International Journal on Perceptive and Cognitive Computing (IJPCC) Vol 12, Issue 1 (2026)
https://doi.org/10.31436/ijpcc.v12i1.636

Interpretable AI for Stroke Prediction: A Structured Approach Using Explainable AI Techniques

Lazeena Tarnim Ranak, Sharyar Wani
Department of Computer Science, Kulliyyah of Information and Communication Technology
International Islamic University Malaysia, Kuala Lumpur, Malaysia

*Corresponding author: sharyarwani@iium.edu.my
(Received: 23rd November 2025; Accepted: 13th December, 2025; Published on-line: 30th January, 2026)

Abstract— The lack of interpretability in AI-driven healthcare diagnostics poses a significant challenge to clinical adoption. This study explores the methodological integration of explainable artificial intelligence (XAI) tools using open clinical prediction dataset, the SCI-XAI pipeline, for stroke risk prediction. We apply multiple machine learning models ranging from white-box approaches (Logistic Regression, Decision Tree, Explainable Boosting Machine) to black-box models (Random Forest, XGBoost, LightGBM, and Multi-Layer Perceptron) and evaluate their trade-offs between predictive accuracy and explainability using techniques such as SHAP, LIME, and ELI5. The study uses a systematic approach involving pre-modeling, modeling, and post-modeling phases, aiming to improve model interpretability for potential use in clinical decision-support contexts. The experimental results show that ensemble models achieve superior accuracy, while traditional models provide inherent transparency. However, the SCI-XAI framework demonstrated that post-hoc explainability tools can extend such transparency to complex models. SHAP-based feature im-portance analysis identifies age, glucose levels, and BMI as the most influential predictors of stroke. The integration of structured explainability into AI based diagnostics helps bridge the gap between algorithmic prediction and clinical interpretability, offering a methodological foundation for more transparent decision-support systems.

Keywords— explainable artificial intelligence (xai), stroke prediction, interpretable machine learning, sci-xai pipeline, clinical decision-making

I. Introduction

In the realm of modern healthcare, artificial intelligence (AI) holds immense potential to revolutionize medical diagnostics and treatment. However, the opacity of AI models presents a significant barrier to their widespread acceptance and utility in healthcare settings. This research project aims to address this challenge by focusing on Explainable AI (XAI) techniques, particularly with the SCI-XAI pipeline, to enhance interpretability and foster actionable insights in AI-driven healthcare diagnostics, specifically using tabular data [1].

Explainable AI (XAI) refers to methods and techniques that make the decision-making pro-cesses of AI models understandable to humans. In healthcare diagnostics, the need for interpret-ability is crucial, as it enables healthcare professionals to effectively utilize AI recommendations. This project integrates several XAI methodologies to ensure transparency and comprehensibility: SHAP (Shapley Additive explanations) provides a unified measure of feature importance, offering both global and local explanations of model predictions; ELI5 (Explain Like I'm 5) simplifies the understanding of model decisions by breaking down complex models into understandable components; and LIME (Local Interpretable Model-agnostic Explanations) generates local explanations for individual predictions, offering insights into how specific features influence model outcomes [2]. Our approach focuses on the SCI-XAI pipeline, a systematic method designed to enhance the interpretability of AI models by focusing on feature selection and extraction [3]. The pipeline involves two main steps – feature selection and feature extraction. The former identifies the most relevant features that significantly impact model predictions, ensuring that the models are not only accurate but also interpretable by highlighting the key factors driving decisions. The latter transforms raw data into a set of informative features that can be used for model training, enhancing the clarity and transparency of the models.

Unlike prior studies that use the SCI-XAI framework or individual explainability tools in isolation, the contribution lies in its single systematic integration of the SCI-XAI pipeline with a broad spectrum of models ranging from interpretable (white-box) to opaque (grey-black-box) architectures and multiple XAI techniques (SHAP, LIME, ELI5). This unified experimental setup enables a structured comparison of interpretability–performance trade-offs across model complexities in a healthcare setting. By applying this

https://doi.org/10.31436/ijpcc.v12i1.636

framework to stroke risk prediction, the study demonstrates how explainability methods can extend transparency from inherently interpretable models to complex ensembles and neural networks, offering methodological rather than clinical novelty in the context of medical AI.

To integrate these techniques into diagnostics, we explore several powerful machine learning algorithms [4], particularly suited for tabular data. These include Adaptive Gradient Boosting, an ensemble technique that improves model performance by iteratively correcting errors from previous models; XGBoost (Extreme Gradient Boosting), known for its efficiency and high performance; and LightGBM (Light Gradient Boosting Machine), optimized for speed and efficiency, effectively handling large datasets and providing quick, accurate predictions [5]. In addition to these ensemble methods, we employ traditional regression models like logistic regression, which establishes relationships between features and outcomes, providing clear insights into data trends, and random forest, an ensemble method that enhances prediction accuracy by averaging the results of multiple decision trees, also offering insights into feature importance.

This paper explores the methodological factors influencing model opacity in healthcare-focused AI and applies XAI-integrated machine learning models for stroke risk pre-diction using tabular data. The focus is on interpretability employing a range of explainable AI (XAI) techniques to provide both quantitative and qualitative insights into the decision-making processes of diverse model architectures.

The objectives of this research are to implement and systematically evaluate a combination of classical machine learning, ensemble, and deep learning models within the SCI-XAI framework, and to quantify interpretability using established XAI tools. This ensures both transparency and measurable insight into model behavior.

It is important to understand the key parameters and factors that influence stroke risk and outcomes before delving into the specific applications of XAI in stroke prediction and other medical diagnoses. These factors serve as the foundation for many AI-driven studies, as they are critical for accurate prediction and diagnosis.

Research has thoroughly explored the multifaceted risk factors and outcomes associated with strokes, highlighting important demographic and health-related considerations. Gender differences in stroke incidence and outcomes are evident, with males having higher incidence and mortality rates between ages 45-74, while females experience higher rates after age 74 due to factors like increased comorbidities [5]. Stroke incidence is higher in women under 30, while men generally have higher incidence during midlife. By age 80, rates equalize or favour women. The lifetime risk of stroke is around 25.1% for women and 24.7% for men, with women typically experiencing strokes 4-6 years later than men [6]. Additionally, hypertension, affecting about 64% of stroke patients, is a major risk factor, linked to ischemic strokes and hemorrhages [7] [8]. Blood pressure thresholds $\geqslant$ 140/90 mmHg are important for stroke detection, and maintaining systolic blood pressure below 140 mmHg is crucial for prevention [7]. Furthermore, heart failure (HF) increases stroke risk by 2-3 times, with a five-fold increase in stroke risk when combined with atrial fibrillation (AF) [8]. Marital status provides a protective effect against stroke outcomes, with married individuals exhibiting lower mortality rates [9]. Job loss and unemployment, particularly in high-stress jobs, are associated with increased stroke risks. Continuous employment, regardless of sector, is linked to lower stroke risks [10].

Further studies have shown nuanced influences like residential areas and associated risks. The impact of residential areas on stroke incidence is explored, showing a slightly higher in rural areas (3.35 per thousand) compared to suburbs (2.90 per thousand) for men and slightly higher in suburbs (2.34 per thousand) than in town centers (2.14 per thousand) for women, although case fatality was lower in rural areas [11]. The study of stroke risk associated with average glucose levels in 12,321 participants indicates a clear correlation between higher glucose levels and in-creased stroke risk. Participants with diabetes had a 3.5% stroke incidence, compared to 2.2% in non-diabetic participants, with higher glucose levels ($\geqslant$ 126 mg/dL) increasing stroke risk by 78% (HR 1.78) compared to those with glucose levels of 90-99.9 mg/dL [12]. Blood pressure, BMI, cholesterol, and glucose levels were all significantly higher in participants with diabetes, reinforcing the importance of managing metabolic factors for stroke prevention [13], [14].

Now, with these key stroke-related parameters in mind, we turn to the application of XAI in medical diagnostics, particularly focusing on stroke and other health conditions. One study ap-plied machine learning models combined with XAI tools such as ELI5 and LIME for stroke pre-diction using EEG signals, achieving around 80% accuracy [15]. Another study enhanced intra-operative decision-making in ovarian cancer surgery by integrating XAI with human factors and domain knowledge, demonstrating the value of explainability in real-time clinical settings [16]. These studies collectively delve into the realm of Explainable AI (XAI) in medical diagnostics and decision support. They explore various methodologies and applications, from interpreting AI-generated clinical decision support systems (CDSS) through human-centered design approaches to evaluating XAI methods like Grad-CAM and Eigen-CAM on medical imaging datasets [17]. Moreover, they introduce innovative solutions such as the SCI-XAI pipeline for clinical prediction models and a geriatric MDSS incorporating XAI elements [3] [18]. The studies emphasize the importance of

https://doi.org/10.31436/ijpcc.v12i1.636

interpretability, transparency, and user involvement in AI systems [19], aiming to enhance diagnostic accuracy, trust, and ultimately patient outcomes. Another study evaluated Grad-CAM and Eigen-CAM visual XAI methods on the VinDrCXR Chest X-ray Abnormalities Detection dataset using YOLO models. It highlighted the limitations of these methods in explaining model decisions and cautioned against sole reliance on their outputs. The authors recommended validating results through sample images, manual evaluations, or automated methods to find the most suitable XAI tools for specific domains. Challenges such as partiality, overfitting, and limited interpretability for complex models were discussed, emphasizing the need for careful application and additional verification in visual XAI [20]. These findings pave the way for further research and development in XAI to address practical challenges and improve medical decision-making processes across different domains [21].

The study in [22] presents a systematic comparison of post-hoc explainability methods i.e. LIME, SHAP, and Anchors which are evaluated on healthcare datasets for fidelity, stability, separability, computational efficiency, and bias detection. Results indicate that LIME provides dis-tinct explanations but with lower fidelity, SHAP demonstrates superior speed and bias detection, while Anchors offers intuitive rule-based outputs. The research highlights critical trade-offs be-tween interpretability and reliability, underscoring the complementary strengths of these approaches. A related investigation in [23] focuses on stroke prediction using an ensemble framework achieving approximately 96% accuracy. Five explainability techniques; SHAP, LIME, ELI5, Anchors, and QLattice were applied to interpret predictions. Across all methods, age, BMI, blood glucose, and hypertension emerged as key contributing factors, consistent with clinical evidence. The integration of multiple XAI tools ensured alignment between model behavior and medical understanding, supporting transparency and clinician trust.

The work in [24] introduces an interpretable ensemble model to differentiate iron-deficiency anemia from aplastic anemia, marking the first use of interpretable AI for this diagnostic task. Employing SHAP, LIME, ELI5, QLattice, and Anchors, the study identified platelet count, mean cell volume, haemoglobin, and white blood cell count as dominant features, enhancing the transparency of model reasoning and clinical reliability. Similarly, research on type 2 diabetes prediction in [25] developed an ensemble framework achieving 92.5% accuracy (AUC $\approx$ 0.98), incorporating SHAP, LIME, EBM, and counterfactual analysis to interpret predictions. Key risk factors such as BMI, age, and physical activity were consistently highlighted, demonstrating both interpretability and clinical coherence. Finally, a comparative analysis in [26] evaluated traditional statistical and modern machine learning approaches for stroke risk prediction using logistic regression, Cox regression, Bayesian networks, EBM, and XGBoost with SHAP-based explanations. XGBoost achieved the highest performance (C-statistic = 0.89; F1-score = 0.80), followed closely by EBM (C-statistic = 0.87). Both models identified atrial fibrillation, hypertension, age, and HDL cholesterol as major predictors, aligning with established clinical knowledge.

Building on prior works that primarily compared interpretability tools, this study extends those efforts by applying a comparative framework for stroke risk prediction that integrates white-box and black-box models within the SCI-XAI pipeline. This integration enables a systematic evaluation of both global and local explanations, offering balanced insights into accuracy and interpretability.

Recent work by Challenging the Performance-Interpretability Trade-Off: An Evaluation of Interpretable Machine Learning Models [27] provides strong empirical evidence that the performance–interpretability trade-off is not inevitable. In a large-scale evaluation of generalized additive models (GAMs) such as the Explainable Boosting Machine (EBM) against black-box baselines across twenty tabular datasets, the authors demonstrated that advanced GAMs could achieve competitive predictive performance while remaining inherently interpretable. This finding supports our selection of EBM as a key component within the SCI-XAI pipeline and motivates our comparative inclusion of both transparent and opaque models to evaluate explainability and performance in stroke prediction.

In closing, this introduction has outlined the motivation, methodological framing and key analytical lenses of our study. With a clear view of the interpretability-performance landscape and the domain-specific foundations of stroke risk factors, the next section details the experimental methodology and SCI-XAI modeling framework.

II. Experimental Setup

A. Dataset Description

The dataset employed in this study is the Stroke Prediction Dataset, publicly available on Kaggle, curated by fedesoriano. It contains a total of 5,110 patient records; each aimed at sup-porting the prediction of stroke occurrence based on various demographic and health-related attributes. The target variable is stroke, a binary classification label where 1 indicates the occurrence of a stroke, and 0 indicates no stroke. A closer examination of the class distribution reveals a strong imbalance, with only 249 instances (approximately 4.87%) indicating stroke events, and the remaining 4,861 instances (approximately 95.13%) representing non-stroke cases. This pronounced imbalance is a crucial factor that informs model development and evaluation strategies.

https://doi.org/10.31436/ijpcc.v12i1.636

The dataset comprises 11 features, encompassing both categorical and numerical variables. These include gender, age, hypertension, heart_disease, ever_married, work_type, Residence_type, avg_glucose_level, bmi, and smoking_status. Together, these features provide a comprehensive foundation for stroke risk modeling but require careful preprocessing due to mixed data types and imbalance.

B. Data Modeling Process

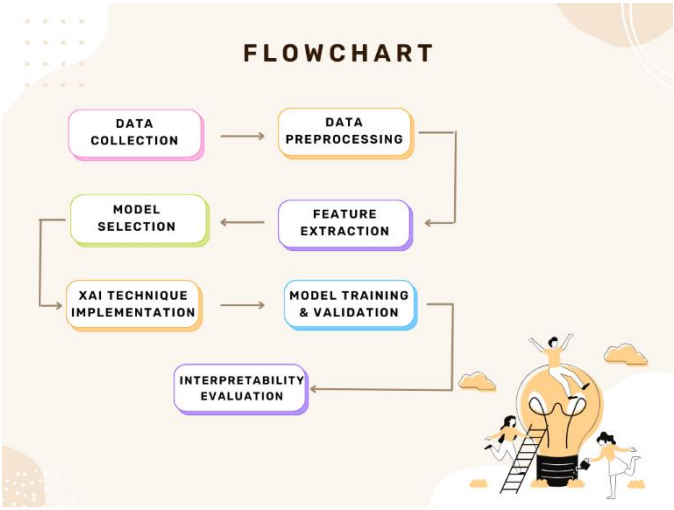


Fig. 1 Data Mining Flowchart.

Figure 1. outlines the high-level progression of this study, starting from data collection to interpretability evaluation. It captures the essential stages including preprocessing, model selection, and pipeline strategies, followed by model training and the integration of explainable AI (XAI) techniques to ensure transparent and accountable predictions in healthcare diagnostics.

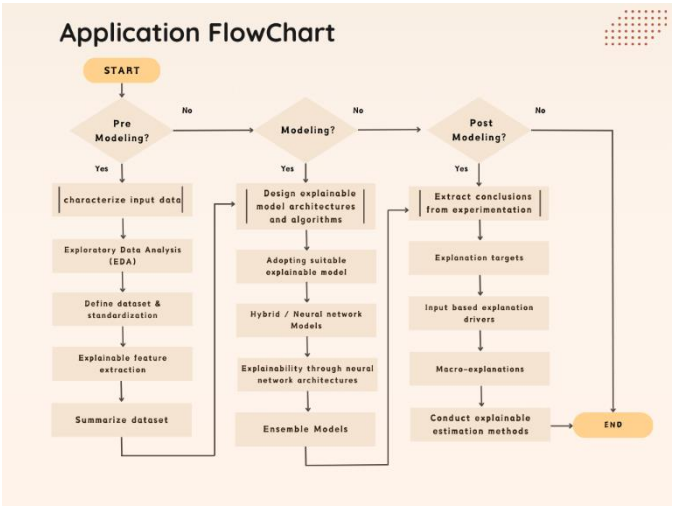


Fig. 2 Modeling Flowchart.

Figure 2. illustrates the comprehensive modelling workflow employed in this study, which encompasses pre-modelling, modelling, and post-modelling stages, with a strong emphasis on interpretability.

The experimental methodology was carefully structured into three key phases: pre-modeling, modeling, and post-modeling, forming a systematic approach to developing a stroke prediction framework that is both interpretable and healthcare relevant. The research began with collecting and curating medical data from openly available sources to ensure a strong foundation for predictive modeling. The dataset included demographic, lifestyle, and clinical attributes such as age, gender, hypertension, heart disease, glucose levels, BMI, marital status, work type, residence type, and smoking status, with stroke occurrence as the target variable. These features were selected to capture key risk factors while ensuring that the model's predictions remain interpretable for healthcare professionals.

In this phase, data preprocessing was a critical step, involving normalization, standardization, and handling of missing values. Missing numerical data, including BMI, were imputed using median values to maintain central tendency without distorting variance, while missing categorical data such as smoking status were labeled as "Unknown" to retain sample completeness. Although advanced imputation strategies (e.g., MICE) may offer higher fidelity, the chosen method provided consistency and interpretability within the comparative experimental framework.

To ensure compatibility across diverse model architectures, preprocessing steps were tailored to each model type rather than applied uniformly. Categorical variables were transformed based on the algorithm's sensitivity: label encoding was applied for tree-based models such as XGBoost and Random Forest, while one-hot encoding was used for distance-based and linear models like Logistic Regression to preserve feature relationships. Numerical features (e.g., age, glucose level, and BMI) were standardized using StandardScaler only for models sensitive to feature magnitude such as Logistic Regression, MLP, and gradient boosting methods to facilitate stable optimization and faster convergence. Tree-based models (Decision Tree, Random Forest, LightGBM) were trained using unscaled numerical inputs, as these algorithms rely on relative thresholds rather than distance metrics. Exploratory Data Analysis (EDA) provided insights into feature distributions, correlations, and key stroke risk factors. Statistical techniques like Pearson correlation, mutual information, and variance inflation factor (VIF) were applied to assess feature relevance and detect multicollinearity.

To refine the dataset, feature selection and extraction were carried out using the SCI-XAI pipeline, incorporating recursive feature elimination (RFE) and correlation-based filtering to retain only the most significant predictors. This

https://doi.org/10.31436/ijpcc.v12i1.636

helped improve model performance while ensuring interpretability. Additionally, feature engineering was applied to incorporate domain knowledge. For example, gender was treated as a binary risk factor due to higher stroke prevalence in females, and age-based stroke risk was introduced using a 45-year threshold, recognizing that stroke risk increases with age. Clinical factors such as hypertension and heart disease were directly included, while marital status was encoded to reflect potential lifestyle stability, which is linked to better health outcomes. Work type was categorized to account for stress levels, distinguishing between high-stress jobs (private/self-employed) and more stable employment (government jobs). Residence type was encoded to capture healthcare access disparities between rural and urban areas.

Metabolic and lifestyle factors were also considered. Individuals with glucose levels in the prediabetic range (126–139.9 mg/dL) were flagged under glucose stroke risk, as elevated glucose levels are associated with cardiovascular events. BMI risk was assigned based on standard classifications for underweight (<18.5) and obesity ($\geqslant$30), recognizing their impact on stroke risk. Smoking status was categorized to distinguish between active smokers and former/non-smokers.

The focus of the modeling phase was to implement explainable machine learning models that balance predictive accuracy with interpretability. Guided by the SCI-XAI framework, both white-box and black-box models were explored to analyze stroke risk factors through multiple methodological perspectives. White-box models, including Logistic Regression, Decision Trees, and the Explainable Boosting Machine (EBM), were first implemented to establish baseline performance and interpretability benchmarks. These models provided transparent insights into in-dividual feature contributions, enabling straightforward interpretation of clinical risk indicators such as age, glucose level, and hypertension. EBM offered additive feature interactions that preserved interpretability while slightly improving accuracy over traditional linear methods. Subsequently, gray and black-box models such as Random Forest, XGBoost, and LightGBM were introduced to capture non-linear interactions and improve predictive robustness. An ensemble-based Adaptive Gradient Boosting model was also experimented with to integrate the complementary strengths of XGBoost and LightGBM, achieving improved stability and predictive performance across evaluation metrics.

To capture complex, non-linear relationships in the data, deep learning models were also explored. A Multi-Layer Perceptron (MLP) neural network was trained to recognize intricate feature interactions. The MLP consisted of two hidden layers (128 and 64 neurons), ReLU activation functions, and dropout regularization (0.3) with batch normalization to prevent overfitting. Although CNN and LSTM architecture were initially explored for potential pattern and sequence modeling, their performance was not sufficiently generalizable for inclusion in the final report due to the dataset's non-sequential nature and limited sample size. Hyperparameter tuning was conducted using a hybrid strategy combining grid search and Bayesian optimization to ensure equitable comparison across models. For the white-box models, hyperparameters were configured based on empirical testing and literature-recommended defaults. Logistic Regression employed ℓ_2 regularization (penalty='l2') with an increased iteration limit (max_iter=1000) to ensure convergence. The Decision Tree classifier used the entropy criterion with a constrained maxi-mum depth of 3 and minimum samples per split set to 2 to prevent overfitting. For the Random Forest model, the number of estimators (100–500), tree depth, and feature subset sizes were adjusted to improve generalization and reduce overfitting. Among the gradient boosting models, XGBoost and LightGBM were fine-tuned for learning rate (0.01–0.3), tree depth (3–10), and number of estimators (100–300) to achieve optimal balance between bias and variance. For the deep learning model, the Multi-Layer Perceptron (MLP) was tuned for learning rate (0.001–0.01), number of neurons per layer (64–256), batch size (32–128), and dropout rate (0.2–0.5). The adaptive Aquila Optimizer was employed to accelerate convergence and prevent stagnation during training, allowing faster stabilization without compromising accuracy. The dataset was divided into training and testing sets using an class–20 stratified split to preserve class balance. Model performance was evaluated using accuracy, precision, recall, F1-score, and ROC-AUC metrics, ensuring a fair and consistent comparison between models differing in complexity and interpretability.

To enhance model reliability and ensure that the evaluation was not biased toward a specific data partition, k-fold cross-validation (k=5) was employed during model training and hyperparameter tuning. This involved dividing the dataset into five equally sized folds, iteratively training on four folds and validating on the remaining one. The process was repeated for all folds, and performance metrics were averaged to obtain a more robust estimate of each model's generalization capability. Cross-validation was particularly applied to the Logistic Regression and Decision Tree, and Random Forest as well as to ensemble approaches such as, XGBoost, and LightGBM, to assess their stability across multiple data splits. For the MLP model, due to computational constraints, an 80–20 stratified split with early stopping and validation monitoring was used instead of full k-fold evaluation.

To reduce overfitting risks arising from the small and imbalanced dataset, stratified sampling and class-balancing via SMOTE were applied. Alternative imbalance-handling strategies such as random under-sampling were also evaluated conceptually but not adopted, as preliminary

https://doi.org/10.31436/ijpcc.v12i1.636

tests indicated a loss of minority-class representation. Future work could further explore hybrid sampling methods to improve generalization. Additionally, validations were employed using re-peated train–test splits. Ensemble models were regularized through depth constraints and learning-rate tuning to prevent excessive fitting to noise as discussed earlier.

The final phase, post-modeling, aimed at making the model's predictions interpretable. Various Explainable AI (XAI) techniques were used to bridge the gap between algorithmic insights and medical decision-making. Feature importance was analyzed using Shapley Additive explanations (SHAP), providing a breakdown of how each variable contributed to stroke predictions. SHAP values offered both global insights (overall feature impact) and local explanations (personalized risk factors), allowing clinicians to understand individual predictions.

Key features such as age, hypertension, glucose level, and BMI were highlighted as primary predictors, aligning with clinical stroke risk assessments. Macro-explanation methods were used to validate decision rules across multiple patient samples, ensuring model consistency. Local interpretability was addressed using Local Interpretable Model-agnostic Explanations (LIME), which generated simplified models to explain individual predictions.

Model evaluation metrics including accuracy, precision, recall, and F1-score were analyzed to assess classification performance. Additionally, the stability of XAI explanations was examined by checking SHAP value variations across different data subsets. This step ensured that the model remained reliable and its interpretations consistent across diverse patient groups.

III. Experimental Results

This section presents the outcomes of the implemented machine learning models across both standard and SCI-XAI pipelines. Emphasis is placed not only on traditional performance metrics such as accuracy, F1-score, and ROC-AUC, but also on the interpretability of the models. Through both global and local explanation techniques, the results are analyzed to evaluate their reliability and clinical decision-making in stroke prediction. To ensure robust evaluation, relevant models were assessed using both hold-out testing and 5-fold cross-validation.

A. Logistic Regression

We first trained a Logistic Regression model as part of a standard machine learning pipeline, including data balancing using SMOTE to address class imbalance. The baseline performance showed promising but nuanced results: the training set achieved an accuracy of 85.6% (AUC = 0.91, weighted F1 = 0.86), while the test set achieved 71.6% accuracy (AUC = 0.84, weighted F1 = 0.79).

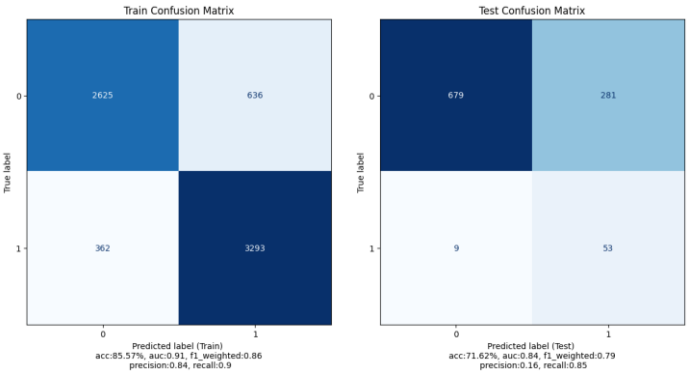


Fig. 3 LogReg Confusion Matrices.

The confusion matrices shown in Figure 3 reveal a familiar pattern for imbalanced datasets: the model achieves high precision (0.99) and F1 (0.82) for the majority class (non-stroke), but low precision (0.16) and moderate F1 (0.27) for the minority class (stroke) despite a strong recall of 0.85. This indicates that while the model successfully detects most stroke cases, it does so at the expense of generating false positives, an expected trade-off in high-sensitivity healthcare screening.

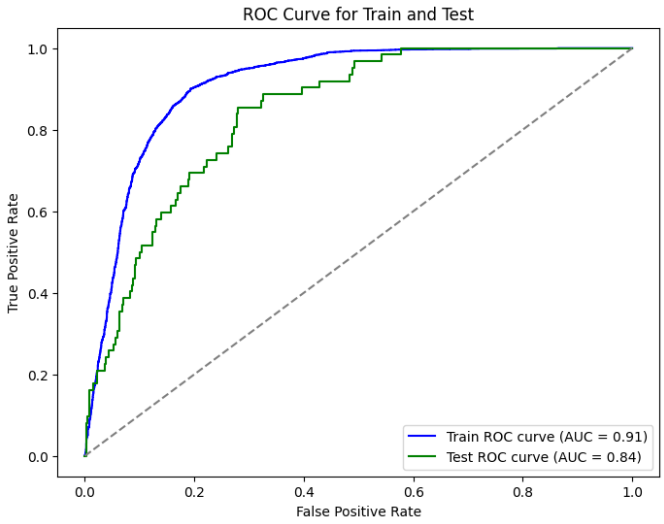


Fig. 4 LogReg ROC-AUC Score.

The ROC curves at Figure 4 confirmed good separability, with AUC scores of 0.91 (train) and 0.84 (test), suggesting minimal overfitting and decent generalization. Cross-validation further validated this stability, yielding a mean AUC of 0.91 ± 0.01, suggesting that the model's predictive performance remained consistent across different data folds. However, these numerical metrics alone do not offer clinical practitioners' insight into why the model makes its predictions, a crucial requirement in sensitive domains like healthcare.

To address this, we integrated the model into the SCI-XAI pipeline, using state-of-the-art interpretability techniques: SHAP, LIME, and ELI5.

https://doi.org/10.31436/ijpcc.v12i1.636

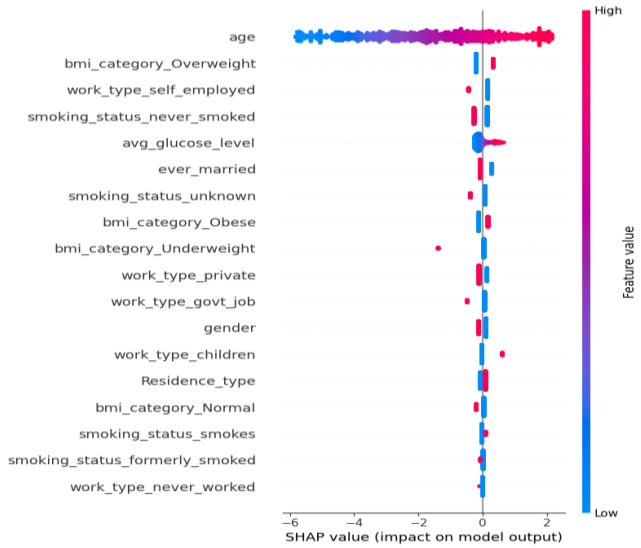


Fig. 5 LogReg SHAP Summary

Figure 5 plot clearly demonstrates that age is the dominant predictor of stroke, with higher age values strongly contributing to positive stroke predictions (as shown by red points on the right). Other impactful features include BMI categories (Overweight, Obese, Underweight) and occupational indicators such as work_type_self_employed and smoking status. This offers clinicians an intuitive way to understand the feature hierarchy and how individual features shift the stroke risk up or down (see Figure 6).

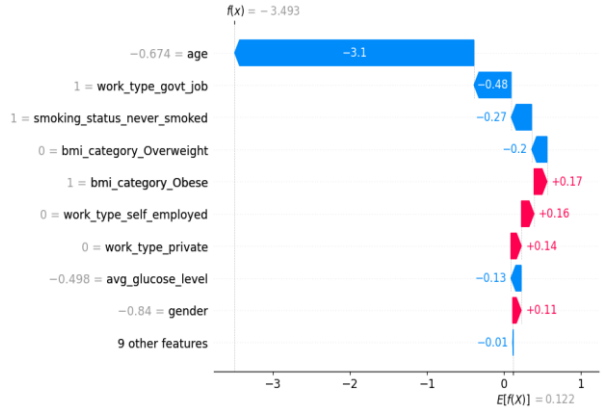


Fig. 6 LogReg SHAP force plot.

Drilling into an individual prediction, we observe that a lower age (-0.674) combined with working in government jobs and a never-smoked status reduces risk (blue bars), while obesity-related BMI and self-employment slightly increased risk (red bars). This granular level of ex-planation is invaluable for personalized risk assessments.

This dependence plot illustrates a strong positive linear relationship between age and stroke risk, with SHAP values increasing steadily from approximately −6 to +2.5 as normalized age rises (see Figure 7).

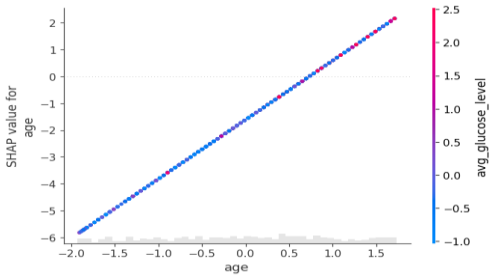


Fig. 7 LogReg SHAP dependence plot.

This trend confirms that higher age substantially elevates the model's predicted stroke risk. The color gradient, representing average glucose level, further suggests that elevated glucose levels amplify this age-related effect, reinforcing the combined influence of metabolic and demographic factors in stroke prediction.

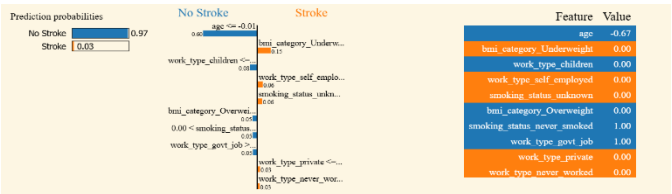


Fig. 8 LogReg LIME Explanations.

Figure 8 emphasize age (coefficient 0.60) as the strongest risk driver, followed by BMI category (Underweight), work_type_children, and work_type_self_employed, which align well with the SHAP findings. The LIME output also includes feature thresholds (e.g., age <= -0.01), making it easy to visualize which side of a decision boundary the patient lies on.

TABLE 1
LOGREG ELI5 EXPLANATIONS.

y=1 top features

Weight[?]	Feature
+2.207	age
+0.645	work_type_children
+0.531	bmi_category_Overweight
+0.291	bmi_category_Obese
	... 3 more positive ...
	... 6 more negative ...
-0.415	smoking_status_never_smoked
-0.465	smoking_status_unknown
-0.549	work_type_govt_job
-0.597	work_type_self_employed
-1.126	<BIAS>
-1.423	bmi_category_Underweight

y=0 (probability **0.970**, score **-3.493**) top features

Contribution[?]	Feature
+1.488	age
+1.126	<BIAS>
+0.549	work_type_govt_job
+0.415	smoking_status_never_smoked
+0.122	ever_married
+0.097	avg_glucose_level
+0.085	Residence_type
-0.096	gender
-0.291	bmi_category_Obese

https://doi.org/10.31436/ijpcc.v12i1.636

Table 1 further strengthens interpretability by breaking down the model's coefficients. For the positive class (stroke), age (+2.207) again tops the list, with BMI and occupational status providing additional weight. Interestingly, smoking_status_never_smoked (-0.415) and work_type_govt_job (-0.549) appear as significant negative contributors, echoing the SHAP force plot. The model's inherent bias term also plays a non-negligible role.

Together, these visualizations and explanations reveal a consistent pattern: age, BMI, and occupational/smoking factors as key stroke risk drivers. The SCI-XAI pipeline enhances the logistic regression model by turning raw predictions into clear, patient-specific explanations, helping interpret and communicate risk effectively by combining solid predictive performance with both global and local interpretability.

B. Decision Trees

The decision tree model achieved a test accuracy of ~77.98%, with a ROC AUC of 0.81 and a weighted F1 score of 0.83. These metrics suggest solid predictive performance, especially the relatively high recall (71%) for the minority class (stroke).

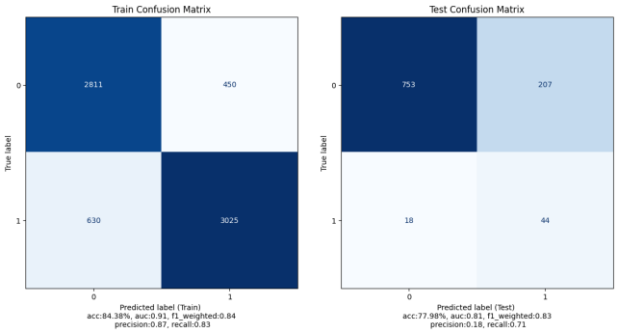


Fig. 9 DT Confusion Matrices.

As demonstrated in Figure 9, the confusion matrices reveal a balanced ability to detect both stroke and non-stroke cases during training, but the test matrix shows some over-prediction of the majority class (non-stroke), which is typical for decision trees even after SMOTE balancing.

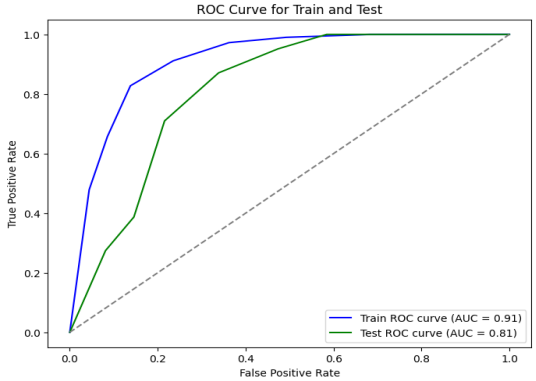


Fig. 10 DT ROC-AUC Scores.

The ROC curves (Figure 10.) demonstrate good class separation capability, comparable to logistic regression but with slightly reduced generalization on the test set (AUC = 0.81 vs. 0.91 for training). Cross-validation reinforced these findings, yielding a mean AUC of 0.91 ± 0.01, indicating that while the model performs consistently across folds, some overfitting tendencies remain due to its hierarchical nature. To further unpack how individual features influence predictions, we applied SHAP analysis to gain both global and local interpretability beyond the raw tree splits.

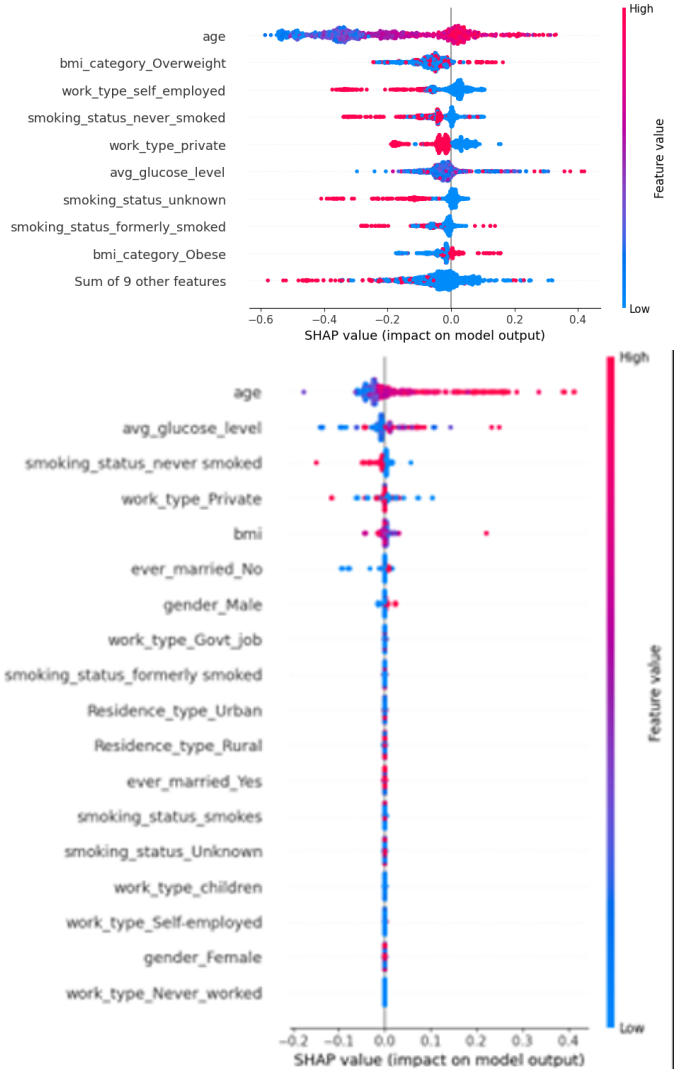


Fig. 11. DT SHAP summary plots.

In Figure 11. the SHAP summary plots emphasizes that age (~0.48) remains the dominant predictor, followed by BMI-related features (~0.14 and ~0.11), smoking status (notably smoking_status, never_smoked, ~0.27), and work type (~0.10) mirroring the findings from the logistic regression but with subtle shifts in feature impact distribution.

https://doi.org/10.31436/ijpcc.v12i1.636

TABLE II
DT LIME EXPLANATIONS

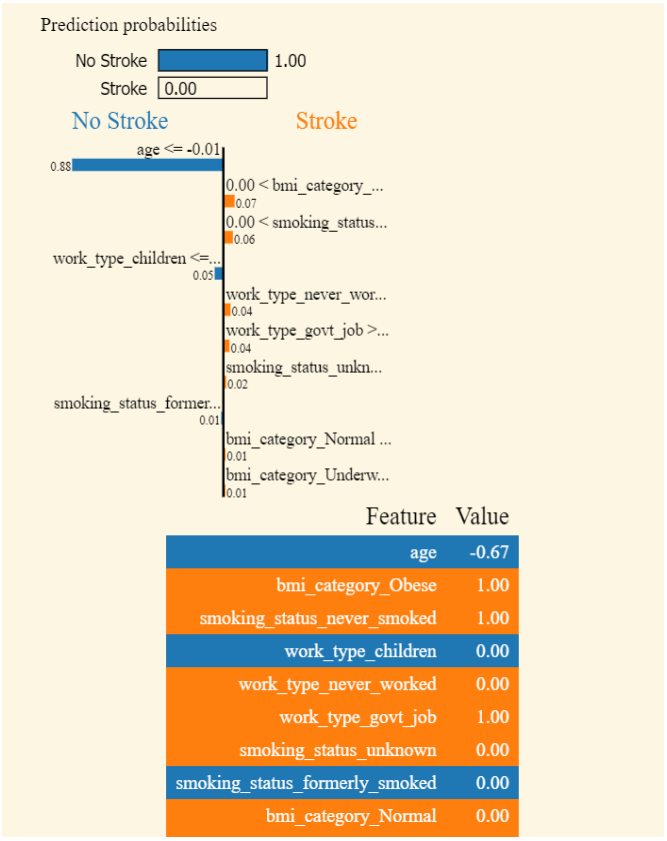


Table 2. reinforces that age and BMI categories are primary drivers of individual predictions, with clear rule-based logic (e.g., "age ⩽ -0.01 contributes 88% toward a 'No Stroke' prediction), reflecting the deterministic nature of decision trees.

Table III
DT ELI5 Explanations.

Explained as: decision tree

Decision tree feature importances; values are numbers 0 <= x <= 1; all values sum to 1.

Weight	Feature
0.4767	age
0.1475	avg_glucose_level
0.0672	smoking_status_never_smoked
0.0390	smoking_status_formerly_smoked
0.0336	smoking_status_unknown
0.0333	bmi_category_Overweight
0.0325	work_type_self_employed
0.0308	bmi_category_Obese
0.0294	work_type_govt_job
0.0250	Residence_type
… 8 more …	

The ELI5 tree breakdown (detailed in the appendix 1 and visualized in Table 3. & Table 4.) provides direct interpretability through the tree structure itself, showing precise splits and thresholds offering clinicians a transparent, step-by-step rationale for each prediction. For example, age contributes +0.494 to the final score (out of a base bias of +0.500), while other features like work_type_self_employed and avg_glucose_level have smaller contributions (+0.004 and +0.002 respectively), reinforcing the dominance of age in prediction. This not only lists feature importance but shows how individual features push predictions higher or lower for a specific case, making it especially useful in medical settings.

Table IV
DT ELI5 Explanations

Prediction Explanation for Predicted Class

Explained as: decision path

Features with largest coefficients.

Feature weights are calculated by following decision paths in trees of an ensemble (or a single tree for DecisionTreeClassifier). Each node of the tree has an output score, and contribution of a feature on the decision path is how much the score changes from parent to child. Weights of all features sum to the output score or proba of the estimator.

Caveats:
1. Feature weights just show if the feature contributed positively or negatively to the final score, and does not show how increasing or decreasing the feature value will change the prediction.
2. In some cases, feature weight can be close to zero for an important feature. For example, in a single tree that computes XOR function, the feature at the top of the tree will have zero weight because expected scores for both branches are equal, so decision at the top feature does not change the expected score. For an ensemble predicting XOR functions it might not be a problem, but it is not reliable if most trees happen to choose the same feature at the top.

y=No Stroke (probability **1.000**) top features

Contribution[?]	Feature
+0.500	<BIAS>
+0.494	age
+0.004	work_type_self_employed
+0.002	avg_glucose_level

Overall, the SCI-XAI pipeline transforms raw model predictions into human-readable, actionable insights. While decision trees are naturally interpretable, SCI-XAI's layered explanations that combine global feature importance (SHAP/ELI5) and individualized prediction paths (LIME) enhance clarity and communication. This makes the tool particularly valuable in clinical contexts, where both performance and explanation are crucial.

C. *Explainable Boosting Machines (EBM)*

Explainable Boosting Machine (EBM) offers a highly interpretable model architecture that combines the power of machine learning with human-understandable outputs. One of its major advantages lies in its interactive nature: through a dropdown interface, users can dynamically inspect global and local explanations for each feature and interaction term. This transparency al-lows clinicians and stakeholders to dissect individual model decisions easily, fostering deeper understanding of the predictive patterns. Notably, the EBM was trained using the SCI-XAI pipeline without requiring additional class imbalance handling, thanks to its inherent robustness in managing such data distributions.

https://doi.org/10.31436/ijpcc.v12i1.636

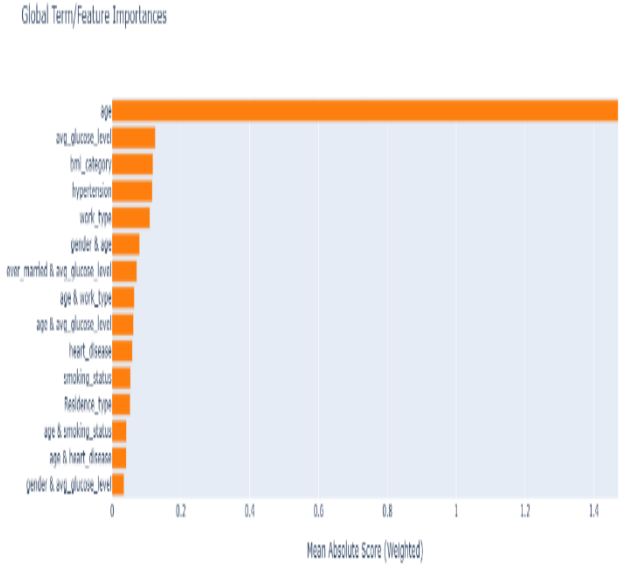


Fig. 12 EBM Explainable Summary.

As shown in Figure 12, the summary plot highlights that age dominates the feature contributions with a mean absolute weighted score around 1.4, followed by average glucose level (~0.15), BMI category (~0.14), and hypertension (~0.13). Other contributing features include work type (~0.12) and notable pairwise interactions such as ever_married & avg_glucose_level (~0.08).

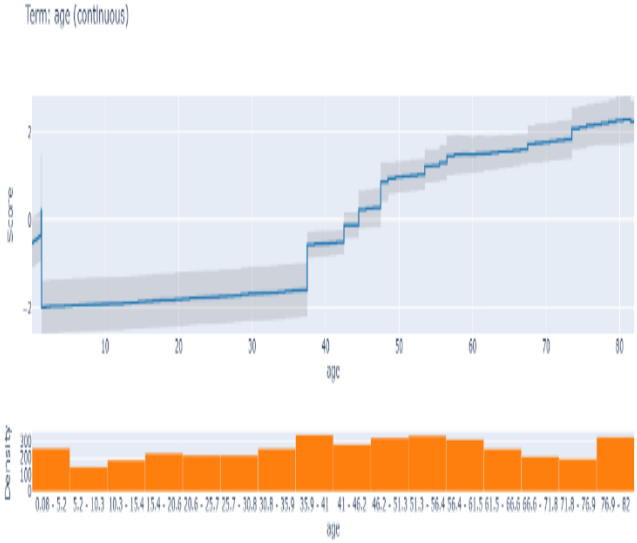


Fig. 13. EBM Global Explanation: Age.

Examining age (Figure. 13) in detail reveals a clear, near-linear increase in risk contribution after age 40, with a sharp inflection between 50–60 years, reaching a maximum contribution of around +2.5 for older patients. These matches established clinical knowledge, reinforcing age as a principal driver of stroke risk.

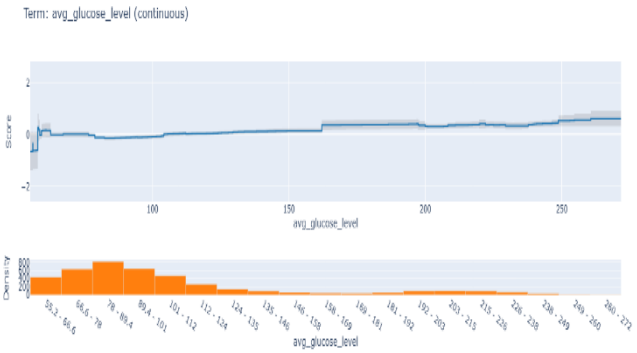


Fig. 14 EBM Global Explanation: Average Glucose Level.

The average glucose level above shows a subtler but still meaningful pattern: risk rises modestly after glucose levels of ~150 mg/dL, peaking at contributions around +1.5, though most density is clustered below 100 mg/dL (see Figure 14).

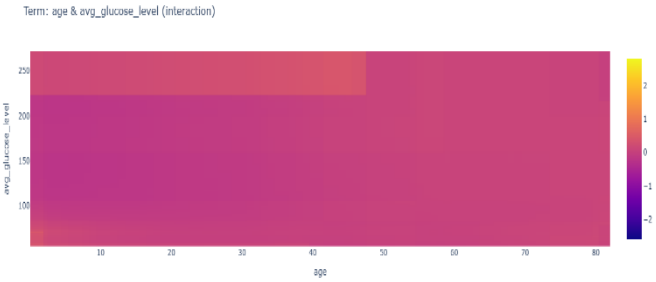


Fig. 15 EBM Global Explanation: Age vs Avg Glucose Lvl.

Figure 15. provides deeper insight into how these two features interplay. The heatmap indicates that higher risk is concentrated in the upper-right quadrant, where both age > 60 and glucose > 200 mg/dL, reflecting compounding effects of these factors.

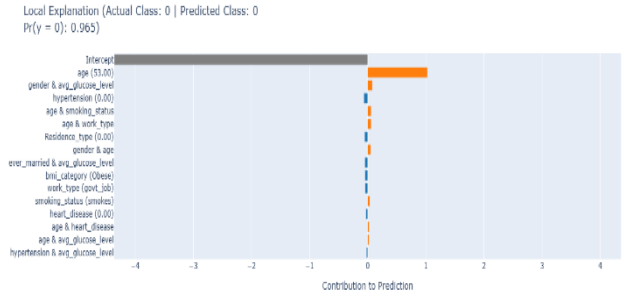


Fig. 16 EBM Local Explanation: (Actual = 0, Predicted = 0, Pr(y=0)=0.965).

Finally, the local explanations at Figure 16. provides us with case-by-case insights, showing exactly how each feature influenced a specific prediction. For instance, in one patient example Actual = 0 (No Stroke), Predicted = 0 (No Stroke), with a predicted probability of 0.965, the model

https://doi.org/10.31436/ijpcc.v12i1.636

was highly confident in its decision. Here, age (53 years) was the dominant factor reinforcing the "No Stroke" prediction, contributing the largest positive weight. Additional, smaller positive contributions came from the combined effect of gender & average glucose level, while features like BMI (Obese) and smoking status (smokes) applied slight opposing pressure (negative contributions), nudging the prediction toward stroke risk but not enough to outweigh the stronger protective signals. This case demonstrates how EBM offers a transparent, numeric breakdown of in-dividual risk profiles, which can be invaluable for clinical decision-making.

D. Random Forest

To extend the exploration beyond white-box models, Random Forest recognized as a grey-box model was applied. While Random Forests are robust and capable of capturing complex, non-linear relationships, they traditionally lack inherent interpretability. To address this gap, we utilized the SCI-XAI pipeline to generate both global and local explanations alongside standard statistical machine learning evaluations (see Figure 17).

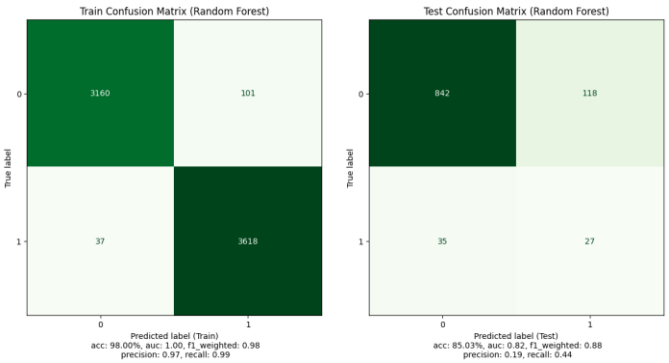


Fig. 17 RF Confusion Matrices.

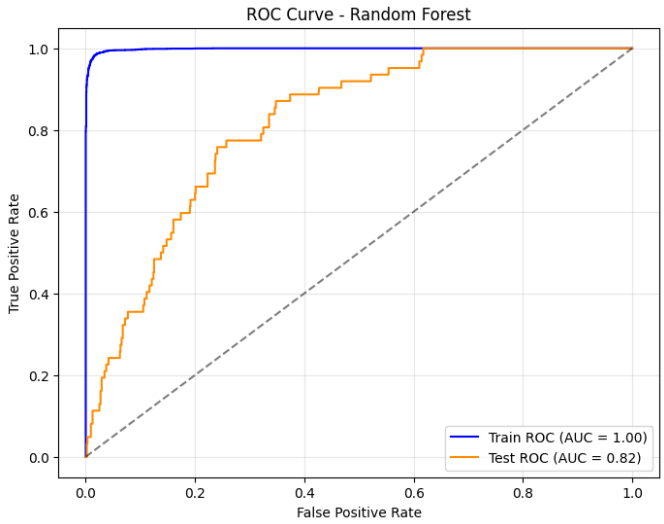


Fig. 18 RF ROC Scores.

The model, trained using a standard machine learning pipeline enhanced with SMOTE for class balancing, yielded solid predictive results: a test accuracy of 85.03% and a ROC AUC score of 0.82, as observed in the classification report. Precision and recall for the minority class (stroke) were 0.19 and 0.44 respectively, reflecting the ongoing challenge of sensitivity in imbalanced medical datasets. To ensure robustness and mitigate potential overfitting, 5-fold cross-validation was performed on the training data. The Random Forest achieved a mean cross-validated ROC-AUC of 0.9952 ± 0.0033, indicating stable and consistent performance across folds despite slight overfitting tendencies observed in single split evaluations (see Figure 18).

To add interpretability, the SCI-XAI pipeline transformed this grey-box model into an explainable one. The SHAP summary plot (see Figure. 19) below highlighted age as the most influential predictor by a wide margin, followed by avg_glucose_level, gender (both male and female), and smoking status. The distribution of SHAP values for age revealed consistent, high-magnitude contributions across many samples, underscoring its pivotal role in the model's decisions.

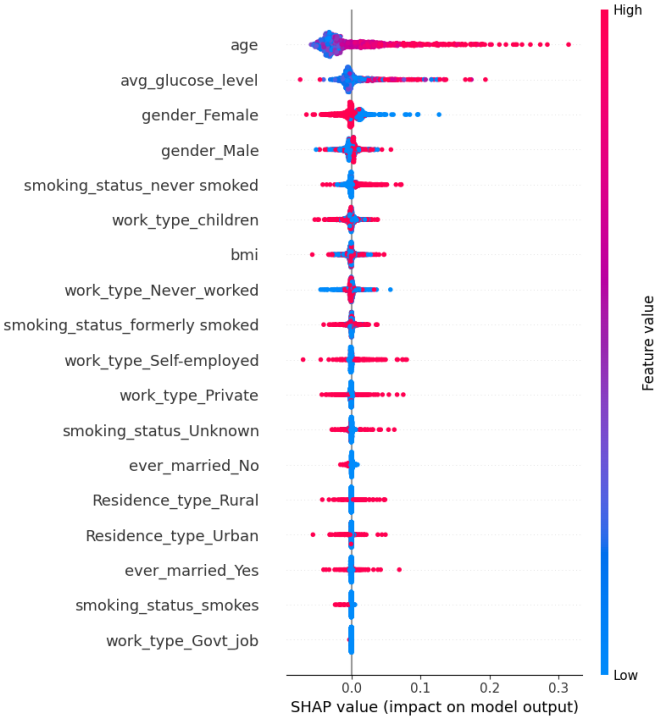


Fig. 19 RF SHAP Summary.

Delving deeper, the dependence plot below for age (Figure 20.) reveals a non-linear relationship with stroke risk. SHAP values remain close to zero for younger individuals (normalized age < 0) but rise sharply beyond approximately 1.0 (around 50 years old), indicating a substantial increase in stroke likelihood with advancing age. The colour gradient represents smoking status, where individuals who have

https://doi.org/10.31436/ijpcc.v12i1.636

never smoked (red) generally exhibit lower SHAP values at comparable ages. This pattern highlights how age and smoking behaviour jointly modulate stroke risk, aligning with established clinical evidence linking aging and lifestyle factors to elevated *cerebrovascular* vulnerability.

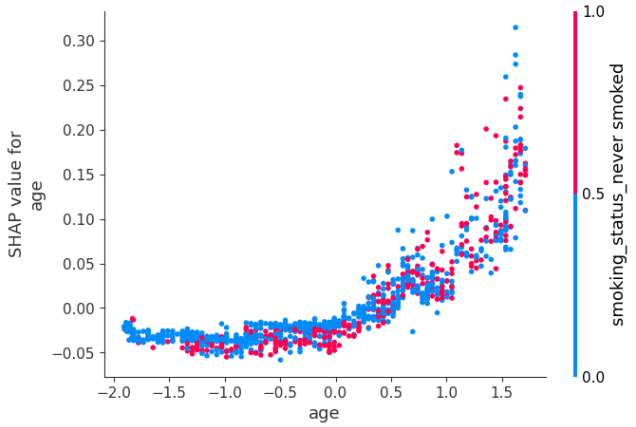


Fig. 20 RF SHAP dependence plot.

E. *Ensemble Model*

The ensemble, comprising XGBoost and LightGBM, classified as a black-box model was trained and evaluated using a standard machine learning pipeline that included class balancing via SMOTE. The model achieved relevantly a high performance on the test set, with an overall accuracy of 88.36%, a ROC AUC score of 0.8048, and a macro-averaged F1-score of 0.58. Notably, for the minority class (stroke cases), the model attained a precision of 0.19 and recall of 0.29, indicating moderate ability to detect stroke instances despite class imbalance challenges.

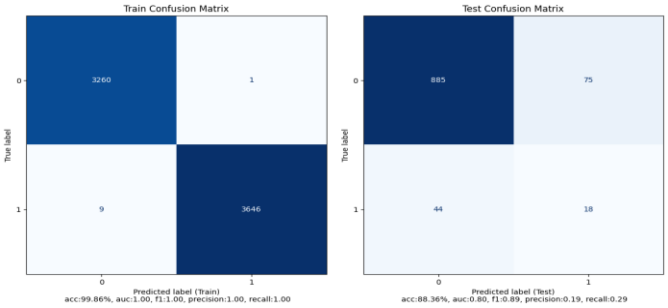


Fig. 21 Ensemble Confusion Matrices.

Figure 21. shows for both training and testing sets that illustrate the model's performance. While the training confusion matrix shows near-perfect classification (accuracy ~99.86%), the test matrix reveals some degradation in performance, particularly in sensitivity (recall) for the minority class. To mitigate potential overfitting, 5-fold cross-validation produced a mean AUC of 0.9965 ± 0.0043, reinforcing the reliability and consistency of the model's predictive performance across data splits.

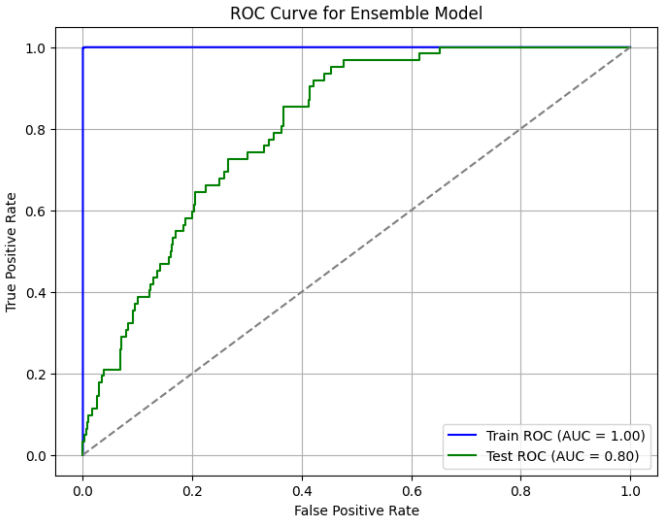


Fig. 22 Ensemble ROC Scores.

The ROC curves further confirm this. The training ROC curve exhibits an AUC of 1.00, indicating perfect separation of classes, while the test ROC curve achieves an AUC of 0.80, suggesting the model maintains good discriminative ability on unseen data but is less ideal than the overly optimistic training performance (see Figure 22).

While these metrics and visualizations (confusion matrix and ROC curve) provide useful quantitative assessments of model performance, they fall short in offering insights into why the model makes certain predictions, an essential factor for clinical decision-making.

To bridge this gap, we applied the SCI-XAI pipeline, employing SHAP (SHapley Additive exPlanations) to deliver both global and local interpretability.

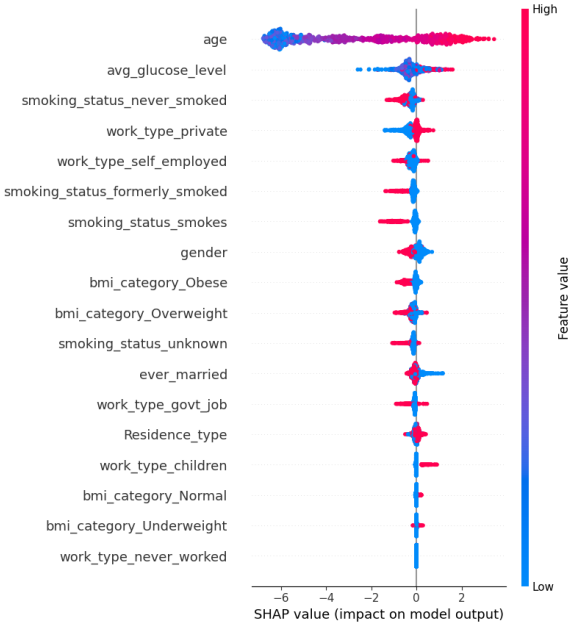


Fig. 23 Ensemble SHAP Summary.

https://doi.org/10.31436/ijpcc.v12i1.636

Figure. 23's summary plot identifies and ranks features based on their overall impact on the model's predictions. Here, age stands out as the most significant predictor, followed by average glucose level, smoking status, and various work type and BMI category variables. The color gradient (red to blue) visually maps each feature's value (e.g., high vs. low age), and its influence on pushing predictions toward stroke or non-stroke. This plot not only validates clinical knowledge (e.g., age and glucose level being critical risk factors) but also helps clinicians understand nuanced associations, such as how certain employment types or smoking history influence stroke likelihood. This global view is crucial for population-level insights and aligns the model's reasoning.

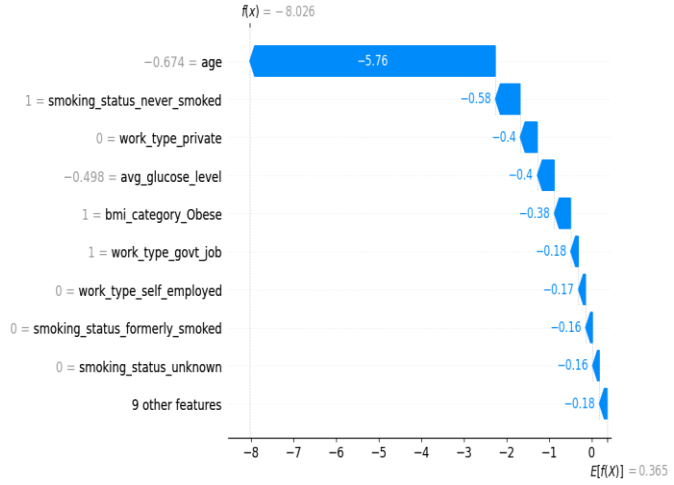


Fig. 24 Ensemble SHAP Waterfall force plot.

The waterfall plot dissects a single prediction to show how individual features cumulatively influence the model's output. For example, in one instance, age contributes a significant negative impact, decreasing the likelihood of stroke, while features like smoking status (never smoked) and BMI (obese) contribute to smaller, nuanced effects. This granular view allows clinicians to trace back the reasoning for specific patients, ensuring that individual predictions are transparent and defensible, a key requirement in high-stakes environments like healthcare (see Figure 24).

F. Multi-Layer Perceptron Model

The model comprised two hidden layers (128 and 64 neurons) with ReLU activation, dropout (0.3), and batch normalization to prevent overfitting. Using the Adam optimizer (learning rate = 0.001) and binary cross-entropy loss, it achieved 79.2% accuracy and an ROC-AUC of 0.78. While slightly less accurate than ensemble models, the MLP captured non-linear feature interactions effectively with stable generalization.

Fig. 25 Training Validation Accuracy & Loss

These graphs at Figure 25. illustrate the MLP's learning dynamics over 50 epochs. The training accuracy (left) steadily increased to around 95%, while validation accuracy plateaued near 80%, indicating the model learned well but began to generalize less effectively after ~20 epochs. Similarly, the training loss (right) consistently declined to about 0.12, whereas validation loss fluctuated between 0.5–0.7, suggesting mild overfitting, the model fits training data very well but shows higher error on unseen data. Overall, the network achieved stable performance but could benefit from stronger regularization or early stopping to improve generalization (see Figure 26).

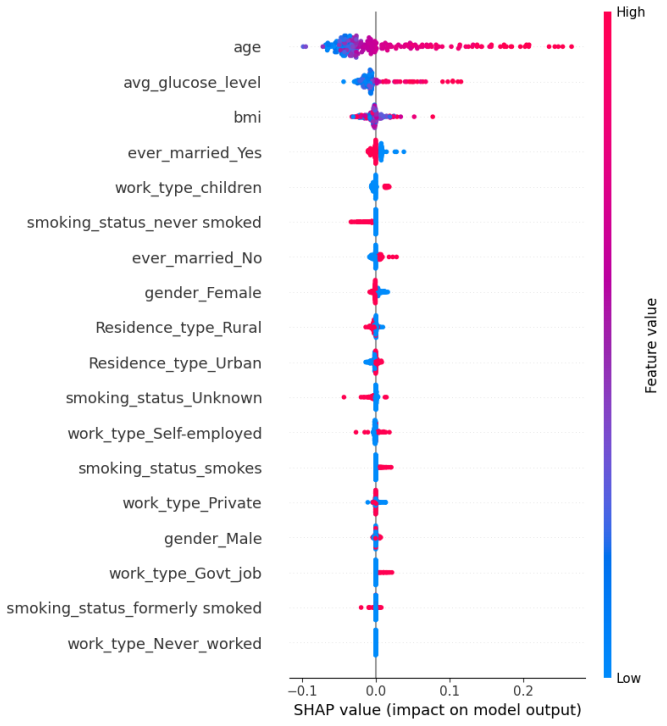


Fig. 26 MLP SHAP Summary Plot.

The summary plot demonstrates that age, average glucose level, and BMI were the most influential predictors in the MLP model, showing the highest positive SHAP values (approximately +0.25, +0.20, and +0.15, respectively). Higher values for these features significantly increased the model's

https://doi.org/10.31436/ijpcc.v12i1.636

predicted stroke risk, consistent with established clinical evidence. Moderate effects were observed for marital status and smoking-related variables, while demographic and occupational features contributed minimally, clustering near zero SHAP values. The colour gradient indicates that high feature values (red) generally pushed predictions toward higher stroke probability, whereas lower values (blue) reduced it. Overall, these findings highlight that metabolic and age-related factors predominantly drive the MLP model's decision-making process.

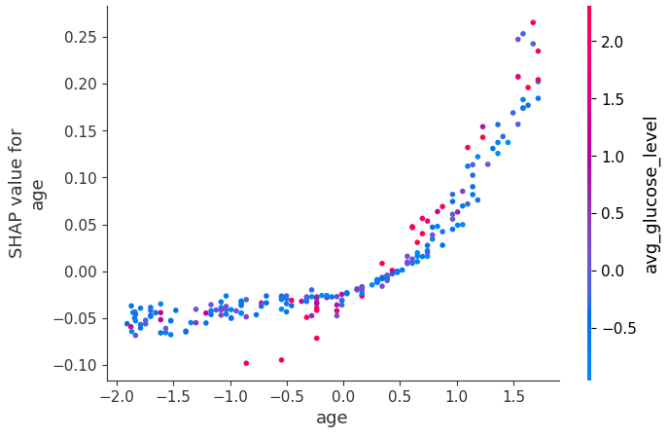


Fig. 27 MLP SHAP Dependence Plot.

Figure 27. plot reveals a strong, non-linear relationship between the feature age and the MLP model's output for stroke prediction. For low, normalized age values (e.g., from -2.0 to 0.0), the SHAP value is consistently negative, ranging from approximately -0.07 to 0.00, indicating that younger ages decrease the model's predicted probability of stroke. Conversely, as age increases beyond 0.0, the SHAP value rises exponentially, peaking at approximately +0.27 for the highest normalized age value of around 1.75. This demonstrates that increasing age is a major contributor to a higher predicted stroke probability. Furthermore, the plot highlights an interaction effect with avg_glucose_level (color-coded). At the highest age values (age > 1.0), the largest positive SHAP contributions (up to +0.27) are predominantly associated with the highest avg_glucose_level values (red, up to 2.0), suggesting that the positive impact of advanced age on stroke risk is amplified by a high average glucose level.

By integrating the SCI-XAI pipeline, we extend all the model's utility beyond raw predictive performance into the realm of Explainable AI (XAI). This is particularly valuable in medical contexts, where black-box predictions are often viewed with scepticism. The SHAP-based interpretability framework not only boosts interpretation confidence but also supports patient-specific consultations, potentially enabling caregivers to explain risk profiles in understandable terms.

IV. Discussion

Our results demonstrate that the SCI-XAI pipeline effectively balances predictive performance and interpretability in stroke prediction. The comparative framework particularly EBM and the SHAP-augmented ensemble achieved ROC-AUC and F1-scores closely matching those of black-box models, confirming that interpretability did not come at the cost of accuracy. This aligns with literature suggesting that for tabular data, advanced interpretable models like EBM can capture nonlinear patterns without compromising performance [26] [27]. For example, in prior work an EBM attained accuracy nearly on par with a random forest while remaining a transparent "glass-box" model [26] In our study, while the XGBoost + LightGBM ensemble achieved the highest metrics, EBM closely rivaled it, falling only marginally short. This finding reinforces the idea that interpretable models can deliver state-of-the-art performance while providing trans-parency an important insight for developing trustworthy AI systems [27] The key predictive fac-tors identified age, hypertension, and heart disease were consistent across EBM and SHAP anal-yses and align with established stroke risk factors reported in the literature (e.g., atrial fibrillation, blood pressure) [26] [27]. This consistency supports the validity of the models' reasoning and suggests that the SCI-XAI pipeline produces explanations that align well with domain knowledge.

A core aspect of the evaluation was comparing global and local interpretability. EBM, as an intrinsically interpretable model, offered clear global insights through its feature effect plots. For instance, stroke risk increased sharply with age beyond a certain threshold and rose consistently in patients with hypertension. Locally, EBM decomposed each prediction into additive contributions, showing precisely how individual features influenced outcomes and strengthened its white-box design [26].

By contrast, the ensemble (XGBoost + LightGBM) is a black-box model that cannot be directly interpreted [27]. Through SCI-XAI, we applied SHAP to this ensemble, enabling post-hoc inter-pretability. SHAP's global summaries highlighted top features such as age, glucose level, and hypertension closely mirroring EBM's insights. Locally, SHAP's force plots broke down individual predictions, showing how specific features (e.g., "Age = 75," "Hypertension = Yes") raised or lowered stroke risk compared to the baseline. The agreement between EBM's intrinsic explana-tions and SHAP's post-hoc outputs strengthens confidence in the reliability of these interpreta-tions.

https://doi.org/10.31436/ijpcc.v12i1.636

Additionally, traditional models like logistic regression and decision trees were examined. Logistic regression provided interpretable coefficients but was limited by its linear assumptions, making it less able to capture nonlinear risk effects. Decision trees were intuitive but, when deeper, became harder to interpret and underperformed compared to EBM and the ensemble. EBM im-proved upon both by modeling complex patterns transparently. Meanwhile, SHAP successfully added interpretability to the ensemble, demonstrating how post-hoc XAI tools can make high-performing black-box models more explainable [26].

While ensemble models such as Random Forest and XGBoost achieved high training perfor-mance, the observed discrepancy between training and testing metrics indicates mild overfitting. This is expected given the limited dataset size and high-class imbalance. Nonetheless, the applied mitigation strategies including tree depth limitation, regularization, and SMOTE-based resampling partially reduced this gap. Importantly, the study's focus is methodological rather than clinical deployment; thus, the results primarily demonstrate feasibility and interpretability rather than clinical readiness.

Overall, these findings highlight the value of combining inherently interpretable models with post-hoc explainability methods to achieve both accuracy and transparency. While this study does not develop a clinical tool, it offers a methodological framework that can inform future work aiming for transparent, interpretable AI in healthcare diagnostics.

V. Limitations & Future Work

While the results are promising, several limitations and avenues for future research must be noted. First, the ensemble model (XGBoost + LightGBM), though highly performant, is complex and risks overfitting particularly given modest sample size and high feature dimensionality. We mitigated this through cross-validation, depth control and hyperparameter tuning, but generalizability to new data cannot be guaranteed. Thus, future studies should validate these findings on larger, external datasets, ideally incorporating real-world clinical records to strengthen reliability. Simpler models (e.g., logistic regression, decision trees) were more robust to overfitting but showed lower predictive accuracy. Future work could explore more regularization techniques, simplified ensemble strategies, or hybrid models to retain performance while reducing complexity. Additionally, stroke was modeled as a binary classification task due to dataset constraints, which overlook the temporal progression of stroke risk. Future studies could incorporate longitudinal or time-to-event data to better capture real-world disease trajectories.

Second, the dataset was highly imbalanced, with stroke cases representing fewer than 5% of the total. While SMOTE oversampling and class-balanced training improved sensitivity, synthetic sampling may introduce artifacts and may not fully represent true population distributions. The dataset also reflects a specific demographic and regional context, which limits external validity. Future research should validate these models on broader, multi-center datasets and test alternative imbalance-handling techniques such as cost-sensitive learning or anomaly detection to enhance generalizability [26].

Third, feature limitations impacted the models. Key risk factors such as diet, exercise, and detailed cardiac history (e.g., atrial fibrillation) were missing, and some variables (like "heart disease") were overly broad. Incorporating richer clinical data including imaging, lab tests, and genetic information could improve both accuracy and interpretability. Additionally, investigating model calibration to ensure that predicted probabilities align with absolute stroke risk remains an important next step.

Lastly, while this study focused on technical interpretability using tools like SHAP, ELI5, and LIME, the assessment of interpretability remains *theoretical*. We did not conduct user studies to evaluate how healthcare professionals perceive and apply these explanations. Future work should include human-factors evaluations to measure usability, trust, and practical value, as well as explore ways to simplify outputs (e.g., streamlined SHAP visuals or natural-language summaries). Expanding interactive feedback loops where clinicians can flag unexpected results could also support continuous model refinement.

Overall, addressing these limitations will make future iterations of this framework more robust, generalizable, and useful as a benchmark for transparent AI in healthcare research. While clinical impact is a long-term goal, this work primarily contributes a methodological foundation for balancing predictive performance and interpretability in medical diagnostics.

VI. Conclusions

This study applied the SCI-XAI pipeline a structured framework combining advanced machine learning with Explainable AI (XAI) techniques to benchmark stroke risk prediction models with an emphasis on interpretability and performance. By integrating both white-box models, such as the Explainable Boosting Machine (EBM), and black-box ensembles (XGBoost + LightGBM) augmented with SHAP, LIME, and ELI5 explanations, we demonstrated that interpretable models could achieve predictive accuracy comparable to complex, opaque models while providing much clearer and more actionable insights.

Our comparative analysis highlighted that EBM, despite being inherently interpretable, closely rivaled the performance of black-box ensembles in terms of ROC-AUC and F1-score, reinforcing that high predictive power and transparency can coexist especially in tabular medical

datasets. SHAP and other post-hoc explainability tools effectively bridged the gap for black-box models, ensuring that even complex algorithms could be made more interpretable and suitable for actionable insights. Crucially, the key predictive features identified such as age, hypertension, and heart disease aligned with established clinical knowledge, lending further credibility to the out-puts and supporting the relevance of XAI in medical diagnostics. [26].

In conclusion, our study highlights the importance and feasibility of bringing interpretability to the forefront of white, gray-black-box models in healthcare. While the models presented are not intended for immediate clinical deployment, the structured application of the SCI-XAI framework provides a robust foundation for future research aiming to make AI in healthcare more trans-parent and reliable. This work contributes valuable insights into the trade-offs between model complexity, interpretability, and performance, offering a roadmap for researchers and practitioners interested in balancing these dimensions.

Declarations: This study did not involve any human participants, and all data was obtained from publicly available open-source datasets. The research was conducted using publicly accessible data in compliance with relevant guidelines and regulations. This work received no specific funding from any public, commercial, or not-for-profit sources.

Acknowledgment

The author expresses sincere gratitude to the Most Gracious and Most Merciful, for His guidance and blessings throughout this work. Deep appreciation is extended to Dr. Sharyar Wani for his invaluable supervision, insightful feedback, and continuous encouragement, which greatly shaped and strengthened this study. The author also thanks peers and colleagues for their constructive discussions and assistance during the research process, and family members for their unwavering support and motivation.

Conflict of Interest

The authors declare that there is no conflict of interest.

Author(s) Contribution Statement

All authors contributed equally to this work.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Ethics Statement

This study did not require ethical approval

References

[1]. H. O'Brien Quinn, M. Sedky, J. Francis, and M. Streeton, "Literature Review of Explainable Tabular Data Analysis," Electronics (Basel), vol. 13, no. 19, p. 3806, Sep. 2024, doi: 10.3390/electronics13193806.

[2]. A. M. Antoniadi et al., "Current Challenges and Future Opportunities for XAI in Machine Learning-Based Clinical Decision Support Systems: A Systematic Review," Applied Sciences, vol. 11, no. 11, 2021, doi: 10.3390/app11115088.

[3]. P. A. Moreno-Sanchez, "An automated feature selection and classification pipeline to improve explainability of clinical prediction models," in Proceedings - 2021 IEEE 9th International Conference on Healthcare Informatics, ISCHI 2021, Institute of Electrical and Electronics Engineers Inc., Aug. 2021, pp. 527–534. doi: 10.1109/ICHI52183.2021.00100.

[4]. U. Pawar, D. O'shea, S. Rea, and R. O'reilly, "Incorporating Explainable Artificial Intelligence (XAI) to aid the Understanding of Machine Learning in the Healthcare Domain."

[5]. J. Ospel, N. Singh, A. Ganesh, and M. Goyal, "Sex and Gender Differences in Stroke and Their Practical Implications in Acute Care," Jan. 01, 2023, Korean Stroke Society. doi: 10.5853/jos.2022.04077.

[6]. K. M. Rexrode, T. E. Madsen, A. Y. X. Yu, C. Carcel, J. H. Lichtman, and E. C. Miller, "The Impact of Sex and Gender on Stroke," Circ Res, vol. 130, no. 4, pp. 512–528, Feb. 2022, doi: 10.1161/CIRCRESAHA.121.319915.

[7]. M. Wajngarten and G. Sampaio Silva, "Hypertension and stroke: Update on treatment," European Cardiology Review , vol. 14, no. 2, pp. 111–115, 2019, doi: 10.15420/ecr.2019.11.1.

[8]. W. Kim and E. J. Kim, "Heart failure as a risk factor for stroke," Jan. 01, 2018, Korean Stroke Society. doi: 10.5853/jos.2017.02810.

[9]. C. Zhu et al., "The association of marital/partner status with patient-reported health outcomes following acute myocardial infarction or stroke: Protocol for a systematic review and meta-analysis," Nov. 01, 2022, Public Library of Science. doi: 10.1371/journal.pone.0267771.

[10]. E. S. Eshak et al., "Changes in the Employment Status and Risk of Stroke and Stroke Types," Stroke, vol. 48, no. 5, pp. 1176–1182, May 2017, doi: 10.1161/STROKEAHA.117.016967.

[11]. O. Grimaud et al., "Stroke incidence and case fatality according to rural or urban residence results from the French Brest Stroke Registry," Stroke, vol. 50, no. 10, pp. 2661–2667, Oct. 2019, doi: 10.1161/STROKEAHA.118.024695.

[12]. X. Peng et al., "Longitudinal Average Glucose Levels and Variance and Risk of Stroke: A Chinese Cohort Study," Int J Hypertens, vol. 2020, 2020, doi: 10.1155/2020/8953058.

[13]. K. Miwa et al., "Clinical impact of body mass index on outcomes of ischemic and hemorrhagic strokes," International Journal of Stroke, Oct. 2024, doi: 10.1177/17474930241249370.

[14]. J. Chen et al., "Impact of Smoking Status on Stroke Recurrence," J Am Heart Assoc, vol. 8, no. 8, Apr. 2019, doi: 10.1161/JAHA.118.011696.

[15]. M. S. Islam, I. Hussain, M. M. Rahman, S. J. Park, and M. A. Hossain, "Explainable Artificial Intelligence Model for Stroke Prediction Using EEG Signal," Sensors, vol. 22, no. 24, Dec. 2022, doi: 10.3390/s22249859.

[16]. A. Laios et al., "Factors Predicting Surgical Effort Using Explainable Artificial Intelligence in Advanced Stage Epithelial Ovarian Cancer," Cancers (Basel), vol. 14, no. 14, Jul. 2022, doi: 10.3390/cancers14143447.

[17]. S. K. Mandala, "XAI Renaissance: Redefining Interpretability in Medical Diagnostic Models," Jun. 2023, [Online]. Available: http://arxiv.org/abs/2306.01668

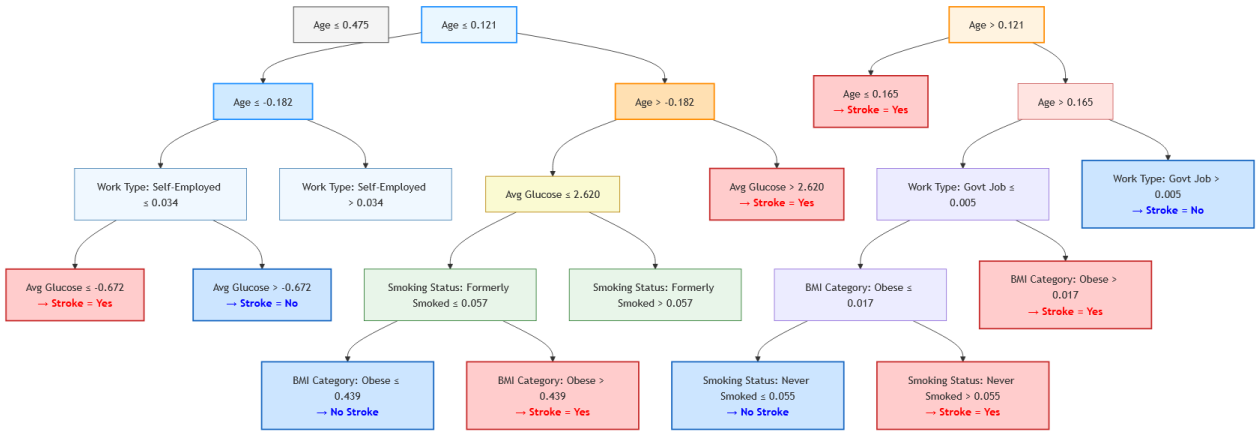
[18]. V. Petrauskas et al., "XAI-based Medical Decision Support System Model," International Journal of Scientific and Research Publications (IJSRP), vol. 10, no. 12, pp. 598–607, Dec. 2020, doi: 10.29322/ijsrp.10.12.2020.p10869.

[19]. T. A. J. Schoonderwoerd, W. Jorritsma, M. A. Neerincx, and K. van den Bosch, "Human-centered XAI: Developing design patterns for explanations of clinical decision support systems," International Journal of Human Computer Studies, vol. 154, Oct. 2021, doi: 10.1016/j.ijhcs.2021.102684.

[20]. J. Stodt, M. Madan, C. Reich, L. Filipovic, and T. Ilijas, "A Study on the Reliability of Visual XAI Methods for X-Ray Images," in Studies in

Health Technology and Informatics, IOS Press BV, Jun. 2023, pp. 32–35. doi: 10.3233/SHTI230416.
[21]. S. Alkhalaf et al., "Adaptive Aquila Optimizer with Explainable Artificial Intelligence-Enabled Cancer Diagnosis on Medical Imaging," Cancers (Basel), vol. 15, no. 5, Mar. 2023, doi: 10.3390/cancers15051492.
[22]. R. El Shawi, Y. Sherif, M. Al-Mallah, and S. Sakr, "Interpretability in HealthCare A Comparative Study of Local Machine Learning Interpretability Techniques," in 2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS), IEEE, Jun. 2019, pp. 275–280. doi: 10.1109/CBMS.2019.00065.
[23]. S. S, K. Chadaga, N. Sampathila, S. Prabhu, R. Chadaga, and S. K. S, "Multiple Explainable Approaches to Predict the Risk of Stroke Using Artificial Intelligence," Information, vol. 14, no. 8, p. 435, Aug. 2023, doi: 10.3390/info14080435.
[24]. B. S. D. Darshan et al., "Differential diagnosis of iron deficiency anemia from aplastic anemia using machine learning and explainable Artificial Intelligence utilizing blood attributes," Sci Rep, vol. 15, no. 1, p. 505, Jan. 2025, doi: 10.1038/s41598-024-84120-w.
[25]. B. Khokhar, V. Pentangelo, F. Palomba, and C. Gravino, "Towards Transparent and Accurate Diabetes Prediction Using Machine Learning and Explainable Artificial Intelligence." [Online]. Available: https://www.kaggle.com/datasets/
[26]. S. Lolak, J. Attia, G. J. McKay, and A. Thakkinstian, "Comparing Explainable Machine Learning Approaches With Traditional Statistical Methods for Evaluating Stroke Risk Models: Retrospective Cohort Study," JMIR Cardio, vol. 7, 2023, doi: 10.2196/47736.
[27]. S. Kruschel, N. Hambauer, S. Weinzierl, S. Zilker, M. Kraus, and P. Zschech, "Challenging the Performance-Interpretability Trade-Off: An Evaluation of Interpretable Machine Learning Models," Business and Information Systems Engineering, 2025, doi: 10.1007/s12599-024-00922-2.

APPENDIX 1: Simplified Decision Tree Visualization for Stroke Prediction



Visualization generated via Mermaid (https://mermaid.js.org/), adapted for academic presentation.

This appendix presents the ELI5-generated decision tree breakdown, corresponding to the interpretability discussion in the main text (referenced in Table 3 and Table 4). The visualization was reproduced using the Mermaid diagram-ming framework to illustrate the internal decision logic of the trained model. Each branch represents a feature-based split (e.g., Age, Average Glucose Level, BMI Category, Smoking Status, Work Type), providing a trans-parent, step-by-step rationale for the model's classification outcomes.

Color coding enhances interpretability: blue nodes indicate pathways leading to No Stroke predictions, while red nodes denote Stroke = Yes outcomes. Intermediate nodes shaded in orange or light blue correspond to decision points based on threshold values that guide the classification process.

Note: Numeric thresholds (e.g., Age ≤ 0.475) represent normalized feature values derived through Min–Max scaling, where all continuous features were transformed to a [0,1] range. These values reflect the relative percentile positions within the dataset (for example, Age ≤ 0.475 corresponds to individuals below approximately the 47.5th per-centile of the observed age range). This normalization ensures consistent feature comparison across different clinical measures and supports stable, interpretable model behavior.