Published online on the page : https://ejournal.uinbukittinggi.ac.id/index.php/ijokid

Knowbase : International Journal of Knowledge in Database

| ISSN (Print) 2798-0758 | ISSN (Online) 2797-7501 |

Enhancing Stroke Diagnosis with Machine Learning and SHAP-Based Explainable AI Models

Galih Hendro Martono[1], Neny Sulistianingsih[2]*

[1,2]Universitas Bumigora, Mataram, Indonesia

Article Information

Article History:

Submitted: November 7, 2024
Revision: December 9, 2024
Accepted: December 29, 2024
Published: December 31, 2024

Keywords

Algorithm Insight,
Explainable AI,
Machine Learning,
SHAP Algorithm,
Stroke Detection

Correspondence

E-mail:
galih.hendro@universitasbumigora.ac.id*

A B S T R A C T

Stroke is a serious illness that needs to be treated quickly to enhance patient outcome. Machine Learning (ML) offers promising potential for automated stroke detection through precise neuroimaging analysis. Although existing research has explored ML applications in stroke medicine, challenges remain, such as validation concerns and limitations within available datasets. The study aims to compare ML models and SHapley Additive exPlanations (SHAP) algorithm insights for stroke detection optimization. The research evaluates classifiers' performance, including Deep Neural Networks (DNN), AdaBoost, Support Vector Machines (SVM), and XGBoost, using data from www.kaggle.com. Results demonstrate XGBoost's superior performance across various data splits, emphasizing its effectiveness for stroke prediction. Utilizing SHAP provides deeper insights into stroke risk factors, facilitating comprehensive risk assessment. Overall, the study contributes to advancing stroke detection methodologies and highlights ML's role in enhancing clinical practice in stroke medicine. Further research could explore additional datasets and advanced ML algorithms to enhance prediction accuracy and preventive measures.

1. Introduction

Stroke is a severe medical illness characterized by an abrupt cessation of blood flow to the brain, potentially leading to significant neurological deficits and enduring disabilities [1]. Precise and efficient stroke diagnosis is crucial for commencing prompt therapy and enhancing patient outcomes. In recent years, machine learning (ML) methodologies have surfaced as essential instruments for stroke detection and diagnosis, providing the capability for automated and accurate interpretation of neuroimages [2]. ML algorithms have been increasingly utilized in healthcare systems, including stroke care, to enhance diagnostic accuracy and streamline patient management processes [3].

A significant body of literature exists on the application of ML in stroke medicine as in [2], [4]–[10], highlighting various approaches and methodologies employed for stroke identification, subtype classification, severity prediction, and outcome assessment. Studies have utilized diverse ML techniques, including natural language processing (NLP), feature selection, and ensemble classifiers, to automate tasks such as classifying ischemic stroke subtypes and predicting stroke severity [11]. Additionally, systematic reviews have been conducted to assess the effectiveness of ML-based patient classification systems and to explore advancements, challenges, and prospects in ML applications in stroke medicine [12].

Despite the advancements in ML-based stroke diagnosis and management, several gaps and limitations

persist in the existing literature. These include a lack of robust external validation and interpretability of ML models [13]. Moreover, many studies have focused on relatively small datasets, potentially limiting the generalizability of their findings to broader patient populations [10, 11]. Additionally, there is a need for further investigation into the scalability, real-world applicability, and potential biases inherent in ML algorithms used for stroke detection and prediction [16].

The study aims to compare machine learning models and SHAP (SHapley Additive exPlanations) algorithm insights for stroke detection optimization. By leveraging data from sources like Kaggle.com, which provides relevant variables such as sex, age, hypertension, heart disease, and smoking status, the research evaluates multiple classifiers' performance, including Deep Neural Networks (DNN) [17], AdaBoost [18], [19], Support Vector Machines (SVM) [20], Logistic Regression [21], XGBoost [22], Gradient Boost Tree [23], and Generalized Additive Models (GAM) [24]. The analysis showcases metrics such as precision, recall, F1 score, and accuracy across different training-testing data splits.

This comparison analysis aims to determine the best effective machine learning model for optimizing stroke detection. The research seeks to augment the interpretability and comprehension of model detection by the integration of SHAP algorithm insights, hence aiding clinical decision-making and enhancing patient outcomes in stroke therapy. Overall, this research contributes to advancing stroke detection methodologies and underscores the importance of machine learning in enhancing clinical practice in stroke medicine.

2. Method

The proposed research aims to address stroke issues by applying machine learning techniques, specifically classification methods. The research methodology encompasses several stages, including data preprocessing, model development, evaluation, and analysis of SHAP (SHapley Additive exPlanations)—fig 1 shows the research methodology used in this research.

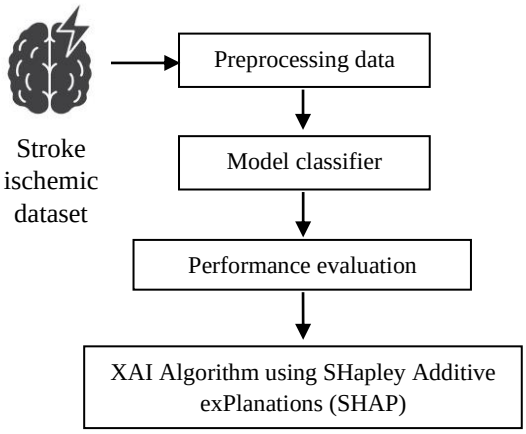

Figure. 1. Research methodology

2.1. Dataset

The dataset under consideration comprises 40,910 entries and encompasses various demographic and health-related attributes. It originates from www.kaggle.com [25], a popular datasets and data science competition platform. Each entry in the dataset represents an individual and includes the following attributes: sex, age, hypertension, heart disease, marital status, work type, residence type, average glucose level, BMI (Body Mass Index), smoking status, and stroke occurrence. A detail of table attributes is shown in Table 1.

Table 1. Dataset Attributes

Attribute	Description
Sex	Gender of the individual (1 for male, 0 for female)
Age	Age of the individual in years
Hypertension	Presence (1) or absence (0) of hypertension
Heart Disease	Presence (1) or absence (0) of heart disease
Ever Married	Marital status (1 for married, 0 for unmarried)
Work Type	Type of work the individual is engaged in
Residence Type	Type of residence (1 for urban, 0 for rural)
Average Glucose Level	The average glucose level in the individual's blood
BMI (Body Mass Index)	Body Mass Index (weight in kg/height in m^2)
Smoking Status	Smoking status of the individual
Stroke	Occurrence of stroke (1 for yes, 0 for no)

2.2. *Preprocessing Data*

Preprocessing data entails finding outliers, which are data points that deviate significantly from the rest of the dataset and may bias the results. The Interquartile Range (IQR) is a regularly used approach for detecting outliers. The IQR is calculated as the difference between the third quartile (Q3) and the first quartile (Q1), denoted as bellow [26].

$$IQR = Q3 - Q1$$

It represents the middle 50% of the data and provides a reliable measure of variability. Analysts can define an upper limit beyond which data points are considered outliers by splitting the dataset into quartiles and calculating the IQR.

The process of removing outliers involves several key steps. Firstly, box plots are generated to visually identify outliers across numerical features concerning the target label, 'stroke'. Subsequently, the upper limit for outliers is determined based on each numerical feature's interquartile range (IQR). The maximum threshold is 1.5 times the interquartile range above the third quartile (P75) [26]. Data points exceeding this upper limit are identified as outliers. Once identified, these outliers are replaced with the upper limit value to mitigate their influence on subsequent analyses. Finally, the updated dataset is visualized again using box plots to confirm the successful removal of outliers and ensure the integrity of the data for further analysis. This recurrent method guarantees the dataset's robustness and reliability, enabling more precise insights into the correlation between numerical features and the target variable, 'stroke'.

2.3. *Model Classifier*

This study utilizes a variety of classifier models. Conventional machine learning methodologies, including Logistic Regression, Support Vector Machine, AdaBoost, and Generalized Additive Model (GAM), were employed. Furthermore, Tree-Based Ensemble Methods such as XGBoost and Gradient Boost Trees, together with Deep Learning Neural Networks, were utilized in this study project. These methods collectively provide a comprehensive framework for analyzing and modelling complex datasets, allowing for a thorough exploration of patterns and relationships within the data. Each method brings unique

strengths and characteristics to the analysis, contributing to a holistic understanding of the studied phenomena.

2.3.1. Logistic Regression

Logistic regression is a supervised learning technique that analyses dependent values by employing independent variables from given datasets. It evaluates output values through logistic regression of the dependent variables, offering solutions in either absolute or differential forms, which may be numerical or binary (e.g., Yes or No, 0 or 1, true or false). While computer-based representations often adopt a binary format (0 or 1), logistic regression inherently represents feasible values between 0 and 1. The logistic function is denoted as below [4].

$$\emptyset(z) = \frac{1}{1 + e^{-z}}$$

Where, $z = b_0 + b_1 \times age + b_2 \times systolicBP + \cdots + b_9 \times cholesterol$.

2.3.2. Support Vector Machine

The Support Vector Machine (SVM) [27] is a supervised learning technique employed for classification and regression applications within machine learning theory. Support Vector Machine (SVM) determines a hyperplane with maximum margin using training data to differentiate data into discrete groups or classes [28]. This hyperplane is optimally positioned to maximize the distance between the nearest data points of any class. Support Vector Machines employ kernel functions to convert data from lower-dimensional spaces, which may not be linearly separable, into higher-dimensional spaces. Diverse kernels, such as polynomial, Gaussian radial, and exponential radial basis, are employed to calculate the scores of each subject in nonlinear contexts. Support Vector Machine (SVM) divides data into categories by utilizing an optimum hyperplane and following the principle of structural risk minimization.

2.3.3. AdaBoost

The AdaBoost algorithm is the most prevalent and often employed ensemble learning approach. The method termed "boosting" transforms each weak classifier into a singular, robust classifier. AdaBoost's core principle lies in its capacity to produce a weak learner from the initial training set and subsequently modify the training set's distribution for each new weak learner iteration based on anticipated performance. The subsequent phase will emphasize samples that had low prediction accuracy in preceding rounds. A robust learner is established by integrating weaker learners with varying weights [17, 18]. A particular method for training a boosted classifier is referred to as AdaBoost. A boosted classifier is represented as below.

$$F_T(x) = \sum_{t=1}^{T} f_t(x)$$

Given an object x as input, each f_t weak learner produces a value indicating the item's class. In a binary classification scenario, the predicted object class is determined by the sign of the weak learner's output, while the absolute value reflects the confidence in that classification. Similarly, if the sample is classified as belonging to a positive class, the t-th classifier is deemed positive; otherwise, it is considered negative.

2.3.4. Generalized Addictive Model (GAM)

For simulating intricate interactions between predictors and a response variable, Generalized Additive Models (GAMs) [31] offer a versatile and potent statistical framework. Because each predictor can have a

smooth, non-parametric effect on the response, GAMs, in contrast to typical linear models, can tolerate non-linear relationships. The exponential function in GAM is denoted as below [31].

$$g(E(Y)) = \beta_0 + f_1(x_1) + f_2(x_2) + \cdots + f_m(x_m)$$

The functions f_i may be characterized non-parametrically, semi-parametrically, or as "smooth functions" to be estimated by non-parametric methods. Alternatively, the functions may be expressed in a parametric form, such as a polynomial or an unpenalized regression spline of a variable.
Smoothing functions that represent the underlying structure of the data without imposing strict assumptions, like splines or kernel functions, are used to do this. Because GAMs show the smooth functions of individual predictors, analysts may more easily comprehend the nature of each predictor's influence on the response and provide results that are easy to interpret. GAMs also have advantages when managing missing values, high-dimensional data, and interactions between predictors. GAMs are especially helpful in fields because of these features.

2.3.5. XGBoost

By using a Newton-Raphson technique in function space instead of gradient descent, XGBoost functions differently from gradient boosting [22]. The Newton-Raphson method employs a second-order Taylor approximation of the loss function for connection. There might not be any revolutionary discoveries in mathematics, but XGBoost meticulously arranges gradient gain options to guarantee precision and efficiency. It combines tree-based and linear approaches, utilizing several AI algorithms to evaluate a new tree's dependability for improving model accuracy [7]. For example, it can learn from parallel learning (bagging) and random forest (impulsive). Furthermore, methods such as data collection are utilized to handle data gaps and expedite the precise processing of sophisticated AI models. These tactics are combined to tackle the issue of data gaps.

2.3.6. Gradient Boost Tree

An ensemble of decision trees is gradually built using the sophisticated machine learning technique known as gradient boosting tree (GBT), with each new tree fixing mistakes caused by the ones that came before it. GBT serially develops trees, with each new tree concentrating on capturing the residual mistakes left by the preceding trees, in contrast to standard decision tree algorithms that generate trees singly [31], [32]. This iterative process continues until the model converges or a certain number of trees are generated. The objective of gradient descent optimization, which entails iteratively adjusting model parameters towards the steepest descent to minimize a loss function, integrates the benefits of decision trees to formulate gradient boosting trees (GBT). Through iterative model refinement based on past iteration mistakes, GBT effectively learns complex relationships within the data and can handle both regression and classification tasks. This approach results in highly accurate predictive models that are robust to overfitting and capable of capturing intricate patterns in the data.

2.3.7. Deep Neural Network

Artificial neural networks utilized in machine learning are termed deep learning. Convolutional neural networks, deep belief networks, recurrent neural networks, and deep neural networks (DNN) exemplify several topologies in deep learning. Neural networks are a class of algorithms designed to identify patterns, loosely inspired by the structure of the human brain [7]. These are extensively employed throughout various fields of study, including computer vision, gaming, audio recognition, speech recognition, and natural language processing. A Deep Neural Network comprises an input layer, several hidden layers, and an output layer. The network is trained by backpropagation, which reduces the divergence between the expected and actual output. The parameters of the DNN utilized in the research are presented in Table 2.

Table 2. Parameter of DNN Used in the Research

Parameter	Value
Architecture	Sequential
Activation	ReLU
Input shape	(Number of features)
Dropout rate	0.5 (after each Dense layer)
Number of Dense layers	3 (including output layer)
Number of units per Dense layer	64, 32, 1 (output layer)
Output activation	Sigmoid
Optimizer	Adam
Loss function	Binary cross-entropy
Metrics	Accuracy
Number of epochs	10
Batch size	32
Validation split	0.1

2.3.8. Shapley Addictive exPlanations (SHAP)

A game-theoretic method known as SHapley Additive exPlanations (SHAP) aims to clarify the predictions generated by any machine learning model. SHAP connects optimal credit distribution with localized explanations by utilizing conventional Shapley values from game theory and their corresponding extensions. SHAP, created by [33], is a technique for elaborating on specific forecasts by utilizing the Shapley game-theoretically optimal values. In essence, SHAP is a framework for analyzing machine learning model results. Its central idea is firmly anchored in SHAP values and cooperative game theory. Unlike other approaches, SHAP gives us a thorough grasp of each feature's contributions to the forecasts, promoting equity and making. The utility of SHAP lies in its ability to elucidate the significance of each feature in shaping predictions. Offering SHAP [34]–[36] values enables a nuanced comprehension of intricate models and how input features influence predictions. SHAP details the justification as follows.

$$g(z') = \phi_0 + \sum_{j=1}^{M} \phi_j z^i{}_j$$

Where ϕ_i is the SHAP value for feature ii, indicating the contribution of that feature to the prediction, S is a subset or coalition of all features is a subset or coalition of all features N, excluding feature i, $|S|$ represents the number of features in subset S, $f(S \cup \{i\})$ is the model's prediction f when feature i is added to subset S.

2.3.9. Evaluation Metric

The study utilized diverse evaluation metrics to analyze the efficacy of various classifiers over many training and testing data partitions. The assessment metrics include Precision, Recall, F1 Score, and Accuracy. Precision measures the ratio of true positive predictions to the total positive predictions generated by the classifier. Precision is denoted as below [2], [15], [37]–[40].

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

True Positives (TP) denote the count of accurately predicted positive cases, whereas False Positives (FP) indicate the count of inaccurately predicted positive instances.

Recall, also known as Sensitivity or True Positive Rate, is the proportion of true positive predictions relative to the total number of actual positive instances in the dataset. Recall is calculated using the formula

[2], [15], [37]–[40].

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

A False Negative (FN) denotes the quantity of positive cases erroneously classified as negative.

The F1 Score represents the harmonic mean of Precision and Recall. It equilibrates Precision and Recall and is especially advantageous when addressing imbalanced datasets.—the F1 score is calculated as below [2], [15], [37]–[40].

$$F1score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Accuracy evaluates the comprehensive correctness of the classifier's predictions. Accuracy measures the ratio of accurately predicted cases (including true positives and true negatives) to the total instances in the dataset. Accuracy is calculated as below [2], [15], [37]–[40].

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

True Negatives (TN) denotes the quantity of accurately predicted negative situations.

3. Results and Discussion

3.1. Stroke Dataset

The examination of the dataset uncovers some significant relationships among its properties. A robust positive association of 0.26 exists between hypertension and heart disease, indicating that those with hypertension are more predisposed to heart disease. Secondly, there exists a robust positive correlation of 0.27 between average glucose levels and the occurrence of stroke, which implies that higher average glucose levels in the blood may increase the risk of stroke. On a different note, a weak negative correlation of -0.12 is identified between gender and age. The issues suggest a slight tendency for younger individuals to be male and older individuals to be female within the dataset. Lastly, a weak positive correlation of 0.053 is noted between work type and smoking status, which indicates that certain types of employment may slightly influence an individual's smoking habits, albeit the effect is not very strong. These correlations provide valuable insights into the interrelationships among various factors within the dataset, shedding light on potential patterns and associations worth further exploration. Figs. 2 and 3 show the correlation matrix between attributes and histograms for each attribute in the dataset used.

3.2. Preprocessing Data

After conducting preprocessing to identify outliers in the dataset, several key observations emerged. Firstly, no outlier records were detected in attributes like age, hypertension, heart disease, ever_married, work_type, residence_type, and average glucose level. This absence of outliers suggests a consistent and reliable distribution of data across these attributes, enhancing the dataset's suitability for various analyses related to cardiovascular health, demographics, and metabolic conditions. However, the attribute BMI revealed many outlier records, totaling 915 instances. These outliers signify potential irregularities or extreme values in BMI measurements within the dataset. It is essential to address these outliers to maintain the accuracy and integrity of analyses concerning obesity, metabolic health, and other health outcomes linked to BMI. Figures 4 and 5 show a visualization of the outlier identified after handling preprocessing data.

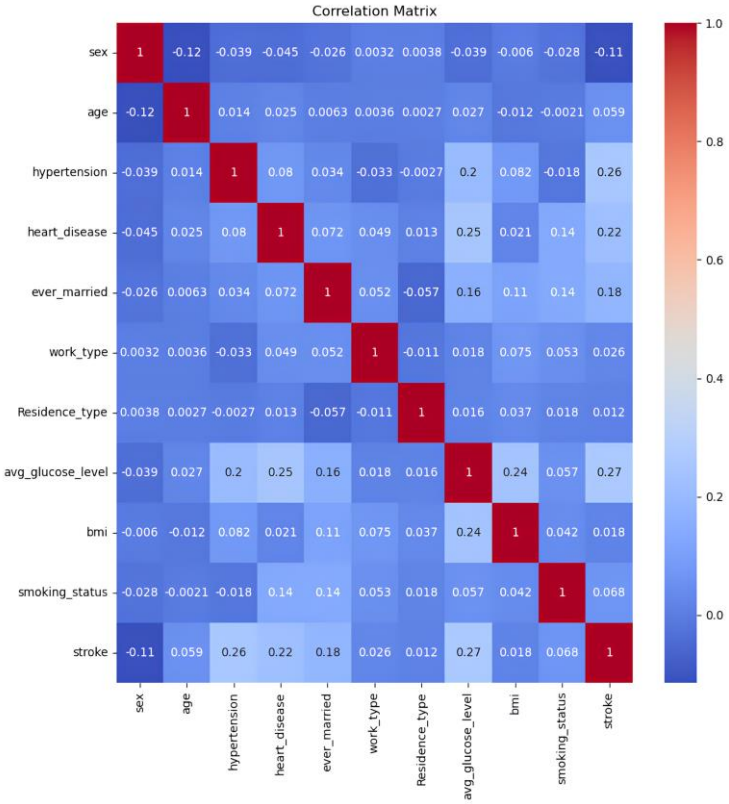

Figure 2. Correlation matrix between attributes in the dataset used

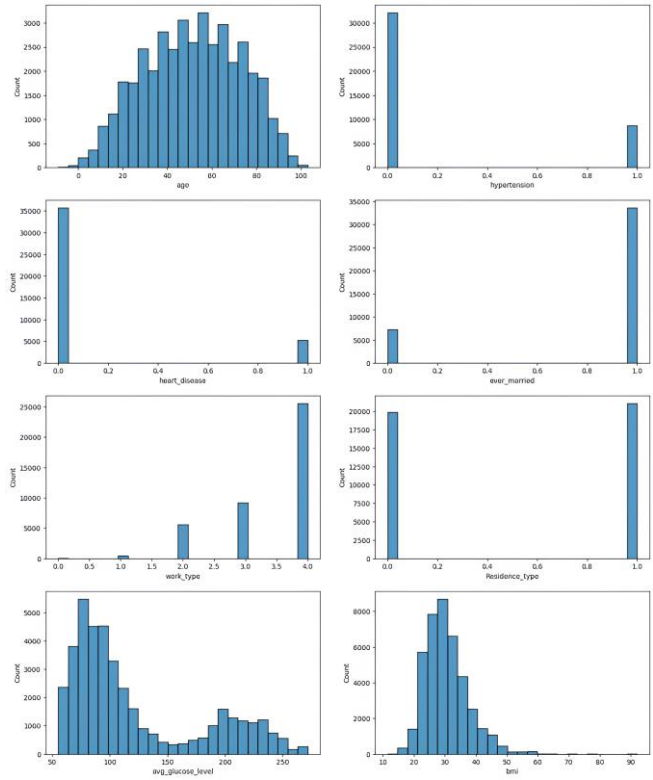

Figure 3. Histogram of each attribute in the dataset used

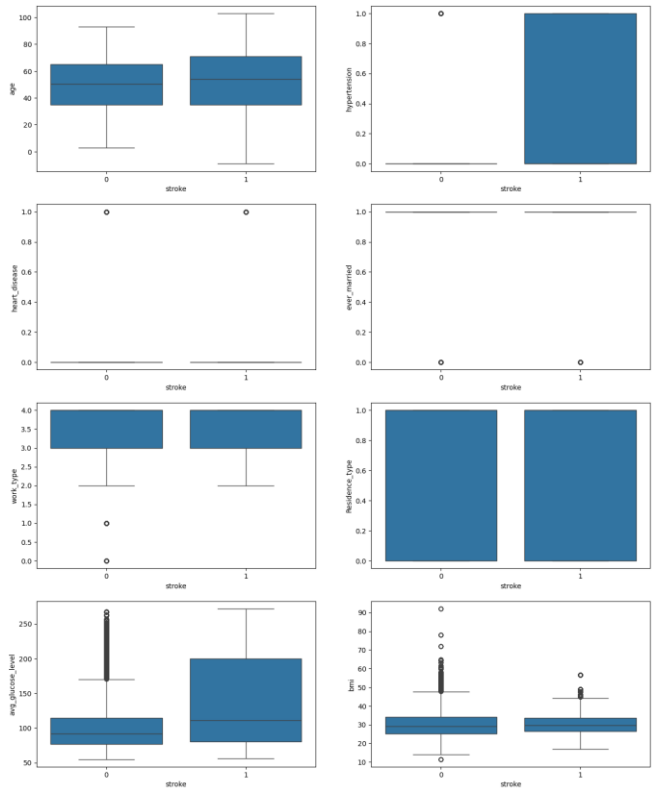

Figure 4. Visualization of outlier identified in preprocessing data

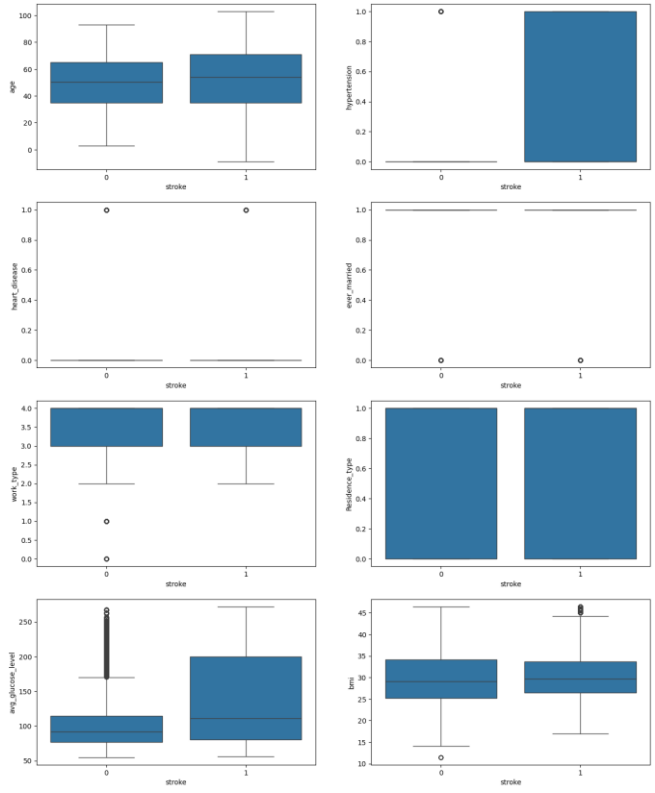

Figure 5. Visualization of outlier after handling in preprocessing data

3.3. Model Classifier using various Machine Learning

Table 3 outlines the results of a study conducted on different classifiers utilizing varying training-to-testing data ratios. Each classifier's performance is evaluated based on precision, recall, F1 score, and accuracy metrics. For instance, in the first scenario with a 90:10 training-to-testing data ratio, diverse classifiers were employed, including Deep Neural Network (DNN), AdaBoost, Support Vector Machine (SVM), Logistic Regression, XGBoost, Gradient Boost Tree, and Generalized Additive Model (GAM). Each classifier's precision, recall, F1 score, and accuracy are presented. This process is repeated for subsequent scenarios with different training-to-testing data ratios (80:20 and 70:30), maintaining the same classifiers. The study aims to assess how classifier performance varies concerning data availability by systematically altering the training-to-testing data ratios as show in Table 3.

Table 3. Performance Analysis and Evaluation

Train-testing data split	Classifier	Precision	Recall	F1score	Accuracy
90:10	DNN	0.701	0.685	0.680	0.685
	AdaBoost	0.744	0.742	0.741	0.742
	SVM	0.654	0.639	0.631	0.639
	Logistic regression	0.684	0.679	0.678	0.679
	XGBoost	0.836	0.836	0.836	0.836
	Gradient Boost Tree	0.803	0.802	0.801	0.802
	GAM	0.701	0.697	0.697	0.697
80:20	DNN	0.694	0.692	0.691	0.692
	AdaBoost	0.736	0.733	0.733	0.733
	SVM	0.658	0.641	0.633	0.641
	Logistic regression	0.681	0.676	0.675	0.676
	XGBoost	0.802	0.800	0.800	0.800
	Gradient Boost Tree	0.841	0.841	0.841	0.841
	GAM	0.699	0.696	0.696	0.696
70:30	DNN	0.682	0.680	0.679	0.680
	AdaBoost	0.736	0.734	0.733	0.734
	SVM	0.655	0.641	0.633	0.641
	Logistic regression	0.681	0.677	0.675	0.677
	XGBoost	0.830	0.830	0.830	0.830
	Gradient Boost Tree	0.804	0.804	0.804	0.804
	GAM	0.697	0.695	0.694	0.695

Across the board, Tabel 3 shows that XGBoost consistently outperforms other classifiers regarding precision, recall, F1 score, and accuracy, showcasing its robustness and effectiveness across different data splits. Mainly, in the 90:10, 80:20, and 70:30 training-testing data splits, XGBoost yields precision, recall, F1 score, and accuracy values of 0.836, 0.802, 0.801, and 0.836; 0.802, 0.841, 0.841, and 0.841; and 0.830, 0.804, 0.804, and 0.830, respectively. Gradient Boost Tree also demonstrates commendable performance, closely following XGBoost, especially in the 80:20 split with precision, recall, F1 score, and accuracy values of 0.841, 0.841, 0.841, and 0.841, respectively. On the other hand, DNN exhibits competitive performance across all

splits, albeit slightly lower than XGBoost and Gradient Boost Tree, indicating its suitability for the task. Meanwhile, AdaBoost consistently performs well across different splits, showing stable and reliable performance. However, SVM, Logistic Regression, and GAM exhibit comparatively lower performance metrics across all splits, suggesting their limitations in capturing the complexities of the underlying data. These findings underscore the importance of selecting an appropriate classifier based on the specific requirements and characteristics of the dataset, with XGBoost emerging as a top-performing choice in this study due to its consistently high performance across various data splits.

The results obtained from the research align closely with findings from [41], which focus on acute stroke detection utilizing machine learning models, particularly XGBoost using brain CT dataset. In both studies, XGBoost emerges as a top-performing classifier, showcasing superior accuracy compared to alternative models. In [41], proposed XGBoost achieved an accuracy of 97%, outperforming existing models such as Random Forest (RF) with 86% accuracy, RF-LTSR with 90% accuracy, RF-HERMES with 89% accuracy, BE-FAST with 76% accuracy, and FAST with 80% accuracy. Similarly, in the present analysis, XGBoost consistently demonstrates high precision, recall, F1 score, and accuracy across different training-testing data splits, indicating its robustness and effectiveness in various machine-learning tasks. This parallelism underscores the reliability and versatility of XGBoost as a classifier across different domains and applications.

Furthermore, both outcomes highlight the importance of feature selection and preprocessing techniques in enhancing model effectiveness. The study highlights the use of hybrid preprocessing techniques and altered segmentation methods to enhance picture quality and feature extraction, while the current research underlines the importance of comprehending the impact of each feature on model predictions using SHAP values. Both methodologies seek to enhance model efficacy by identifying the most pertinent features and minimizing noise and extraneous information. The alignment of these studies' findings highlights the importance of utilizing advanced machine learning techniques like XGBoost and implementing thorough feature selection and preprocessing methods to improve model accuracy and reliability across various applications [41].

3.4. Shapley Addictive exPlanations (SHAP) for XGBoost and Gradient Boosted Tree

The next phase of this project concentrates on employing SHAP for XGBoost and Gradient Boosted Tree models. The SHAP Summary graphic offers critical insights into the factors influencing stroke prediction, clarifying the impact of each parameter on model output via SHAP values displayed on the X-axis. Elevated SHAP values indicate significant contributions to forecasts, whilst the Y-axis illustrates essential elements for stroke prediction. The caption differentiates these qualities by color, with red signifying increased stroke risk and blue showing reduced risk. This work establishes a foundation for a more comprehensive investigation of stroke risk factors utilizing SHAP for XGBoost and Gradient Boosted Tree models. The next SHAP Summary Plot for both models examine the complex correlation between several features and stroke risk. It emphasizes the importance of parameters including average blood glucose, BMI, hypertension, and heart disease in assessing stroke risk. Increased blood glucose and BMI levels, in conjunction with hypertension and cardiovascular disease, are identified as main risk factors.

Moreover, age, marital status, occupation, gender, smoking habits, and residential area influence stroke risk. Increased age, male sex, and smoking status are linked to a heightened risk, but marital status and participation in non-manual labor occupations are correlated with a reduced risk. Significantly, living in rural regions is associated with an increased risk of stroke relative to metropolitan areas. These findings emphasize the intricate interaction of multiple factors in determining stroke risk and highlight the necessity of thorough risk evaluation and focused interventions to reduce stroke occurrence. Figures 6 and 7 illustrate the SHAP Summary Plot for XGBoost and Gradient Tree Boosting, respectively.

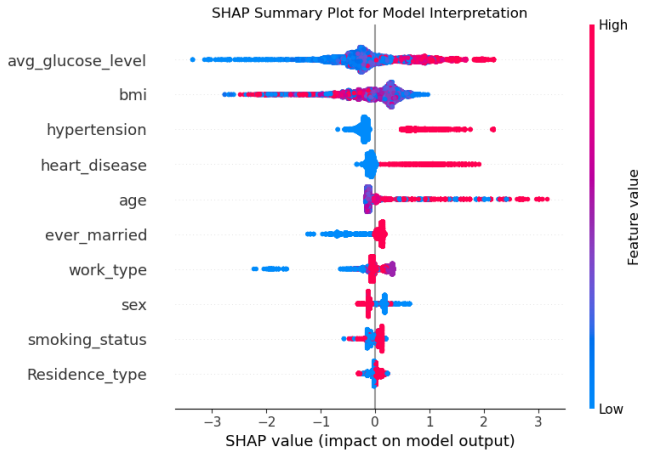

Figure 6. SHAP Summary Plot for XGBoost model

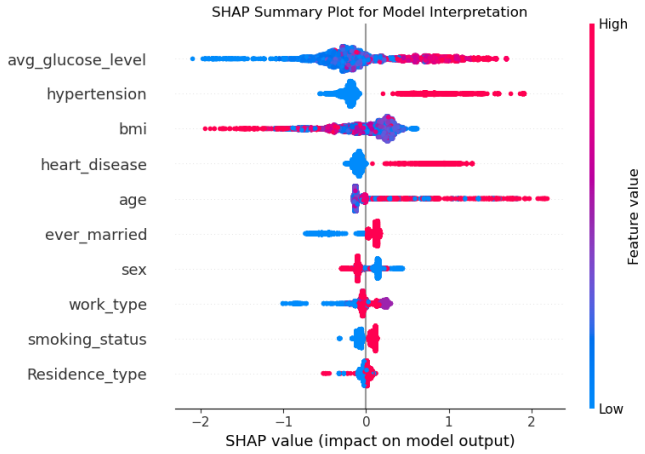

Figure 7. SHAP Summary Plot for Gradient Tree Boosting model

4. Conclusion

The study reveals substantial correlations in the Stroke Dataset, indicating potential risk variables such as average glucose levels and heart disease, along with moderate negative relationships between age and gender. Furthermore, a slight positive correlation is observed between employment type and smoking status. Preprocessing detects outliers, especially in the BMI attribute, highlighting the significance of data integrity in analysis concerning metabolic health. XGBoost regularly surpasses competing classifiers across diverse data splits, underscoring its efficacy in stroke prediction challenges. These findings underscore the importance of feature selection and preprocessing techniques in enhancing model effectiveness. Moreover, employing SHAP for XGBoost and Gradient Boosted Tree models yields profound insights into stroke risk variables, underscoring the necessity for thorough risk evaluation and focused interventions to reduce stroke occurrence. SHAP Summary Plots clearly depict the influence of features on model predictions, providing essential insights for subsequent study.

The research offers valuable insights into stroke prediction through comprehensive analysis of datasets, preprocessing techniques, machine learning classifiers, and SHAP interpretations. Future research may examine supplementary datasets, optimize preprocessing techniques, and delve further into sophisticated machine learning algorithms to improve stroke prediction precision and enable more efficient preventive strategies.

References

[1] S. A. Sheth, L. Giancardo, M. Colasurdo, V. M. Srinivasan, A. Niktabe, and P. Kan, "Machine learning and acute stroke imaging," *J. Neurointerv. Surg.*, vol. 15, no. 2, pp. 195–199, 2022, doi: 10.1136/neurintsurg-2021-018142.

[2] M. A. Saleem *et al.*, "Innovations in Stroke Identification: A Machine Learning-Based Diagnostic Model Using Neuroimages," *IEEE Access*, vol. 12, no. February, 2024, doi: 10.1109/ACCESS.2024.3369673.

[3] K. H. Yeh, "A Secure IoT-Based Healthcare System with Body Sensor Networks," *IEEE Access*, vol. 4, pp. 10288–10299, 2016, doi: 10.1109/ACCESS.2016.2638038.

[4] V. Bandi, D. Bhattacharyya, and D. Midhunchakkravarthy, "Prediction of brain stroke severity using machine learning," *Rev. d'Intelligence Artif.*, vol. 34, no. 6, pp. 753–761, 2020, doi: 10.18280/RIA.340609.

[5] K. Mridha, S. Ghimire, J. Shin, A. Aran, M. M. Uddin, and M. F. Mridha, "Automated Stroke Prediction Using Machine Learning: An Explainable and Exploratory Study With a Web Application for Early Intervention," *IEEE Access*, vol. 11, no. April, pp. 52288–52308, 2023, doi: 10.1109/ACCESS.2023.3278273.

[6] S. Mainali, M. E. Darsie, and K. S. Smetana, "Machine Learning in Action: Stroke Diagnosis and Outcome Prediction," *Front. Neurol.*, vol. 12, no. December, 2021, doi: 10.3389/fneur.2021.734345.

[7] S. Rahman, M. Hasan, and A. K. Sarkar, "Prediction of Brain Stroke using Machine Learning Algorithms and Deep Neural Network Techniques," *Eur. J. Electr. Eng. Comput. Sci.*, vol. 7, no. 1, pp. 23–30, 2023, doi: 10.24018/ejece.2023.7.1.483.

[8] G. Singh *et al.*, "Role of Machine Learning in Acute Stroke Imaging : A Technical Review," no. Ml, 2019.

[9] P. Yogendra Prasad, M. Ramu, K. Anitha, K. Lalasa, D. Hasritha, and B. A. Reddy, "Brain Stroke Detection Through Advanced Machine Learning and Enhanced Algorithms," *2024 Int. Conf. Recent Adv. Electr. Electron. Ubiquitous Commun. Comput. Intell. RAEEUCCI 2024*, 2024, doi: 10.1109/RAEEUCCI61380.2024.10547987.

[10] M. Lee *et al.*, "Machine learning-based prediction of post-stroke cognitive status using electroencephalography-derived brain network attributes," *Front. Aging Neurosci.*, vol. 15, no. September, pp. 1–10, 2023, doi: 10.3389/fnagi.2023.1238274.

[11] R. Garg, E. Oh, A. Naidech, K. Kording, and S. Prabhakaran, "Automating Ischemic Stroke Subtype Classification Using Machine Learning and Natural Language Processing," *J. Stroke Cerebrovasc. Dis.*, vol. 28, no. 7, pp. 2045–2051, 2019, doi: 10.1016/j.jstrokecerebrovasdis.2019.02.004.

[12] S. Ruksakulpiwat *et al.*, "Machine learning-based patient classification system for adults with stroke: A systematic review," *Chronic Illn.*, vol. 19, no. 1, pp. 26–39, 2023, doi: 10.1177/17423953211067435.

[13] F. Kremers *et al.*, "Outcome Prediction Models for Endovascular Treatment of Ischemic Stroke: Systematic Review and External Validation," *Stroke*, vol. 29, no. 2, pp. 825–836, 2022, doi: 10.1161/STROKEAHA.120.033445.

[14] T. I. Shoily, T. Islam, S. Jannat, S. A. Tanna, T. M. Alif, and R. R. Ema, "Detection of Stroke Disease using Machine Learning Algorithms," *2019 10th Int. Conf. Comput. Commun. Netw. Technol. ICCCNT 2019*, pp. 6–11, 2019, doi: 10.1109/ICCCNT45670.2019.8944689.

[15] E. Dritsas and M. Trigka, "Stroke Risk Prediction with Machine Learning Techniques," *Sensors*, vol. 22, no. 13, 2022, doi: 10.3390/s22134670.

[16] Nikita and G. Parashar, "Brain Stroke Detection and Prediction Using Machine Learning Approach: A Cloud Deployment Perspective," *Proc. Int. Conf. Circuit Power Comput. Technol. ICCPCT 2023*, no. August 2023, pp. 1705–1714, 2023, doi: 10.1109/ICCPCT58313.2023.10245699.

[17] H. Cheng, M. Zhang, and J. Q. Shi, "A Survey on Deep Neural Network Pruning: Taxonomy, Comparison, Analysis, and Recommendations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 14, no. 8, pp. 1–30, 2024, doi: 10.1109/TPAMI.2024.3447085.

[18] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 904, pp. 23–37, 1995, doi: 10.1007/3-540-59119-2_166.

[19] L. Wang, M. Sugiyama, C. Yang, Z. H. Zhou, and J. Feng, "On the margin explanation of boosting algorithms," *21st Annu. Conf. Learn. Theory, COLT 2008*, pp. 479–490, 2008.

[20] T. He, T. Wang, R. Abbey, and J. Griffin, "High-Performance Support Vector Machines and Its Applications," in *ICDATA 2018*, 2019, pp. 1–7, [Online]. Available: http://arxiv.org/abs/1905.00331.

[21] H. A. Park, "An introduction to logistic regression: From basic concepts to interpretation with particular attention to nursing domain," *J. Korean Acad. Nurs.*, vol. 43, no. 2, pp. 154–164, 2013, doi: 10.4040/jkan.2013.43.2.154.

[22] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, vol. 13-17-Augu, pp. 785–794, 2016, doi: 10.1145/2939672.2939785.

[23] Á. Delgado-Panadero, B. Hernández-Lorca, M. T. García-Ordás, and J. A. Benítez-Andrades, "Implementing local-explainability in Gradient Boosting Trees: Feature Contribution," *Inf. Sci. (Ny).*, vol. 589, pp. 199–212, 2022, doi: 10.1016/j.ins.2021.12.111.

[24] C.-H. Chang, R. Caruana, and A. Goldenberg, "NODE-GAM: Neural Generalized Additive Model For Interpretable Deep Learning," in *ICLR 2022*, 2022, pp. 1–25.

[25] P. Chuks, "Diabetes, Hypertension and Stroke Prediction," 2015. https://www.kaggle.com/datasets/prosperchuks/health-dataset.

[26] H. P. Vinutha, B. Poornima, and B. M. Sagar, "Detection of outliers using interquartile range technique from intrusion dataset," *Adv. Intell. Syst. Comput.*, vol. 701, pp. 511–518, 2018, doi: 10.1007/978-981-10-7563-6_53.

[27] C. Cortes and V. Vapnik, "Support Vector Networks," *Mach. Learn.*, vol. 20, pp. 273–297, 1995, [Online]. Available: http://image.diku.dk/imagecanon/material/cortes_vapnik95.pdf.

[28] F. F. Sherif and K. S. Ahmed, "A Machine Learning Approach for Stroke Differential Diagnosis by Blood Biomarkers," *J. Adv. Inf. Technol.*, vol. 15, no. 1, pp. 1–9, 2024, doi: 10.12720/jait.15.1.1-9.

[29] D. C. Feng *et al.*, "Machine learning-based compressive strength prediction for concrete: An adaptive boosting approach," *Constr. Build. Mater.*, vol. 230, no. January, 2020, doi: 10.1016/j.conbuildmat.2019.117000.

[30] T. Raj Ojha and A. Kumar Jha, "Analyzing the Performance of the Machine Learning Algorithms for Stroke Detection," *Int. J. Educ. Manag. Eng.*, vol. 13, no. 2, pp. 27–35, 2023, doi: 10.5815/ijeme.2023.02.04.

[31] T. J. Hastie and R. J. Tibshirani, "Hastie T.J., Tibshirani R.J. - Generalized Additive Models (CRC,1990)(175d)0412343908.pdf," *New York: Chapman and Hall.*, vol. 1, no. 3. pp. 297–310, 1990.

[32] S. M. Piryonesi and T. E. El-Diraby, "Data Analytics in Asset Management: Cost-Effective Prediction of the Pavement Condition Index," *J. Infrastruct. Syst.*, vol. 26, no. 1, 2020, doi: 10.1061/(asce)is.1943-555x.0000512.

[33] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," *Adv. Neural Inf. Process. Syst.*, vol. 2017-Decem, no. Section 2, pp. 4766–4775, 2017.

[34] F. Fumagalli, M. Muschalik, P. Kolpaczki, E. Hullermeier, and B. Hammer, "SHAP-IQ: Unifed Approximation of any-order Shapley Interactions," in *37th Conference on Neural Information Processing Systems (NeurIPS 2023)*, 2023, pp. 1–37.

[35] R. O. Alabi, M. Elmusrati, I. Leivo, A. Almangush, and A. A. Mäkitie, "Machine learning explainability in nasopharyngeal cancer survival using LIME and SHAP," *Sci. Rep.*, vol. 13, no. 1, pp. 1–14, 2023, doi: 10.1038/s41598-023-35795-0.

[36] A. Temenos, N. Temenos, M. Kaselimi, A. Doulamis, and N. Doulamis, "Interpretable Deep Learning Framework for Land Use and Land Cover Classification in Remote Sensing Using SHAP," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, pp. 0–4, 2023, doi: 10.1109/LGRS.2023.3251652.

[37] Z. G. Al-Mekhlafi *et al.*, "Deep Learning and Machine Learning for Early Detection of Stroke and Haemorrhage," *Comput. Mater. Contin.*, vol. 72, no. 1, pp. 775–796, 2022, doi: 10.32604/cmc.2022.024492.

[38] M. R. Fahlevvi, "Sentiment Analysis And Topic Modeling on User Reviews of Online Tutoring Applications Using Support Vector Machine and Latent Dirichlet Allocation," *Knowbase Int. J. Knowl. Database*, vol. 2, no. 2, p. 142, 2022, doi: 10.30983/knowbase.v2i2.5906.

[39] Y. Lizar, A. S. Firrizqi, A. Guci, and J. Sunadi, "Data Mining Analysis to Predict Student Skills Using Naïve Bayes Method," *Knowbase Int. J. Knowl. Database*, vol. 3, no. 2, p. 150, 2023, doi: 10.30983/knowbase.v3i2.7481.

[40] N. Sulistianingsih and G. H. Martono, "Enhancing Predictive Models: An In-depth Analysis of Feature Selection Techniques Coupled with Boosting Algorithms," *MATRIK J. Manajemen, Tek. Inform. dan Rekayasa Komput.*, vol. 23, no. 2, pp. 353–364, 2024, doi: 10.30812/matrik.v23i2.3788.

[41] S. K. UmaMaheswaran, F. Ahmad, R. Hegde, A. M. Alwakeel, and S. Rameem Zahra, "Enhanced non-contrast computed tomography images for early acute stroke detection using machine learning approach," *Expert Syst. Appl.*, vol. 240, no. November 2023, 2024, doi: 10.1016/j.eswa.2023.122559.