International Multilingual Journal of Science and Technology (IMJST)
ISSN: 2528-9810
Vol. 8 Issue 5, May - 2023

An Explainable Machine Learning Framework for Early Detection of Stroke

Md. Jamal Uddin[1], Prodip Kumar Sarker[2], Md. Nesarul Hoque[1], Sakifa Aktar[1], Md. Martuza Ahmad[1]
[1]Department of Computer Science and Engineering
[1]Bangabandhu Sheikh Mujibur Rahman Science and Technology University, Gopalganj-8100, Bangladesh
[2]Department of Computer Science and Engineering
Begum Rokeya University, Rangpur, Bangladesh

***Abstract*— A stroke is a critical neurological defect of the brain's blood vessels that occurs when the blood supply to a portion of the brain struggles or stops depriving brain cells of oxygen. It yields various forms of physical imbalance. It is one of the leading causes of illness and mortality worldwide. 20–25% of stroke survivors have substantial impairment, which has been linked to an increased mortality risk. Recognizing the numerous stroke warning signs early can prevent a stroke from occurring. In this study, we developed an ensemble learning-based machine learning architecture capable of analyzing stroke patient datasets and precisely predicting and recognizing stroke features. At first, a stroke dataset is collected, and then the Synthetic Minority Oversampling Technique (SMOTE) is used to balance it. Then, we implemented several machine learning techniques, such as Decision Tree, Naive Bayes, K-Nearest Neighbors, Random Forest, Extreme Gradient Boosting, Multilayer Perceptron, Ada Boost, and our proposed Ensemble framework. After optimizing hyperparameters, our proposed framework demonstrated the highest accuracy (99.90%) among all machine learning classifiers. We identified Age, BMI, and Average Glucose Level, Heart Disease as significant stroke indicators using machine learning (information gain, correlation, and Relief F) and statistical feature selection techniques. The SHapley Additive exExplanations (SHAP) method is utilized to determine the influence of each attribute on the model outcome. We believe that our proposed framework can assist physicians and clinicians in prescribing and detecting a potential stroke early on.**

Index Terms—Stroke; Machine Learning; Feature Selection; Feature Importance; SHAP;

I. INTRODUCTION

A stroke occurs when the regular supply of blood to the brain is interrupted or blocked suddenly, or the blood vessels are leaked or ruptured in the brain. Without sufficient blood circulation, brain cells progressively perish, resulting in varying degrees of complications. Even though some patients can recover after suffering a stroke, depending on the severity of the stroke, a significant number of patients continue to struggle with difficulties. These problems can include complications with attention, concentration, and memory; struggling to speak or understand speech; emotional issues such as depression; loss of balance or capability to walk; loss of sensation on the affected side of the body; and having trouble eating food [1, 2]. Following the World Stroke Organization report, one in four adults over 25 will experience a stroke in their lifespan [3]. In addition, 12.2 million persons will suffer their first stroke in 2023, resulting in 6.5 million deaths. The report also notes that over 110 million individuals have suffered a stroke worldwide. It also negatively impacts the patients' families, friends, workplaces, and social circumstances [4]. Furthermore, contrary to prevalent opinion, it can occur at any stage of life, gender, or physical condition.

Patients may be at risk for stroke for multiple causes. Following to the National Heart, Lung, and Blood Institute, high blood pressure, heart and blood vessel diseases, diabetes, smoking, brain aneurysms, family history or genetics, and other complications may be the leading causes of stroke [5]. To reduce the risk of stroke, it is essential to routinely track blood pressure, exercise consistently, maintain a healthy weight, stop smoking and taking alcohol, and consume a healthy, low-fat diet [6, 7].

Typically, the medical dataset includes patient symptoms and physical conditions. Recent years have seen the emergence of machine learning (ML) as a cutting-edge approach for healthcare prognosis and diagnosis that can classify medical information into specific class labels, such as sick or non-sick [8]. Such a strategy has enabled the successful implementation of ML for increased diagnostic accuracy and efficiency [9–11]. A growing number of studies over the past decade have investigated the applicability of the ML models for predicting stroke [12, 13]. The authors of [14–16] experimented on the same dataset containing 5,110 patients with eleven clinical features and one target attribute. However, Dritsas and Trigka [14] considered 3,254 participants over 18 years old. In the preprocessing phase, they handled missing values and removed noisy contents. Then, they manipulated SMOTE (synthetic minority over-sampling technique) method to balance the significant (non-stoke) and minor (stoke) classes. Finally, they applied various ML models, where the stacking of Logistic Regression

(LR), Naive Bayes (NB), Random forest (RF), RepTree, and J48, outperformed with 98% accuracy and 97.4% F-measure. The authors did not consider any deep learning models in this prediction system. In [15], the authors removed an irrelevant feature and substituted the missing values with the mean value of a column. They used the under sampling method to balance the dataset, where the total rows reached 498 (249 for the stroke class and 249 for the non-stroke class). They got the highest accuracy of 82% by employing the NB classifier. The authors did not apply neural network based models in this study. Furthermore, they were concerned about the detection performance because of limited trained and test data. Rahman et al. [16] filled the missing values with the most frequent values of the feature column, applied the MinMaxScaler method to normalize the features, utilized Principal Component Analysis (PCA) technique to reduce the feature dimensions, and employed a random over-sampling strategy for balancing the dataset in the pre-processing stage. After that, they applied various ML and deep neural network (DNN) models, where the RF showed excellent accuracy of 99% than the other models. In this investigation, the authors performed less work to analyze the most significant factors strongly related to stroke occurrence. Dev et al. [17] worked on a comparatively large dataset. The authors investigated various factors presented in the Electronic Health Record (EHR) records of 29,072 patients, where only 548 entries are associated with stroke condition, while the remaining 28,524 are not of to stroke nature. They sorted out four significant factors, including age, average glucose status, hypertension, and heart disease, using Learning Vector Quantization (LVQ) model for stroke prediction. They employed a random down sub-sampling approach to handle the bias of the majority class (not stroke). They achieved an optimistic output for their four selected features of 78% accuracy with a low miss rate of 19% by utilizing a neural network (NN) model. However, the performance score is still lacking for treatment and precluding measures for an individual. Liu et al. [18] worked on another bigger dataset for predicting cerebral stroke containing eleven input features and 43,400 samples, of which only 783 patients experienced stroke. Besides the highly imbalanced nature, the dataset is incomplete because of the missing value of some fields. As the outliers and noisy samples, the authors filtered those patient instances that have aged below 25 and a BMI (body mass index) value is more than 60%. Then they employed the random forest regression (RFR) method to attribute the missing values. Using an automated hyperparameter optimization (AutoHPO)-based DNN model, they obtained the lowest false negative rate of 19.1% with an accuracy of 71.6%. They used XGBoost (XGB) and RF models to get significant factors regarding the occurrence of strokes. The authors did not utilize DNN methods in the feature important analysis. Additionally, there is still enough space to enhance the stroke prediction system. Govindarajan et al. [19] developed a stroke classification system to identify two types of stroke: ischemic stroke (IS) and hemorrhage stroke (HE). At first, the authors collected 507 distinct patient records between the age group 35 and 90, where 91.52% of them are infected by IS, whereas 8.48% are by HE. Then, they extracted 22 significant features from those records using a base-form generator and a stemmer algorithm. Lastly, they exploited several ML models like Decision Tree (DT), Support Vector Machine (SVM), Artificial Neural Networks (ANN), RF, and others, where ANN outperformed the others with 95.3% accuracy and a lower standard deviation (14.69). This study reveals that stroke is more common in men than women and those aged 40 to 60. The authors did not employ any technique to handle the highly imbalanced dataset.

Thought out the analysis of the existing research works, we have detected several issues, such as dataset completeness and balancing, feature selection analysis, detection performance, and others. In this research, first, we collect a renowned dataset from a popular data repository platform Kaggle.com. After that, we made the following contributions by focusing on all the above issues:

- To handle the completeness and the highly imbalanced nature of the dataset.
- To figure out the important features of stroke risk factors.
- To propose an ensemble ML model for automated stroke prediction with outstanding performance.
- To rank and analyze, the risk factors of stroke using ML techniques.

II. Materials and Methods

In this research, using a stroke dataset, we employed statistical and ML techniques to determine

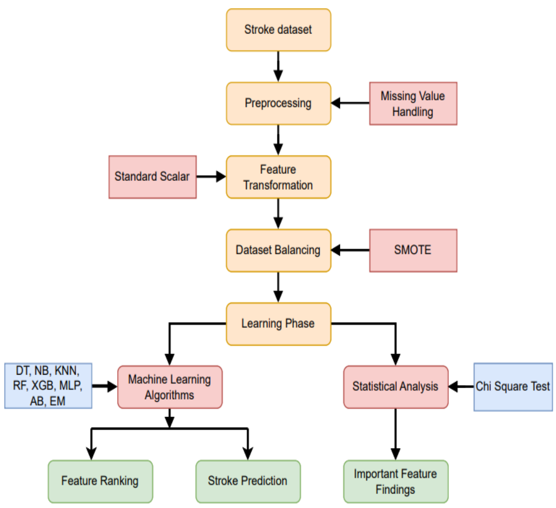

Fig. 1: The workflow of stroke prediction

the most prominent indicators connected to stroke. Next, ML models were utilized to identify early-stage strokes. Figure 1 depicts a comprehensive pictorial representation of the workflow.

A. *Data Collection and Description*

We obtained the dataset from the publicly accessible Kaggle data repository, and it contains twelve features: ID, Gender, Age, Hypertension, Heart disease, Marital Status, Work Type, Residence Type, Average Glucose Level, BMI, Smoking Status, and Stroke. This dataset includes 5,110 records, of which 249 (4.88%) are stroke patients and 4,861 (95.12%) are not. The average age of patients was 43.21 years, ranging from 82 years to 8 months. Male patients were 2,115, and females were 2,994. The mean BMI was 28.85, the maximum was 97.6, and the minimum was 10.3. Hypertension and cardiovascular disease prevalence were extremely low (9.75% and 9.74%, respectively). The average blood glucose level was 106.14, with a maximum of 271.74 and a minimum of 55.12. Table I provides a detailed description of the dataset. Figure 2 depicts the correlation between each attribute. We did not eliminate any features because there is no significant correlation between them.

B. *Dataset Balancing Technique*

The SMOTE creates samples of the minority population that are artificial. It trains a classifier by synthesizing a balanced synthetic training set across classes. With SMOTE, the distribution of the dataset is more uniform as more instances are added to the minority class. Consequently, ML models can profit from acquiring knowledge from a more specific data set. It has found pervasive application in fields where imbalanced datasets are standard, such as identifying fraud, health care diagnosis, and text classification [20].

C. *Feature Transformation Method*

Standard Scaler [21] rescales the characteristics with a mean of zero and a variance of one. It is used when the ranges of the characteristics of the input dataset are extensive. The standard normal distribution is the consequence of subtracting attribute values by their mean and dividing by standard deviation (σ).

$$Standard\ Scalar(x) = \frac{x - \bar{x}}{\sigma} \qquad (1)$$

Where x is the original feature value, $\bar{x}$ is the mean of the feature values. Standard scalar is frequently employed in numerous ML algorithms to prepare and normalize the input features, including linear regression, LR, SVM, and NN.

D. *Feature Ranking Method*

Feature ranking is a method for evaluating the significance or relevance of a dataset's features. It attempts to evaluate the features based on their contribution to an ML model's predictive ability. In our investigation, we used information gain, person correlation, and Relief F techniques for ranking features.

- Information Gain (IG) depends on the entropy concept, which measures the impureness or uncertainty of a data set [22].
- Pearson's correlation (PC) calculates its value for the variable in the class to determine the value of an attribute [23].
- Relief F determines the value of a feature by continually sampling an instance and evaluating the given attribute's value for the nearest instances of the same and different classes [24].

E. *Machine Learning Model*

In this study, we employed a number of ML algorithms, including Decision-Tree (DT), Naive Bayes (NB), K-Nearest Neighbors (KNN), Random Forest (RF), Extreme Gradient Boosting Machine (XGB), Multilayer Perceptron (MLP), Ada Boost (AB), and Ensemble Method (EM). We searched for the best-performing models to predict stroke. We illustrated each model below:

Decision-Tree (DT) technique has rapidly become a popular ML tool for classification and regression problems. The algorithm's pervasive approval and use can be attributed to its remarkable resemblance to human thought. It can be used as a non-parametric supervised learning classifier, expanding its versatility. A limited number of "nodes" within a decision tree represent potential steps, while "leaf nodes" define the outcomes of those actions.

Naive Bayes (NB) is a simple yet effective classifier that can effectively forecast the outcomes of many challenging problems in a brief period. It employs Bayes' theorem and operates under the assumption of a distinct variable [25].

K-Nearest Neighbors (KNN) is a straightforward, nonparametric supervised learning technique that looks for similar features in the training set [26]. It frequently employs the Euclidean, Manhattan, and Minkowski distance procedures to differentiate between new input and preexisting knowledge.

Random Forest (RF) [27] is a compilation of decision trees derived from multiple samples that are employed to enhance the accuracy of a dataset. Each decision tree was trained on a sample of data selected at random with replacement using the bagging method. The output of a RF is a composite prediction equation that is derived from the outputs of multiple decision trees. It can be applied to both regression and classification problems [28].

eXtreme Gradient Boosting (XGBoost) technique combines gradient enhancement and boosting to produce exceptional results. It is an effective and scalable variant of the Gradient Boosting Method (GBM) that can perform a variety of tasks, including regression, classification, and ranking [29].

TABLE I: Dataset Description

ID	Attribute Name	Feature Description	Count/Average Value
1	Gender	Gender of the patient	Male: 2115(41.39%) Female: 2994(58.60%) Other: 1(00.01%)
2	Age	Age of the patient	Average: 43.21 Max: 82 Min: 0.08 Median: 45
3	Hypertension	Whether or not the respondent has hypertension	Yes: 498(9.75%) No: 4612(90.25)
4	Heart disease	Whether the respondent has heart disease or not	Yes: 498(9.74%) No: 4612(90.26)
5	Ever married	Whether the respondent is married or single	Yes: 3353(65.61%) No: 1757(34.39%)
6	Work type	It can be govt. job, never worked, private or self-employed	Govt. job: 657(12.85%) Never worked: 709(13.88%) Private: 2925(57.24%) Self-employed: 819(16.03%)
7	Residence type	Residence can be rural or urban	Rural: 2514(49.19%) Urban: 2596(50.81%)
8	Average glucose level	Average glucose level in blood	Average: 106.14 Max: 271.74 Min: 55.12 Median: 91.88
9	BMI	Body mass index	Average: 28.85 Max: 97.6 Min: 10.3 Median: 27.78
10	Smoking status	Smoking status can be formerly smoked, never smoked, smokes or unknown	Formerly smoked: 885(17.32%) Never smoked: 1892(37.03%) Smokes: 789(15.45%) Unknown: 1544(30.22%)
11	Stroke	The patient has a stroke or not	Yes: 249(4.88%) No: 4861(95.12%)

Multilayer Perceptron (MLP) is a popular form of artificial neural network (ANN) in ML. It is a neural network with multiple layers of interrelated nodes called neurons or units. It is renowned for its ability to discover intricate data relationships and patterns. They can handle a vast array of challenges in domains, such as classification, regression, and even sequence related tasks.

Adaptive Boosting (AB) is a technique popularized by a boosting algorithm [30]. This method aims to integrate multiple weak classifiers into a single robust classifier.

Ensemble Method (EM) combines XGB and RF and incorporates the advantages of both algorithms to enhance the overall predictive performance. By integrating the assets of multiple models, ensemble techniques can offer improved generalization, increased robustness, and enhanced precision. It can aid in mitigating individual model biases and enhancing model performance overall.

F. *Hyperparameter Optimization Technique*

Grid search is a straightforward but exhaustive technique for tuning the hyperparameters of each classifier. It investigates all possible hyperparameter value combinations within the specified grid. The ML algorithm is trained with altered parameters to identify the feature with the highest degree of precision. This entire process is a cycle within which assessments are continued.

G. *Performance Evaluation Metrics*

In this research, we evaluated the effectiveness of classifiers using a variety of assessment metrics [31], including accuracy [32], kapa statistics, precision, recall, F1-score, AUC, and log- loss. Using the

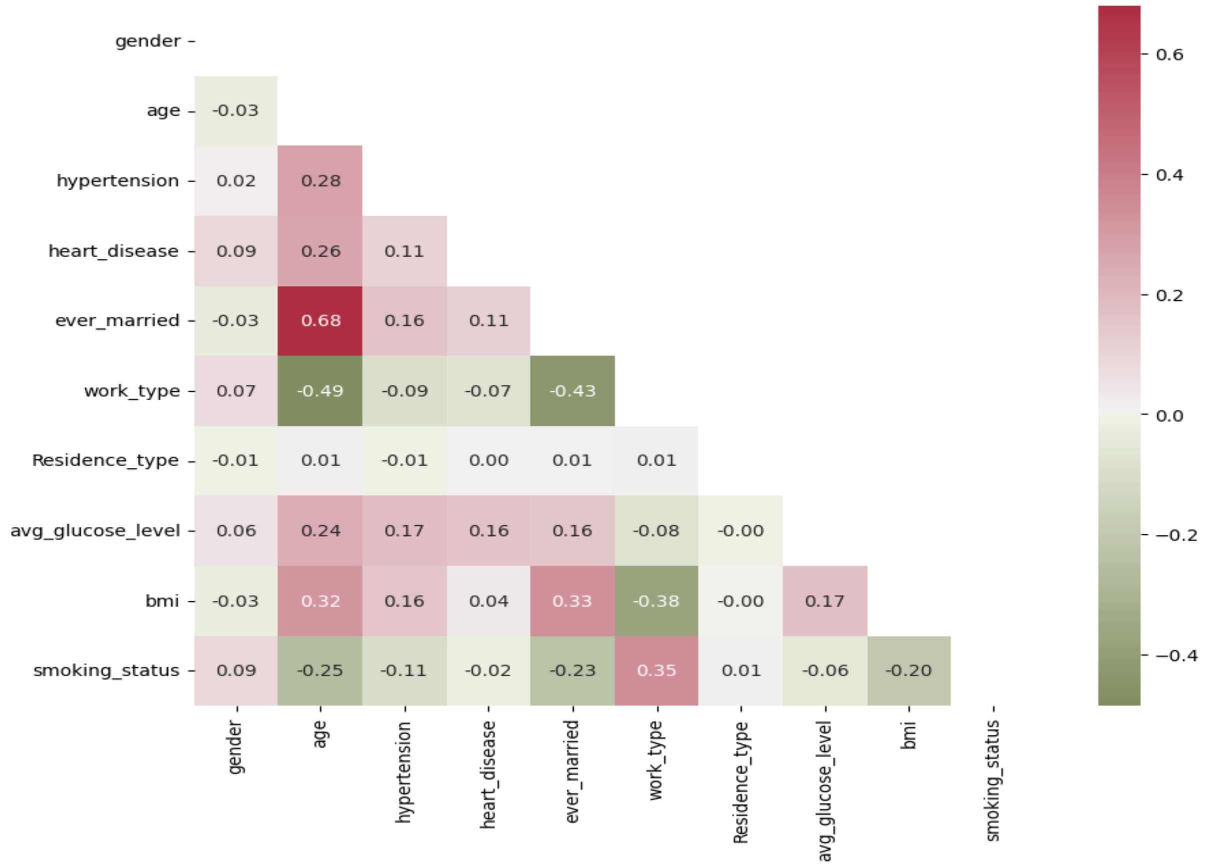

Fig. 2: *The correlation of each feature of stroke dataset*

following equations, evaluation metrics are calculated.

Accuracy: Accuracy is the proportion of accurate forecasts relative to the total number of forecasts. This metric performs optimally when each class contains approximately the same number of samples. The accuracy is expressed as

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (2)$$

Kapa Statistics: It is a measure of how well two evaluators agree on a value. In the context of ML model evaluation metrics, this value is the difference between the predicted and observed output.

$$Kapa\ statistics = \frac{1-(1-p_0)}{1-p_e} \quad (3)$$

Precision: Precision is defined as the proportion of accurately predicted outcomes to actual outcomes. Precision is expressed as:

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

Recall: Recall is the proportion of accurate predictions to all the actual positive outcomes. The equation of recall is:

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

AUC-ROC: The efficacy of our model can be demonstrated using the Receiver Operating Characteristic (ROC) Curve's Area Under the Curve (AUC). It is a crucial rating criterion.

F1-score: Precision and recall are weighted averages to provide the F1 score. The mathematical equation of the F1-score is

$$F1 - Score = 2\frac{Precision*Recall}{Precision+Recall} \quad (6)$$

Log Loss: It is a reliable indicator of classification work-place productivity. The Log Loss probability-based metric. Lower values for Log Loss indicate more precise predictions. An ideal classifier would exhibit zero Log Loss. The mathematical equation of Log Loss is:

$$Log\ Loss = -\frac{1}{N}\sum_{i=1}^{N} y_i \log\big(p(y_i)\big) + (1-y_i)log\big(1 - p(y_i)\big) \quad (7)$$

Here, TP is True Positive, FP is False Positive, TN is True Negative and TP is True Positive. In log loss, y represents the level of the target variable, p(y) represents the projected probability, and q represents the actual log loss. And in the Kappa-Statistics p_0 is among raters agreement of relative observed, and p_e is a chance agreement for hypothetical probability.

III. Results

We implemented the ML classifiers DT, NB, KNN, RF, XGB, MLP, AB, LR, and EM in our work. At Google Collaboratory, the experimental task was performed with Python sci-kit-learn. In this study, prediction models were developed using a 10-fold cross-validation procedure. We used the IG, PC, and Relief F feature selection methodologies to determine the significance of each feature. The SHAP summary graph was generated using Python's shap module. Several evaluation metrics, including precision, recall, AUC-ROC, F1-score, and log loss, are applied to validate the experimental results.

A. *Identifying crucial traits of stroke using statistical and machine learning techniques*

The stroke dataset utilized the Chi-square test to identify the most significant factors that caused the stroke. Figure 3 depicts our findings. The most significant indicators in descending order are Age, Heart Disease, Average Glucose Level, Hypertension, and Marital Status.

Using the IG, PC, and Relief F methods, we calculated the feature importance for the stroke dataset to identify the risk factors for stroke prediction. Table II displays the numerical outcomes.

IG determined that the highest feature importance value is 0.5129 for age, 0.4467 for BMI, and 0.0836 for Work Type. The most significant attributes, Age (0.2453), Heart Disease (0.1349), and Average Glucose Level (0.1319), were manipulated by PC techniques. Age (0.2818), Work Type (0.0972), and BMI (0.0927) were determined to be the essential characteristics by the Relief F method. We also calculated the average value of the IG, PC, and Relief F methodologies, as shown in Figure 4. Age, Body Mass Index, and Average Glucose Level were observed as the most significant characteristics.

B. *Exploring Discriminatory Stroke Identification Factors*

Figure 5 illustrates the order of importance of SHAP values within stroke datasets. The evaluation of these values was conducted using the XGB, which performed excellently. Age, BMI, Average Glucose Level, and Smoking Status were the most critical discriminatory characteristics for detecting stroke at an early stage. In contrast, Heart Disease, Hypertension, and Work Type constituted the least significant discriminatory factors.

C. *Classification of Stroke Using Machine Learning Algorithms*

In the main dataset (Table III), EM calculated the maximum accuracy (95.07%) and minimal log loss (1.7775). The KNN, RF, XGB, MLP, and AB all demonstrated approximately 95.00% accuracy. NB demonstrated the highest kappa statistics (17.01%), recall (42.17%), auc-roc (65.29%), and F1-score (22.91%). XGB performed with the highest accuracy (20.29%). Overall, the NB classifier demonstrated the most remarkable performance on the primary data set.

In the case of the balanced dataset (Table IV), XGB calculated the highest accuracy (95.60%), kappa statistics (91.20%), auc-roc (95.60), F1-score (95.59%), and lowest log loss (1.5868). The EM and KNN had the highest precision (96.08%) and recall (98.17%). Similarly, the RF and EM produced well across all evaluation metrics. Overall, XGM performed better than competing classifiers.

In relation to hyperparameter tuning of classifiers (Table V), the EM determined the maximum accuracy (99.90%), kappa statistics (99.79%), recall (99.90%), auc-roc (99.90%), F1-score (99.90%), and the least amount of log loss (0.0371). The KNN demonstrated the utmost precision (one hundred percent). The XGB and KNN additionally showed excellent performance across all evaluation metrics. Overall, EM performed better than competing classifiers.

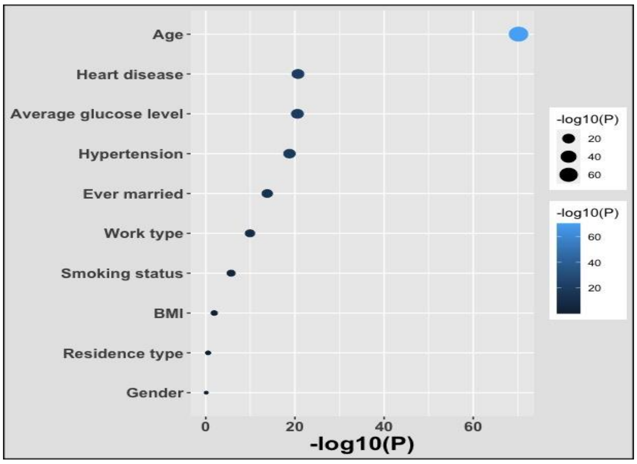

Fig. 3: *The feature significance in which a larger bubble represents a greater importance*

TABLE II: Feature Importance using machine learning technique

ID	Attribute Name	Info Gain	Correlation	Relief F
1	Gender	0.0270	0.0089	0.0180
2	Age	0.5129	0.2453	0.2818
3	Hypertension	0.0066	0.1279	0.0022
4	Heart disease	0.0000	0.1349	0.0046
5	Ever married	0.0085	0.1083	0.0262
6	Work type	0.0836	0.0671	0.0972
7	Residence type	0.0253	0.0155	0.0119
8	Average glucose level	0.0709	0.1319	0.0831
9	BMI	0.4467	0.0350	0.0927
10	Smoking status	0.0715	0.0198	0.0769

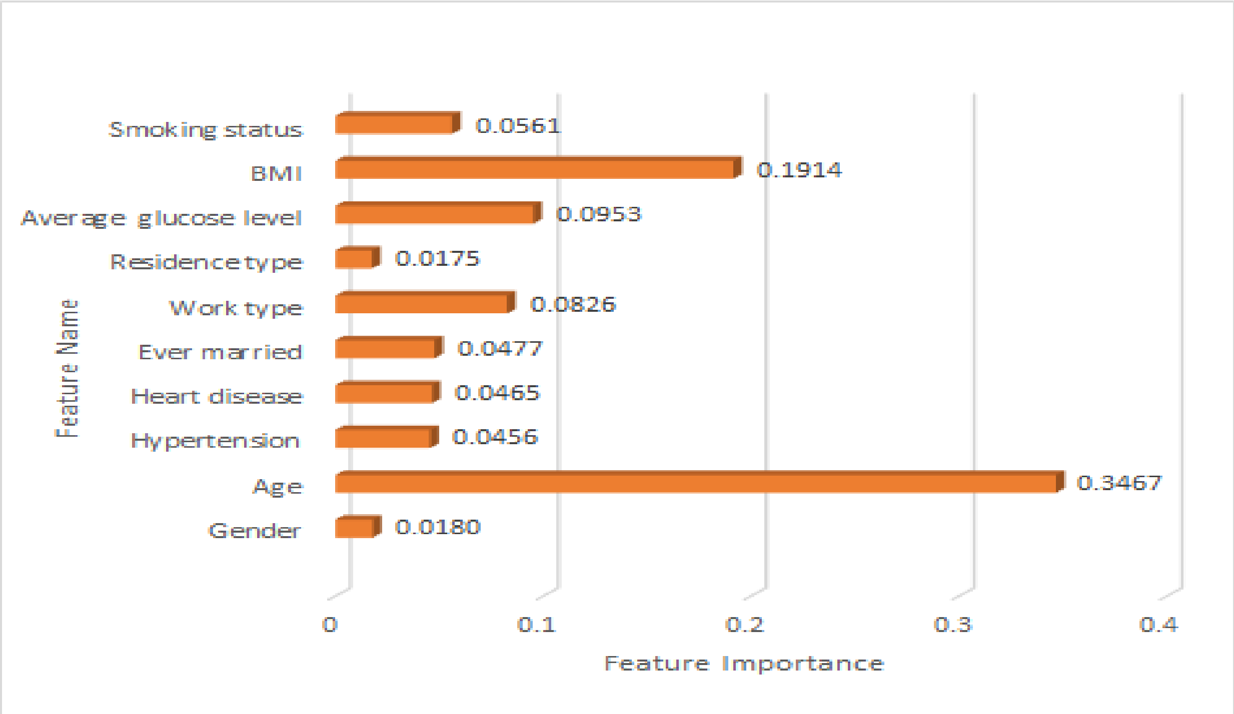

Fig. 4: *Feature ranking using machine learning techniques*

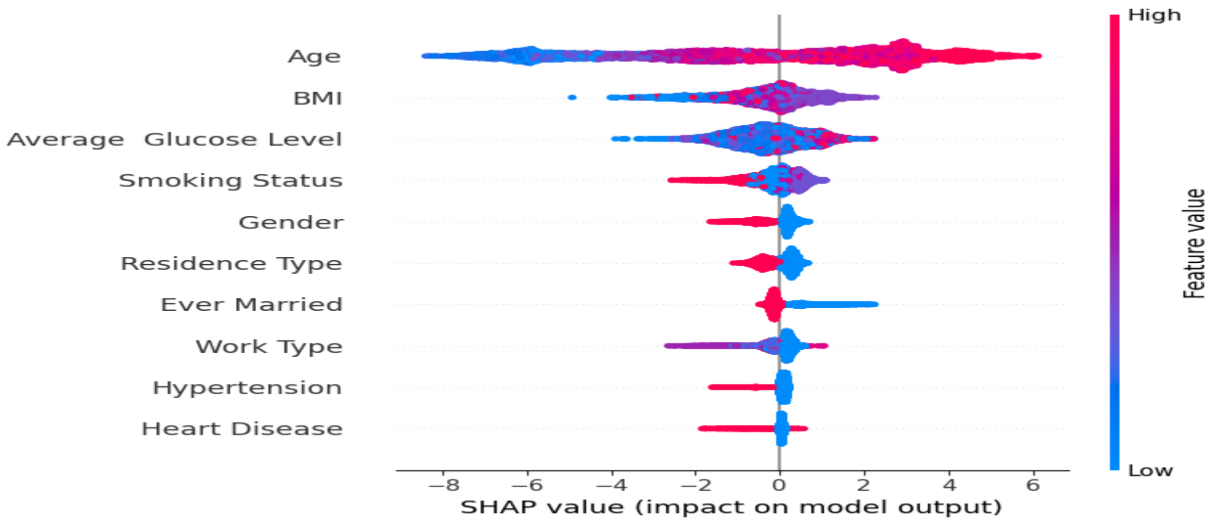

Fig. 5: *Analysis of Shapley values for the stroke dataset*

TABLE III: Performance Analysis of Different Classifiers in Main Dataset

Evaluation Metrics	Classifiers							
	DT	NB	KNN	RF	XGB	MLP	AB	Ensemble
Accuracy	0.9088	0.8616	0.9474	0.9491	0.9432	0.9485	0.9499	0.9507
Kapa Stat.	0.0956	0.1701	0.0061	0.0097	0.0683	0.0152	0.0044	0.0060
Precision	0.1322	0.1572	0.0833	0.1333	0.2029	0.1500	0.1111	0.2000
Recall	0.1566	0.4217	0.0080	0.0080	0.0562	0.0120	0.0040	0.0040
AUC-ROC	0.5520	0.6529	0.5018	0.5027	0.5225	0.5043	0.5012	0.5016
F1-score	0.1434	0.2290	0.0147	0.0152	0.0881	0.0223	0.0078	0.0079
Log Loss	3.2870	4.9869	1.8974	1.8339	2.0455	1.8551	1.8057	1.7775

TABLE IV: Performance Analysis of Different Classifiers in Balanced Dataset

Evaluation Metrics	Classifiers							
	DT	NB	DT	RF	DT	MLP	DT	Ensemble
Accuracy	0.8948	0.7974	0.8926	0.9343	0.9560	0.8362	0.8437	0.9522
Kapa Stat.	0.7895	0.5947	0.7852	0.8685	0.9120	0.6725	0.6873	0.9043
Precision	0.8792	0.7414	0.8332	0.9118	0.9567	0.8073	0.8224	0.9608
Recall	0.9152	0.9134	0.9817	0.9615	0.9552	0.8834	0.8766	0.9428
AUC-ROC	0.8948	0.7974	0.8926	0.9343	0.9560	0.8362	0.8437	0.9522
F1-score	0.8969	0.8184	0.9014	0.9360	0.9559	0.8436	0.8486	0.9517
Log Loss	3.7927	7.3036	3.8706	2.3690	1.5868	5.9022	5.6353	1.7240

TABLE V: Performance Analysis of Different Classifiers Based on Hyperparameter Tunning

Evaluation Metrics	Classifiers							
	DT	NB	DT	RF	DT	MLP	DT	Ensemble
Accuracy	0.8524	0.7997	0.9840	0.9001	0.9930	0.8921	0.8736	0.9990
Kapa Stat.	0.7048	0.5995	0.9679	0.8002	0.9860	0.7842	0.7472	0.9979
Precision	0.8154	0.7450	1.0000	0.8583	0.9940	0.8674	0.8603	0.9990
Recall	0.9111	0.9115	0.9679	0.9584	0.9920	0.9257	0.8920	0.9990
AUC-ROC	0.8524	0.7997	0.9840	0.9001	0.9930	0.8921	0.8736	0.9990
F1-score	0.8606	0.8199	0.9837	0.9056	0.9930	0.8956	0.8759	0.9990
Log Loss	5.3202	7.2184	0.5784	3.5999	0.2521	3.8891	4.5564	0.0371

TABLE VI: Comparative analysis of the proposed model with the other prior studies

Approach	Accuracy	Kapa Stat.	Precision	Recall	AUROC	F1-score	Log Loss
Stacking [14]	0.9800		0.9740	0.9740	0.9890	0.9740	
NB [15]	0.8200		0.7920	0.8570		0.8230	
Ensemble (Proposed Model)	0.9990	0.9979	0.9990	0.9990	0.9990	0.9990	0.0371

IV. Discussion

Several studies utilizing stroke datasets have been conducted, but stroke prediction still requires substantial refinement. In our investigation, we collected a stroke dataset and used the SMOTE method to balance it. After transforming the features using Standard Scalar, we applied DT, NB, KNN, RF, XGB, MLP, AB, and Ensemble classifiers. The grid search approach was then utilized for tuning the hyperparameter of each classifier. Here, we found that NB, KNN, XGB, MLP, AB, and Ensemble classifiers improved performance. In contrast to other classifiers, the ensemble method produced the most accurate results. Additionally, we used SHAP to explain the ML model's output. We also ranked the features using IG, PC, and Relief F feature ranking techniques.

Our findings imply several crucial and relevant characteristics for early stroke diagnosis. Depending on the log-based relationship, the essential characteristics are Age, Heart Disease, Average Glucose Level, Hypertension, and Marital Status. Age, BMI, average glucose level, work type, and smoking status are the most significant characteristics in the case of ML models. In addition, we identified critical metrics, such as Age and Mean Glucose Level, that are the same for both log-based relationships and ML techniques. Our research indicates that important characteristics are adequate for identifying a stroke, which will make more accessible the execution of stroke diagnosis.

Table VI compares the proposed model with pertinent prior findings. Dritsas et al. [14] employed stacking approach to achieve the highest levels of accuracy (98.00%), precision (97.40%), recall (97.40%), AUCROC (98.90%), and F1-score (97.40%). In a separate study, Sailasya et al. [15] utilized NB to obtain the best accuracy (82.0%), precision (79.20%), recall (85.70%), and F1-score (82.30%). In our proposed framework, however, the implementation of trait balancing, transformation, and hyperparameter optimization yielded the highest accuracy (99.90%), KS (99.70%), precision (99.90%), recall (99.90%), AUCROC (99.90%), F1-score (99.90%), and log loss (0.0371).

V. Conclusion

A stroke is a life-threatening condition that demands immediate care. The dataset was preprocessed in this study, and ML and statistical methods were used to identify critical stroke patient diagnostic characteristics. Age, BMI, and average glucose level are the three most important risk factors for stroke. It was also found that our proposed ensemble framework exhibited a high level of classification accuracy (99.90%), KS (99.70%), precision (99.90%), recall (99.90%), AUCROC (99.90%), F1-score (99.90%), and log loss (0.0371), which suggests that our findings can be utilized for computer-assisted medical diagnosis to assist healthcare professionals and physicians in examining stroke in a cost-effective manner. Our research enables the early identification of patients with a high risk of stroke who require additional examinations and treatment prior to the progression of the disease. This investigation's ultimate objective is to enhance the ML architecture using deep learning techniques. In order to assess the predictive potential of deep learning algorithms for stroke incidence, we will also acquire image data from CT and MRI imaging of the brain.

Data Availability

The dataset is publicly available.

Conflicts Of Interest

The authors declare that they have no conflicts of interest.

References

[1] B. Delpont, C. Blanc, G. Osseby, M. Hervieu B`egue, M. Giroud, and Y. B´ejot, "Pain after stroke: a review," Revue neurologique, vol. 174, no. 10, pp. 671–674, 2018.
[2] S. Kumar, M. H. Selim, and L. R. Caplan, "Medical complications after stroke," The Lancet Neurology, vol. 9, no. 1, pp. 105–118, 2010.
[3] "Learn about stroke," World Stroke Organization, accessed: 20 Mar 2023. [Online]. Available: https://www.world-stroke.org/ world-stroke-day campaign/why-stroke-matters/learn-about-stroke.
[4] T. Elloker and A. J. Rhoda, "The relationship between social support and participation in stroke: a systematic review," African Journal of Disability, vol. 7, no. 1, pp. 1–9, 2018.
[5] "Causes and risk factors," National Heart, Lung, and Blood Institute, accessed: 20 Mar 2023. [Online]. Available: https://www.nhlbi.nih.gov/ health/stroke/causes
[6] J. D. Pandian, S. L. Gall, M. P. Kate, G. S. Silva, R. O. Akinyemi, B. I. Ovbiagele, P. M. Lavados, D. B. Gandhi, and A. G. Thrift, "Prevention of stroke: a global perspective," The Lancet, vol. 392, no. 10154, pp. 1269–1278, 2018.
[7] V. L. Feigin, B. Norrving, M. G. George, J. L. Foltz, G. A. Roth, and G. A. Mensah, "Prevention of stroke: a strategic global imperative," Nature Reviews Neurology, vol. 12, no. 9, pp. 501–512, 2016.
[8] I. Yoo, P. Alafaireet, M. Marinov, K. Pena Hernandez, R. Gopidi, J.- F. Chang, and L. Hua, "Data mining in healthcare and biomedicine: a survey of the literature," Journal of medical systems, vol. 36, pp. 2431– 2448, 2012.
[9] A. N. Richter and T. M. Khoshgoftaar, "A review of statistical and machine learning methods for modeling cancer risk using structured clinical data," Artificial intelligence in medicine, vol. 90, pp. 1–14, 2018.
[10] C. R. Pereira, D. R. Pereira, S. A. Weber, C. Hook, V. H. C. De Al- buquerque, and J. P. Papa, "A survey on computer-assisted parkinson's disease diagnosis," Artificial intelligence in medicine, vol. 95, pp. 48–63, 2019.
[11] A. Kaya, "Cascaded classifiers and stacking methods for classification of pulmonary nodule characteristics," Computer Methods and Programs in Biomedicine, vol. 166, pp. 77–89, 2018.
[12] O. R. Shishvan, D.-S. Zois, and T. Soyata, "Machine intelligence in healthcare and medical cyber physical systems: A survey," IEEE Access, vol. 6, pp. 46 419–46 494, 2018.
[13] C. Colak, E. Karaman, and M. G. Turtay, "Application of knowledge discovery process on the prediction of stroke," Computer methods and programs in biomedicine, vol. 119, no. 3, pp. 181– 85, 2015.
[14] E. Dritsas and M. Trigka, "Stroke risk prediction with machine learning techniques," Sensors, vol. 22, no. 13, p. 4670, 2022.
[15] G. Sailasya and G. L. A. Kumari, "Analyzing the performance of stroke prediction using ml classification algorithms," International Journal of Advanced Computer Science and Applications, vol. 12, no. 6, 2021.
[16] S. Rahman, M. Hasan, and A. K. Sarkar, "Prediction of brain stroke using machine learning algorithms and deep neural network techniques," European Journal of Electrical Engineering and Computer Science, vol. 7, no. 1, pp. 23–30, 2023.
[17] S. Dev, H. Wang, C. S. Nwosu, N. Jain, B. Veeravalli, and D. John, "A predictive analytics approach for stroke prediction using machine learning and neural networks," Healthcare Analytics, vol. 2, p. 100032, 2022.
[18] T. Liu, W. Fan, and C. Wu, "A hybrid machine learning approach to cerebral stroke prediction based on imbalanced medical dataset," Artificial intelligence in medicine, vol. 101, p. 101723, 2019.
[19] P. Govindarajan, R. K. Soundarapandian, A. H. Gandomi, R. Patan, P. Jayaraman, and R. Manikandan, "Classification of stroke disease using machine learning algorithms," Neural Computing and Applications, vol. 32, pp. 817–828, 2020.
[20] M. J. Uddin, M. M. Ahamad, P. K. Sarker, S. Aktar, N. Alotaibi, S. A. Alyami, M. A. Kabir, and M. A. Moni, "An integrated statistical and clinically applicable machine learning framework for the detection of autism spectrum disorder," Computers, vol. 12, no. 5, p. 92, 2023.
[21] M. M. Ahamad, S. Aktar, M. J. Uddin, T. Rahman, S. A. Alyami, S. Al-Ashhab, H. F.

Akhdar, A. Azad, and M. A. Moni, "Early-stage detection of ovarian cancer based on clinical data using machine learning approaches," Journal of Personalized Medicine, vol. 12, no. 8, p. 1211, 2022.
[22] T. Akter, M. S. Satu, M. I. Khan, M. H. Ali, S. Uddin, P. Lio, J. M. Quinn, and M. A. Moni, "Machine learning-based models for early stage detection of autism spectrum disorders," IEEE Access, vol. 7, pp. 166 509–166 527, 2019.
[23] G. Fang, P. Xu, and W. Liu, "Automated ischemic stroke subtyping based on machine learning approach," IEEE Access, vol. 8, pp. 118 426– 118 432, 2020.
[24] S. M. Hasan, M. P. Uddin, M. Al Mamun, M. I. Sharif, A. Ulhaq, and G. Krishnamoorthy, "A machine learning framework for early-stage detection of autism spectrum disorders," IEEE Access, 2022.
[25] K. P. Murphy et al., "Naive bayes classifiers," University of British Columbia, vol. 18, no. 60, pp. 1 8, 2006.
[26] L. E. Peterson, "K-nearest neighbor," Scholarpedia, vol. 4, no. 2, p. 1883, 2009.
[27] M. J. Vowels, "Trying to outrun causality with machine learning: Limitations of model explainability techniques for identifying predictive variables," arXiv preprint arXiv:2202.09875, 2022.
[28] S. Wan, Y. Liang, Y. Zhang, and M. Guizani, "Deep multi-layer perceptron classifier for behavior analysis to estimate parkinson's disease severity using smartphones," IEEE Access, vol. 6, pp. 36 825–36 833, 2018.
[29] A. Natekin and A. Knoll, "Gradient boosting machines, a tutorial," Frontiers in neurorobotics, vol. 7, p. 21, 2013.
[30] R. Rojas et al., "Adaboost and the super bowl of classifiers a tutorial introduction to adaptive boosting," Freie University, Berlin, Tech. Rep, 2009.
[31] M. M. Ahamad, S. Aktar, M. J. Uddin, M. Rashed-Al-Mahfuz, A. Azad, S. Uddin, S. A. Alyami, I. H. Sarker, A. Khan, P. Li`o et al., "Adverse effects of covid-19 vaccination: machine learning and statistical approach to identify and classify incidences of morbidity and postvaccination reactogenicity," in Healthcare, vol. 11, no. 1. MDPI, 2022, p. 31.
[32] T. Akter, M. H. Ali, M. I. Khan, M. S. Satu, M. J. Uddin, S. A. Alyami, S. Ali, A. Azad, and M. A. Moni, "Improved transfer-learning-based facial recognition framework to detect autistic children at an early stage," Brain Sciences, vol. 11, no. 6, p. 734, 2021.