STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **15**, June 2026, pp 5406–5421.
Published online in International Academic Press (www.IAPress.org)

Integrating Tuned Ensemble Learning and Explainable AI for Reliable Stroke Prediction

Fatematuj Zohura Yasrin, Most. Jannatul Ferdous*, Fatema, Tamanna Akter Shanta, Azmol Hasan Ratul

Department of CSE, Bangladesh University of Business and Technology(BUBT), Dhaka, Bangladesh

Abstract A stroke occurs due to a sudden interruption of blood flow to a specific region of the brain. Although extensive research has already been conducted using this dataset, previous studies have not effectively addressed the issue of data leakage during testing. The primary contribution of this study is to mitigate both the data leakage and class imbalance problems by applying three different balancing techniques. We employed nine classification algorithms in this study: Random Forest (RF), AdaBoost (AB), Logistic Regression (LR), Gradient Boosting (GB), K-Nearest Neighbors (KN), Decision Tree (DT), Naive Bayes (NB), Voting (VT), and Stacking (ST). A key aspect of our approach involved using hyperparameter tuning to determine the optimal configuration for each model. Additionally, we proposed a novel ensemble method, named Voting, which combines hyperparameter-optimized LR, DT, GB, and RF classifiers. The performance of this method was compared with other models using various evaluation metrics, including accuracy, precision, ROC curve, PR-AUC curve, and more. To enhance the interpretability of important features within the dataset, we also applied two explainable AI (XAI) techniques: Local Interpretable Model-Agnostic Explanations (LIME) and Shapley Additive Explanations (SHAP).

Keywords Stroke, Machine Learning, Ensemble Technique, Voting, XAI, LIME, SHAP.

AMS 2010 subject classifications 97P50, 97R40

DOI: 10.19139/soic-2310-5070-3462

1. Introduction

A stroke is a sudden and severe neurological event that impacts the brain's blood vessels. It happens when the blood supply to a specific area of the brain is disrupted, cutting off oxygen to brain cells. Although rare, a stroke can also result from a ruptured blood vessel causing cerebral hemorrhage. The most frequent type, however, is ischemic stroke, which occurs when an artery becomes narrowed or blocked, limiting blood flow to parts of the brain. Risk factors for stroke include a history of myocardial infarction, previous stroke, heart failure, atrial fibrillation, age over 55, high blood pressure, smoking, high blood cholesterol, and narrowing or calcification of the carotid arteries. Stroke symptoms may appear suddenly or develop gradually and often include paralysis or numbness on one side of the body, difficulty speaking or walking, dizziness, blurred vision, headache, and vomiting. In severe cases, it can lead to loss of consciousness or coma. The initial 24 hours after a stroke are particularly crucial for the patient's prognosis.

This study presents a machine learning-based binary classification approach to effectively predict stroke occurrences. To address class imbalance, we applied three synthetic balancing techniques—SMOTETENN, SMOTE and ADASYN exclusively on the training dataset.After preprocessing, various classification models were developed and evaluated on the balanced dataset. We experimented with several algorithms, including Random

*Correspondence to: Most. Jannatul Ferdous (Email: jannatul.083035@gmail.com). Department of CSE, Bangladesh University of Business and Technology (BUBT)

ISSN 2310-5070 (online) ISSN 2311-004X (print)

Forest, K-Nearest Neighbors (KNN), Logistic Regression, Decision Tree, and Naive Bayes. Additionally, ensemble techniques such as stacking and majority voting were implemented, with majority voting being the primary contribution of this study.

To enhance the transparency and trustworthiness of our predictive models, we employed Explainable AI (XAI) techniques. These tools help stakeholders and developers understand how the models make decisions. The integration of ensemble learning with XAI significantly improved the effectiveness and interpretability of the stroke prediction models.Experimental results demonstrated that the voting ensemble method outperformed individual models in terms of performance metrics such as the Precision-Recall curve and ROC curve.

1.1. Research Contribution

- To prevent data leakage and maintain the integrity of the evaluation process, balancing techniques were applied strictly to the training set, ensuring the test set remains an unbiased indicator of real-world model performance.
- To address the issue of highly imbalanced data, we used three different oversampling techniques: SMOTE, SMOTEENN and ADASYN. The impact of each method on classifier performance was thoroughly evaluated.
- To ensure optimal performance and robust generalization, all models were fine-tuned using GridSearchCV with 10-fold cross-validation.
- SHAP and LIME were employed to enhance model interpretability by breaking down predictions into individual feature contributions, helping us understand the reasoning behind specific outcomes.

2. Literature Review

Several researchers have explored the use of machine learning techniques [1] for stroke prediction. This section outlines the key contributions of specific studies in the field. Sung et al. [2] reported that the K-Nearest Neighbors (KNN) algorithm outperformed other models, including regression tree modeling and multiple linear regression (MLR), in predicting stroke severity. Cheng et al. [3] employed a dataset from Sugam Multispecialty Hospital in Kumbakonam, Tamil Nadu, India, to predict ischemic stroke using two artificial neural network (ANN) models. The reported accuracies were 95.1% and 79.2%, respectively.

Nataliia [4] used a variety of models, and the random forest classifier consistently outperformed them, with around 90% precision, recall, F1-score, and accuracy. When improved using grid search on balanced data, the accuracy increased to 96%, demonstrating the difficulties of relying just on accuracy to evaluate imbalanced datasets. These findings highlight the importance of improved preprocessing and ensemble learning in creating dependable, data-driven tools for stroke risk assessment.

Kayola, G. et al. [5] studies have identified multiple barriers to post-stroke rehabilitation in low- and middle-income countries (LMICs), including limited national guidelines, low health system prioritization, workforce shortages, financial hardship, and poor health literacy. These problems contribute to higher mortality, slower recovery, increased caregiver load, and economic losses. Establishing stroke facilities, expanding specialist training, introducing task shifting, and promoting rehabilitation and community-based treatments are all proposed strategies to increase access and care quality.

Singh et al. [6] applied artificial intelligence to forecast strokes using data from the Cardiovascular Health Study (CHS). They utilized principal component analysis (PCA) in conjunction with a decision tree algorithm for feature extraction, followed by a classification neural network, which achieved 97% accuracy. Amini et al. [7] analyzed data from 807 participants, both healthy and stroke-affected, and identified 50 potential risk factors such as diabetes, heart disease, tobacco use, hypertension, and alcohol consumption. Among the algorithms used, the C4.5 decision tree achieved 95% accuracy, while KNN achieved 94%.

Adam et al. [8] also used data from the Sugam Multispecialty Hospital to evaluate the performance of decision tree and KNN classifiers. Their findings indicated that decision trees outperformed the KNN model. Jeena et al.

[9] investigated various risk factors to better understand the likelihood of stroke occurrence. Singh and Choudhary [10] utilized a decision tree algorithm on the CHS dataset to predict stroke cases. Sudha [11] proposed a stroke prediction framework using neural networks, decision trees, and Naïve Bayes classifiers. The dataset used, sourced from the Institute of Medicine, contained patient demographic data, medical history, and genetic indicators associated with stroke symptoms.

Kansadub [12] presented a stroke risk prediction methodology based on demographic features such as gender, age, and education level. Due to severe class imbalance in the dataset (250 stroke cases and 67,897 non-stroke cases), resampling was applied to create a balanced set of 250 stroke and 500 non-stroke records. Tazin [13] utilized physiological indicators along with machine learning algorithms for stroke prediction. Among the algorithms tested, Random Forest achieved the highest performance, with a precision of approximately 96%.A study in [14] trained models using Random Forest, Decision Tree, and Multilayer Perceptron (MLP). The performance of the models was closely matched, with Decision Tree achieving 74.31%, Random Forest 74.53%, and MLP 75.02% accuracy. Another study [15] demonstrated the use of machine learning algorithms—Decision Tree, Naïve Bayes, and SVM—for predicting cardiac events. However, the highest accuracy achieved was only 60%, indicating limited model performance.In contrast, the trial results reported in [16] demonstrated that classification stacking significantly outperformed other approaches, achieving an AUC of 98.9%, and precision, recall, and F1-score of 97.4%, with an overall accuracy of 98%.

Le, N. B., Pham [17] used machine learning methods such Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbors (KNN), and XGBoost on medical datasets to predict strokes and assess risk. The XGBoost classifier outperformed the others, with an accuracy, precision, recall, and F1-score of around 87.5%. Comparative evaluations show that ensemble-based models [18, 19, 20], such as Random Forest and XGBoost, regularly outperform classical classifiers across numerous assessment metrics.

It is important to note that most of the above studies applied balancing techniques to both training and test sets. While this improved model accuracy, it may have led to data leakage and over-optimistic evaluations. In our work, balancing techniques are applied strictly only to the training set, ensuring that the test set remains completely unseen by the model and serves as a reliable indicator of real-world performance.

3. Material and Approach

3.1. Dataset Descriptions

Our study utilized a publicly available dataset from Kaggle as its foundation [21]. The dataset contains 5,110 observations and 12 attributes—11 features used for machine learning modeling and 1 target variable representing the stroke class.

1. ID: a unique identifier.
2. gender: ”Male”, ”Female”, or ”Other”
3. Patient’s age
4. Hypertension: 0 (no hypertension) or 1 (hypertension).
5. Heart illness: 0 (no heart disease) or 1 (heart disease).
6. Ever married? ”No” or ”Yes”.
7. BMI: Body Mass Index.
8. Smoking status options include ”formerly smoked,” ”never smoked,” ”smokes,” or ”unknown”*.
9. Stroke: 1 if the patient had a stroke, 0 otherwise.
10. Work type: ”children,” ”government jobs,” ”never worked,” ”private,” or ”self-employed.”
11. Residential type: ”Rural” or ”Urban”.
12. Avg glucose level: The average blood glucose level.

3.2. Machine Learning Classifier and Evaluation Matrices

This section describes the machine learning (ML) classifiers employed to develop stroke prediction models. The selected classifiers include Voting, Stacking, Boosting, Naïve Bayes, Random Forest, K-Nearest Neighbors (KNN), Logistic Regression, and Decision Tree. These models were chosen due to their widespread use and proven effectiveness in vulnerability and disease prediction research, as supported by previous studies [22], [23].To ensure reliable performance evaluation, we used stratified 10-fold cross-validation, maintaining the class distribution in each fold. All models were trained and optimized using GridSearchCV from scikit-learn version 1.5.2, which allows systematic tuning of hyperparameters. The specific hyperparameters for each classifier are reported in detail. In addition to standard performance metrics, confusion matrices were analyzed to gain insights into each model's classification behavior, particularly for imbalanced class distributions. These classifiers served as the foundation for developing accurate and interpretable stroke prediction systems.

3.3. Dataset Preprocessing

3.3.1. Feature Encoding: Feature encoding is the process of transforming categorical variables into a format suitable for machine learning algorithms. In this study, the values in the 'gender' column were encoded by converting the categorical strings—'Male', 'Female', and 'Other'—into corresponding numerical values: 0, 1, and 2, respectively.

3.3.2. Dataset Balance: The dataset contains 5,110 entries, with 4,861 representing non-stroke cases and 249 representing stroke cases, highlighting a notable class imbalance. To tackle this problem, we employed three widely used oversampling methods: SMOTEENN, SMOTE, and ADASYN. These techniques were applied only to the training data to avoid data leakage and maintain an unbiased evaluation. Of the methods tested, SMOTEENN demonstrated the best results in our experiments.

3.3.3. Feature Scaling: Feature scaling standardizes the range of data values, facilitating a clearer understanding of data distribution and relationships, which ultimately enhances the accuracy and reliability of machine learning models. In this study, we used Standard scaling to normalize the features in the dataset.

3.4. Applying Models

3.4.1. Random forest classifier: This highly powerful ensemble learning technique [24] is sensitive to a variety of hyperparameters that can greatly affect its performance. To maximize its effectiveness, we focused on tuning the following hyperparameters:

- n estimators (200): Total number of estimators.
- max depth (20): The greatest depth of forest trees.
- Criteria (gini): The divided quality evaluation function.
  Although the model achieved an accuracy of 81%, its overall performance showed room for improvement. Optimizing a predictive machine learning model for stroke necessitates careful hyperparameter tuning to enhance accuracy and reliability.

The confusion matrices for the Random Forest Classifier are shown in Fig. 1.

3.4.2. LogisticRegressionClassifier: Logistic Regression is a widely used method for binary classification tasks, including stroke prediction. Its performance depends on several hyperparameters that control the model's behavior. Below, we review the hyperparameters we tuned in our study:

- C (10): This parameter establishes the potential effectiveness of the punishment.
- Penalty (L2): 'L1' denotes L1 regularisation and 'L2' denotes L2. regularisation are the regularisation terms that are employed.
  The accuracy of the model was 72%.

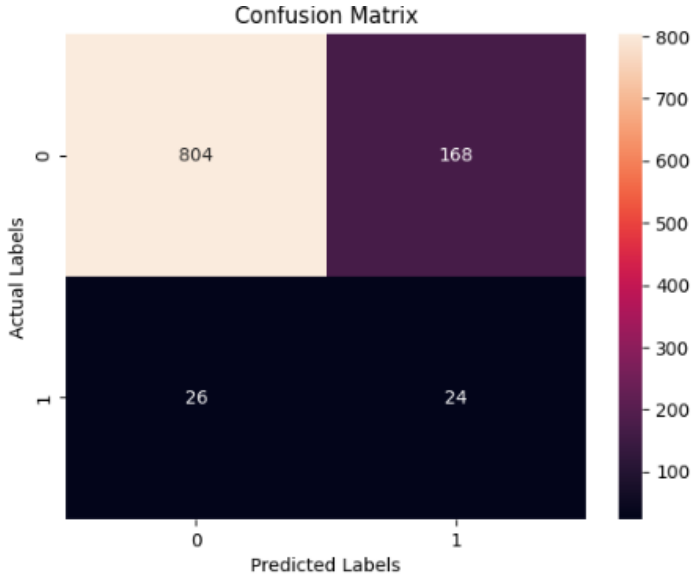

Figure 1. Confusion Matrix For RandomForestClassifier

The confusion matrices for the Logistic Regression classifier are presented in Fig. 2.

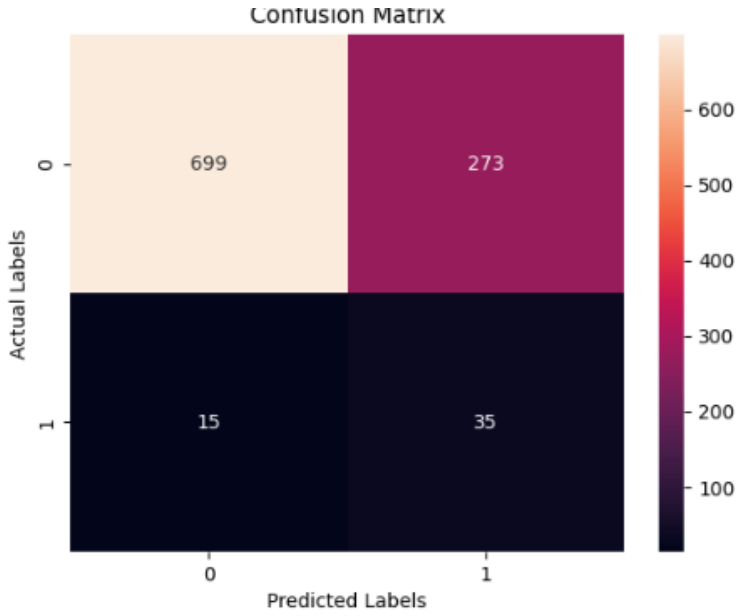

Figure 2. Confusion Matrix For LogisticRegressionClassifier

3.4.3. Gradient Boost: A Gradient Boosting Classifier [25] is a machine learning technique for classification tasks that is part of the ensemble learning method family. It creates a strong predictive model by sequentially integrating numerous "weak learners," which are often decision trees. It operates as follows:

- Train a weak model, such as a shallow decision tree, on the data.
- Calculate the model's errors (differences between predicted and actual values).

-A new, weak model is trained to anticipate these errors.

- This new model is added to the ensemble, and the process continues, with each new model attempting to remedy the mistakes of the previous ensemble.

- The final forecast combines the projections of all the weak learners.

Below, we discuss the hyperparameters we considered:

- n estimators (150): The highest count at which boosting is terminated.
- learning rate (0.2): At each boosting step, weight is added to each classifier. Every classifier contributes more when the learning rate is higher.
- max depth (5): The maximum depth of the individual regression estimators. The maximum depth restricts the amount of nodes in the tree.

In our study, the Gradient Boost model achieved an accuracy of 83%. The confusion matrices for the Gradient Boost classifier are shown in Fig. 3.

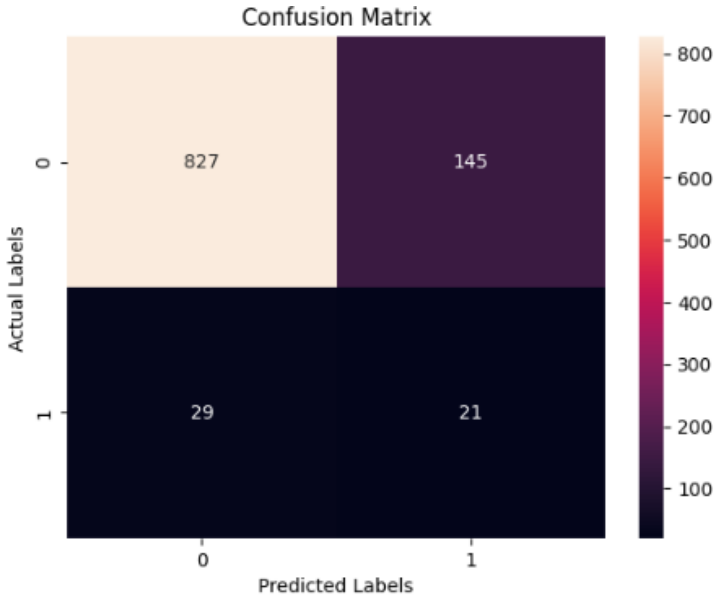

Figure 3. Confusion Matrix Of Gradient Boost

3.5. Proposed Model

Voting is an ensemble method that determines the class label of an input by combining the predictions of several base models, often called K base models [26]. Voting can be categorized into two primary types: hard voting and soft voting.

In hard voting, each base model predicts a class label, and the class receiving the majority of votes is chosen as the final prediction. In soft voting, each base model outputs a probability distribution over the classes, and the class with the highest average probability across all base models is selected. In our study, we employed both hard and soft voting methods, with soft voting demonstrating superior accuracy.

Voting combines the predictions of ensemble members in an effective manner, while stacked generalization (stacking) aggregates outputs from base estimators (Logistic Regression, Decision Tree, Random Forest, Gradient Boosting) and uses a meta-classifier (Logistic Regression) to produce the final prediction. To address class imbalance, we applied three different balancing techniques exclusively on the training data. After hyperparameter tuning, different accuracy levels were achieved for each balanced dataset. This approach ensures that the model can generalize well and make accurate predictions on unseen external data. For hyperparameter optimization

of classifiers such as Random Forest, Decision Tree, Logistic Regression and Gradient Boost, we employed GridSearchCV with stratified 10-fold cross-validation, which provides unbiased performance estimates.

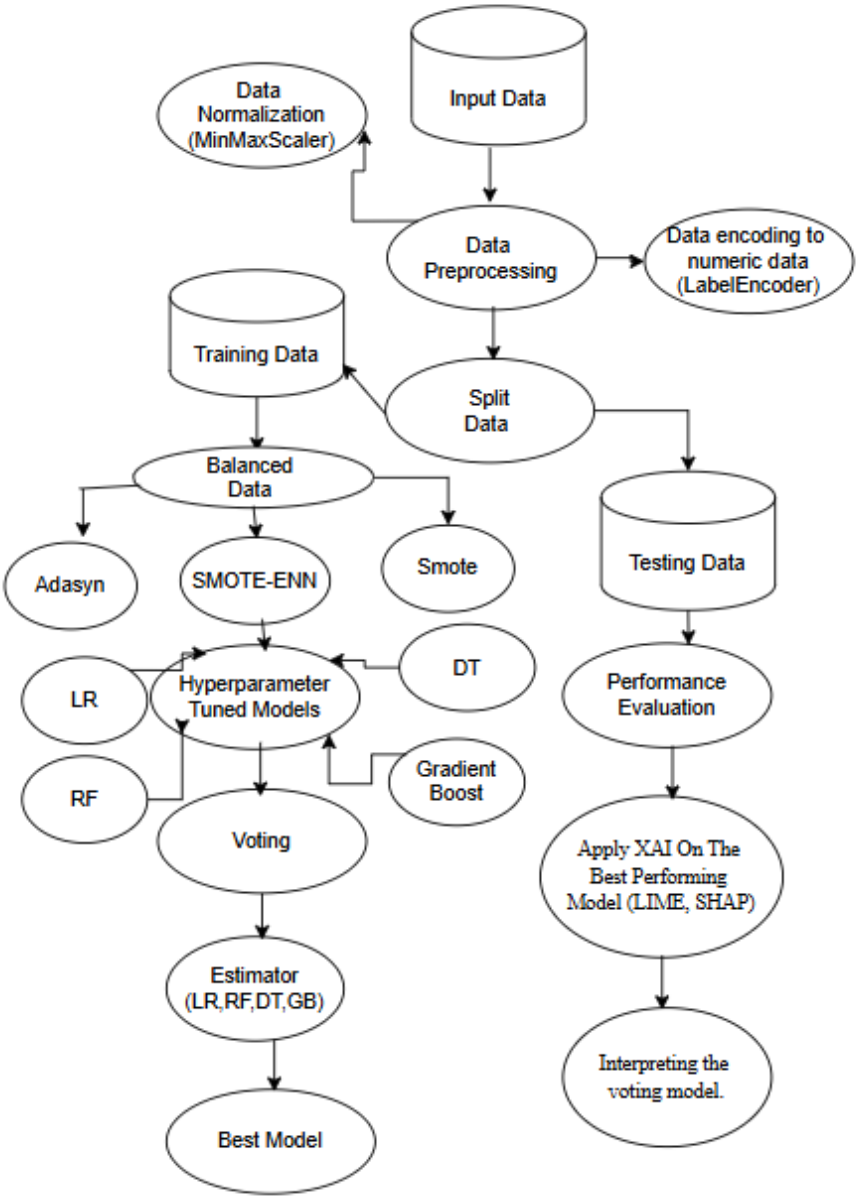

Figure 4. The Proposed Model

The datasets used for training and testing the Voting Classifier, along with the data processing steps, are illustrated in Fig. 4. This figure demonstrates how the preprocessed dataset—which includes encoding, normalization, and balancing—is employed to train the proposed voting model using the training set. The test data is subsequently used to evaluate the model's performance. The base estimators consist of Logistic Regression, Random Forest, Decision Tree and Gradient Boost, all of which were optimized through hyperparameter tuning. The entire implementation pipeline, developed using Python and scikit-learn libraries, is summarized in Fig. 4.

- Input data: We utilize an openly accessible Kaggle stroke dataset containing 5,110 records and 12 features. The binary target variable distinguishes stroke cases (1) from non-stroke cases (0).
- Formulation of the problem: Stroke prediction is framed as a binary classification task. The dataset exhibits significant class imbalance, with only 249 positive (stroke) instances. To ensure generalizability and prevent data leakage, resampling techniques are applied exclusively to the training set after data splitting.
- Preprocessing of the data: Initial steps include identifying and handling duplicate and missing values. Missing values, notably in the BMI attribute, are imputed using the mean of available data. Redundant columns are removed, and label encoding is applied to categorical variables. All features are converted into numerical form. Subsequently, the Standard Scaler is used to normalize the dataset. For interpretability, we employ the SHAP (Shapley Additive Explanations) method [27], which assigns importance scores to each feature based on Shapley values from cooperative game theory.
- Split Data: The data is split into training and testing subsets to enable the processes of model training and performance evaluation.
- Estimator: Four classifiers—RandomForestClassifier, LogisticRegressionClassifier, DecisionTreeClassifier and GradientBoost—are trained and tested. Hyperparameter tuning is performed for each algorithm, with predefined ranges for each hyperparameter.
- Voting: After training individual classifiers, a voting classifier combines their predictions to enhance overall accuracy. Each model's performance is assessed through confusion matrices, ROC curves, and Precision-Recall (PR) curves.
- Best Model: The voting classifier aggregates the strengths of the four base algorithms to determine the most accurate predictive model.
- Model Interpretation: Two explainable AI (XAI) techniques were applied to interpret the model's predictions:
  -SHAP [28] assigns Shapley values to features, representing their average marginal contribution across all possible feature subsets. This provides both global and local interpretability of model predictions.
  - LIME [29] generates interpretable local surrogate models by perturbing input data and observing prediction changes. It effectively explains predictions on tabular, text, and image data by identifying key contributing features.

Although voting ensembles are a well-established technique, our study offers added value through:
- Domain-specific design choices tailored for stroke prediction.
- A rigorous evaluation pipeline that avoids common pitfalls such as data leakage.
- A reproducible workflow applicable to similar healthcare prediction problems.

4. Outcome and Discussion

4.1. Results of Correlation

Figure 5 depicts the correlation between stroke occurrence and several other variables. While the graph suggests that no single attribute solely determines stroke risk, factors such as age, heart disease, hypertension, and average blood sugar level appear to play a notably influential role.

4.2. Evaluating Classifiers

A confusion matrix is a commonly used tool for assessing the effectiveness of a classification model on a test dataset [30]. It provides a summary of the model's correct and incorrect predictions, clearly distinguishing between different prediction outcomes. Table 1 outlines the layout of a confusion matrix, which consists of four prediction categories: True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN).

These values serve as the basis for calculating key performance metrics, including accuracy, precision, recall, and the F-measure (F-score):

- According to accuracy, the classification method operates like this :

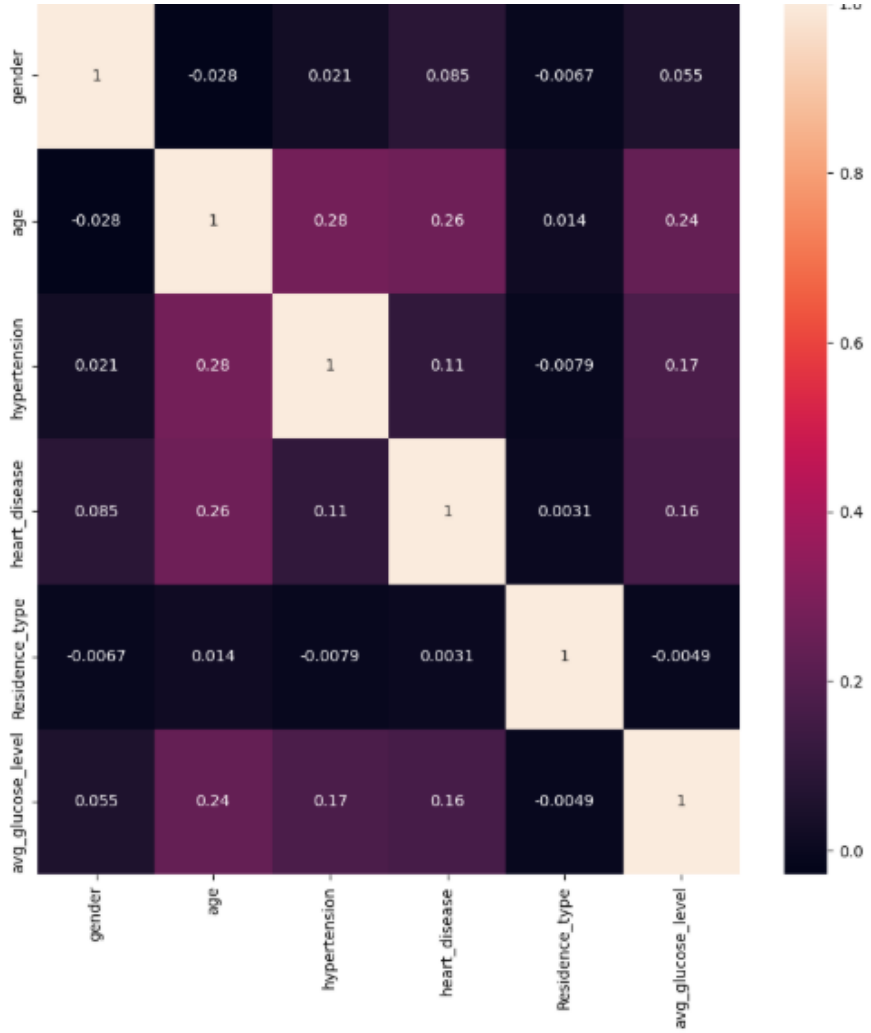

Figure 5. Correlation Matrices Among the Lifestyle Status and Disease.

Table 1. Confusion Matrices

	PredictedClass(0)	PredictedClass(1)
ActualClass(0)	TN	FP
ActualClass(1)	FN	TP

$$\left(\frac{TP+TN}{\text{TP+TN+FP+FN}}\right) \tag{1}$$

- Precision quantifies how many of the instances predicted as positive are actually correct. It is determined using the following formula:

$$\left(\frac{TP}{\text{TP+FP}}\right) \tag{2}$$

- The following recall equation is given:

$$\left(\frac{TP}{\text{TP+FN}}\right) \tag{3}$$

- The F-measure, also known as the F1-score, is the harmonic mean of precision and recall, offering a unified metric that balances the two. Since it emphasizes both metrics equally, the F1-score typically leans toward the lower of the two values—precision or recall. It is computed using the following formula:

$$\left(\frac{2 * precision * recall}{\text{precision+recall}}\right) \tag{4}$$

4.3. Performance Analysis

In the results section, we examine the test dataset used to evaluate the performance and classification accuracy of the machine learning models. From a total of 5,110 data points, 1,556 were allocated for testing.

Tables 2, 3, and 4 summarize the performance of the classifiers employed in this study for stroke prediction using a highly imbalanced test dataset. It is crucial to highlight that the test data remained unchanged to prevent data leakage and ensure an unbiased evaluation, preserving the integrity of the model assessment. To handle class imbalance, we applied balancing techniques—SMOTE-ENN, SMOTE, and ADASYN—solely to the training set.

Table 2. Assessment outcome for test data, where training data is balanced by applying SMOTE-ENN

CN	Acc.	Cls.Lable	Precision	Recall	F1
LR	0.72	1	0.11	0.70	0.20
DT	0.77	1	0.12	0.60	0.20
KNN	0.77	1	0.11	0.52	0.18
RF	0.81	1	0.12	0.48	0.20
NB	0.72	1	0.12	0.74	0.20
AdaBoost	0.72	1	0.11	0.70	0.20
GraBoost	0.83	1	0.13	0.42	0.19
Voting	0.80	1	0.13	0.58	0.22
Stacking	0.82	1	0.12	0.42	0.19

Table 3. Assessment outcome for test data, where training data is balanced by applying SMOTE

CN	Acc.	Cls.Lable	Precision	Recall	F1
LR	0.76	1	0.14	0.63	0.23
DT	0.80	1	0.12	0.40	0.19
KNN	0.82	1	0.09	0.22	0.12
RF	0.85	1	0.07	0.13	0.09
NB	0.73	1	0.14	0.67	0.22
AdaBoost	0.74	1	0.14	0.67	0.23
GraBoost	0.91	1	0.22	0.18	0.20
Voting	0.82	1	0.12	0.33	0.18
Stacking	0.86	1	0.08	0.13	0.10

Figure 6 presents a SHAP beeswarm plot, which is functionally equivalent to the SHAP summary plot. This visualization displays the distribution of SHAP values for each feature across all data samples, offering insight into both the importance of each feature and the direction of its influence on the model's predictions. We use SHAP to

Table 4. Assessment outcome for test data, where training data is balanced by applying ADASYN

CN	Acc.	Cls.Lable	Precision	Recall	F1
LR	0.75	1	0.14	0.63	0.23
DT	0.79	1	0.12	0.42	0.19
KNN	0.80	1	0.08	0.23	0.12
RF	0.84	1	0.06	0.12	0.08
NB	0.68	1	0.15	0.90	0.25
AdaBoost	0.73	1	0.13	0.67	0.22
GraBoost	0.91	1	0.22	0.18	0.20
Voting	0.82	1	0.13	0.37	0.19
Stacking	0.83	1	0.08	0.17	0.10

interpret how individual features impact the predictions of the VotingClassifier model trained on the dataset, with applicability to any classifier. Key insights from the plot include:

- Identification of the most influential features.
- Understanding whether high or low values of a feature increase or decrease the predicted outcome.
- The consistency with which each feature impacts the prediction.

Age and average blood sugar are two of the most significant predictors of stroke, according to the SHAP summary and interaction graphs. Positive SHAP contributions are strongly correlated with higher age values (highlighted in red), suggesting a higher risk of stroke. Aging is a significant risk factor for cerebrovascular illnesses because of vascular degradation and hypertension, which is in line with established medical facts. Despite having a smaller SHAP magnitude, gender seems to have a significant influence. Gender-specific risk modification is implied by the age-gender interaction plot, which suggests that older males may have somewhat greater SHAP levels. This is somewhat consistent with clinical research that indicates gender differences in stroke risk patterns, frequently as a result of lifestyle and hormonal factors.

Figure 7 shows a SHAP waterfall plot, which illustrates how each feature contributes to a specific prediction. The waterfall plot resembles a detailed breakdown, indicating how much each feature either adds to or subtracts from the final prediction for an individual data point—in this case, row 77. The plot begins with a baseline value and then visualizes:

- Positive SHAP values, representing features that increase the predicted likelihood.
- Negative SHAP values, representing features that decrease the predicted likelihood.
- The final bar reflects the model's predicted probability (or score) for that specific instance.

Age (-0.32) is the most important factor in lowering the risk of stroke, while avg. glucose level (+0.0) greatly raises it, according to the SHAP waterfall plot for a single case. Patients with cardiovascular comorbidities are more likely to have a stroke; other factors, such as hypertension, heart disease, and type of habitation, show lower impacts but are nonetheless consistent with known medical hazards.

Figure 8 displays the results of LIME (Local Interpretable Model-agnostic Explanations) for a single stroke prediction. LIME is a model-agnostic approach that explains individual predictions by approximating the complex model locally with a simpler, interpretable surrogate model. This method reveals which features were most influential in making a particular prediction by analyzing the model's behavior around that data point.

Stroke forecasts are significantly positively impacted by characteristics including heart disease, hypertension, and average blood sugar levels. The local interpretability of LIME shows how certain thresholds (such as avg. glucose level ¿ 120 mg/dL) have a substantial impact on categorization results. When taken as a whole, different XAI approaches show that while the model's predictive performance is mediocre, its decision-making is clinically rational.

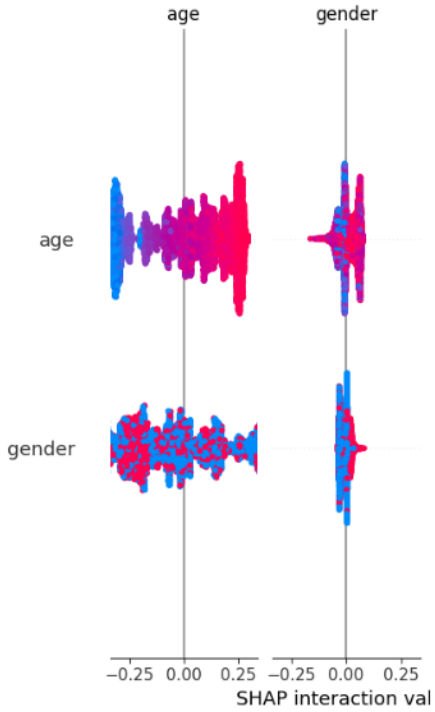

Figure 6. SHAP Beeswarm.

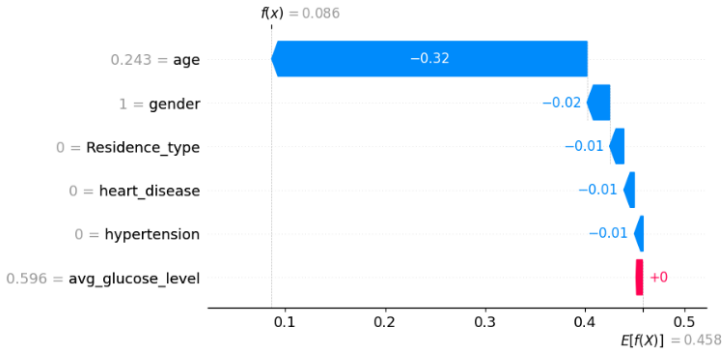

Figure 7. SHAP Waterfall.

Figure 8. LIME on a Single Instance.

4.4. The ROC Curve and The Precision-Recall Curve

In this study, the Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC) metric were used to evaluate the performance of various machine learning classifiers for binary classification tasks. The ROC curve illustrates the trade-off between sensitivity (true positive rate) and specificity (1 – false positive rate) across different decision thresholds, providing a comprehensive measure of classifier discrimination ability. The AUC quantifies the overall performance, with higher values indicating better discriminatory power.

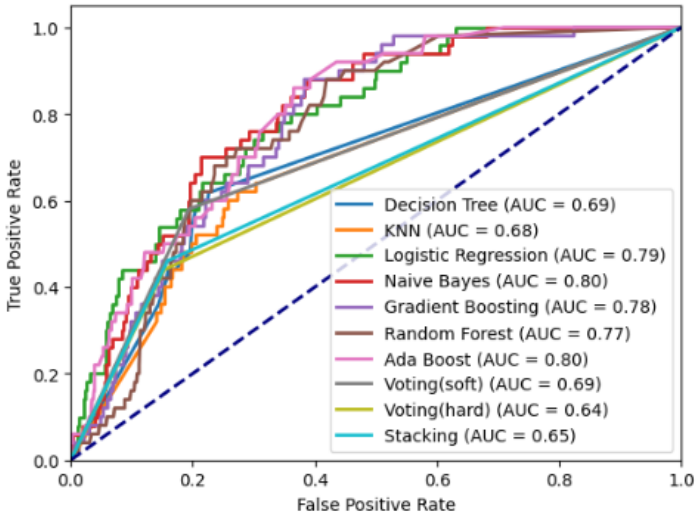

Figure 9. ROC Curve(SMOTE-ENN On Training Data) That Displays Each Classification Model's Performance.

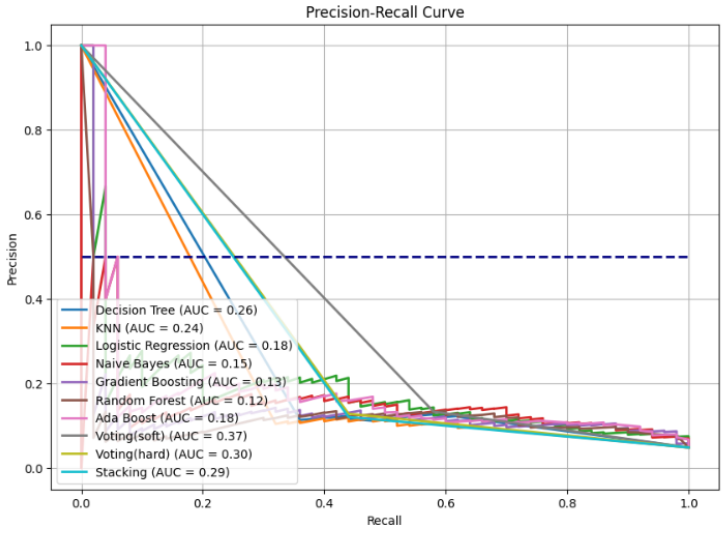

Figure 10. PR Curve(SMOTE-ENN On Training Data) That Displays Each Classification Model's Performance.

Precision is defined as the ratio of correctly predicted positive instances (true positives) to all instances predicted as positive (true positives + false positives). Recall (also known as sensitivity) measures the proportion of actual positive instances that are correctly identified (true positives divided by true positives + false negatives). A Precision-Recall (PR) curve plots precision values against recall values and is particularly informative for

imbalanced datasets. The model detects some real stroke episodes but also produces a large number of false positives, as seen by the low PR-AUC of the Precision-Recall (PR) curve. This is important in the healthcare industry since many ”at-risk” forecasts might not be accurate, leading to unnecessary testing or concern. Excellent recall, however, is more important than precision in clinical settings because it is more riskier to miss a real stroke case. Consequently, recall should take precedence above precision in the model. A better balance between these two metrics can be achieved by using cost-sensitive training or modifying thresholds.

By applying three different oversampling techniques—SMOTE-ENN, SMOTE, and ADASYN—to balance the training data, the SMOTE-ENN yielded the best results. Figures 9 and 10 display the ROC and Precision-Recall curves, respectively, illustrating the performance of each classification model under these conditions.

4.5. Discussion

The datasets used in this study are highly imbalanced. While balancing both training and testing data can lead to near-perfect performance, this approach is unrealistic in real-world applications and introduces the risk of data leakage. By balancing only the training data, our model avoids this issue, resulting in a more practical and reliable evaluation—though performance on the test data may appear lower. Despite this, our proposed voting model demonstrates superior performance compared to existing methods when assessed across multiple evaluation metrics.

SMOTE, SMOTE-ENN, ADASYN resampling were used within the modeling pipeline to solve class imbalance; however, it did not result in a significant improvement in performance. One key problem is that the dataset has a limited number of features. When only a few predictors are known (e.g., age, average glucose level, hypertension, heart disease), the feature space becomes highly compact, and minority-class samples (stroke cases) are located near majority-class boundaries. This finding demonstrates that resampling is insufficient when the dataset has low feature dimensionality and overlapping classes. Improving model performance may necessitate adding more clinically relevant information (e.g. blood pressure readings, cholesterol levels, lifestyle markers) or implementing cost-sensitive algorithms that prioritize recall for stroke cases.

Looking ahead, we plan to extend this research by incorporating various types of medical image data, such as ECG signals, brain MRI, and CT scans. Combining clinical variables with imaging data will enable more robust model training and improve prediction accuracy. Ultimately, this model has the potential to serve as an effective preliminary tool for stroke prediction based on clinical factors.

5. Conclusion

In this study, we employed four hyperparameter-tuned classifiers combined through voting to predict stroke occurrence. The dataset was balanced exclusively on the training data to ensure that the model makes accurate predictions on unseen external data. Our evaluation focused primarily on the ROC and Precision-Recall (PR) curves. The weighted voting ensemble outperformed other commonly used machine learning algorithms, achieving an ROC AUC of 69% and a PR AUC of 37%. These results suggest that voting ensembles are a promising approach for stroke prediction. Additionally, by integrating explainability methods—LIME and SHAP—we enhanced the transparency of the model’s decision-making process. We also analyzed the relationships between various medical conditions and stroke risk, highlighting that early management of these diseases can help reduce stress and improve health outcomes.

In future work, we plan to conduct ablation studies to evaluate the contribution of each component model’s to the ensemble and refine performance through precision–recall trade-off optimization, guided by clinical consultation to prioritize early stroke detection. We will integrate explainable AI (XAI) insights into feature engineering and collaborate with clinicians to validate and interpret model explanations, strengthening interpretability and trust. To assess real-world applicability, the model will be deployed in a pilot clinical workflow to observe clinician interaction and its influence on patient care, supported by actionable outputs such as “High SHAP risk score triggers neurology referral.” We will also audit demographic biases across age, gender, and ethnicity, applying fairness-enhancing techniques such as re-sampling or model recalibration, and establish continuous monitoring to

detect drift over time. Furthermore, confidence intervals and statistical significance tests (e.g., McNemar's or paired bootstrap) will be used to improve result reliability, while enhanced feature engineering—such as incorporating temporal glucose trends and BMI categories—will capture complex clinical patterns. Collectively, these initiatives aim to enhance the model's clinical credibility, robustness, and transparency for real-world deployment.

Declaration of Competing Interest

The authors declare that they have no competing interests.

Data Availability

Information will be provided upon request.

REFERENCES

1. M. J. Ferdous and R. Shahriyar, "A comparative analysis for stroke risk prediction using machine learning algorithms and convolutional neural network model," pp. 1–6, 2023.
2. C.-Y. H. S.-F. Sung, "Developing a stroke severity index based onad ministrative data was feasible using data mining techniques," *Journalof Clinical Epidemiology, vol. 68, no. 11, pp. 1292–1300*, Nov. 2015.
3. Y.-C. L. C.-A. Cheng and H.-W. Chiu, "Prediction of the prognosis ofischemic stroke patients after intravenous thrombolysis using artifi cial neural networks," *Studies in Health Technology and Informatics, vol.202, pp. 115–118*, 2014.
4. Y. P. Nataliia Melnykova, "Machine learning for stroke prediction using imbalanced data," *Scientific Reports*, 2025.
5. G. e. a. Kayola, "Stroke rehabilitation in low- and middle-income countries," *American Journal of Physical Medicine Rehabilitation*, February, 2023.
6. M. S. Singh and P. Choudhary, "Stroke prediction using artificialin telligence," *in 8th Annual Industrial Automation and Electrome chanical Engineering Conference (IEMECON)*, 2017.
7. M. T. F. S. A. M. F. J. F. K. R. N. L. Amini, R. Azarpazhouh and N. Toghianfar, "Prediction and control of stroke by data mining," *International Journal of Preventive Medicine,vol. 4, no. Suppl 2, pp. S245–249*, May 2013.
8. A. Y. Adam, S.Y. and M. Bashir, "Classification of ischemic stroke using machine learning algorithms," *Int J Comput Appl, 149(10): p. 26-31*, 2016.
9. S. K. R.S. Jeena, "Stroke prediction using svm," *in: Proc. InternationalConference on Control, Instrumentation, Communication and ComputationalTechnologies (ICCICCT), pp. 600–602*, 2016.
10. P. C. M.S. Singh, "Stroke prediction using artificial intelligence."
11. P. G. A. Sudha and N. Jaisankar, "Effective analysis and predictive model of stroke disease using classification methods," *International Journal of Computer Applications, vol. 43, no. 14, pp. 26–31*, 2012.
12. S. K. T. Kansadub, S. Thammaboosadee and C. Jalayondeja, "Classification of brain cancer using artificial neural network," *presented at the 2015 Biomedical Engineering International Conference*, 2015.
13. N. N. D. M. S. B. T. Tazin, M. N. Alam, "Stroke disease detection and prediction using robust learning approaches," *J. Healthc. Eng., vol. 2021, doi: 10.1155/2021/7633381*, 2021.
14. P. D. S. B. J. D. Nwosu, Bhardwaj, "to predict stroke risk, using electronic health records," *Biology Society's 41st Annual International Conference*.
15. F. S. Alotaibi, "Predicting heart failure using a machine learning model in practice," *issue of International Journal of Advanced Computer Science and Applications (IJACSA)*, 2019.
16. C. G. C. E. Verbakel JY, Steyerberg EW, "For clinical prediction models, a comprehensive review demonstrates that machine learning does not outperform logistic regression in terms of performance," *J Clin Epidemiol 110:12–22*, 2019.
17. P. T. T. H. Le, N. B., "Ai-powered predictive model for stroke and diabetes diagnostic," 8 Feb. 2024, https://doi.org/10.5815/ijisa.2024.01.03 https://doi.org/10.5815/ijisa.2024.01.03.
18. M. S. Alam, M. Ferdous, N. S. Neera *et al.*, "Enhancing diabetes prediction: An improved boosting algorithm for diabetes prediction." *International Journal of Advanced Computer Science & Applications*, vol. 15, no. 5, 2024.
19. M. J. Ferdous, "A more effective ensemble ml method for detecting breast cancer," pp. 171–184, 2017.
20. M. J. Ferdous and R. Shahriyar, "An ensemble convolutional neural network model for brain stroke prediction using brain computed tomography images," *Healthcare Analytics*, vol. 6, p. 100368, 2024.
21. A. T. Clerk Maxwell, *on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73*, stroke Prediction Dataset: https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset.
22. S. H. H.-B. L. H. Lee, E.-J. Lee, "Machine learning approach to identify strokewithin 4.5 hours," *Stroke, vol. 51, no. 3, pp. 860–866(pubMed)*, 2020.
23. S. K. T. Kansadub, S. Thammaboosadee, "Stroke risk prediction model based on demographic data," *8thBiomedical Engineering International Conference (BMEiCON). IEEE,2015, pp. 1–3*, 2015.
24. Random forest classification from scikit-learn. Available: https://scikit-learn.org/dev/modules/generated/sklearn.ensemble.RandomForestClassifier.html.
25. Gradient Boost classification from scikit-learn. Available: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html.

26. D. Dogan, A.; Birant, "A weighted majority voting ensemble approach to classification," *Fourth International Conference on Computer Science and Engineering (UBMK), doi:10.1109/ubmk.2019.8907028*, 2019.
27. XAI: https://www.datacamp.com/tutorial/explainable-ai-understanding-and-trusting-machine-learning-models.
28. S. Lundberg, "A unified approach to interpreting model predictions," *arXiv preprint arXiv:1705.07874*, 2017.
29. C. M. T. Ribeiro, S.S.; Guestrin, "What makes me believe you? describing the predictions of any classifier." *in the proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016.
30. S. P. Sirage Zeynu, "Prediction of chronic kidney disease using data mining feature selection and ensemble method," *Medicine, Computer Science, WSEAS Transactions on Information Science and Applications archive*, 2018.