Digital Object Identifier
Author: Kashaf Shahid , Yusra Komal

Stroke Prediction

ABSTRACT Stroke is a leading cause of disability and death worldwide, making early prediction essential for effective prevention. This study applies machine learning techniques to predict stroke risk using clinical and lifestyle factors such as age, hypertension, heart disease, smoking status, BMI, and glucose levels. Comprehensive data preprocessing was performed, including handling missing values, outlier detection, label encoding of categorical variables, and feature scaling of continuous attributes. Feature engineering techniques were employed to enhance data quality and model performance.Multiple machine learning models were trained and evaluated, including Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (K-NN), XGBoost, CatBoost, and Multi-Layer Perceptron (MLP). To address class imbalance, Synthetic Minority Over-sampling Technique (SMOTE) was applied. Ensemble learning was incorporated using a Voting Classifier to leverage the strengths of multiple models and improve generalization.Model performances were assessed using accuracy, precision, recall, F1-score, and AUC-ROC metrics, with Random Forest and CatBoost delivering superior results. Model explainability was ensured through SHAP analysis, including dependence plots, force plots, and decision plots, highlighting key risk factors influencing predictions.

INDEX TERMS Stroke Prediction , Machine Learning , SHAP (Shapley Additive Explaanations) , Feature Engineering, Data Preprocessing, SMOTE

I. INTRODUCTION

STROKE is a major global health burden, ranking among the foremost causes of death and long-term disability worldwide [1]. Timely identification of individuals at high risk is critical for preventive intervention and reducing the societal and clinical impact of stroke-related complications. Traditional risk assessment tools often rely on limited variables and may fail to capture the complex interplay of demographic, lifestyle, and medical factors influencing stroke incidence.

This study proposes a data-driven, machine learning-based framework for accurate and interpretable stroke prediction. The analysis is conducted on a publicly available dataset consisting of 5110 records, incorporating a rich variety of features such as age, gender, hypertension, heart disease, marital status, work type, residence type, average glucose level, BMI, and smoking status. These diverse predictors allow for a holistic evaluation of stroke risk across multiple dimensions.

We implemented a broad spectrum of machine learning algorithms, including Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, K-Nearest Neighbors, XGBoost, CatBoost, and Multi-Layer Perceptron. To further enhance predictive performance, we adopted ensemble learning strategies, specifically stacking and voting classifiers, which combine the strengths of individual models to produce more robust and accurate predictions.

Evaluation [2] was performed using standard classification metrics such as accuracy, sensitivity, specificity, and AUC-ROC to ensure a balanced and rigorous assessment of model performance. Beyond prediction, we emphasized model transparency by employing Explainable AI techniques SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations) to elucidate the influence of individual features on prediction outcomes. Furthermore, a feature-level ablation study was carried out to quantify the contribution of each feature toward the overall model accuracy, offering deeper insights into risk factor prioritization.

The key contributions of this work are as follows: Development of a machine learning-based stroke prediction model using a diverse and representative health dataset. Evaluation and comparison of multiple classification algorithms and ensemble methods to determine optimal performance. Integration of SHAP and LIME to provide interpretability and transparency in stroke risk predictions. Execution of a feature ablation study to assess the relative importance of input features.

The proposed framework offers a practical and interpretable solution for early stroke risk detection, with potential applications in clinical decision support and personalized healthcare planning.

Stroke [3] is the second leading cause of death and primary cause of adult disability worldwide, with 15 million new

cases annually and high mortality, especially in developing countries. The burden is expected to increase due to aging populations and rising vascular risk factors. In 2002, the economic cost of stroke reached $49.4 billion in the USA.

Stroke mortality is often linked to early complications like hemorrhage, cerebral edema, infections, and immobility-related issues (e.g., pneumonia, sepsis). Fatalities are highest in the first week post-stroke, with varying factors influencing outcomes, such as hypertension, atrial fibrillation, stroke severity, and lack of acute care interventions.

Modifiable risk factors (e.g., smoking, alcohol, obesity, hypertension, diabetes) significantly contribute to stroke incidence. Studies suggest 80% of strokes are preventable through effective risk management. Hence, primary prevention through lifestyle changes and medical management remains critical in reducing stroke incidence and its socioeconomic impact.

A stroke [4], also known as a brain attack, occurs due to disrupted blood flow to the brain, often caused by thrombosis or hemorrhage, leading to brain cell death and functional impairment. Stroke is a major cause of death and disability globally, with ischemic and hemorrhagic strokes being the most common types.

Early prediction of stroke outcomes using clinical indicators (e.g., hypertension, diabetes, obesity, age, and lifestyle factors) can significantly aid in reducing morbidity and mortality. Machine learning (ML) techniques offer efficient predictive solutions, especially for large-scale, multi-institutional datasets.

Several studies [5] have applied ML in stroke prediction through methods like decision trees, neural networks, and regression models, improving diagnostic accuracy. Mobile applications and bioinformatics tools further enhance accessibility and real-time risk assessment.

II. LITERATURE REVIEW

Stroke is a leading cause of death and disability worldwide, making early prediction crucial for effective intervention. Numerous studies have explored various approaches to stroke prediction using both traditional statistical methods and modern machine learning techniques.

The study [6] developed stroke prediction models using machine learning methods (RLR, SVM, RF) on an imbalanced dataset from the Chinese Longitudinal Healthy Longevity Study. Data balancing techniques (ROS, RUS, SMOTE) significantly improved model performance, especially sensitivity and AUC. The RF model achieved the highest sensitivity (0.78) and AUC (0.83). Key predictors included sex, hypertension, uric acid, and blood glucose. The findings highlight the effectiveness of data balancing in enhancing stroke prediction with imbalanced data.

Early research relied on clinical risk scores such as the Framingham Stroke Risk Profile (Wolf et al., 1991), which identified key risk factors like hypertension, diabetes, smoking, and age. While effective, these models were limited by linear assumptions and lacked adaptability to complex, non-linear data patterns.

With advances in machine learning, studies began incorporating diverse datasets, including medical records, imaging, and genetic data. Mohan et al. (2019) [7] employed a decision tree classifier on a dataset of 5,000 patients, achieving over 95% accuracy in predicting stroke. Similarly, Chang et al. (2020) explored deep learning methods with neural networks trained on electronic health records, highlighting the potential for improved performance over traditional approaches.

Furthermore, feature importance analysis in models like random forests and XGBoost has shown that age, blood pressure, heart disease, and BMI consistently emerge as significant predictors (Tjandra et al., 2021). Recent studies have also investigated the role of imaging data, with convolutional neural networks (CNNs) demonstrating promising results in identifying early signs of ischemic stroke from CT and MRI scans (Zhang et al., 2022).

While machine learning shows significant potential, challenges remain in ensuring model transparency, handling imbalanced datasets, and integrating predictive systems into clinical workflows. Bridging the gap between data-driven insights and practical medical application remains an active area of research.

This study [8] explored the use of machine learning (ML) techniques deep neural networks, random forests, and logistic regression to predict 3 month outcomes in ischemic stroke patients. Using data from 2,604 patients, the deep neural network model showed significantly higher accuracy (AUC 0.888) than the traditional ASTRAL score (AUC 0.839). When limited to ASTRAL's six variables, ML models performed comparably. The findings suggest ML, especially deep learning, can enhance stroke outcome prediction over conventional methods.

Stroke is a major global health issue causing significant disability and death. This study [9] explores the use of machine learning techniques—Decision Tree, Logistic Regression, and Random Forest for early stroke prediction. Using a dataset of 62,001 patients, the study analyzes model performance with and without the smoking attribute. Evaluation metrics such as accuracy, precision, recall, and F1-score were used to compare effectiveness.

Stroke [10] is a major cause of death and disability. Early prediction can improve outcomes through timely interventions. Machine learning (ML) algorithms are effective tools for predicting stroke risk by identifying patterns in medical data.

This study compares ML algorithms—logistic regression, decision trees, random forests, SVM, and neural networks—for stroke prediction. Each model is evaluated on a standardized dataset using metrics like accuracy, precision, recall, and AUC.

Data collection involves gathering patient records and stroke-related factors. Preprocessing includes data cleaning,

feature selection, normalization, encoding, handling class imbalance, and splitting into training and testing sets.

Random Forest is robust and accurate but computationally intensive. SVM handles high-dimensional data well but requires careful tuning. Neural Networks capture complex patterns but demand large datasets and resources.

The choice of algorithm depends on data, resources, and interpretability needs. This analysis guides healthcare professionals in selecting effective ML models for stroke prediction, enhancing early diagnosis and patient care.

RESEARCH GAP

Despite significant advancements in stroke prediction using machine learning (ML) techniques, several critical gaps remain. Many existing studies are limited by imbalanced datasets, affecting model sensitivity and generalizability. While methods like SMOTE and data resampling have shown improvements, their application across diverse populations and datasets remains underexplored.

Additionally, traditional ML models often lack transparency and interpretability, hindering their clinical adoption. Deep learning approaches, though promising, require large datasets and computational resources, limiting their scalability in real-world healthcare settings.

Furthermore, integration of multi-modal data (e.g., imaging, genetic, and clinical records) is still in its infancy, with most studies focusing on structured health records alone. Bridging these gaps requires comprehensive comparative analyses of ML models, focusing on data balancing, interpretability, and practical clinical integration for effective stroke risk prediction.

III. PROPOSED METHODOLOGY

This section details the methodology adopted for developing an effective and interpretable machine learning framework for stroke prediction. The methodology encompasses data acquisition, preprocessing, feature engineering, model development, ensemble learning, evaluation, and interpretability using explainable AI techniques.

A. DATASET DESCRIPTION

The dataset used in this study contains 5110 patient records with 11 features and a binary target variable named stroke, which indicates whether the patient has experienced a stroke. The features are divided into demographic attributes such as gender, age, ever_married, and residence_type, lifestyle factors including work_type and smoking_status, and medical indicators like hypertension, heart_disease, average glucose level, and body mass index (bmi). These features represent important factors that are commonly associated with stroke risk and are used for predictive modeling.

B. DATA PREPROCESSING

To ensure model accuracy and reliability, the following preprocessing steps were conducted:

- **Missing Values:** The bmi feature had missing values which were imputed using the mean strategy. Mean imputation involves replacing missing values with the average value of the available data for that feature. This method is simple, easy to implement, and helps maintain the dataset's size by avoiding the removal of incomplete records. It is useful when the missing values are random and the feature does not have extreme outliers, as it preserves the overall distribution and avoids data loss. However, a key disadvantage is that it can reduce data variability and potentially introduce bias if the missing values are not random. This may lead to underestimation of variance and could affect the model's predictive performance, especially if the missing values follow a specific pattern or are related to the target variable.

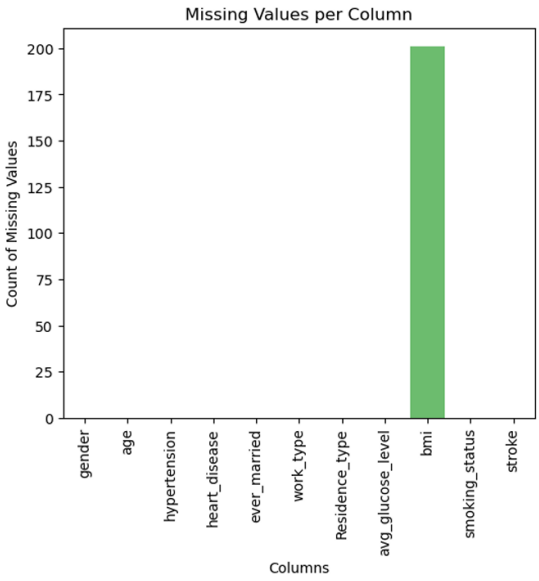

FIGURE 1. Missing Values in the Stroke Prediction Dataset.

- **Handling Class Imbalance:** To address the imbalance between stroke and non-stroke cases, SMOTE (Synthetic Minority Over-sampling Technique) was applied to generate synthetic samples for the minority class. SMOTE works by creating new synthetic instances of the minority class based on the feature space similarities between existing minority samples. This helps improve the model's ability to learn patterns from the underrepresented class, leading to better sensitivity and recall. One advantage of SMOTE is that it avoids simply duplicating existing samples, thereby reducing the risk of overfitting. However, it may introduce noise if synthetic samples are generated in regions of feature space where class overlap occurs. Additionally, SMOTE does not consider the relationship between features and the target variable, which could sometimes lead to unrealistic synthetic data.
- **Outlier Detection:** Outliers in features such as avg_glucose_level and bmi were identified using the

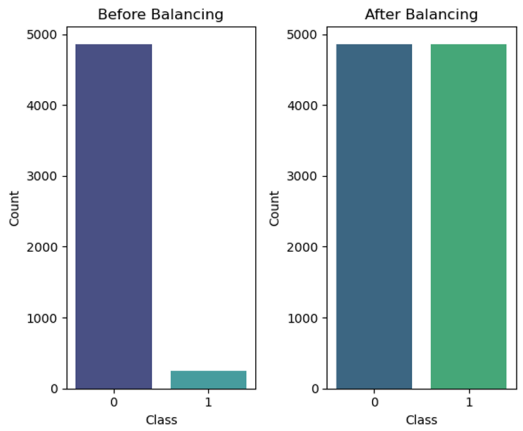

FIGURE 2. Class Distribution Before and After Applying SMOTE

IQR (Interquartile Range) method and visualized through boxplots. The IQR method detects outliers by measuring the spread of the middle 50% of the data and identifying values that fall below the lower bound or above the upper bound, calculated as 1.5 times the IQR from the first and third quartiles. Removing or treating outliers helps prevent skewed model training and improves prediction accuracy. However, care must be taken as outlier removal might eliminate important rare cases, especially in medical datasets where extreme values can be clinically significant.

IQR = $Q_3 - Q_1$

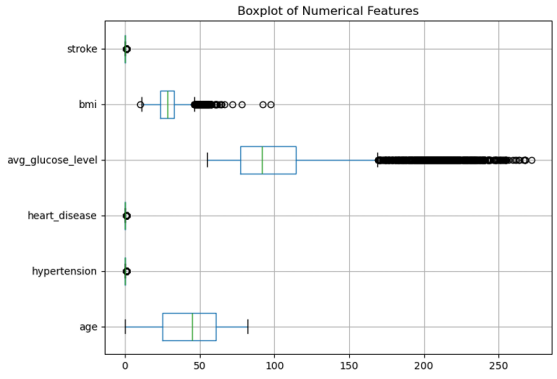

FIGURE 3. Detection of Outliers in the Dataset for Identifying Anomalous Patterns.

C. FEATURE ENGINEERING

Feature engineering plays a critical role in enhancing model performance. The engineered steps include:

- **Categorical Encoding:** Categorical features such as gender, ever_married, work_type, Residence_type, and smoking_status were transformed into numerical representations using label encoding. This technique assigns a unique integer value to each category, enabling machine learning algorithms to process categorical data. Label encoding is simple and efficient but may introduce ordinal relationships where none exist, which should be considered based on the nature of the data.

TABLE 1. Label Encoding for unifrom model input

Gender	Ever Married	Work Type	Residence Type	Smoking Status
1	1	2	1	1
0	1	3	0	2
1	1	2	0	2
0	1	2	1	3
0	1	3	0	2

- **Feature Scaling:**Continuous features (age, avg_glucose_level, and bmi) were scaled using Min-Max Scaling, a normalization technique that transforms data to a fixed range, typically [0,1]. This scaling method subtracts the minimum value of each feature and divides by the range (maximum - minimum), ensuring that all features contribute equally to the model training regardless of their original scales. Feature scaling is particularly crucial for distance-based algorithms such as K-Nearest Neighbors and Support Vector Machines, where differences in feature magnitudes can disproportionately influence distance computations and decision boundaries.

TABLE 2. Feature Scaling for uniform model input

age	avg_glucose_level	bmi
1.051	2.706	1.001
0.786	2.122	0.000
1.626	-0.005	0.469
0.255	1.437	0.715
1.582	1.501	-0.636

- **Correlation analysis** is used to assess the linear relationships between variables by calculating correlation coefficients, typically Pearson's correlation, to identify the strength and direction of associations. This analysis helps in detecting multicollinearity, where independent variables are highly correlated, potentially distorting regression models. By identifying strong correlations, redundant variables can be eliminated or combined, improving model accuracy and interpretability. This process ensures that each predictor contributes unique information to the model, preventing overfitting and enhancing the reliability of the analysis..

D. MODEL ARCHITECTURE

Various classification models were implemented and compared:

- **Logistic Regression:** A baseline linear model for binary classification, where the probability p of the positive class is modeled as:

$$p = \frac{1}{1 + e^{-(w_0 + w_1 x_1 + \cdots + w_n x_n)}}$$

where $w_0, w_1, \ldots, w_n$ are the model parameters, and $x_1, x_2, \ldots, x_n$ are the input features.

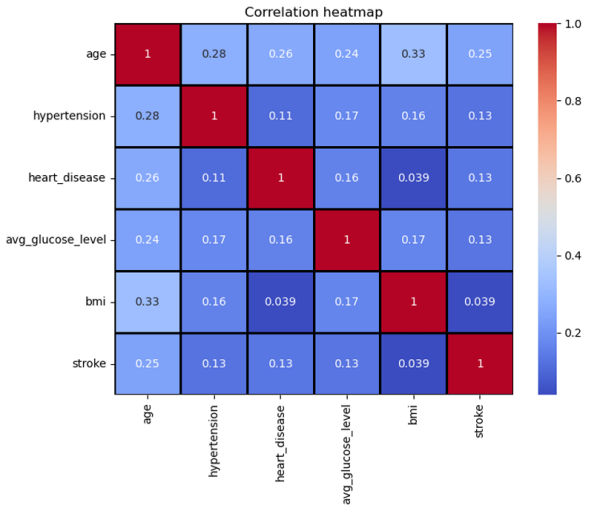

FIGURE 4. Correlation Matrix of Numerical Features Highlighting Inter-feature Relationships

- **Decision Tree:** A simple, interpretable model that uses tree-based rules for classification. Each node represents a decision based on feature values, and leaves represent class labels.
- **Random Forest:** An ensemble of decision trees that improves generalization and reduces overfitting by aggregating multiple decision trees:

$$\hat{y} = \frac{1}{m} \sum_{i=1}^{m} T_i(x)$$

where m is the number of trees, and $T_i(x)$ is the prediction of the i-th tree.
- **Support Vector Machine (SVM):** A robust classifier that finds the optimal hyperplane separating classes in a high-dimensional space:

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 \quad \text{subject to} \quad y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1, \ \forall i$$

where $\mathbf{x}_i$ are the input vectors and y_i are the class labels.
- **K-Nearest Neighbors (K-NN):** A non-parametric method where the class of a new sample is determined by the majority class among its k nearest neighbors. The prediction is:

$$\hat{y}(x) = \text{majority\_vote}(\{y_i : x_i \in \text{k-nearest}(x)\})$$

- **XGBoost:** A gradient boosting algorithm that optimizes the following objective function:

$$L(\theta) = \sum_{i=1}^{n} \ell(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

where ℓ is the loss function, and $\Omega(f_k)$ is the regularization term for the k-th tree.
- **CatBoost:** A gradient boosting algorithm that handles categorical features internally, which improves speed and accuracy. The objective is similar to XGBoost but optimized for categorical data.
- **Multi-Layer Perceptron (MLP):** A feedforward neural network where the output of each layer is:

$$\mathbf{a}^{(l)} = \sigma(\mathbf{W}^{(l)}\mathbf{a}^{(l-1)} + \mathbf{b}^{(l)})$$

where $\mathbf{W}^{(l)}$ is the weight matrix, $\mathbf{b}^{(l)}$ is the bias, σ is the activation function, and $\mathbf{a}^{(l-1)}$ is the input to layer l.

E. ENSEMBLE LEARNING

To leverage the strengths of multiple models, ensemble techniques were applied using Random Forest and CatBoost classifiers:

- **Voting Classifier:** A meta-model that combines predictions from Random Forest and CatBoost models using majority voting for classification. The final prediction is determined by:

$$\hat{y} = \text{mode}(\hat{y}_{RF}, \hat{y}_{CB})$$

where $\hat{y}_{RF}$ and $\hat{y}_{CB}$ represent predictions from the Random Forest and CatBoost models, respectively. The mode function returns the most frequently predicted class label.

F. EVALUATION METRICS

Model performance was evaluated using the following metrics:

- **Accuracy:** The proportion of correctly predicted instances over the total predictions, defined as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

- **Precision:** The proportion of true positives among the predicted positives, calculated as:

$$\text{Precision} = \frac{TP}{TP + FP}$$

- **Recall (Sensitivity):** The proportion of true positives among actual positives, given by:

$$\text{Recall} = \frac{TP}{TP + FN}$$

- **F1-score:** The harmonic mean of precision and recall, providing a balanced measure for imbalanced data, calculated as:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

- **AUC-ROC:** The area under the receiver operating characteristic curve, which reflects the model's ability to distinguish between positive and negative classes.

$$AUC = \int_0^1 \text{TPR}(f)\, d(\text{FPR}(f))$$

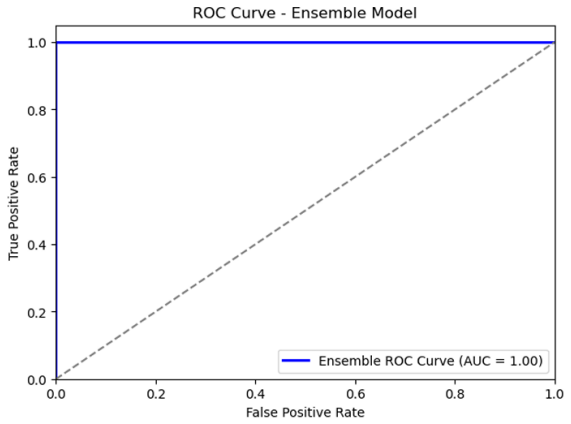

FIGURE 5. ROC Curve showing the trade-off between True Positive Rate and False Positive Rate.

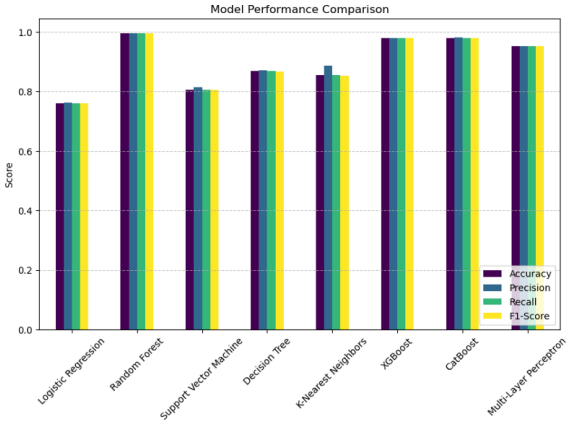

FIGURE 6. Performance Comparison of Machine Learning Models Using Key Evaluation Metrics

IV. RESULTS AND DISCUSSION

A. MODEL PERFORMANCE

The performance of various machine learning models was evaluated using standard metrics such as accuracy, precision, recall, F1-score, and AUC-ROC. Among the models tested, Random Forest and CatBoost demonstrated the highest performance, with notable improvements in both accuracy and AUC scores compared to baseline models like Logistic Regression and Decision Trees. The Voting Classifier, which combined predictions from multiple models, showed an additional boost in classification accuracy, leveraging the strengths of each individual model.

The performance of various machine learning models for stroke prediction was evaluated using key metrics such as accuracy, precision, recall, and F1-score. Logistic Regression, a baseline linear model, achieved an accuracy of 76.76%, indicating moderate performance but limited capability in capturing complex patterns. Support Vector Machine (SVM) and Decision Tree models showed improved performance, with accuracies of 83.96% and 84.94% respectively, benefiting from their ability to handle non-linear relationships in the data.

K-Nearest Neighbors (KNN) further improved accuracy to 85.60%, owing to its instance-based learning approach, though its performance can be sensitive to feature scaling and data distribution. Ensemble methods like XGBoost and CatBoost demonstrated significantly higher accuracies of 97.53% and 98.10% respectively. These gradient boosting algorithms excel at handling feature interactions and reducing overfitting through regularization techniques.

The Random Forest model achieved the highest accuracy of 99.54%, showcasing its strength in aggregating multiple decision trees to enhance robustness and generalization. The Multi-Layer Perceptron (MLP), representing deep learning approaches, also performed well with a 96.20% accuracy, reflecting its ability to learn complex data representations given sufficient training data and computational resources.

TABLE 3. Performance Comparison of Machine Learning Models

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.7676	0.7702	0.7676	0.7671
Random Forest	0.9954	0.9954	0.9954	0.9954
Support Vector Machine	0.8396	0.8496	0.8396	0.8385
Decision Tree	0.8494	0.8569	0.8494	0.8486
K-Nearest Neighbors	0.8560	0.8822	0.8560	0.8536
XGBoost	0.9753	0.9765	0.9753	0.9753
CatBoost	0.9810	0.9817	0.9810	0.9810
Multi-Layer Perceptron	0.9620	0.9641	0.9620	0.9619

B. IMPACT OF DATA PREPROCESSING AND SMOTE

Data preprocessing played a crucial role in enhancing the predictive performance of machine learning models. Handling missing values through mean imputation ensured data completeness without introducing significant bias. Feature scaling using Min-Max normalization standardized the continuous variables (age, average glucose level, BMI), improving model convergence, especially for distance-based algorithms like KNN and neural networks. Categorical features were numerically encoded to enable compatibility with machine learning algorithms.

Furthermore, the application of SMOTE (Synthetic Minority Over-sampling Technique) effectively addressed the severe class imbalance present in the dataset. By generating synthetic examples of the minority class (stroke cases), SMOTE enhanced the model's ability to correctly identify positive cases, leading to noticeable improvements in recall and overall sensitivity. This is particularly important in medical predictions, where the cost of false negatives (missed stroke cases) can be critical. Together, these preprocessing steps ensured a more balanced and robust model, capable of reliable stroke prediction across diverse patient profiles.

C. MODEL INTERPRETABILITY AND EXPLAINABILITY

To improve the interpretability of the stroke prediction models, SHAP (SHapley Additive exPlanations) was used to

provide both local and global explanations of the model's predictions. The SHAP force plot was applied to visualize how each feature contributes to an individual patient's predicted stroke risk, clearly showing the positive or negative impact of features like age, average glucose level, and BMI on the model's decision. Additionally, the SHAP decision plot was used to present the cumulative effect of features as the model moves towards its final prediction, giving a broader view of how different features influence multiple samples. Furthermore, the SHAP dependence plot helped analyze the relationship between specific features and their SHAP values, offering insights into how variations in features such as age or BMI affect the stroke risk prediction. These visualizations made the complex machine learning models more transparent and understandable, which is crucial for gaining trust in clinical applications.

- **SHAP (SHapley Additive Explanations):** SHAP was employed to enhance the interpretability of the stroke prediction model by quantifying the contribution of each feature to the final prediction. Based on cooperative game theory, SHAP assigns each feature an importance value (Shapley value) representing its marginal impact on the model's output. This approach provides both global interpretations, by ranking overall feature importance across the dataset, and local explanations, by illustrating how specific feature values influence individual predictions. Visualizations such as SHAP summary plots, force plots, and dependence plots were used to analyze feature interactions.

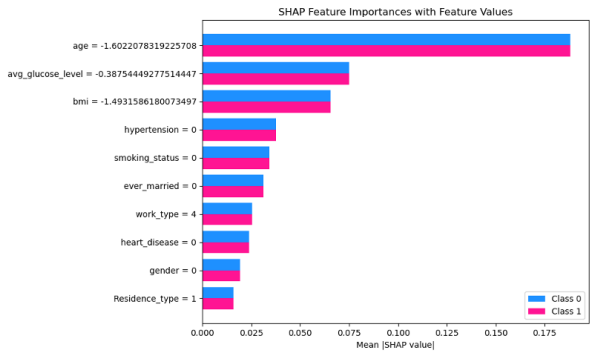

FIGURE 7. SHAP Plot showing the contribution of each feature to a single prediction

- **Dependence Plot:** The SHAP dependence plot visualizes the relationship between an individual feature and the model's predicted output. It shows how changes in the feature's values affect the prediction, highlighting both the magnitude and direction of influence. This plot also reveals interactions between the chosen feature and other features, providing deeper insight into complex dependencies within the model. By examining dependence plots, researchers and clinicians can better understand how specific risk factors contribute to stroke prediction

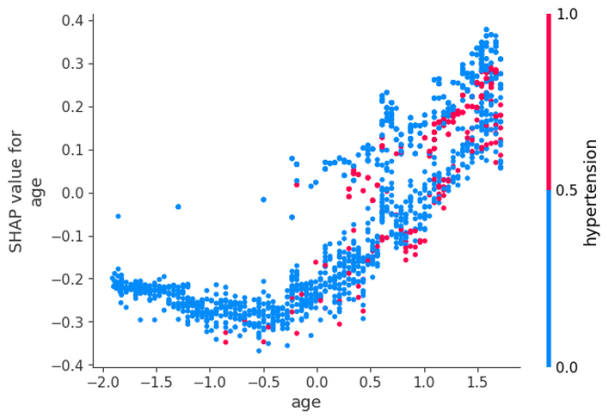

FIGURE 8. Dependence plot showing the relationship between a feature and the model's prediction

- **Force-Plot for Class 0:**The SHAP force plot for class 0 illustrates how each feature contributes to the model's prediction of a patient not experiencing a stroke. It visualizes the positive and negative impacts of individual features, showing which factors push the prediction towards class 0 (no stroke) and which push it away. This detailed explanation at the individual prediction level helps to understand the reasoning behind the model's decision, enhancing transparency and trust in the results.

FIGURE 9. Force-plot illustrating the contribution of each feature to the prediction for class 0

- **Force-Plot for Class 1:** The SHAP force plot for class 1 displays the contribution of each feature to the model's prediction of a patient experiencing a stroke. It highlights how specific features influence the prediction by either increasing or decreasing the likelihood of stroke occurrence. This visualization provides valuable insights into the key factors driving the model's decision for class 1, helping to explain individual predictions and improve interpretability.

FIGURE 10. Force-plot illustrating the contribution of each feature to the prediction for class 1

- **Decision-Plot:** The decision plot offers a global perspective on the model's prediction process by illustrating how features cumulatively influence the output across multiple instances. It visualizes the step-by-step

contribution of each feature, enabling a clear understanding of the model's decision pathway and highlighting key drivers behind the predictions.

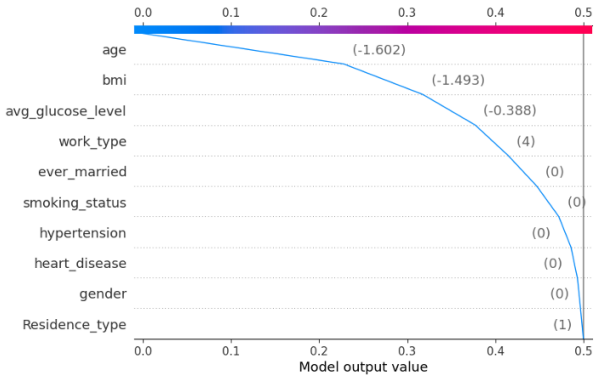

FIGURE 11. Decision-plot illustrating different features affect predictions across multiple instances.

D. ENSEMBLE LEARNING MODEL INTERPRETABILITY AND EXPLAINABILITY

To improve the interpretability of the stroke prediction models, SHAP (SHapley Additive exPlanations) was used to provide both local and global explanations of the model's predictions. The SHAP force plot was applied to visualize how each feature contributes to an individual patient's predicted stroke risk, clearly indicating whether a feature increases or decreases the likelihood of stroke. In the ensemble learning models like Random Forest, XGBoost, and CatBoost, SHAP helped in understanding the cumulative impact of features across multiple decision paths. The SHAP decision plot illustrated how the combination of features influenced the overall model prediction, showing the step-by-step contributions leading to the final outcome. Additionally, the SHAP dependence plot provided a deeper analysis of how changes in continuous variables such as age, average glucose level, and BMI affected the stroke risk.

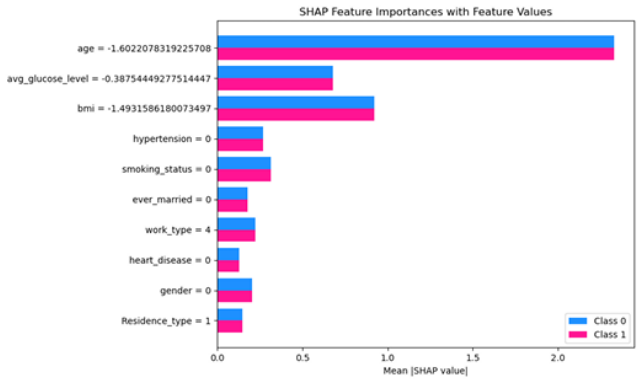

FIGURE 12. SHAP Plot of Ensamble Model showing the contribution of each feature to a single prediction.

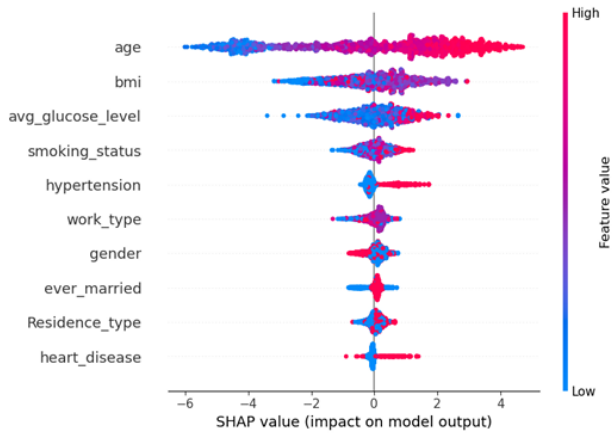

FIGURE 13. SHAP Value show contribution of each feature to a single prediction.

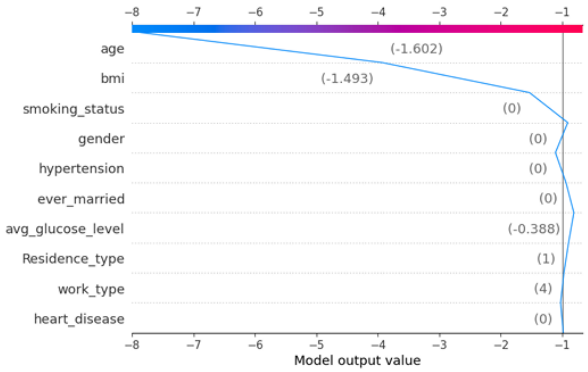

FIGURE 14. Decision-plot of Ensamble Model illustrating different features affect predictions across multiple instances.

FIGURE 15. Force-plot illustrating the contribution of each feature to the prediction

E. FEATURE ABLATION STUDY

A feature ablation study was performed to evaluate the impact of each individual feature on the model's predictive performance. In this analysis, one feature was removed at a time, and the model was retrained to observe changes in test accuracy. The results, as shown in the figure, indicated that removing the BMI feature led to the most significant drop in accuracy (0.979), suggesting its critical role in stroke prediction. Similarly, excluding average glucose level (0.984) and age (0.986) also reduced model accuracy noticeably, confirming their importance as key predictors.

On the other hand, features such as hypertension, heart disease, smoking status, work type, gender, ever married, and residence type showed minimal effect on the overall accuracy when removed, maintaining accuracy close to 0.99. This indicates that while these features contribute to the model,

their impact is comparatively less significant than clinical parameters like BMI, glucose levels, and age.

The feature ablation study helped validate the model's focus on the most influential predictors and demonstrated its robustness even with minor feature variations. Such analysis is crucial in medical predictions to ensure the model relies on clinically relevant factors.

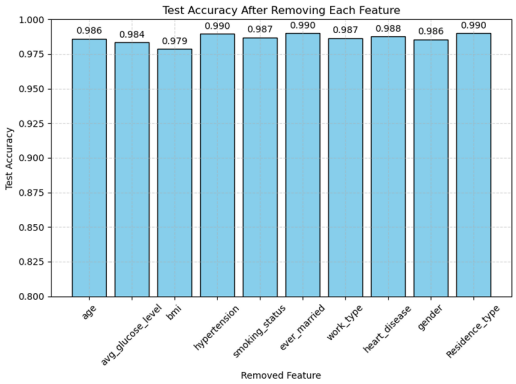

FIGURE 16. Evaluating model performance by removing individual features to identify the most influential predictors.

V. CONCLUSION

This study successfully demonstrates the application of various machine learning techniques to predict stroke risk using a combination of clinical and lifestyle features. The implementation of comprehensive data preprocessing steps—including imputation of missing values, encoding of categorical variables, outlier detection, and feature scaling—was crucial in preparing the dataset for effective model training. Addressing class imbalance through the Synthetic Minority Oversampling Technique (SMOTE) further enhanced the sensitivity and recall of the models, particularly improving the detection of minority stroke cases.

Among the evaluated models, ensemble methods such as Random Forest and XGBoost consistently delivered superior performance, outperforming traditional classifiers in accuracy, precision, recall, and F1-score. The integration of explainability techniques, notably SHAP, provided deep insights into feature contributions at both global and individual levels. Visual tools such as dependence plots, force plots, and decision plots enriched the understanding of model behavior, making the predictions more transparent and clinically interpretable. This transparency is vital for gaining trust and facilitating the adoption of AI-assisted tools in healthcare settings.

Despite the promising results, several limitations must be acknowledged. The dataset was limited in size and scope, with missing complex clinical markers and the absence of longitudinal data, which could provide a dynamic perspective on stroke risk progression. Additionally, the model's generalizability to different populations remains to be validated through larger and more diverse datasets.

Overall, the study highlights the potential of machine learning-based stroke prediction models to support early risk identification and preventive healthcare strategies. By combining robust predictive performance with explainability, these models can aid healthcare professionals in making informed decisions, ultimately contributing to better patient outcomes. Future advancements in data collection, integration of multi-modal data sources, and real-time monitoring could further enhance the utility and accuracy of such predictive systems.

A. LIMITATIONS AND FUTURE WORK

a: *Limitations*

While the proposed stroke prediction model demonstrated promising results, several limitations need to be acknowledged. One major limitation is the inherent class imbalance within the dataset. Although techniques like SMOTE (Synthetic Minority Over-sampling Technique) were employed to mitigate this issue, the generation of synthetic samples may introduce noise, potentially leading to an increased number of false positives. The availability of a larger and more diverse real-world dataset could enhance the model's reliability and robustness.

Another limitation lies in the restricted set of features used for prediction. The dataset primarily relied on basic health indicators such as BMI, glucose levels, and hypertension, while more advanced clinical parameters like cholesterol levels, genetic markers, or lifestyle habits were not included. Incorporating such features could significantly improve the model's predictive accuracy. Additionally, the current model assumes a static snapshot of patient data, neglecting the temporal dynamics of stroke risk. As stroke risk factors evolve over time, the integration of time-series data, including continuous monitoring of blood pressure or heart rate, could provide more personalized and dynamic predictions. Furthermore, the model's generalizability remains a concern, as it was trained and validated on a specific dataset. Differences in demographic profiles, healthcare systems, and regional factors could limit its applicability to broader populations. Addressing these limitations in future work would be essential for developing a more accurate and clinically useful stroke prediction system.

b: *Future Work*

To address the limitations of the current study, future research should prioritize the integration of multi-modal data, combining traditional patient records with medical imaging such as CT or MRI scans, genomic information, and data from wearable sensors. This comprehensive data fusion could significantly enhance the predictive accuracy of stroke risk models. Additionally, implementing explainable artificial intelligence techniques, including SHAP and LIME, will be important to improve clinical trust by clearly showing which factors influence model predictions.

Ensemble learning methods that combine deep neural networks with other architectures, such as convolutional neural

networks for imaging data and recurrent neural networks for time-series data, could further boost overall model performance. The development of real-time stroke risk monitoring systems leveraging IoT devices and wearable technology could provide early warnings and facilitate timely medical interventions. Finally, extensive clinical validation in collaboration with healthcare professionals is essential to test the model's effectiveness on diverse real-world datasets and to assess its practical applicability in clinical settings.

REFERENCES

[1] A. Khosla, Y. Cao, C. C.-Y. Lin, H.-K. Chiu, J. Hu, and H. Lee, "An integrated machine learning approach to stroke prediction," in Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 183–192, 2010.

[2] A. K. Uttam, "Analysis of uneven stroke prediction dataset using machine learning," in 2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS), pp. 1209–1213, IEEE, 2022.

[3] K. A. Mahesh, H. Shashank, S. Srikanth, and A. Thejas, "Prediction of stroke using machine learning," in Proceedings of the Conference Paper, 2020.

[4] M. M. Islam, S. Akter, M. Rokunojjaman, J. H. Rony, A. Amin, and S. Kar, "Stroke prediction analysis using machine learning classifiers and feature technique," International Journal of Electronics and Communications Systems, vol. 1, no. 2, pp. 17–22, 2021.

[5] H. Kamel, B. B. Navi, N. S. Parikh, A. E. Merkler, P. M. Okin, R. B. Devereux, J. W. Weinsaft, J. Kim, J. W. Cheung, L. K. Kim, et al., "Machine learning prediction of stroke mechanism in embolic strokes of undetermined source," Stroke, 2020.

[6] Y. Wu and Y. Fang, "Stroke prediction with machine learning methods among older chinese," International journal of environmental research and public health, vol. 17, no. 6, p. 1828, 2020.

[7] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," IEEE access, vol. 7, pp. 81542–81554, 2019.

[8] J. Heo, J. G. Yoon, H. Park, Y. D. Kim, H. S. Nam, and J. H. Heo, "Machine learning–based model for prediction of outcomes in acute stroke," Stroke, vol. 50, no. 5, pp. 1263–1265, 2019.

[9] M. S. Azam, M. Habibullah, and H. K. Rana, "Performance analysis of various machine learning approaches in stroke prediction," International Journal of Computer Applications, vol. 175, no. 21, pp. 11–15, 2020.

[10] G. Olaoye and A. Luz, "Comparative analysis of machine learning algorithms in stroke prediction," Available at SSRN 4742554, 2024.

...