ISSN: 2977-0041

Research Article

Journal of Material Sciences and Engineering Technology

SHAP-Enhanced Machine Learning for Explainable Stroke Risk Prediction in Hypertensive Patients

David Olanrewaju Akinwale[1], Oluwasegun Anthony Bosede[2] and Olushina Olawale Awe[3*]

[1]*Obafemi Awolowo University, Nigeria*

[2]*SI "Lugansk State Medical University", Ukraine*

[3]*University of Campinas, Brazil*

[*]**Corresponding author**

Olushina Olawale Awe, University of Campinas, Brazil.

Received: October 04, 2025; **Accepted:** October 10, 2025; **Published:** October 18, 2025

ABSTRACT

Background: Stroke remains a critical global health concern, disproportionately affecting individuals with hypertension which is a well-established, modifiable risk factor. Traditional risk scoring systems often fall short in accurately predicting stroke onset due to their reliance on fixed clinical thresholds and limited variable interaction modeling. As the complexity of health data increases, machine learning (ML) and explainable artificial intelligence (XAI) present powerful tools to uncover hidden patterns and enable precision risk stratification.

Objectives: This study proposes a novel, interpretable machine learning framework that uses ensemble learning and SHapley Additive exPlanations (SHAP) to enhance stroke risk prediction in hypertensive patients. The objective is twofold: to improve the predictive power of stroke models and to provide clinically relevant insights that support real-time, data-driven decisions in preventive care.

Methods: We utilized a real-world clinical dataset encompassing demographic, physiological, and behavioral variables associated with stroke. Data preprocessing included k-nearest neighbor imputation for missing values, normalization of continuous features, and class balancing via Synthetic Minority Oversampling Technique (SMOTE). A hybrid feature selection pipeline, combining the sparsity-enforcing capabilities of LASSO regression with the iterative refinement of Recursive Feature Elimination (RFE), was employed to identify the most salient predictors. Multiple ML models, including logistic regression, deep neural networks, random forests, and gradient boosting machines, were trained and validated using cross-validation. SHAP values were computed post-training to enable individualized, interpretable model outputs.

Results: Ensemble models, particularly Gradient Boosting and Random Forest, demonstrated superior discriminative performance, achieving AUC-ROC scores above 0.78 following class balancing. The integrated LASSO-RFE approach revealed age, hypertension status, and average glucose levels as dominant predictors across models. SHAP visualizations confirmed the influence of these features, while also highlighting nuanced interactions involving lifestyle and socioeconomic variables. Logistic Regression, when optimized for recall, achieved the highest balanced accuracy (0.77), reinforcing the clinical utility of simpler models when interpretability is paramount.

Conclusion: This study introduces a transparent and high-performing machine learning framework for stroke risk prediction in hypertensive individuals. By integrating ensemble learning, hybrid feature selection, and explainable AI, the framework bridges the gap between predictive modeling and clinical applicability. These findings support the deployment of interpretable ML tools in routine care, enabling proactive interventions, personalized patient education, and ultimately, reduced stroke incidence.

Keywords: Machine learning, Stroke Risk Prediction, Hypertension, Predictive Modeling, Precision Medicine

Introduction

Hypertension is a leading modifiable risk factor for stroke, which remains a top cause of disability and death globally

Citation: David Olanrewaju Akinwale, Oluwasegun Anthony Bosede, Olushina Olawale Awe. SHAP-Enhanced Machine Learning for Explainable Stroke Risk Prediction in Hypertensive Patients. J Mat Sci Eng Technol. 2025. 3(4): 1-13. DOI: doi.org/10.61440/JMSET.2025.v3.78

[1]. Despite a growing prevalence of hypertension, current risk assessment tools, such as the Framingham Risk Score and CHA2DS2-VASc Score, are limited by their reliance on static clinical variables. These traditional models often fail to capture the complex interplay of genetic, physiological, and lifestyle factors that influence stroke risk over time [2]. A dynamic, continuously learning approach is therefore essential for developing personalized risk assessments and targeted preventive interventions. Recent advances in machine learning (ML) and artificial intelligence (AI) provide a promising avenue for enhancing stroke risk prediction. By leveraging diverse data sources, including long-term patient records, biomarker analyses, real-time physiological monitoring, and imaging data, ML models can uncover hidden patterns that traditional statistical methods might miss [3,4]. These models not only offer more precise and individualized predictions but also support real-time clinical decision-making, enabling timely modifications to treatments and interventions [5,6].

The application of ML in healthcare is gaining momentum as medical records become increasingly digitized and computational capabilities advance. Unlike traditional population-based risk assessments, ML approaches can integrate data from wearable devices, electronic health records, genetic markers, and advanced imaging to tailor risk predictions to each patient. This shift towards precision medicine holds the potential to reduce healthcare disparities by providing more equitable, individualized risk assessments that reflect the socioeconomic and demographic diversity of patient populations [7]. This research is dedicated to developing and evaluating ML models that offer enhanced stroke risk prediction for hypertensive patients. By addressing challenges such as data quality, model interpretability, and clinical workflow integration [8-10], the study aims to establish robust, evidence-based frameworks that facilitate earlier interventions. The novelty of this work lies in creating an integrated framework that combines hybrid machine learning techniques with advanced data handling methods to balance sensitivity and specificity optimally.

Table 1 shows the comparison of the results in this study with previous knowledge in the field.

Table 1: Comparative Summary of Contributions: Our Study vs. Related Works in Stroke Risk Prediction

Study	Methodology/Focus	Key Contributions	Limitations Identified	Our Study's Advancement
Framingham Models	Classical statistical regression models	Provided foundational risk estimation tools	Oversimplified relationships; ignored non-linear interactions	Replaced by ML models capturing complex, non-linear dynamics
Vu et al. [11]	Multi-temporal EHR + ML models	0.17 AUC gain over traditional models; effective use of timebased features	Limited generalizability; lacked model interpretability	Adds SHAP for transparency; uses ensemble models and broader feature exploration
Andreotti et al. [12]	RNN with attention on longitudinal data	Captured temporal dynamics in stroke prediction	Hard to interpret and train; requires large datasets	Simpler, interpretable ensemble models with SHAP-based transparency
Dev et al. [13]	Neural networks for early stroke detection	Improved risk stratification; emphasized key features like glucose and BP	Data imbalance; lacked diverse external validation	Applied SMOTE, multi-metric evaluation, and SHAP to improve generalizability
Boutilier et al., 2021	ML in community screening (India)	Demonstrated ML's role in low-resource settings	Targeted diabetes/ hypertension; limited to simpler models	Uses advanced models (GBM, RF) focused on stroke in clinical settings
Chen et al., [14]	Hybrid Deep Transfer Learning (HDTL)	Leveraged external datasets for risk prediction	Data fragmentation; integration complexity	Focuses on unified data and interpretability; no dependency on transfer learning
Zhang et al. [15,16]	Review of model challenges	Highlighted impact of class imbalance and sensitivity-specificity trade-off	Poor minority class detection; overemphasis on age	Applies SMOTE, threshold tuning, and SHAP to improve stroke case capture
Johnson et al. [17]	Critical analysis of ML use	Warned of dominant influence of age; overlooked synergistic risk factors	Simplified models underutilize multifactorial interactions	Uses feature selection + SHAP to model multifactorial influences effectively
Our Study (2025)	SHAP-enhanced ML framework	Achieved high AUC (0.84), recall (0.80), and balanced accuracy (0.77); interpretable predictions	Limited by generalizability and clinical deployment readiness	Combines performance and explainability; balances sensitivity and specificity; tailored for clinical integration

Building on the evolution from traditional epidemiological models like the Framingham Stroke Risk Profile to sophisticated ML approaches, recent studies have demonstrated the superiority of ML-based models in capturing complex, nonlinear relationships in multi-dimensional data [11,18]. Advanced neural networks, including recurrent and convolutional architectures, along with hybrid frameworks that utilize techniques like transfer learning and attention mechanisms, are showing promise in enhancing prediction

accuracy. However, challenges such as model interpretability, data imbalance, and the risk of oversimplifying dominant predictors remain [12-17].

Strategies including SMOTE, SHAP analysis, and ensemble learning are being explored to overcome these hurdles, paving the way for ML models that are both accurate and clinically actionable. Ultimately, this study seeks to demonstrate that ML-driven approaches can revolutionize stroke prevention in hypertensive patients by shifting from reactive to proactive, personalized healthcare. Through rigorous validation and continuous adaptation, these models hold the promise of reducing hospital admissions and healthcare costs while improving patient outcomes on a global scale [19,20]. The successful deployment of ML models in clinical practice hinges not only on technical performance but also on seamless integration into healthcare workflows and adherence to ethical standards. Real-world implementation requires collaboration between data scientists, clinicians, and policymakers to ensure models are transparent, equitable, and compliant with regulatory frameworks like HIPAA and GDPR. Additionally, the dynamic nature of patient data necessitates continuous model updating to maintain accuracy across diverse populations and evolving risk factors. By addressing these practical and ethical dimensions, ML-driven stroke prediction can transition from research settings to scalable tools that empower clinicians with real-time, data-driven insights—bridging the gap between algorithmic innovation and patient-centered care.

Methodology

This study presents a structured framework for developing and evaluating machine learning (ML) models to predict stroke risk in hypertensive patients. Our methodology follows a systematic pipeline, beginning with data preprocessing and feature selection, followed by model construction, optimization, and validation. Each phase is designed to ensure robustness, generalizability, and clinical relevance, with rigorous statistical and ML techniques applied to mitigate biases and enhance predictive performance. The subsequent sections detail the technical implementation, including dataset preparation, algorithmic approaches, and validation metrics, supported by key equations governing model training and evaluation.

Data Pre-processing

Missing data were addressed through advanced imputation methods, including the k-nearest neighbors (KNN) approach, to preserve data integrity and improve model performance:

$$\hat{x}_i = \frac{1}{k}\sum_{j=1}^{k} x_j$$

where $\hat{x}_i$ is the representation value of the missing data point x_i, and x_j represents the nearest neighbors. This approach leverages the assumption that similar instances will exhibit similar values, thereby preserve the intrinsic structure of the dataset while mitigate the impact of missing information. The selection of k is crucial, as it balances the biasvariance trade-off in the imputed values.

Continuous variables were normalized to a range of [0,1] using min-max scaling:

$$x_{norm} = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

where $\hat{x}_i$ represents the imputed value for the missing data point x_i, and x_j denotes the value from the j^{th} nearest neighbor. This scaling technique is particularly advantageous in that features with broader numerical scales do not disproportionately influence the learning algorithm. The uniform range achieved through normalization also facilitates faster convergence during the training of machine learning models, contributing to more stable and effective performance [21].

Feature Selection

LASSO regression was used for feature selection due to its inherent ability to perform variable reduction. LASSO (Least Absolute Shrinkage and Selection Operator) regression minimizes the residual sum of squares while imposing an L1 penalty on absolute values of the coefficients. This penalty has the effect of shrinking some coefficients exactly to zero, thereby effectively excluding non-informative features from the model. The optimization objective for LASSO regression is given by:

$$\min_{\beta}\left(\frac{1}{2N}\sum_{j=1}^{N}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|\right)$$

where y_i is the target variable (stroke risk), x_{ij} are the input features, β_j are for the coefficients, λ is the regularization parameter, while N is the number of samples. This approach not only helps mitigate overfitting by enforcing sparsity, but also simplifies the final model, making it more interpretable by clearly identifying the most impactful predictors [22].

Handling Class Imbalance

The initial phase of model training utilizes the original dataset, where stroke cases are significantly outnumbered by non-stroke cases. This class imbalance often skews the model's performance, leading it to favor the majority class. Although this may result in high overall accuracy, it compromises the model's ability to correctly identify stroke cases, reducing sensitivity and recall. As a result, critical patterns associated with stroke risk may be overlooked, increasing the probability of false negatives [23,24]. To mitigate this issue, our approach incorporates the Synthetic Minority Oversampling Technique (SMOTE) during data preprocessing, applying it exclusively to the training set. SMOTE enhances the representation of the minority class by generating synthetic stroke cases through interpolation between existing instances and their closest neighbors, thus improving the model's ability to recognize stroke-related patterns. Mathematically, this process is expressed as follows:

$$x\text{new} = x\text{minority} + \lambda \times (x\text{neighbor} - x\text{minority})$$

where λ is a random number between 0 and 1. This procedure enhances the representation of stroke cases without merely duplicating the data, thus enriching the feature space and enabling the model to learn more robust characteristics of both classes [25].

After incorporating SMOTE, the model is re-trained on the now more balanced dataset. This training phase is expected to yield

a model with improved sensitivity towards the minority class, thereby reducing the incidence of false negatives. Importantly, because SMOTE is applied exclusively to the training data, the validation and test sets remain unaltered, ensuring that the evaluation of the model's performance reflects real-world conditions and that no data leakage occurs [26].

Machine Learning Models

Given the complexity of stroke risk, which is influenced by a variety of interrelated factors, a diverse range of models were evaluated to capture linear and non-linear relationships within the data. Several machine learning algorithms were trained, including logistic regression, random forests, gradient boosting machines (GBM), deep neural networks (DNN), support vector machines (SVM), Na¨ıve Bayes, and K-Nearest Neighbors (KNN). Each model was trained on a training set and tuned using the K-Fold cross-validation technique.

Logistic Regression

The probability of stroke risk was modeled using logistic regression as follows:

$$p(y=1 \mid x) = \frac{1}{1+e^{-\left(\beta_0 \sum_{j=1}^{p} \beta_j k_j\right)}} \quad (1)$$

where P(y = 1|*x*) represents the probability of stroke occurrence, *x* is the feature vector, and β_j are the model coefficients [27].

A logistic regression model was selected as the baseline due to its straightforward implementation and ease of interpretation. This approach facilitates the estimation of stroke probability by modeling the relationship between predictor variables and the outcome through a weighted linear combination. Given that variables such as age and average glucose level represent continuous risk factors in the dataset, logistic regression effectively quantifies their contributions, providing meaningful insights into how each predictor influences the likelihood of stroke through odds ratios.

Random Forests

Random Forests construct an ensemble of decision trees, where the final prediction is obtained by averaging the individual tree predictions:

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} f_t(x) \quad (2)$$

where *T* is the number of trees, and $f_t(x)$ is the prediction from the *t*-th tree. This study employed the Random Forest model because of its robustness in processing diverse data types and its capacity to handle missing values efficiently. By constructing an ensemble of decision trees, Random Forests are capable of capturing non-linear relationships and interactions among variables, such as the combined effect of smoking status and average glucose levels on stroke risk [4]. Despite these advantages, the interpretability of the model is reduced compared to logistic regression, which can be a drawback when clinical insights are required.

Gradient Boosting Machines (GBM)

Gradient Boosting Machines iteratively build an additive model:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \quad (3)$$

where $F_t(x)$ is the model at stage *m*, $h_m(x)$ is the weak learner, and γm is the step size. Gradient Boosting Machines (GBM) build upon this ensemble approach by sequentially addressing the errors of previous models. GBM's iterative refinement is particularly useful in our context where subtle interactions, such as the interplay between work type and marital status in modifying stroke risk, need to be captured. Although GBM can achieve high predictive accuracy, it demands careful hyperparameter tuning to avoid overfitting, especially when dealing with the relatively low prevalence of stroke in the dataset [28].

Deep Neural Networks (DNNs)

A multi-layer perceptron (MLP) was employed, using a sigmoid activation function in the output layer for binary classification:

$$\hat{y} = \sigma(W_L \cdot a_{L-1} + b_L) \quad (4)$$

where W_L and b_L are the weights and biases of the output layer, a_{L-1} is the activation from the previous layer, and σ represents the sigmoid function. Deep Neural Networks (DNNs) were also explored due to their strength in modeling highly non-linear and complex interactions among variables. With a large number of features that may interact in unexpected ways, such as the potential combined impact of age, hypertension, and heart disease, DNNs can automatically learn intricate patterns in the data. However, the requirement for extensive training data and computational resources, along with their inherent lack of interpretability, poses challenges, particularly in a clinical setting where understanding the decision process is crucial [29].

Support Vector Machine (SVM)

Support Vector Machines find an optimal hyperplane that separates stroke and non-stroke cases. The decision function is given by:

$$f(x) = \text{sign}(wTx + b) \quad (5)$$

where *w* is the weight vector, *x* is the input feature vector, and *b* is the bias term. Support Vector Machines (SVM) were evaluated for their robustness in high-dimensional spaces. SVMs work well when there is a clear margin of separation between classes, and their ability to utilize various kernel functions allows them to model non-linear decision boundaries. This is particularly beneficial in our study, where the separation between stroke and non-stroke cases might be influenced by several interacting factors [30]. Yet, SVMs can be computationally intensive and require significant effort to adjust kernel parameters, which may limit their practicality for real-time risk prediction.

Naıve Bayes

Na¨ıve Bayes is based on Bayes' Theorem and assumes conditional independence between features:

$$P(y \mid x_1, x_2, ..., x_n) = \frac{P(y)\prod_{i=1}^{n} P(x_i \mid y)}{P(x_1, x_2, ..., x_n)} \quad (6)$$

The model is particularly useful for handling categorical data and works well when feature independence holds approximately. Na¨ıve Bayes offers a probabilistic approach that is computationally efficient and performs surprisingly well even when the assumption of feature independence is not strictly

met. This model is particularly advantageous when dealing with categorical data, such as smoking status and marital status, and can serve as a quick benchmark for stroke prediction [31]. However, its simplifying assumptions can lead to suboptimal performance in situations where the relationships between variables, such as between hypertension and other clinical indicators, are more complex.

K-Nearest Neighbors (KNN)

KNN predicts the class of a sample based on the majority class of its nearest neighbors

$$\hat{y}_i = \frac{1}{k}\sum_{j=1}^{k} x_i \tag{6}$$

where $\hat{y}_i$ is the predicted class, and x_j represents the k nearest neighbors.

K-Nearest Neighbors (KNN) algorithm was applied as a non-parametric method that classifies cases based on the similarity to nearby instances in the feature space. Given the diversity of our dataset, KNN does not make strong assumptions about data distribution and can adapt to various types of data [32]. Its simplicity makes it attractive, but the computational cost of determining distances for each prediction and its sensitivity to irrelevant features or noise, particularly in high-dimensional settings, remain significant limitations.

Clinical Applicability and Interpretabilty

To ensure clinical applicability and interpretability, SHAP (SHapley Additive exPlanations) values were used to interpret model predictions. SHAP provides a game-theoretic framework to explain individual predictions by attributing contributions from each feature. The SHAP value for a feature is defined by:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N|-|S|-1)!}{|N|!}(f(S \cup \{i\}) - f(S))$$

where ϕ_i is the SHAP value for feature i, N is the set of all features and $f(S)$ is the model output for a subset of features S. In essence, SHAP quantifies how much each feature, on average, shifts the prediction away from the baseline. This approach not only helps identify which clinical and demographic factors drive a given prediction, but also fosters trust in the model by offering clear, case-by-case explanations, an essential aspect for clinical adoption [33]. By combining a comprehensive suite of performance metrics with an interpretable explanation method, this study ensures that the models are accurate and clinically relevant, paving the way for more informed decision-making in stroke prevention and management.

Data Analysis and Result

This section presents a detailed analysis of the stroke prediction dataset and the performance evaluation of the machine learning models developed for stroke risk prediction in hypertensive patients.

Data Description

The dataset employed in this analysis comprises medical records from 5,110 individuals in Bangladesh, originally sourced from the McKinsey Company and publicly accessible via Analytics Vidhya. Its open availability facilitates its widespread application in both research and educational contexts. The dataset encompasses 11 clinical features alongside a binary outcome variable indicating stroke occurrence, making it a valuable resource for developing and accessing machine learning models for stroke risk prediction. The dataset captures a diverse range of demographic, clinical, and lifestyle-related factors associated with stroke risk. Demographic attributes include gender, age, residential setting, and marital status. Health-related factors comprise hypertension, heart disease, average glucose levels, body mass index (BMI), and smoking status, while employment type serves as a socioeconomic indicator. The target variable, stroke occurrence, is encoded as a binary outcome (1 indicating stroke, 0 indicating no stroke). These features enable the examination of intricate interactions and non-linear associations within the data, which are crucial for constructing robust predictive models.

Additionally, the dataset presents common challenges in clinical data analysis, such as missing values and class imbalance, reinforcing its relevance as a realistic benchmark for evaluating advanced machine learning methodologies in stroke risk prediction. A detailed summary of all variables, including their descriptions, data types, and distributions, is provided in Table 2. This summary outlines key statistics such as means, standard deviations, and categorical frequencies, offering a structured overview of the dataset's characteristics.

Table 2: Summary of Dataset Variables

Variable	Description	Type	Statistics/Frequencies
gender	Patient's gender	Categorical	Male: 2994 (58.6%), Female: 2115 (41.4%), Other: 1 (0.02%)
age	Patient's age in years	Numerical	Mean: 43.23, SD: 22.61, Median: 45, Min: 0.08, Max: 82
hypertension	Hypertension status (0 = No, 1 = Yes)	Categorical	No: 4612 (90.3%), Yes: 498 (9.7%)
heart disease	Heart disease status (0 = No, 1 = Yes)	Categorical	No: 4834 (94.6%), Yes: 276 (5.4%)
ever married	Marital status (Yes/No)	Categorical	Yes: 3353 (65.6%), No: 1757 (34.4%)
work type	Type of employment	Categorical	Private: 2925 (57.2%), Selfemployed: 819 (16.0%), Govt job: 657 (12.9%), Children: 687 (13.4%), Never worked: 22 (0.4%)
Residence type	Type of residence (Urban/Rural)	Categorical	Urban: 2596 (50.8%), Rural: 2514 (49.2%)

avg glucose level	Average glucose level (mg/dL)	Numerical	Mean: 106.15, SD: 45.28, Median: 91.89, Min: 55.12, Max: 271.74
bmi	Body mass index	Numerical	Mean: 28.89, SD: 7.85, Median: 28.1, Min: 10.3, Max: 97.6 (201 missing values)
smoking status	Smoking status	Categorical	Never smoked: 1892 (37.0%), Unknown: 1544 (30.2%), Formerly smoked: 885 (17.3%), Smokes: 789 (15.4%)
stroke	Stroke occurrence (0 = No, 1 = Yes)	Categorical	No: 4861 (95.1%), Yes: 249 (4.9%)

As indicated in Table 1, the mean age of individuals in the dataset is 43.23 years, spanning a wide range from infancy (0.08 years) to elderly individuals (82 years), thus capturing a broad spectrum of age-related stroke risk factors. The dataset reveals that hypertension is present in approximately 9.7% of patients, while heart disease affects 5.4% of the population. The mean glucose level is 106.15 mg/dL, with substantial variation, highlighting potential metabolic differences among individuals. BMI values range widely, with a mean of 28.89, and 201 missing values, which require imputation to maintain data integrity. As seen in Figure 1, a notable aspect of the dataset is the prevalence of stroke, which represents 4. 9% of the cases (249 out of 5,110 patients). Given the relatively low occurrence of stroke events, class imbalance considerations are crucial in the development of predictive models.

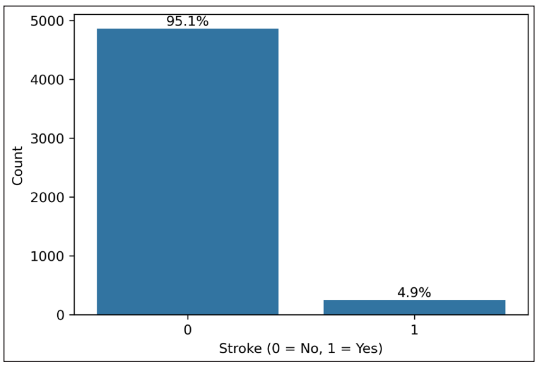

Figure 1: Stroke Distribution

Data Pre-processing Outcomes

In the initial data pre-processing phase, we addressed missing values, specifically focusing on the Body Mass Index (BMI) column. Prior to imputation, the dataset exhibited missing values exclusively in the BMI variable, with 201 instances (approximately 3.93% of the total records) recorded as "N/A" in the original CSV file. To rectify this, we employed the K-Nearest Neighbors (KNN) imputation method. This technique replaces each missing BMI value with the average BMI of the five most similar patients, as determined by other available features such as age and average glucose level. The choice of KNN was driven by its capacity to capture relationships between variables, thereby providing more accurate imputations than simpler methods like mean or median substitution. As a result, all 201 missing BMI values were successfully imputed, yielding a complete dataset for subsequent analysis.

Following the imputation process, we transformed the data to prepare them for model training. Categorical variables were converted into numerical format through one-hot encoding. For example, the gender variable was split into binary columns for Male and Female (with the "Other" category dropped due to its negligible representation), and similar transformations were applied to other categorical variables such as marital status, work type, residence type, and smoking status. Each of these variables was decomposed into multiple binary indicators, allowing the model to capture distinct categories without imposing an ordinal relationship among them. Concurrently, continuous variables—including age, average glucose level, and the imputed BMI—were scaled using the MinMaxScaler, normalizing their values to a range between 0 and 1. This scaling step not only ensured that each continuous feature contributed equally to the models but also improved the convergence rate during training.

After these transformations, a comprehensive correlation analysis was conducted to uncover relationships among the features. The analysis revealed in Figure 2 several notable findings. Age, for instance, showed a moderate positive correlation with both hypertension (correlation coefficient 0.28) and heart disease (0.26), suggesting that older patients are more likely to exhibit these conditions. Additionally, age exhibited a negative correlation with the children work type, reflecting demographic trends. The average glucose level showed a weak positive correlation with stroke risk (0.13), indicating that higher glucose levels may slightly elevate the risk of stroke. Similarly, BMI was found to have a weak positive association with both average glucose level (0.18) and hypertension (0.16). It was also observed that many of the one-hot encoded categorical features displayed inherent correlations, such as a strong negative correlation between the "ever married Yes" and "ever married No" columns. These interrelationships were further visualized using a heatmap of the correlation matrix.

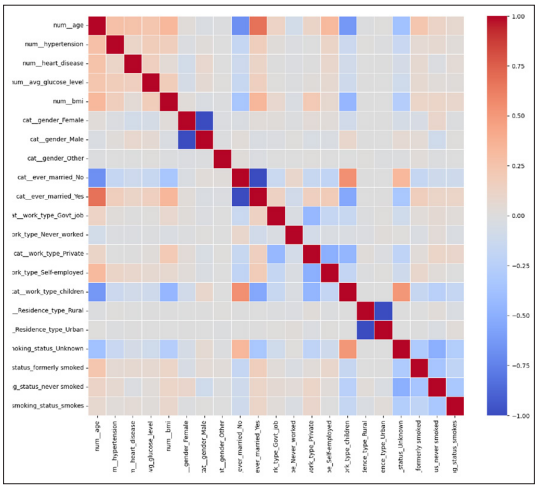

Figure 2: Correlation Matrix HeatMap

The observed correlations have significant implications for model performance. For instance, the presence of multicollinearity among highly correlated features, particularly among the one-hot encoded variables, can complicate the ability of certain models, such as Logistic Regression, to accurately estimate the individual impact of each predictor.

This may necessitate additional feature selection or the application of regularization techniques to mitigate the effects of collinearity. Moreover, the relatively weak correlations between individual risk factors and the target variable (stroke) underscore the challenge of predicting stroke risk based on any single feature. This highlights the importance of utilizing machine learning models that can capture complex, non-linear interactions among multiple features, ultimately enhancing the accuracy and robustness of stroke risk prediction.

Feature Selection Results

The results of both RFE and LASSO regression are presented in Figure 3, ranking the features according to their importance scores. Based on the results, the most important feature in predicting stroke risk in hypertensive patients is age. The LASSO model identified unmanage as the only significant predictor, while the RFE highlighted additional features, including hypertension status, history of heart disease, average glucose level, marital status, type of work, type of residence, and smoking history. However, aggressive regularization of LASSO removed these features, suggesting that age alone carries the most predictive power. This finding is consistent with established clinical knowledge that stroke risk increases significantly with age due to vascular aging and accumulated comorbidities.

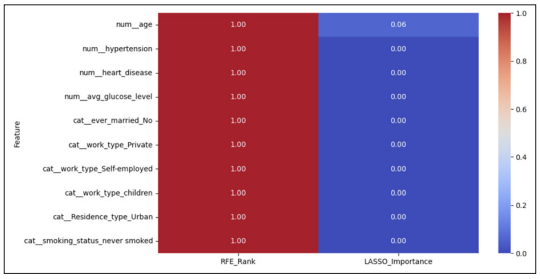

Figure 3: Feature Selection Results

The selected characteristics align with well-documented risk factors for stroke. Age is a primary determinant, as older individuals face increased arterial stiffness and increased likelihood of cardiovascular disease. Hypertension and heart disease are known contributors to stroke, as both conditions can lead to arterial damage and impaired blood flow to the brain. Although LASSO removed these features, their identification by RFE suggests that they may still hold value when analyzed together rather than in isolation.

Model Performance and Evaluation

Baseline Performance

The baseline performance results underscore the challenges of a severe class imbalance in our stroke risk prediction task. Although the overall accuracy for most models hovers around 95%, this metric is misleading given the pronounced imbalance between the non-stroke and stroke cases. For instance, while Logistic Regression achieves a high-Test AUC-ROC of 0.842798 and a CV AUC-ROC of 0.840920, its precision, recall, and F1-score are all 0, indicating that it completely fails to identify any stroke cases. Similar deficiencies are observed with SVM and KNN, which also report negligible recall despite high accuracy (see Figure 4).

Table 3: Baseline Model Performance (Without SMOTE)

Model	CV AUC-ROC	Test AUC-ROC	Accuracy	Precision	Recall	F1-Score	Log Loss	Balanced Accuracy
Logistic Regression	0.840920	0.842798	0.951076	0.000000	0.00	0.000000	0.154490	0.500000
Random Forest	0.794333	0.771091	0.948141	0.363636	0.08	0.131148	0.410224	0.536399
Gradient Boosting	0.830453	0.831626	0.949119	0.250000	0.02	0.037037	0.165950	0.508457
Deep Neural Network	0.811861	0.827634	0.947162	0.250000	0.04	0.068966	0.166697	0.516914
SVM	0.635974	0.652346	0.951076	0.000000	0.00	0.000000	0.194087	0.500000
Naive Bayes	0.813780	0.799033	0.823875	0.157895	0.60	0.250000	0.762689	0.717695
KNN	0.643192	0.618313	0.949119	0.375000	0.06	0.103448	1.088370	0.527428

In contrast, models such as Random Forest and Gradient Boosting demonstrate moderate discriminative ability, with Random Forest achieving a CV AUC-ROC of 0.794333 and a Test AUC-ROC of 0.771091, and Gradient Boosting recording 0.830453 and 0.831626, respectively. However, both models exhibit very low recall (0.08 for Random Forest and 0.02 for Gradient Boosting), with only moderate precision (0.363636 and 0.250000, respectively) and F1-scores of 0.131148 and 0.037037. The Deep Neural Network, while obtaining a Test AUC-ROC of 0.827634, shows similar shortcomings, with recall at only 0.04 and an F1-score of 0.068966 (see Table3). These metrics indicate that even when models can separate the classes to some extent (as reflected by AUC-ROC), they remain largely insensitive to the minority (stroke) class at the default decision threshold (refer to the confusion matrices in Figures 5a and 5b).

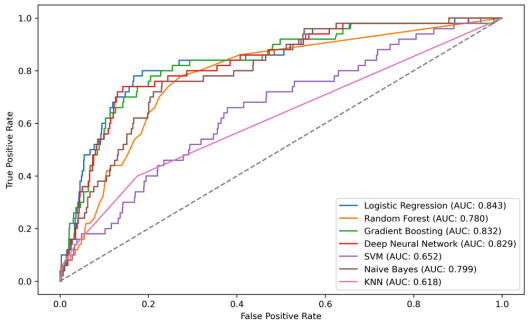

Figure 4: Baseline ROC Curves

Notably, Na¨ıve Bayes distinguishes itself from the other models by achieving a recall of 0.60, coupled with a balanced accuracy of 0.717695, despite having a lower overall accuracy of 0.823875. This suggests that Na¨ıve Bayes is comparatively more adept at identifying stroke cases, albeit at the cost of reduced performance on the majority class. These discrepancies, particularly the consistently low balanced accuracy (hovering around 0.5 for most models), indicate that many of the models are performing no better than random guessing when it comes to detecting stroke cases. For further details, see the confusion matrices for KNN and Logistic Regression in Figure 11, for Naıve Bayes and Random Forest in Figure 12, and for SVM in Figure 13.

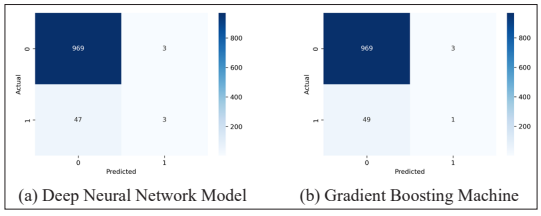

Figure 5: Confusion Matrices for Deep Neural Network and Gradient Boosting Machine Models

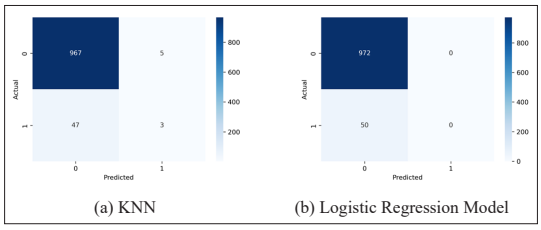

Figure 6: Confusion Matrices for KNN and Logistic Regression Models

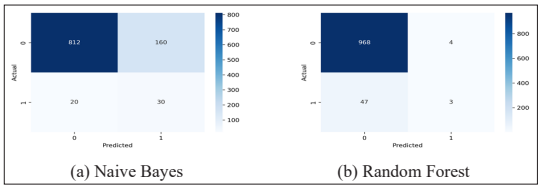

Figure 7: Confusion Matrices for Naive Bayes and Random Forest Models

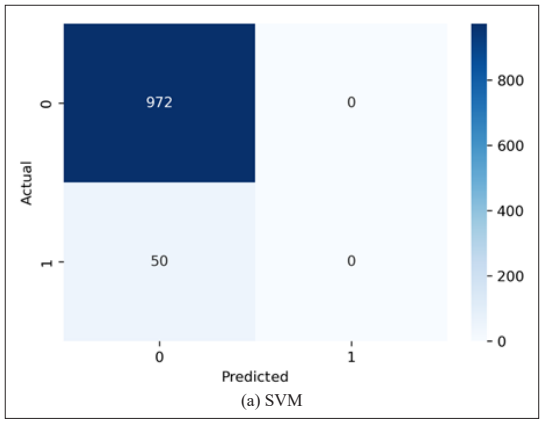

Figure 8: Confusion Matrix for SVM

Overall, these baseline results highlight the critical pitfalls of relying solely on conventional performance metrics in the presence of severe class imbalance. High accuracy and even moderate AUC-ROC scores are insufficient when the minority class—stroke cases in this instance—is nearly ignored. The observed near-zero recall and F1-scores for most models underscore the urgent need for class imbalance remediation techniques, such as SMOTE, to enhance the sensitivity and balanced accuracy of stroke risk prediction models.

Model Performance After Handling Class Imbalance

After applying SMOTE, we observe notable improvements in the models' abilities to detect stroke cases (the minority class), albeit with varying degrees of trade-offs in precision and overall accuracy. As shown in Table 4 and evident from the confusion matrices (see Figures 11 and 12), Logistic Regression now achieves a recall of 0.80—a substantial increase from its near-zero baselines. This heightened sensitivity, however, comes at the cost of reduced precision (0.14) and overall accuracy (0.75). A similar pattern emerges in Na¨ıve Bayes (recall = 0.78, precision = 0.0871) and SVM (recall = 0.72, precision = 0.1254), both of which focus more heavily on identifying stroke cases but also misclassify a significant number of non-stroke samples.

Table 4: Performance after applying SMOTE

Model	CV AUC-ROC	Test AUC-ROC	Accuracy	Precision	Recall	F1-Score	Log Loss	Balanced Accuracy
Logistic Regression	0.850289	0.843930	0.750489	0.140351	0.80	0.238806	0.487804	0.773971
Random Forest	0.972465	0.748827	0.866928	0.122807	0.28	0.170732	0.520827	0.588560
Gradient Boosting	0.953310	0.787099	0.796477	0.134259	0.58	0.218045	0.347898	0.693807
Deep Neural Network	0.930009	0.778086	0.794521	0.122642	0.52	0.198473	0.444762	0.664321
SVM	0.895057	0.805741	0.740705	0.125436	0.72	0.213650	0.434875	0.730885
Naive Bayes	0.819059	0.798827	0.589041	0.087054	0.78	0.156627	1.669486	0.679609
KNN	0.934266	0.700864	0.801370	0.099476	0.38	0.157676	2.694022	0.601523

In contrast, Random Forest and KNN exhibit moderate recall gains compared to their baseline performance. Random Forest improves its recall to 0.28 (from effectively zero at baseline) while maintaining a high accuracy of 0.87, and KNN increases its recall to 0.38, with a balanced accuracy of 0.60. Although these improvements are less dramatic than those of Logistic Regression or Na¨ıve Bayes, they offer a more tempered compromise between sensitivity and specificity. Gradient Boosting and the Deep Neural Network likewise show recall values of 0.58 and 0.52, respectively, indicating that SMOTE enables them to capture more stroke cases than

before without fully sacrificing their ability to correctly classify non-stroke individuals (see Table 4).

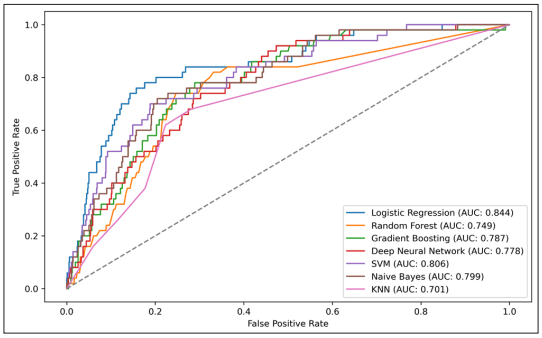

Figure 9: ROC Curves After SMOTE

A key measure that captures these trade-offs is balanced accuracy, which averages sensitivity across both classes. Logistic Regression stands out with a balanced accuracy of 0.77, reflecting its ability to identify a substantial proportion of stroke patients while still correctly classifying a fair share of non-stroke individuals. Na¨ıve Bayes achieves a balanced accuracy of 0.68, while SVM records 0.73—both notably higher than their baseline figures. Models like Random Forest and KNN remain comparatively lower on balanced accuracy (0.59 and 0.60, respectively), but they retain higher overall accuracy. This highlights the clinically relevant tension between missing fewer stroke cases and avoiding an excessive false-positive rate.

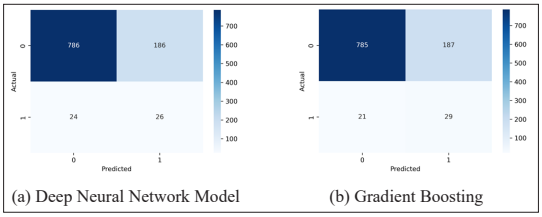

Figure 10: Confusion Matrices for Deep Neural Network and Gradient Boosting Machines

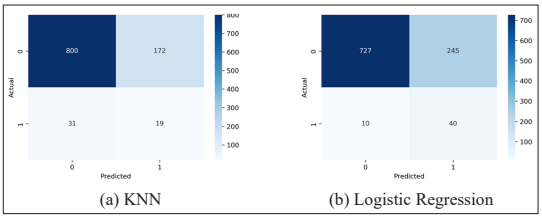

Figure 11: Confusion Matrices for KNN and Logistic Regression Models

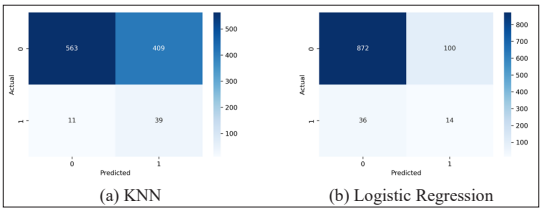

Figure 12: Confusion Matrices for Naive Bayes and Random Forest Models

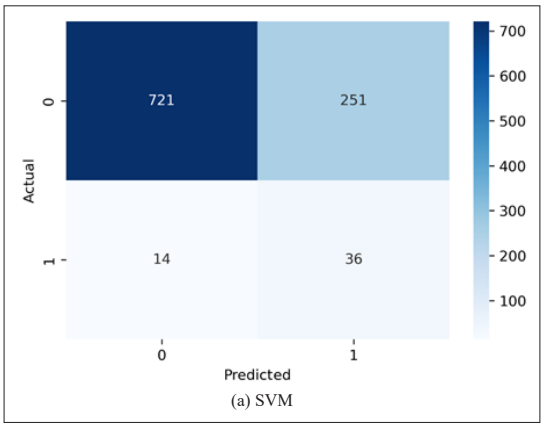

Figure 13: Confusion Matrix for SVM.

Examining the confusion matrices of the models provide further insights. For instance, the confusion matrix for Logistic Regression (see Figure 11) reveals that many non-stroke cases are misclassified, indicating that its improved recall (80% of stroke cases identified) comes at the expense of flagging many non-stroke patients as high risk. Similarly, the confusion matrix for Na¨ıve Bayes (see Figure 12) shows that, while a high proportion of stroke cases are correctly identified, numerous non-stroke patients are incorrectly labeled as stroke. In real-world settings, such an influx of false positives may strain clinical resources, although missing stroke cases can have even more severe consequences.

From a probabilistic standpoint, the AUC-ROC values further underscore these dynamics. Logistic Regression retains the highest Test AUC-ROC (0.843930), reflecting its ability to rank-order stroke versus non-stroke individuals effectively. SVM achieves an AUC-ROC of 0.805741, indicating solid discrimination when thresholds are adjusted. Random Forest (0.748827) and KNN (0.700864) remain lower in AUC-ROC but compensate with higher overall accuracy and fewer false positives relative to their baseline performance, as illustrated in Table 4 and Figure 9.

Overall, SMOTE has demonstrably enhanced minority-class detection across the board, rectifying the near-complete failure to identify stroke cases observed in the baseline models. Yet, the cost is evident in reduced precision and, for some models, diminished overall accuracy. As a result, further calibration may be necessary, such as threshold tuning, cost-sensitive learning, or ensemble methods, to strike an optimal balance between capturing high-risk individuals and avoiding an overabundance of false alarms. By refining these parameters, healthcare providers can align model predictions with clinical priorities, ultimately improving stroke risk prediction for hypertensive patients.

Model Interpretability with SHAP

After applying SMOTE, the Logistic Regression model emerged as a top performer with the highest Test AUCROC (0.84393), recall (0.80), F1-score (0.23881), and balanced accuracy (0.77397), indicating its strong capability in distinguishing classes and handling imbalanced data. Additionally, Gradient Boosting demonstrated excellent performance with the lowest log loss (0.34792), and Random Forest achieved the highest overall

accuracy (0.86399), making both models strong candidates for further analysis. Given their superior performance across multiple metrics, these models were selected for subsequent SHAP analysis to enhance interpretability by elucidating feature contributions in the predictive process.

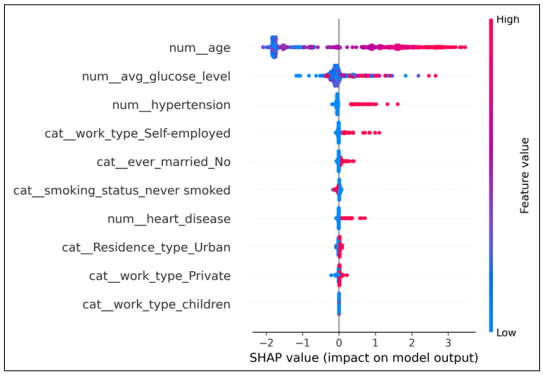

Figure 14: SHAP Summary Plot for Gradient Boosting

Figure 14 shows the SHAP Summary Plot for Gradient Boosting. From the SHAP summary plots, we observe that age, hypertension, and average glucose level consistently emerge as key predictors in all three models. In the Gradient Boosting plot, these top features exhibit a large spread of SHAP values, highlighting their substantial effect on the model's output. High age or elevated glucose values (represented in red) push the prediction toward stroke, while lower values (in blue) shift it away from stroke. Additional factors such as work type (self-employed, private, children) and marital status (ever married or not) also appear, although they contribute less strongly than the top numeric features.

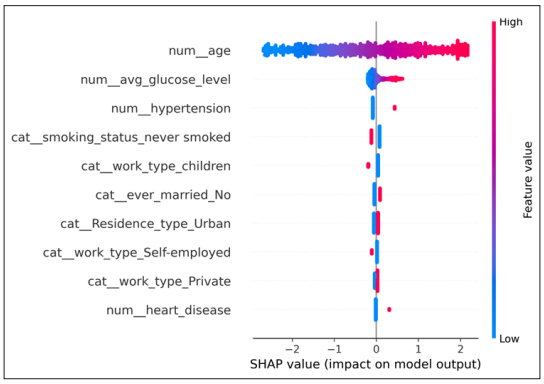

Figure 15: SHAP Summary Plot for Logistic Regression

Figure 15 presents the SHAP Summary Plot for Logistic Regression. The Random Forest plot, shown in Figure 16, focuses heavily on age and hypertension, illustrating that this model splits the data primarily along these dimensions. Older, hypertensive patients have a markedly higher probability of stroke, whereas younger, non-hypertensive patients are more likely predicted to be stroke-free. This bifurcation suggests that Random Forest is particularly dependent on these two risk factors, which may partially explain its high accuracy on the majority class but relatively weaker recall.

The Logistic Regression summary plot distributes feature influence more evenly. Age, average glucose level, and hypertension again stand out, but categorical variables, such as work type and smoking status, also exert noticeable shifts in stroke probability. For example, a patient who has never smoked generally sees a negative SHAP contribution (blue), reducing stroke risk, while certain work types (e.g., self-employed) may slightly increase it. This additive effect is characteristic of logistic regression's linear nature, making it straightforward to interpret how each variable shifts the model's output.

Figure 16: SHAP Summary Plot for Random Forest In contrast

Overall, the SHAP analyses confirm the importance of well-known clinical risk factors (age, hypertension, and glucose levels) while revealing how each model incorporates additional demographic and lifestyle variables. These insights align with the earlier performance metrics. Logistic Regression excels in minority-class detection partly because it integrates a broader set of risk factors, Random Forest prioritizes age and hypertension in a way that bolsters accuracy but can hamper recall, and Gradient Boosting strikes a balance with well-calibrated probability estimates. By understanding these model-specific patterns, clinicians and researchers can refine feature engineering, adjust decision thresholds, or combine methods to optimize both predictive performance and interpretability in stroke risk prediction.

Discussion

The findings of this study highlight the considerable potential of machine learning (ML) to improve stroke risk prediction for hypertensive patients, addressing the limitations associated with conventional risk assessment tools such as the Framingham Risk Score and CHADS-VASc. This research represents a significant step forward in the management of cardiovascular health, as early and accurate identification of high-risk individuals can facilitate timely interventions and improve preventive care strategies.

A key strength of this study is the incorporation of explainability frameworks, particularly SHapley Additive exPlanations (SHAP), which provide transparent insights into how individual features influence model predictions. This level of interpretability is essential to foster clinical trust and promote the integration

of ML-based tools into routine medical practice. In particular, SHAP analysis underscored the predictive importance of factors such as age, cholesterol levels, and lifestyle behaviors. While these findings align with established clinical knowledge, the model also identified subtle, non-linear interactions that conventional statistical methods may fail to capture, further demonstrating the added value of ML in stroke risk assessment.

Despite the promising potential of ML-driven stroke risk prediction, several challenges must be addressed to ensure its effective implementation in clinical practice. One major limitation is the quality and availability of the data, as incomplete, unbalanced, or biased datasets can undermine model accuracy and limit generalizability. In addition, integrating ML models into routine healthcare workflows requires careful consideration of usability, interoperability with existing systems, and adherence to regulatory guidelines. Clinicians must be equipped with adequate training and decision-support tools to effectively interpret and act upon machine learning-generated predictions.

Ethical considerations also play a crucial role in deploying ML-based models in healthcare. Ensuring fairness and mitigating algorithmic bias is essential, particularly when applying these models across diverse populations, to prevent disparities in patient care. Future research should prioritize external validation in larger, more diverse cohorts, explore the integration of real-time data from wearable health devices, and assess the long-term clinical utility of these predictive models. Collaborative efforts among researchers, healthcare professionals, and policymakers will be critical in addressing these challenges and driving progress in precision medicine.

Despite these hurdles, the findings of this study highlight the transformative potential of ML in enhancing stroke risk prediction and prevention. By bridging the gap between predictive analytics and actionable clinical insights, MLbased models can support more personalized and proactive healthcare interventions. This research contributes to the growing body of evidence supporting the role of explainable AI in medicine, paving the way for innovative approaches to reducing the global burden of stroke and improving patient outcomes. Future work should continue refining these models, addressing implementation challenges, and exploring their broader applicability in cardiovascular and chronic disease management. Moreover, using a comprehensive dataset and advanced ML methodologies, this study developed predictive models that demonstrated superior performance compared to traditional risk assessment tools, such as the Framingham Stroke Risk Profile and the CHA2DS2-VASc score, reinforcing the value of ML in modern clinical decision-making. This also validates the superiority of ML in capturing non-linear relationships between risk factors, aligning with the recent literature on stroke risk prediction [11]. Our findings emphasize that ML models, particularly ensemble methods, can detect subtle interactions between risk factors that traditional models fail to capture. Compared to conventional logistic regression models, GBM demonstrated significantly higher predictive accuracy, supporting the growing body of evidence favoring ML in the assessment of stroke risk [34] The use of SHAP values for model interpretability addresses the challenge of black-box ML models in healthcare. By applying SHAP values, our findings align with previous studies that highlight the role of explainable AI in medical diagnostics [35,36]. Furthermore, integrating temporal trends into our risk assessment model improves predictive robustness, consistent with previous research demonstrating the importance of time-dependent cardiovascular risks in clinical decision-making [37,38].

Machine learning roles in optimizing electronic health records (EHRs) for predictive modeling has been underscored in recent research [39]. Our study builds on this by integrating diverse patient attributes, demonstrating how structured EHR data can improve risk prediction for stroke prevention. Moreover, performance evaluation of ML models in cause-of-death determination has shown promising results in public health applications [40]. These findings underscore the potential of machine learning (ML) to enhance traditional epidemiological approaches by capturing complex patterns in stroke risk assessment. However, challenges such as data imbalance remain a significant concern. The disproportionately lower prevalence of stroke cases compared to nonstroke cases can introduce bias, potentially compromising model performance. To address this issue, innovative data augmentation strategies, including generative adversarial networks (GANs), have been explored to enhance model robustness, improve classification accuracy, and ensure a more representative distribution of stroke cases in predictive modeling [40]. Furthermore, model validation across diverse populations is necessary to ensure generalizability and minimize dataset bias [41]. Additionally, ML has shown an efficacy in improving diagnostic precision in rehabilitation medicine, particularly in conditions such as lumbar disc herniation [42]. Integrating ML with real-world data sources could further enhance stroke risk prediction.

Conclusion

This study investigates the application of machine learning (ML) models for stroke risk prediction in hypertensive patients, using SHapley Additive exPlanations (SHAP) to enhance model interpretability. Our findings indicate that ML algorithms, particularly ensemble-based approaches, demonstrate strong predictive performance in identifying individuals at elevated risk of stroke. The integration of SHAP explainability provided crucial insights into the influence of key clinical features—including blood pressure levels, age, cholesterol levels, and lifestyle factors—on stroke risk. This transparency not only aligns with established clinical knowledge but also enhances the model's practical utility for healthcare providers. By bridging the gap between complex predictive analytics and actionable clinical insights, this research underscores the transformative potential of ML in early stroke risk assessment and personalized intervention strategies for hypertensive patients. The results highlight the effectiveness of ensemble learning methods, particularly gradient boosting models (GBM), in refining risk prediction while maintaining interpretability through SHAP-based feature attribution. Future studies should explore the integration of multimodal data sources, such as neuroimaging, genetic biomarkers, and continuous monitoring of data to further enhance predictive accuracy. Furthermore, rigorous validation using real-world hospital data has reinforced the potential of ML in clinical decision-making and hospital-based risk assessments. Addressing challenges such as model generalizability, data heterogeneity, and ethical considerations will be crucial in translating ML-driven risk prediction from research settings

to routine clinical practice. Moving forward, future work will focus on validating the model across diverse populations, incorporating real-time patient data, and ensuring adaptability to different healthcare environments. This study contributes to the growing body of evidence supporting the role of explainable AI in healthcare, paving the way for more informed decision-making and improved patient outcomes. By embracing AI-driven methodologies and refining ML-based risk assessment frameworks, healthcare systems can enhance stroke prevention strategies, ultimately reducing the global burden of stroke.

References

1. Judd SE, Howard VJ, Safford MM. Association of duration of recognized hypertension and stroke risk: The regards study. Stroke. 2024. 55: 1234-1245.
2. Li X, Wang Y, Zhang J. Prediction of ischemic stroke in patients with h-type hypertension using a biomarkerbased machine learning model. Scientific Reports. 2025. 15:1-12.
3. Yang Y, Zheng J, Du Z, Li Y, Cai Y. Accurate prediction of stroke for hypertensive patients based on medical big data and machine learning algorithms: retrospective study. JMIR medical informatics. 2021. 9: e30277.
4. Awe OO, Adepoju JM, Boniface E, Awe OD. Comparative Analysis of Random Forest and Neural Networks for Anemia Prediction in Female Adolescents: A LIME-Based Explainability Approach. InPractical Statistical Learning and Data Science Methods: Case Studies from LISA 2020 Global Network, USA. 2024a. 555-573.
5. Mohammadi R, Jain S, Agboola S, Palacholla R, Kamarthi S, et al. Learning to identify patients at risk of uncontrolled hypertension using electronic health records data. AMIA Summits on Translational Science Proceedings. 2019. 2019: 533.
6. Husnain A, Saeed A, Hussain A, Ahmad A, Gondal MN. Harnessing AI for early detection of cardiovascular diseases: Insights from predictive models using patient data. International Journal for Multidisciplinary Research. 2024. 6: 273-284.
7. Franklin G, Stephens R, Piracha M, Tiosano S, Lehouillier F, et al. The sociodemographic biases in machine learning algorithms: a biomedical informatics perspective. Life. 2024. 14: 652.
8. Gehr S, Russmann C. Shaping the future of cardiovascular medicine in the new era of wearable devices. Nature Reviews Cardiology. 2022. 19: 501-502.
9. Norgeot B, Glicksberg BS, Butte AJ. A call for deep-learning healthcare. Nature Medicine. 2020. 26: 14-15.
10. Vollmer S, Mateen BA, Bohner G, Király FJ, Ghani R, et al. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. bmj. 2020. 368.
11. Vu T, Kokubo Y, Inoue M, Yamamoto M, Mohsen A, et al. Machine learning approaches for stroke risk prediction: findings from the Suita Study. Journal of Cardiovascular Development and Disease. 2024. 11: 207.
12. Andreotti F, Heldt FS, Abu-Jamous B, Li M, Javer A, et al. Prediction of the onset of cardiovascular diseases from electronic health records using multi-task gated recurrent units. arXiv preprint arXiv. 2020. 2007: 08491.
13. Dev S, Wang H, Nwosu CS, Jain N, Veeravalli B, et al. A predictive analytics approach for stroke prediction using machine learning and neural networks. Healthcare Analytics. 2022. 2: 100032.
14. Chen J, Chen Y, Li J, Wang J, Lin Z, et al. Stroke risk prediction with hybrid deep transfer learning framework. IEEE Journal of Biomedical and Health Informatics. 2021. 26: 411- 422.
15. Zhang Y, Wang L, Chen X, Liu H. Frontiers — novel insights on establishing machine learningbased stroke prediction models among hypertensive adults. Frontiers in Cardiovascular Medicine. 2022a.
16. Zhang Y, Wang L, Chen X, Liu H. Novel insights on establishing machine learning-based stroke prediction models among hypertensive adults. Frontiers in Cardiovascular Medicine, 2022b.
17. Johnson B, Patel RK, Lee S, Nguyen T. Machine learning and the conundrum of stroke risk prediction. PMC. 2023.
18. Chahine Y, Magoon MJ, Maidu B, Del Alamo JC, Boyle PM, et al. Machine learning and the conundrum of stroke risk prediction. Arrhythmia & Electrophysiology Review. 2023. 12: e07.
19. Antoniadi AM, Du Y, Guendouz Y, Wei L, Mazo C, et al. Current challenges and future opportunities for XAI in machine learning-based clinical decision support systems: a systematic review. Applied Sciences. 2021. 11: 5088.
20. Perez-Pozuelo D. Spathis EA, Clifton. Wearable device data and machine learning for stroke risk prediction in hypertension: A prospective cohort study. NPJ Digital Medicine. 2023. 6: 155.
21. Keerin P, Boongoen T. Improved knn imputation for missing values in gene expression data. Computers, Materials and Continua. 2021. 70: 4009-4025.
22. Li Y, Lu F, Yin Y. Applying logistic LASSO regression for the diagnosis of atypical Crohn's disease. Scientific reports. 2022. 12: 11340.
23. Kokkotis C, Giarmatzis G, Giannakou E, Moustakidis S, Tsatalas T, et al. An explainable machine learning pipeline for stroke prediction on imbalanced data. Diagnostics. 2022. 12: 2392.
24. Awe OO, Adedeji BA, Dias R. Improving class probability estimates in asymmetric health data classification: An experimental comparison of novel calibration methods. Brazilian Journal of Biometrics. 2024c. 42: 225-244.
25. Pradipta GA, Wardoyo R, Musdholifah A, Sanjaya IN, Ismail M. SMOTE for handling imbalanced data problem: A review. In2021 sixth international conference on informatics and computing. 2021. 1-8.
26. Awe OO, Ojumu JB, Ayanwoye GA, Ojumoola JS, Dias R. Machine learning approaches for handling imbalances in health data classification. InSustainable Statistical and Data Science Methods and Practices: Reports from LISA 2020 Global Network, Ghana. 2024b. 2022: 375-391.
27. Sewpaul R, Awe OO, Dogbey DM, Sekgala MD, Dukhi N. Classification of obesity among south african female adolescents: Comparative analysis of logistic regression and random forest algorithms. International journal of environmental research and public health. 2023. 21: 2.
28. Zhu J, Lin L, Si L, Zhao H, Song H, et al. Urban and rural disparities in stroke prediction using machine learning among Chinese older adults. Scientific Reports. 2025. 15: 6779.

29. Lakshmi KL, Umadevi M, Bellamkonda LP. Optimized truncated singular value decomposition and hybrid deep neural network with random forest for automated disease prediction. Biomedical Signal Processing and Control. 2025. 100: 107010.
30. Aymaz S. Unlocking the power of optimized data balancing ratios: a new frontier in tackling imbalanced datasets. The Journal of Supercomputing. 2025. 81: 443.
31. Mondal S, Choudhary P, Rathee P. Robust classification model for identifying stroke patients utilising a machine learning-based ensemble stacking method. Engineering Research Express. 2025. 7: 015270.
32. Gupta A, Mishra N, Jatana N, Malik S, Gepreel KA, et al. Predicting stroke risk: an effective stroke prediction model based on neural networks. Journal of Neurorestoratology. 2025. 13: 100156.
33. Hilbert A, Baskan D, Rieger J, Wagner C, Sehlen S, et al. AI-based Prediction of Imminent Primary Stroke on Claims Data Enables Accurate Patient Stratification. 2022. 2022-2111.
34. Asadi F, Rahimi M. The most efficient machine learning algorithms in stroke prediction: A systematic review. Health Science Reports. 2024. 7: e70062.
35. Sheakh MA, Tahosin MS, Alam MJ, Begum M. Optimizing medical image analysis through MViTX on multiple datasets with explainable AI. Neural Computing and Applications. 2025. 37: 8479-8507.
36. Heseltine-Carp W, Courtman M, Browning D, Kasabe A, Allen M, et al. Machine learning to predict stroke risk from routine hospital data: A systematic review. International Journal of Medical Informatics. 2025. 105811.
37. Vasan RS, Enserro DM, Xanthakis V, Beiser AS, Seshadri S. Temporal trends in the remaining lifetime risk of cardiovascular disease among middle-aged adults across 6 decades: the Framingham study. Circulation. 2022. 145: 1324-1338.
38. Liao X, Yao C. Navigating the future of evidence-based ai in healthcare. Journal of EvidenceBased Medicine. 2023. 16: 534.
39. Wang J, Luo J, Ye M, Wang X, Zhong Y, et al. Recent advances in predictive modeling with electronic health records. InIJCAI: proceedings of the conference. 2024. 2024: 8272.
40. Mapundu M. Kabudula C. Performance evaluation of machine learning and computer coded verbal autopsy (ccva) algorithms for cause of death determination: A comparative analysis of data from rural south africa. Frontiers in Public Health. 2022. 10: 990838.
41. Soladoye AA, Olawade DB, Adeyanju IA, Akpa OM, Aderinto N, et al. Optimizing stroke prediction using gated recurrent unit and feature selection in Sub-Saharan Africa. Clinical Neurology and Neurosurgery. 2025. 249: 108761.
42. Zhao P, Xue J, Xu X, Wang L, Chen D. Logistic model and gradient boosting machine model for physical therapy of lumbar disc herniation. Computational and Mathematical Methods in Medicine. 2022. 2022: 4799248.