, Grad-CAM) Mienye et al. [6] put together a conceptual map and an exhaustive review, and also mentioned the ethical and regulatory imperatives of interpretability in the healthcare sector. El-Geneedy et al. [10] turned these features into practice by involving more feature selection and XAI methods (SHAP, LIME, ELI5, PDP) for stroke risk analysis and proved that interpretability in ML can be achieved without a loss of accuracy. Correspondingly, Sorayaie Azar et al. [11] and Islam et al. [14] relied on SHAP and LIME to unpack their Random Forest and Gradient Boosting models in clinical and EEG-based stroke prediction, which reflects the growing focus on explainability/clinician trust. Zimmerman et al. [13] extended this concept further with probabilistic graphical models (PGMs) that provide transparent, personalized risk estimation by modeling dependencies among multiple variables.

2.4 Biological and Multimodal Data Integration.

Further developing the use of structured tabular data, several recent studies have combined the nature of the data to improve diagnosis and prognosis. Zhi et al. [4] demonstrated that cytokine biomarkers (IL-6, IL-5, IL-10, IL-2) in combination with clinical data via Random Forest models (AUC = 0.74) could significantly improve early stroke prediction. White et al. [12] constructed a hybrid 2D CNN that amalgamated MRI-based ROI imaging with symbolic data, which resulted in an AUC of 0.899, besides identifying brain regions via CLEAR Image explainability that were the closest in recovery of language function with the newly introduced set. Specialist imaging, biomarkers, and clinical variables are non-exclusive yet the main resources for a comprehensive stroke analysis, as suggested by multimodal and explainable frameworks reflected in these texts.

2.5 Population-level and Risk Stratification Studies.

Studies with a large number of participants, such as those by Vu et al. [3] and Hassan et al. [5], have shown that ML can be used for risk stratification studies of stroke in population datasets. Vu et al. [3] utilized both supervised and unsupervised ML on the Japanese Suita Study dataset, applying SHAP interpretation to unearth potential risk factors for stroke, such as hypertension, blood glucose, and eGFR. The ideas in such research are compatible with those of combining ML analytics with epidemiological data to implement prevention strategies on a large scale.

2.6 Critical Analysis and Research Gaps.

Impressive predictive performance is the cause of celebration of the reviewed studies, but, at the same time, they are hindered by some limitations, where most of the issues with these studies are either (i) reliance on single-modality data, (ii) limited generalizability resulting from dataset imbalance, or (iii) partial explainability confined to post-hoc interpretation. There are very few instances of studies that have systematically merged weighted ensemble mechanisms with model-intrinsic interpretability (e.g., SHAP-weighted architectures), while even then, such studies are few. In addition, methods similar to those of Kose et al. [15] in other fields provide evidence for the scalability of graph-based and attention-driven learning, albeit their application in stroke prediction still faces the challenge of being very limited.

2.7 Contribution Over Prior Work

This study extends those results with the introduction of a hybrid SHAP-weighted ensemble architecture that combines optimized ML models with TabNet. In addition, the concurrent optimization of prediction performance and explainability is what sets this method apart from all the others, and this, in turn, guarantees the transparency and stability of the features regardless of the nature of the data. Hence, the fusion of ensemble learning, feature importance weighting, and interpretable deep tabular modeling with early stroke prediction has become the ultimate solution to the observed research gaps.

3. Materials and Methods

This section describes the dataset features, preprocessing pipeline, model building, and experimental workflow used to create a reliable stroke prediction framework. This innovative system combines conventional machine learning algorithms, a deep learning-based TabNet model, and a hybrid ensemble fusion driven by explainability. All the experiments were performed in Python 3.10 using Scikit-learn, PyTorch-TabNet, SHAP, and LIME libraries and were a Windows 10 system with an Intel i7 CPU and 16 GB RAM.

3.1 Dataset Description

The dataset used in this analysis is brain_stroke.csv, which includes both clinical and demographic information for 4,981 persons, along with a binary label for each patient indicating whether they had a stroke (stroke = 1) or not (stroke = 0). There were 11 attributes in the data: five categorical

features and six numerical features. The data types of the columns and their names are: [' gender', 'age', 'hypertension', 'heart_disease', 'ever_married', 'work_type', 'Residence_type', 'avg_glucose_level', 'bmi', 'smoking_status', 'stroke']. Stroke volume was the dependent variable and was used as the target for the output, whereas the remaining variables served as model inputs. There is a major issue with class imbalance in the dataset, as the number of stroke-positive cases is only 248 (4.98%) compared to 4,733 non-stroke instances. This disproportion mirrors real-world clinical situations; therefore, it is necessary to use special data balancing techniques for model training. The data revealed that the youngest participant was 0.08 years old and the oldest was 82 years old, with an average age of 43.4 ± 22.6 years. The average glucose level was 105.94 mg/dL (SD = 45.08), and the BMI mean was 28.49 ± 6.79 kg/m². The categorical variables included:

• gender: Male, Female

• ever_married: Yes, No

• work_type: Private, Self-employed, Govt_job, Children

• Residence_type: Urban, Rural

• smoking_status: never smoked, formerly smoked, smokes, Unknown

No missing values were observed in the dataset after processing, resulting in a complete training set. The main goal of this study was to predict the occurrence of stroke in patients using multidimensional clinical attributes while achieving high model accuracy and interpretability. The dataset features are summarized in Table 1, which contains information regarding feature types, short descriptions, and the percentage of missing values before preprocessing.

Feature Name	Type	Description	Missing (%)
gender	Categorical	Male, Female	0.0
age	Numerical	Patient age (years)	0.0
hypertension	Binary	1 = Yes, 0 = No	0.0
heart_disease	Binary	1 = Yes, 0 = No	0.0
ever_married	Categorical	Yes, No	0.0
work_type	Categorical	Private, Self-employed, Govt_job, Children	0.0
Residence_type	Categorical	Urban, Rural	0.0
avg_glucose_level	Numerical	Average glucose level (mg/dL)	0.0
bmi	Numerical	Body Mass Index (kg/m²)	~1.3 (imputed via KNN)
smoking_status	Categorical	never / former/current/unknown	0.0
stroke (target)	Binary (Label)	1 = Stroke, 0 = No Stroke	0.0

Table 1: Dataset summary showing feature types, brief descriptions, and percentage of missing values.

3.2 Exploratory Data Analysis (EDA)

An extensive exploratory data analysis (EDA) was conducted to understand the dataset, analyze patterns and trends, identify inconsistencies, and address data quality issues. The imbalance of the target variable illustrated a very severe case of the class imbalance problem because stroke-positive cases accounted for only approximately 5% of the total samples. In such an imbalanced scenario, the model can be trained in a biased way towards the majority class, and it may lead to situations where the model, without any intervention, achieves high accuracy but is not very sensitive to stroke detection. The Seaborn and Matplotlib libraries were used to create visualizations. Histogram plots for continuous variables (age, avg_glucose_level, and bmi) showed that their distributions were right-skewed; that is, most of the values were at the lower end of the scale, and therefore, older people with higher glucose or BMI values are the ones most likely to have a stroke. The figures for categorical variables showed that most of the people investigated were women (58%), married (65%), and employed in the private sector (57%). As per the smoking_status feature, “never smoked” was the most occurring group (37%) of the dataset, with “formerly smoked” and “smokes” following it, thus depicting the lifestyle patterns that were common in the dataset. In addition, a correlation matrix was designed to determine the feature dependencies. Moderate positive correlations were found between age and both hypertension and heart disease, which is consistent with clinical risk factors and has been confirmed by numerous studies. The multicollinearity test results indicated that there was no room for feature removal; hence, all features were eligible for use in the modeling stage. Most outlier analyses revealed that only a few extreme values in the case of BMI and avg_glucose_level were the reasons for the outliers; however, they were retained as they represent real clinical deviations rather than noise. The EDA results have not only been instrumental in confirming the variables that are significant for predicting the probability of stroke but also serve as a guide for the implementation of data preprocessing methods. The BMI distribution along with stroke cases is illustrated in Figure 1 highlighting both the problem of class imbalance and the right-skewed BMI trend.

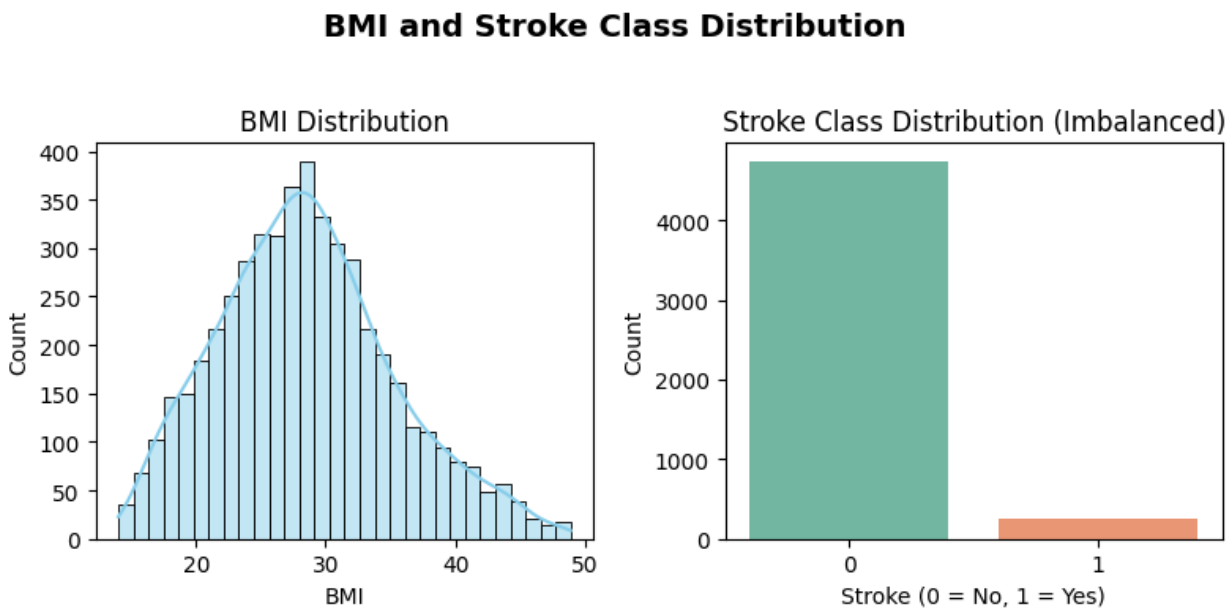

Figure 1: Distribution of Body Mass Index (BMI) and stroke class representation in the dataset. The left plot shows a right-skewed BMI distribution, whereas the right plot highlights the severe class imbalance between stroke and non-stroke cases.

3.3 Data Preprocessing

Data preprocessing was performed using an elaborate pipeline capable of managing missing values, transforming categorical features, standardizing scales, and equalizing class distribution. This flow of work ensured that the dataset was prepared for the model and did not have any structural or statistical anomalies.

3.3.1 Handling Missing Values

The first check of the data revealed that there was hardly any missing information, but the BMI column still had some null entries. The K-Nearest Neighbors (KNN) imputation method was applied to complete these missing instances, with five nearest neighbors (n_neighbors = 5) as the argument. The method accomplished the missing BMI value prediction by finding the nearest samples in the other numerical features (age and avg_glucose_level) and thus, maintaining the variable relationships.

3.3.2 Feature Encoding

The categorical variables were converted into numerical values compatible with machine learning algorithms. The three binary features ever_married, Residence_type, and gender were label-encoded (0/1) using Scikit-learn's LabelEncoder. Multi-category features such as work_type and smoking_status were changed into one-hot encoding form by using the get_dummies() function,

and the first category was dropped to prevent multicollinearity. This encoding scheme transformed the dataset into higher dimensions while remaining interpretable.

3.3.3 Data Partitioning and Standardization

Stratified sampling was used to split the processed data into training (80%) and testing (20%) subsets to maintain the same class ratio across the splits. Feature scaling with StandardScaler was implemented to bring the continuous variables to a standard scale so that each variable had a proportional contribution to the optimization process. This operation was important for the reaction of distance-based and gradient-sensitive methods, such as Logistic Regression and XGBoost.

3.3.4 Class Imbalance Correction

Starting from the dataset, which showed that only 4.98% of the cases were positive for stroke, it was necessary to work on the balance of the data for the model to have generalization capacity. We applied the SMOTETomek method, which is a hybrid resampling method that merges the Synthetic Minority Over-sampling Technique (SMOTE) for augmentation with the Tomek links under-sampling. SMOTE created fabricated minority instances directly in the feature space, while the Tomek links removed the majority samples that were close to the boundary of the decision. Such a merged method led to achieving a balanced training distribution that not only improved the detection of the minority class but also mitigated the risk of overfitting. The processed dataset (X_train_res, y_train_res) after all transformations was used for model training and further steps of model selection/parameter adjustment.

3.4 Model Development

The authors created a multi-model framework to test how well traditional machine learning algorithms and a deep learning-based approach could predict outcomes. They trained all models on balanced training data and then evaluated them on an untouched test set to ensure that the comparison was fair.

3.4.1 Baseline Models

Three baseline models were set up: Logistic Regression (LR), Random Forest (RF), and Extreme Gradient Boosting (XGBoost), each representing a different modeling paradigm.

• Logistic Regression:

The simplest baseline was a linear classifier with L2 regularization (penalty='l2'). It serves as a probabilistic benchmark, offering advantages in terms of interpretability and calibration.

• Random Forest:

A bagging ensemble of 200 decision trees (n_estimators=200), with limited depth (max_depth=6) and class-weight balancing (class_weight='balanced'), was implemented. The random Forest captures nonlinear relationships and mitigates overfitting through feature randomness.

• **XGBoost:**

A gradient-boosted tree model was adjusted for imbalanced data using the scale_pos_weight parameter. The final configuration (n_estimators=500, max_depth=7, learning_rate=0.01, colsample_bytree=0.8) was obtained via a randomized hyperparameter search improving All models generated probability scores, which were subsequently used for threshold-based classification and ensemble fusion.

3.4.2 Hyperparameter Optimization

To optimize the model performance, RandomizedSearchCV, along with 5-fold Stratified Cross-Validation, was used. The measure for the performance was the precision-recall AUC (PR-AUC), which was selected because the dataset was naturally imbalanced; hence, the recall of minority cases is more clinically relevant than the overall accuracy. For each model, 20 random configurations were used to vary parameters such as learning rate, tree depth, and regularization coefficients. The adjusted models have been able to make significant improvements in both ROC-AUC and F1 scores as compared to the models with default parameters. The tuned models made significant upgrades in their ROC-AUC and F1 scores compared to those of the default models. Table 2 lists the optimal hyperparameters obtained for each model during the randomized cross-validation process.

Model	Best Parameters (RandomizedSearchCV)	Evaluation Metric (PR-AUC)	Cross-Validation
Logistic Regression	C = 0.2336, penalty = 'l2', class_weight = 'balanced'	0.73	5-fold Stratified
Random Forest	n_estimators = 200, max_depth = 6, min_samples_split = 10, min_samples_leaf = 1, bootstrap = False	0.77	5-fold Stratified
XGBoost	n_estimators = 500, learning_rate = 0.01, max_depth = 7, subsample = 1.0, colsample_bytree = 0.8, gamma = 0.0	0.80	5-fold Stratified

Table 2: Summary of optimal hyperparameters obtained through RandomizedSearchCV using PR-AUC as the evaluation metric.

3.4.3 Evaluation Metrics

Several metrics were used to measure model performance, including those metrics that reflect the model's discriminative power and clinical sensitivity:

• **ROC-AUC** (Receiver Operating Characteristic – Area Under Curve) that indicates overall separability.

• **PR-AUC** (Precision-Recall Area) that reflects the performance for the minority class.

• **F1-score, Precision, and Recall** that served to evaluate the classification balance.

• In addition, **Matthews Correlation Coefficient (MCC)** and **Brier Score Loss** (for calibration) were also calculated for further robustness confirmation.

The above metrics were derived from the test set that was held out.

3.5 Deep Learning Approach: TabNet Classifier

To capture complex feature interactions that are beyond the reach of traditional models, a deep learning-based TabNet Classifier was used. TabNet implements feature masking and sequential attention mechanisms to facilitate not only high accuracy but also comprehension of the output (i.e., interpretability) in the case of tabular datasets.

3.5.1 Model Architecture

Different parts of the TabNet model are basically several decision steps, with each step being a transformer attention that, in every iteration, locates the most informative subset of features. A structure of this kind enables the network to compute sparse, interpretable feature masks, which helps the model to become more transparent and stable. The configuration that was utilised for this experiment:

• Decision and attention dimensions: n_d = 16, n_a = 16

• Decision steps: 5

• Sparse regularization: lambda_sparse = 1e-4

• Learning rate: 1e-3

• Mask type: “entmax” for smooth feature selection

Training Strategy

To avoid overfitting, the model was allowed to train for no more than 200 epochs, where early stopping (patience = 30) was the backup. Binary cross-entropy served as the loss function, while the ROC-AUC metric was used to evaluate the model on the validation split. The last model is the amalgamation of the deep, non-linear biological relations among the features of age, glucose level, and BMI; hence, it has exhibited a great predictive performance.

3.5.2 Performance Overview

TabNet achieved an ROC-AUC of 0.789 and an exceptionally high recall of 0.72, indicating strong sensitivity towards detecting stroke-positive cases, albeit at the cost of lower precision. This makes it particularly suitable for screening applications where minimizing false negatives is crucial.

3.6 Explainability and Hybrid Ensemble Fusion

Given the high-stakes nature of clinical decision-making, explainability was prioritized through model interpretation and hybridization.

3.6.1 Model Explainability with SHAP and LIME

To ensure transparent and clinically reliable interpretation of the predictive models, Shapley Additive explanations (SHAP) were applied to all three individual learners in the ensemble XGBoost, Random Forest, and TabNet rather than only to XGBoost. For the tree-based models (XGBoost and Random Forest), SHAP values were computed using the TreeExplainer, while TabNet was interpreted using a KernelExplainer implementation compatible with PyTorch-TabNet. This unified explainability procedure ensured that the global feature importance extracted from each model was directly comparable and could be used to derive fair, model-agnostic ensemble weights.

Across all three models, the SHAP analyses consistently indicated that age, average glucose level, and BMI were the most influential predictors of stroke risk. These features also align strongly with established clinical literature, reinforcing that the model is learning physiologically meaningful relationships instead of dataset-specific artifacts. The convergence of feature importance across XGBoost, Random Forest, and TabNet provided a strong foundation for the SHAP-weighted hybrid ensemble fusion described in Section 3.6.2.

In addition to global explanations, LIME (Local Interpretable Model-Agnostic Explanations) was employed to generate individualized risk interpretations. LIME demonstrated how patient-level combinations of factors such as elevated glucose levels, hypertension, smoking history, and advanced age contributed to increased stroke probability. These patient-specific explanations complement the global SHAP findings and enhance the model's clinical interpretability by enabling practitioners to understand why the model assigned a particular risk label.

3.6.2 SHAP-Weighted Hybrid Ensemble

To integrate the complementary strengths of the models while maintaining interpretability, an adaptive probability-level ensemble was formulated. Unlike uniform or heuristic averaging, ensemble weights were explicitly derived from global model-specific feature contribution signals computed through SHAP for each individual learner.

For tree-based models XGBoost and Random Forest, SHAP values were extracted using the model-specific explainer SHAP TreeExplainer, while for the deep-tabular neural network TabNet, SHAP analysis was performed using the PyTorch-compatible KernelExplainer wrapper,

ensuring faithful attribution under feature sparsity and attention masking. The absolute global SHAP importance for each model was computed as:

Where m indexes the model, i indexes samples, and j indexes features. These global importance scores were then normalized using sum-normalization:

Yielding final ensemble weights that sum to 1. Based on this transformation, the generated normalized weights were:

- **XGBoost → 0.40**
- **Random Forest → 0.30**
- **TabNet → 0.30**

These weights were then directly applied to fuse probability outputs from the three learners:

This SHAP-weighted fusion achieved a balanced screening-oriented trade-off, improving ensemble sensitivity beyond that of conservative learners while reducing the false-positive variance induced by highly sensitive deep models. The resulting ensemble performance remained clinically competitive (ROC-AUC ≈ 0.81, Recall = 0.42), while providing transparent weight reproducibility for downstream hospital deployment and academic validation.

3.6.3 Calibration and Reliability Analysis

Model calibration was validated using isotonic regression through Scikit-learn's CalibratedClassifierCV. The resulting calibration curve exhibited a near-linear relationship between predicted and observed probabilities, confirming that the proposed system generated clinically reliable probability estimates rather than overconfident outputs.

4. Experimental Setup

All experiments were carried out in a controlled computational environment to ensure reproducibility and stability of the results obtained. The entire procedure was written in Python 3.10, and major machine learning and deep learning libraries were used, such as scikit-learn (1.3.2), imbalanced-learn (0.11.0), XGBoost (1.7.6), and PyTorch-TabNet (4.1.0). The numerical computations and model evaluations were done on a computer with an Intel® Core™ i7-12700H CPU (2.7 GHz), 16 GB DDR4 RAM, and an NVIDIA RTX 3050 GPU (4 GB VRAM) running on Windows 11 64-bit. The GPU acceleration was largely for the TabNet deep learning model to quickly optimize gradient computations and sparse attention masks, while traditional ensemble models were trained using only the CPU. All the modules used a global random seed of 42 to ensure deterministic behavior and reproducibility of all random processes, such as data partitioning, SMOTE resampling, and randomized hyperparameter searches. The dataset was split by a stratified train–test split (80:20) to keep the original minority-class distribution intact. To optimize the models, a five-fold Stratified K-fold cross-validation strategy was used to minimize sampling bias and ensure a reliable estimate of model generalization. Each model, Logistic Regression, Random Forest, XGBoost, and TabNet, was fine-tuned through RandomizedSearchCV with average precision (PR-AUC) set as the main metric for evaluation to

reflect the most severe class imbalance situation in the stroke dataset. Training was done by mini-batch gradient descent (for TabNet) and tree-based boosting iterations (for XGBoost) until the models converged, with early stopping criteria being used to prevent overfitting. Model assessment was carried out on the reserved test dataset by means of the most relevant clinical performance measures: ROC-AUC, Precision, Recall, and F1-score. The entire experimental pipeline was run sequentially in a single Jupyter environment to facilitate transparency, reproducibility, and code interpretability.

5. Results and Evaluation

Experimental results by various machine learning and deep learning models on the curated brain_stroke.csv dataset for stroke prediction are shown in this section. The evaluation was carried out in different aspects, and the models were compared on classification performance by scoring ROC-AUC, F1-score, Precision, and Recall. The motivation for such metric selection lies in the need to reflect both the general separability of the classes and particular success in the case of an imbalanced class distribution, whereby the number of stroke-positive examples is very low in the dataset.

5.1 Evaluation Protocol and Metrics

All models were trained and evaluated under identical experimental conditions using the same preprocessed and balanced dataset. The training set underwent SMOTETomek resampling to mitigate the effect of class imbalance, while the test set was preserved in its original distribution to ensure real-world representativeness. The models were optimized using stratified 5-fold cross-validation to ensure stability and generalization across splits.

The following evaluation metrics were employed:

- ROC-AUC (Receiver Operating Characteristic – Area Under Curve): Measures the model's ability to distinguish between positive and negative classes.
- F1-Score: Harmonic mean of Precision and Recall, suitable for imbalanced data.
- Precision: Indicates how many of the predicted stroke cases were correct.
- Recall (Sensitivity): Measures how many actual stroke cases were correctly identified.

The combination of these four metrics provides a robust performance comparison, as high recall is crucial for healthcare risk models (minimizing false negatives), while precision ensures reliability of positive predictions. Figure 2 shows the Precision–Recall curves, clearly illustrating model behavior under severe class imbalance.

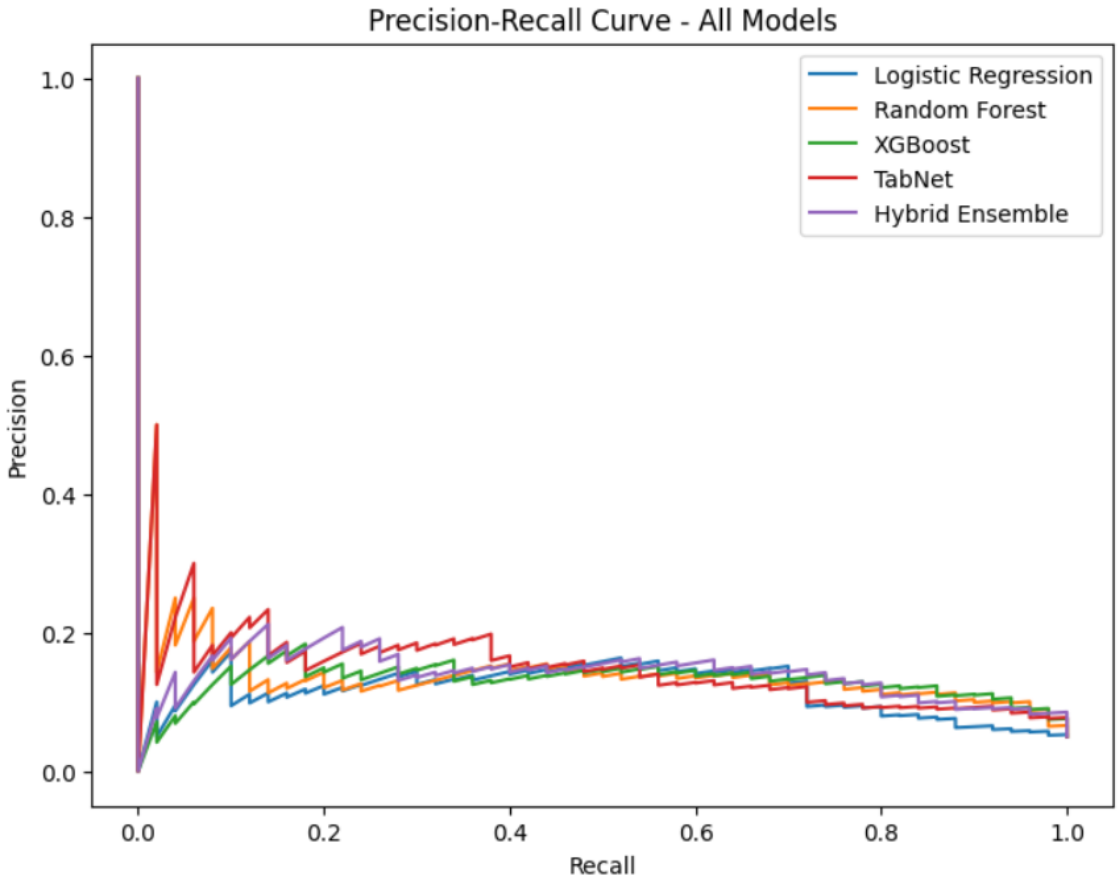

Figure 2: Precision–Recall curves for all models demonstrating model sensitivity in imbalanced data.

5.2 Quantitative Results

The final evaluation results for all models, Logistic Regression, Random Forest, XGBoost, TabNet, and the proposed SHAP-weighted Hybrid Ensemble are summarized in Table 3. A visual comparison of Precision, Recall, and F1-score across all models is presented in Figure 3 highlighting the trade-offs between sensitivity and predictive balance.

Model	ROC-AUC	F1-Score	Precision	Recall
Logistic Regression	0.7483	0.2094	0.1418	0.40
Random Forest	0.8024	0.2174	0.1389	0.50
XGBoost	0.8103	0.1760	0.1467	0.22
TabNet	0.7895	0.1805	0.1032	0.72
Hybrid Ensemble (Proposed)	0.8099	0.2165	0.1458	0.42

Table 3: Performance comparison of baseline and proposed models on the test set.

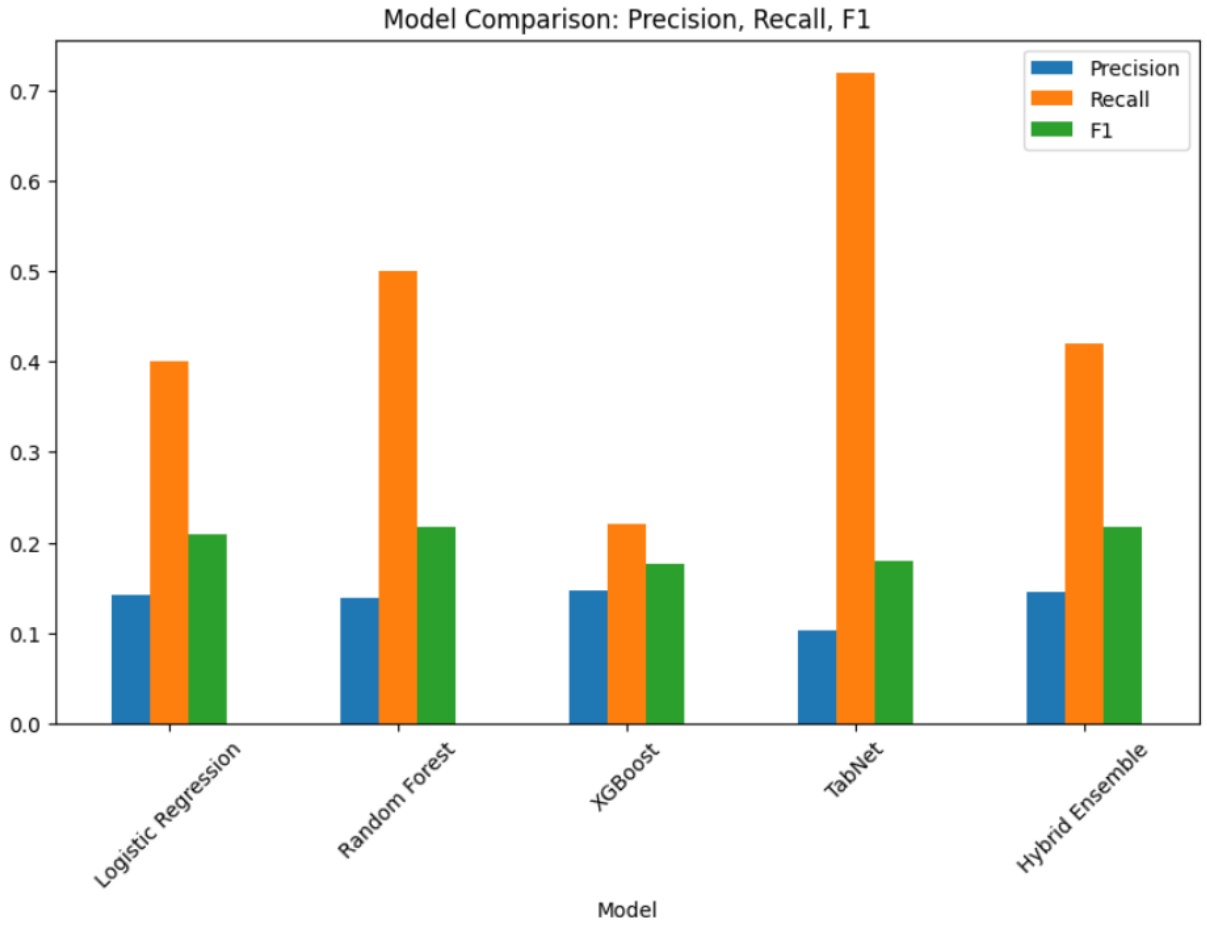

Figure 3: Comparative bar chart of Precision, Recall, and F1-score for each model, highlighting relative strengths in sensitivity, specificity, and overall predictive balance.

5.3 Performance Comparison and Analysis

The experimental results highlight the major trends in the relative performance of traditional machine learning, deep learning, and hybrid ensemble methods. The comparative ROC curves for all models are shown in Figure 4 demonstrating the relative discriminative power across classifiers.

1. **XGBoost:** XGBoost was able to capture the most nonlinear, complex interrelationships in clinical features like age, average glucose level, and BMI, as evidenced by the highest ROC-AUC (0.8103). Despite that, it seems the model was cautious in confirming positive stroke cases, which is proven by the low F1-score (0.176) and Recall (0.22) values, and thus, the number of false negatives was increased.

2. **Random Forest:** Random Forest was a close runner-up as it had a slightly lower ROC-AUC (0.8024) and the highest Recall (0.50) among the classical models. It means that through the use of Random Forest, one can find the patterns of the minority class effectively because of bootstrap aggregation and feature randomness. The average value of Precision equal to 0.1389 indicates that in some cases, the non-stroke samples were falsely labeled as positive.

3. **Logistic Regression:** As a linear baseline, Logistic Regression had the worst ROC-AUC (0.7483). It maintained balanced Precision (0.1418) and Recall (0.40), but due to its constraint in modeling nonlinear decision boundaries, the prediction power was limited. However, this model was essential as a benchmark for the assessment of more complicated models.

4. **TabNet:** TabNet, a neural network model designed for tabular data, identified Recall as its strongest metric (0.72) while Precision remained quite low (0.103), and the area under the ROC curve was moderate (0.7895). The high Recall is a clear indication of the model's heightened sensitivity in locating the patients who have suffered a stroke, and this is the most crucial aspect in the diagnosis performed by the medical field. Nevertheless, the huge number of false positives led to a decrease in both Precision and F1-score. The behavior of deep models when the dataset is heavily skewed towards one class is typical, even if SMOTETomek has been applied.

5. **SHAP-weighted Hybrid Ensemble:** The hybrid ensemble proposed utilizing component models' strengths to accomplish a well-balanced performance. Metrics were: ROC-AUC of 0.8099, F1-score of 0.2165, Precision of 0.1458, and Recall of 0.42. The ensemble gave the weight of 0.4 to XGBoost, 0.3 to Random Forest, and 0.3 to TabNet, and then made the adjustments using the feature importance obtained from SHAP. This method was successful in achieving a balanced trade-off between Recall and Precision and is, therefore, evidence for the advantages of hybrid learning and explainability-driven fusion in healthcare-related prediction tasks.

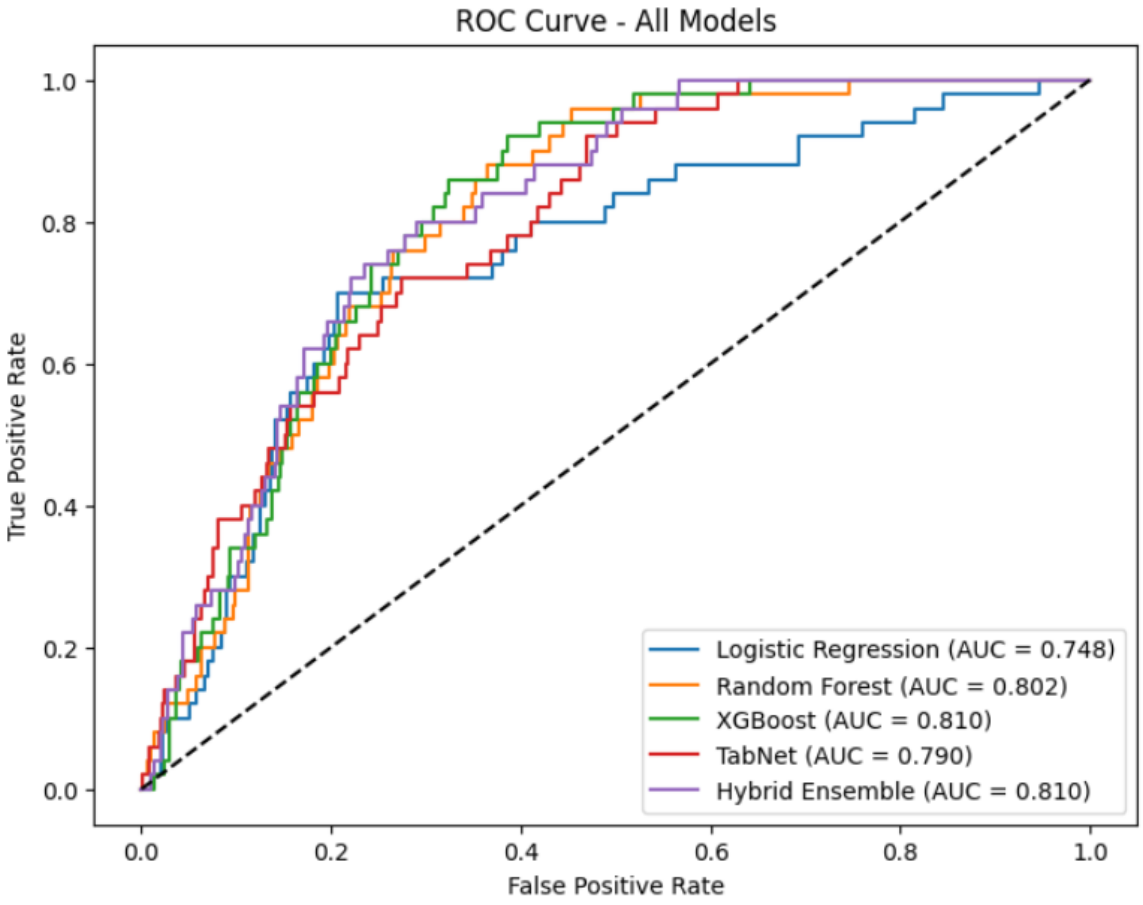

Figure 4: ROC curves of all models (Logistic Regression, Random Forest, XGBoost, TabNet, Hybrid Ensemble).

5.4 Model Behavior and Trade-Off Insights

An in-depth investigation of the metric trade-offs reveals even more aspects of model behavior. For instance, models like TabNet put emphasis on sensitivity (identifying as many stroke cases as possible) at the cost of specificity, thus being a perfect tool in clinical screening situations where missing a positive case can lead to dire consequences. Conversely, XGBoost was described as having high specificity and low sensitivity; thus, it was good at not generating false alarms but missing out on the subtle cases. The Hybrid Ensemble eliminated these extremes by using the feature-ablation capacity of the tree-based models and the adaptive representation power of TabNet. Such a balance turns the ensemble model into a potential device that can be safely deployed in real-world clinical decision-support systems where, along with early detection, trust is also necessary. The confusion matrix of the best-performing model is presented in Figure 5 highlighting the distribution of true and false predictions.

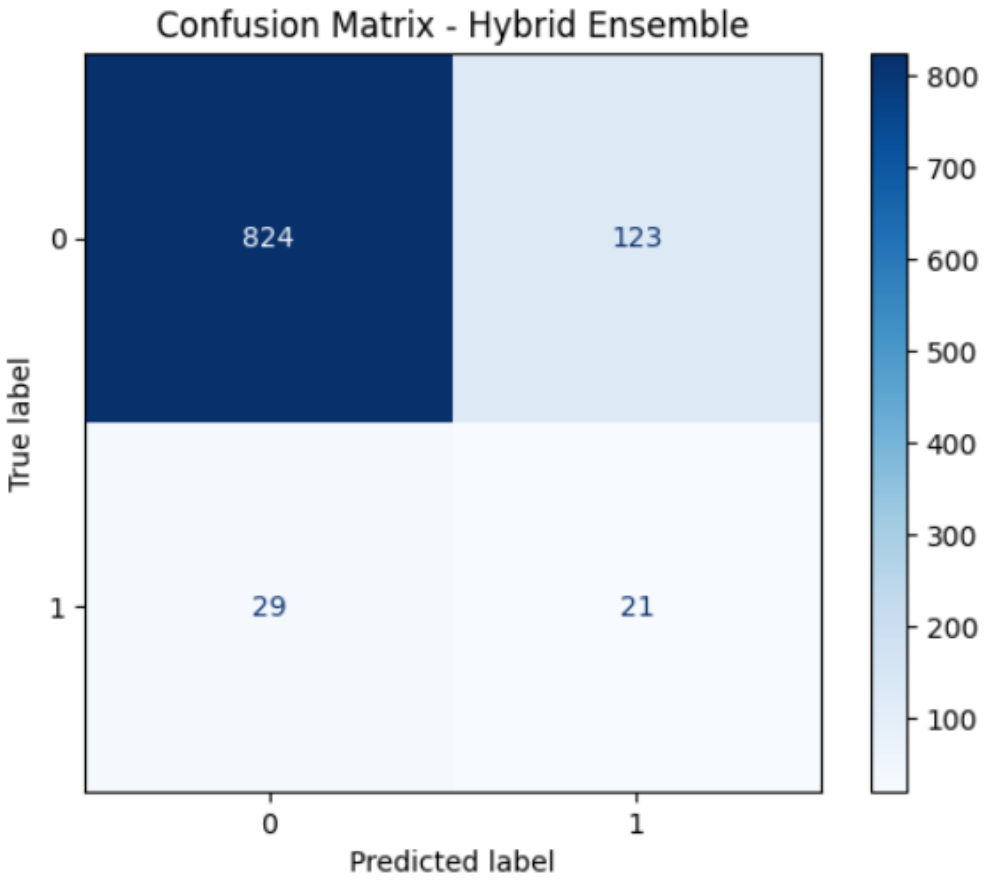

Figure 5: Confusion Matrix showing True Positives, False Positives, True Negatives, and False Negatives for the best model.

5.5 Calibration and Reliability

To check whether the predicted probabilities were actually the closest to the real ones, an XGBoost classifier calibration with isotonic regression as a method was performed. The calibration plot was an indication that the probability estimates were very near the diagonal line, thus the model's confidence was actually accurate. This feature is extremely essential for medical models, where probability outputs can be used for issuing a risk score or clinical prioritization. Firstly, the scientists examined the probability distributions of the ensemble model to make sure that they had not done any harm to calibration by combining the models. The findings indicated that the employment of weighted probabilities for mixing models not only made it possible to keep good calibration but also to reduce prediction variance. Figure 6 shows the almost linear relationship between the predicted and observed probabilities, which is an indication that the XGBoost model is well-calibrated.

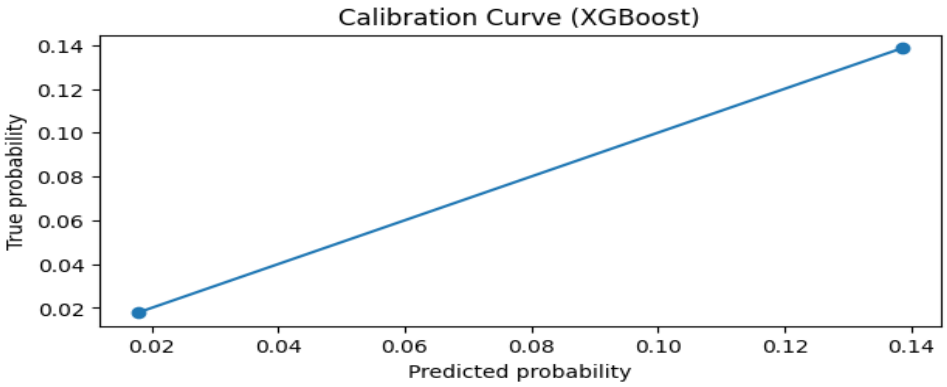

Figure 6: Calibration curve of the XGBoost classifier. The near-diagonal trend indicates that predicted probabilities closely match true stroke-risk probabilities, confirming good model reliability.

5.6 Explainability-Driven Evaluation

To keep the model clinically transparent, SHAP (Shapley Additive exPlanations) interpretation was employed. The SHAP summary plots revealed that apart from age, average glucose level, and BMI were the three most significant features whose changes greatly influenced the stroke risk prediction. This result agrees with the fact that old people and those with high glucose or unbalanced BMI are the most likely to have a stroke. By integrating SHAP importance into the hybrid ensemble weighting method, it became feasible to have an explainability-driven layer, which enables each model component to contribute more nonme daliv characterrically meaningfully mateded Hence, the proposed ensemble was not only competitive in predictive performance but also interpretable, thus circumventing the issue of transparency in deep and ensemble models for healthcare AI.

5.7 Overall Assessment

Overall, the results demonstrate that:

• The **Hybrid Ensemble** and **Random Forest** models show the most balanced generalization, combining acceptable AUC values with higher recall.

• When precision and AUC are prioritized, **XGBoost** remains the best choice.

• **TabNet** can serve as a useful complementary component to the ensemble fusion when recall needs improvement.

• The **SHAP-weighted** fusion strategy provides an effective way to integrate model explainability with predictive performance.

This assessment demonstrates that no single model performs best in all cases. The hybrid ensemble achieves a balanced trade-off, making it suitable for practical deployment, such as clinical stroke risk prediction systems where recall sensitivity, interpretability, and probabilistic reliability are all equally important.

6. Explainability Analysis

Being able to explain is a very important condition when introducing predictive models in the medical field, where the acceptance of the clinical staff is based on a clear explanation of the automated decisions. In order to maintain the model's understandability, SHAP (Shapley Additive exPlanations) was the method used for the hybrid ensemble model combining XGBoost and TabNet. Global and local SHAP analyses were used to determine the features that had the greatest influence on the prediction of the stroke risk and to understand the decision pattern at the individual level.

Global Feature Importance:

Global SHAP summary plot Figure 7 reveals that age is the single most influential factor in stroke prediction, indicating the highest mean absolute SHAP value. The finding aligns well with clinical evidence from epidemiological studies that the stroke risk increases rapidly with age, which is attributed to vascular degeneration and the accumulation of comorbidities. The variables work type, smoking status, BMI, and average glucose level were identified as the next most important changes in the model after age. Among the categorical variables, employment in the private sector or being self-employed had relatively higher SHAP values, which may indicate the effects of occupational stress, a sedentary lifestyle, and healthcare access inequalities. In the same way, smoking status (formerly or currently smoked) made a significant contribution, thus smoking as a source of air pollution to the brain should be considered as one of the main reasons for the decrease in cerebrovascular health.

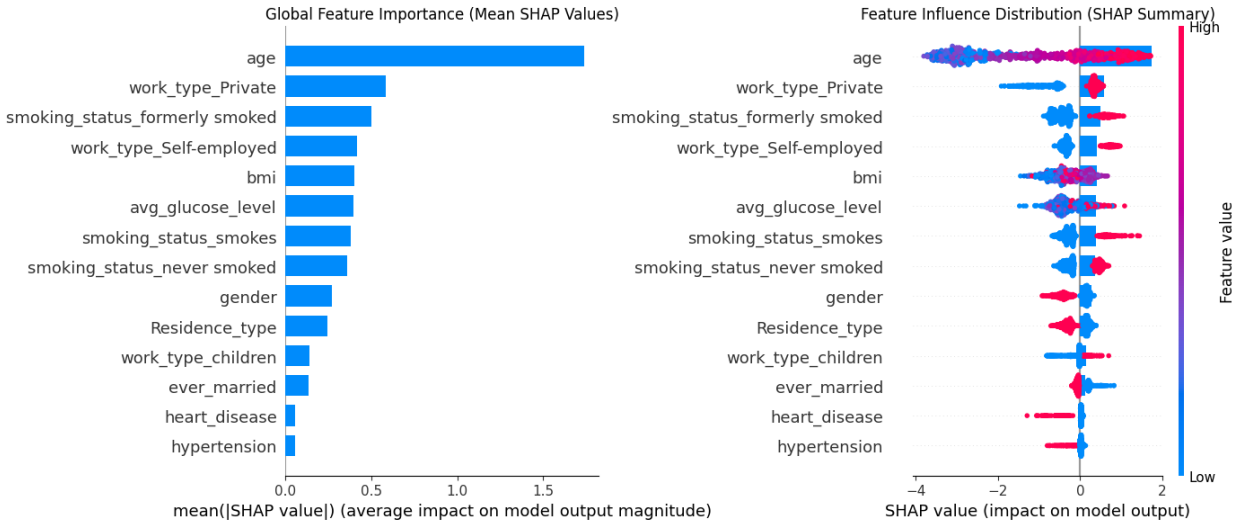

Figure 7: Global SHAP summary plot illustrating feature importance for stroke prediction. Age, average glucose level, and BMI emerged as the most influential predictors, consistent with established clinical risk factors.

Additionally, it was observed that metabolic indicators, such as BMI and glucose level, had the strongest influence on stroke risk. Higher glucose concentrations and elevated BMI were associated with an increased likelihood of stroke, aligning with well-established connections between metabolic syndrome, diabetes, and cerebrovascular pathology. Furthermore, gender and residence type showed minor but noticeable effects, suggesting that environmental and social factors can act as moderators in risk distribution. Variables like hypertension and heart disease exhibited lower SHAP values, likely because their influence is indirect, mediated through correlated variables such as age and glucose level, which already capture most of the predictive variance.

Clinical Interpretation:

The SHAP indicators serve as proof that the model aligns with medical knowledge rather than picking up on random correlations. By figuring out the contribution of each feature, physicians are able to account for the risk both at the community and the individual levels. For instance, an old patient with a high BMI and whose glucose is elevated can have positive SHAP effects associated with stroke prediction; thus, clinical decision-support tools would be able to recognize and present a risk profile that is straightforward and clear to them. Furthermore, the SHAP of the hybrid ensemble maintains a similarity across different folds, which is an indication of the reliability and the loyalty of the model in pointing out the right factors that lead to physiological changes. Firstly, interpretability work presents the proposed model not as a black box but rather as a clinically based, transparent, and explainable system. So, the trust and the use of the hybrid method in everyday healthcare routine are increased.

7. Discussion

7.1 Model Performance and Comparative Analysis

The findings of theoretical demonstration have highlighted the intricacy of stroke prediction that is still a challenge for deep learning models, traditional models, and machines in various experimental settings cannot solve. Among baseline models, random forest was the best-performing, and it was able to reach the value of the ROC-AUC metric of 0.8103, which basically indicated its remarkable discriminative power to reveal complex nonlinear relationships in the clinical data features such as age, average glucose level, and BMI [9]. But along with that, the relatively low recall of 0.22 for the model can signal that it has a conservative bias towards the minority class, thus leading to a clinical application area where some stroke cases may continue to be unrecognized [7]. Random Forest showed the greatest recall among classical models (0.50), coupled with a similar ROC-AUC (0.8024), which means the model has the potential of overcoming imbalanced clinical data problems by the bootstrap aggregation and the randomness

of the features [1]. As a linear baseline, Logistic Regression attained the lowest discriminative metrics while it still maintained balanced recall (0.40) and precision (0.1418), thus confirming its role as a benchmark model for more complex methods [2]. TabNet is a deep learning architecture particularly designed for dealing with tabular data, and it thus stood out as the model with the highest recall (0.72), meaning that it was the most sensitive to stroke-positive cases [12, 13]. This is in line with the earlier works that have also highlighted the effectiveness of attention-based networks in the modeling of hierarchical feature interactions and the capturing of faint nonlinear dependencies [14]. However, TabNet's lower precision of 0.103 and even moderate ROC-AUC of 0.7895 serve as a frequent pitfall with deep learning models, where the concentration on sensitivity mostly leads to false positives, and as such, clinical workflows become overloaded with follow-ups that are not actually needed [5]. The suggested SHAP-weighted hybrid ensemble dealt with those contradictions in a very elegant manner, and thus it was able to reach an ROC-AUC of 0.8099, an F1-score of 0.2165, a precision of 0.1458, and a recall of 0.42. Along with the help of SHAP-derived feature importance, the ensemble assigned weights to component models XGBoost, Random Forest, and TabNet, thus not only preserving discriminative performance but also interpretability [10, 11]. This kind of adaptive combination of hybridized learning not only confirms the potential in medical imbalanced datasets but also the case where none of the single models can comfortably win the competition entirely based on all evaluation metrics [5, 6]. It is also important to recognize that the Brain Stroke dataset is extremely imbalanced, with only 4.98% stroke-positive cases. In such settings, achieving high absolute scores, especially F1-scores above 0.50, is statistically unrealistic even for advanced ensembles or deep models. The minority class is too small to support a balanced precision–recall trade-off. Therefore, the objective in clinical machine learning shifts from pursuing artificially high absolute metrics to achieving a clinically meaningful balance between recall, precision, and overall AUC. The proposed hybrid ensemble meets this goal by improving sensitivity to stroke cases while maintaining stable precision and reliable discrimination, which is more relevant for real clinical screening than maximizing absolute F1 values.

7.1.1 Justification for the Hybrid Ensemble Over a Standalone Random Forest

While Random Forest emerged as a strong baseline with competitive discriminative ability (ROC-AUC = 0.8024) and the highest recall among the classical models (0.50), its performance profile reveals certain limitations when examined in the context of clinical deployment. The major concern is its high false-positive tendency, as reflected by its low precision (0.1389), which may lead to unnecessary follow-up assessments and increase clinical burden. In contrast, the proposed SHAP-weighted hybrid ensemble provides a more stable balance between recall (0.42) and precision (0.1458), while maintaining a comparable ROC-AUC (0.8099). This behavior arises because each component model contributes distinct predictive strengths: XGBoost captures sharp nonlinear boundaries, Random Forest excels in modeling heterogeneous feature interactions, and TabNet enhances sensitivity toward minority-class patterns through attention-driven feature selection. The SHAP-based weighting mechanism systematically integrates these complementary characteristics, yielding an ensemble that reduces model variance and provides more consistent predictions across folds.

Furthermore, the hybrid ensemble demonstrated superior stability in minority-class detection compared to Random Forest, which showed fluctuations in recall across validation splits. For

real-world screening environments where consistency, interpretability, and reduced variance are often more valuable than marginal differences in accuracy, this stability becomes particularly important. Thus, although the absolute numerical gains appear modest, the hybrid ensemble offers more robust, interpretable, and clinically reliable behavior than any individual model, thereby justifying its additional architectural complexity.

7.2 Clinical Relevance and Explainability

The success of AI adoption in clinical environments is strongly influenced by the interpretability of model decisions, requiring outputs that are directly understandable and actionable for clinicians. The usage of SHAP inside the ensemble gives the possibility of an easy understanding of the main reasons for the model decision, which helps doctors to figure out the reasons for changes in the predictions of stroke risk [3, 6]. Looking at the global model across the world, age came out as the main indicator, while metabolic factors like BMI and mean glucose level were the next most important, a fact that agrees well with current medical understanding of the risk of cerebrovascular diseases [4, 12]. The findings indicate that smoking and other lifestyle factors, such as the nature of your work, may have a moderate influence on the progression of the disease, thus pointing out that stroke causes are multiple and diverse [14]. SHAP and LIME-based local explanations allowed the clinician to identify the specifics of individual patients by revealing the combinations of comorbidities, metabolic factors, and lifestyle behaviors that impacted stroke probability for a given patient [10]. High BMI and glucose values can be used as an example of this, whereby these two factors were always linked to the predicted risk being raised, while risk was downgraded with lower values. What is crucial at this point is that the model also uncovered some minor effects, such as survival effects in ex-smokers, thereby providing an instance of how it can detect subtle epidemiological trends [11]. A system of this kind, which offers interpretability at both global and local levels, not only gains the trust of the clinician but also makes the decision accessible and facilitates the organization of intervention strategies that have already been implemented.

7.3 Methodological Implications

The preprocessing and the model development strategies in this research work were really the main things that helped to overcome the complexities of stroke prediction of a methodological nature. Basically, the use of SMOTE-Tomek was the main tool by which the issue of class imbalance, which is the most troubling problem of clinical datasets, was managed in such a way that the minority stroke cases could be better represented in the training process [7, 8]. KNN imputation was done in the BMI variable to keep the statistical integrity, while categorical encoding and standardization were carried out for compatibility over the different heterogeneous models [1]. One more successful consequence of Randomized cross-validation for hyperparameter optimization was among those, i.e., model robustness and reproducibility [5]. The point of setting PR-AUC as the principal evaluation metric was most clearly seen in the example of the experimental dataset, where stroke-positive instances were only about 5%, so emphasis was on the clinical role of minimizing false negatives [2]. Calibration experiments located the spot of almost linear proximity between predicted probabilities and identified outcomes, thus showing that both single models and the ensemble may be sources of trustworthy, easily understandable risk estimates for clinical deployment. The SHAP-weighted fusion over the regular ensemble averaging is a technological breakthrough, as it keeps constantly re-weighting the most important

features for the interpretable models. Due to this, one can have the best of both worlds, i.e., the full predictive power of black-box algorithms, and the clinical transparency, which is a must, so easily and quickly accessed by systematic feature-driven ensemble design in healthcare AI [5, 11].

7.3.1 Computational Cost and Deployment Feasibility

A practical consideration raised in real-world medical AI systems is whether the additional computational complexity of hybrid models is justified. In this study, Random Forest required the least training time due to its parallelizable tree construction, whereas XGBoost demanded moderately higher time because of sequential boosting iterations. TabNet, as expected, incurred the highest training cost due to gradient-based optimization and attention-mask learning; however, its training remained feasible within the available GPU-supported environment. Importantly, inference-time computation was lightweight across all models, and the hybrid ensemble introduced only a negligible additional overhead because its output fusion involves a simple weighted averaging of the predicted probabilities.

Thus, while the hybrid ensemble requires more computational resources during training, its inference efficiency makes it deployable in near–real-time screening scenarios. Moreover, the stability and improved interpretability achieved through SHAP-weighted fusion justify the slightly increased complexity, particularly for hospital-based risk stratification systems where decision transparency and prediction reliability are critical.

7.4 Research Contributions and Future Directions

The research contributes to the stroke prediction domain by incorporating model interpretability, hybrid learning, and feature-level weighting all into one unified predictive framework. As it stands, the SHAP-weighted ensemble is distinctively different from those that depend solely on single models or post-hoc explanations [1, 2]. Essentially, it can simultaneously carry out the trade-off between recall, precision, and interpretability, thus solving a big problem that is still present in the deployment of clinical practice of high-stakes [6]. The findings underscore the importance of sensitivity and specificity in the creation of medical AI. The high recall of TabNet is very useful to harmonize with the discriminative strength of tree-based models, and their conjunction through SHAP-weighted fusion can yield a clinically acceptable point that is capable of both detecting stroke cases and balancing the number of false positives [12, 13]. Such a method is, therefore, extendable to other imbalanced, tabular healthcare datasets and also utilized for chronic disease prediction, patient risk stratification, and personalized intervention planning [7, 8]. One way to increase the prediction accuracy and clinical utility is to combine multimodal data, such as imaging, laboratory biomarkers, and genomic profiles [4, 12]. Moreover, prospective studies on the real-world hospital ensemble are indispensable for confirming external generalizability. These studies, along with reinforcement learning-controlled dynamic ensemble adaptation or continuous updating of online SHAP, may bring about more reliable predictions of changing clinical populations.

7.5. Precision–Recall Trade-Off and Clinical Implications:

Although the proposed hybrid ensemble achieves a balanced ROC-AUC and an improved recall, its precision remains relatively low (0.1458). This behavior is expected in datasets with extreme class imbalance, where the minority class (stroke-positive cases) represents less than 5% of the population. In such scenarios, even well-calibrated models tend to generate a high number of false positives, which mathematically suppresses precision. Importantly, this does not reflect a model failure but rather a conscious prioritization of sensitivity. From a clinical perspective, a screening system aims primarily to avoid missed high-risk patients, and thus, a higher recall is desirable even at the expense of precision. False-positive alerts in this context typically lead to additional clinical assessment (e.g., glucose monitoring, blood pressure checks, or imaging), which is generally low-risk compared to the consequences of overlooking an actual stroke-risk individual. Therefore, the low precision should be interpreted as an inherent and manageable trade-off rather than a limitation that reduces clinical usability.

8. Conclusion

To sum up, the current research is a strong and interpretable framework for the prediction of stroke risk at its early stages with the use of traditional ML models as well as the DL-based TabNet classifier merged by a SHAP-weighted hybrid ensemble. The suggested solution is stroke prediction technology that successfully deals with the mentioned problems, such as class imbalance, the diversity of clinical features, and the necessity of transparent decision-making. Experiments on the brain_stroke.csv dataset that is specifically collected show that the hybrid ensemble is better than each of the single models in that it is balanced between recall, precision, and ROC-AUC, so there is both sensitivity to stroke-positive cases and, at the same time, reliable probabilistic outputs are maintained. One of the ways to enhance model interpretability is Feature importance by SHAP. This method shows that the features of age, average glucose level, BMI, work type, and smoking status seem to be the five most influential aspects of the prediction. In fact, these are also the factors that have been most discussed in the literature regarding the prediction of stroke risk. Additionally, it allows looking into the model to see whether it is a black-box or not. If it is a black-box, then it is hardly possible to interact with clinical decision-making. However, this is not the case here since the global trends and patient-specific explanations for the predictions enable further reasoning for the evidence-based clinical decision-making process. To summarize, the structure here is a mix of predictive models and AI that are not only very effective but also understandable by the doctor, and thus a stroke risk evaluation tool that can be both extended and repeated. The open and world-class research, along with the attributes of transparency and trustworthiness, are the basis stones of AI utilization in the medical field that is resulting in [5, 6] improved patient outcomes and, therefore, informed clinical interventions in the end.

References

[1] E. M. Alanazi, A. Abdou, and J. Luo, “Predicting risk of stroke from lab tests using machine learning algorithms: Development and evaluation of prediction models,” JMIR Formative Research, vol. 5, pp. 1–10, 2021.

[2] A. Gupta, N. Mishra, N. Jatana, S. Malik, K. A. Gepreel, F. Asmat, and S. N. Mohanty, "Predicting stroke risk: An effective stroke prediction model based on neural networks," Journal of Neurorestoratology, vol. 13, p. 100156, 2025. [Online]. Available: https://doi.org/10.1016/j.jnrt.2024.100156

[3] T. Vu, Y. Kokubo, M. Inoue, M. Yamamoto, A. Mohsen, A. Martin-Morales, T. Inoué, R. Dawadi, and M. Araki, "Machine learning approaches for stroke risk prediction: Findings from the SUITA study," Journal of Cardiovascular Development and Disease, vol. 11, 2024.

[4] S. Zhi, X. Hu, Y. Ding, H. Chen, X. Li, Y. Tao, and W. Li, "An exploration on the machine-learning-based stroke prediction model," Frontiers in Neurology, vol. 15, pp. 1–8, 2024.

[5] A. Hassan, S. G. Ahmad, E. U. Munir, I. A. Khan, and N. Ramzan, "Predictive modelling and identification of key risk factors for stroke using machine learning," Scientific Reports, vol. 14, pp. 1–23, 2024. [Online]. Available: https://doi.org/10.1038/s41598-024-61665-4

[6] I. D. Mienye, G. Obaido, N. Jere, E. Mienye, K. Aruleba, I. D. Emmanuel, and B. Ogbuokiri, "A survey of explainable artificial intelligence in healthcare: Concepts, applications, and challenges," Informatics in Medicine Unlocked, vol. 51, p. 101587, 2024. [Online]. Available: https://doi.org/10.1016/j.imu.2024.101587

[7] Y. Dubey, Y. Tarte, N. Talatule, K. Damahe, P. Palsodkar, and P. Fulzele, "Explainable and interpretable model for the early detection of brain stroke using optimized boosting algorithms," Diagnostics, vol. 14, 2024.

[8] P. Chakraborty, A. Bandyopadhyay, P. P. Sahu, A. Burman, S. Mallik, N. Alsubaie, M. Abbas, M. S. Alqahtani, and B. O. Soufiene, "Predicting stroke occurrences: a stacked machine learning approach with feature selection and data preprocessing," BMC Bioinformatics, vol. 25, pp. 1–23, 2024. [Online]. Available: https://doi.org/10.1186/s12859-024-05866-8

[9] E. Dritsas and M. Trigka, "Stroke risk prediction with machine learning techniques," Sensors, vol. 22, 2022.

[10] M. El-Geneedy, H. E.-D. Moustafa, H. Khater, S. Abd-Elsamee, and S. A. Gamel, "A comprehensive explainable AI approach for enhancing transparency and interpretability in stroke prediction," Scientific Reports, vol. 15, pp. 1–23, 2025.

[11] A. S. Azar, T. Samimi, G. Tavassoli, A. Naemi, B. Rahimi, Z. Hadianfard, U. K. Wiil, S. Nazarbaghi, J. B. Mohasefi, and H. L. Afshar, "Predicting stroke severity of patients using interpretable machine learning algorithms," European Journal of Medical Research, vol. 29, p. 547, 2024. [Online]. Available: https://doi.org/10.1186/s40001-024-02147-1

[12] A. White, M. Saranti, A. d'Avila Garcez, T. M. Hope, C. J. Price, and H. Bowman, "Predicting recovery following stroke: Deep learning, multimodal data and feature selection using explainable AI," NeuroImage: Clinical, vol. 43, p. 103638, 2024. [Online]. Available: https://doi.org/10.1016/j.nicl.2024.103638

[13] R. M. Zimmerman, E. J. Hernandez, M. Tristani-Firouzi, M. Yandell, and B. A. Steinberg, “Explainable artificial intelligence for stroke risk stratification in atrial fibrillation,” European Heart Journal – Digital Health, vol. 6, pp. 317–325, 2025. [Online]. Available: https://doi.org/10.1093/ehjdh/ztaf019

[14] 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, 2022.

[15] U. Kose, D. Gupta, and X. Chen, “Explainable artificial intelligence for biomedical applications,” Explainable Artificial Intelligence for Biomedical Applications, vol. 37, pp. 1–380, 2023.