Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 08, No. 03, September 2022.

RISK FACTOR IDENTIFICATION FOR STROKE PROGNOSIS USING MACHINE-LEARNING ALGORITHMS

Tanvir Ahammad

(Received: 16-May-2022, Revised: 6-Jul.-2022, Accepted: 7-Jul.-2022)

ABSTRACT

Stroke is a life-threatening condition causing the second-leading number of deaths worldwide. It is a challenging problem in the public-health domain of the 21st century to healthcare professionals and researchers. So, proper monitoring of stroke can prevent and reduce its severity. Risk-factor analysis is one of the promising approaches for identifying the presence of stroke disease. Numerous researches have focused on forecasting strokes in patients. The majority had a good accuracy ratio, around 90%, on the publicly available datasets. Combining several pre-processing tasks can considerably increase the quality of classifiers, an area of research need. Additionally, researchers should pinpoint the major risk factors for stroke disease and use advanced classifiers to forecast the likelihood of stroke. This article presents an enhanced approach for identifying the potential risk factors and predicting the incidence of stroke on a publicly available clinical dataset. The method considers and resolves significant gaps in previous studies. It incorporates ten classification models, including advanced boosting classifiers, to detect the presence of stroke. The performance of the classifiers is analyzed on all possible subsets of attribute/feature selections concerning five metrics to find the best-performing algorithms. The experimental results demonstrate that the proposed approach achieved the best accuracy on all feature classifications. Overall, this study's main achievement is obtaining a higher percentage (97% accuracy using boosting classifiers) of stroke prognosis than state-of-the-art approaches to stroke dataset. Hence, physicians can use gradient and ensemble boosting-tree-based models that are most suitable for predicting patients' strokes in the real world. Moreover, this investigation reveals that age, heart disease, glucose level, hypertension and marital status are the most significant risk factors. At the same time, the remaining attributes are also essential to obtaining the best performance.

KEYWORDS

Stroke prediction, Machine learning, Classification, Feature selection, Stroke risk factors, Healthcare.

1. INTRODUCTION

Stroke is the second biggest life-threatening disease in the world. It has caused about 11% of deaths worldwide from 2000-2019 [1]-[2]. According to WHO's classifications[1], it is the fourth leading cause of death in low-income countries, the second in lower-middle-income and upper-middle-income countries and the third in high-income countries. In the United States, a stroke happens every 40 seconds, killing one person every (i.e., 1 of every 20 deaths) 3 minutes and 33 seconds [3]. In addition, more than 795,000 people have a stroke, approximately 610,000 of these are the first cases and even a stroke is expected to have a severe long-term disability.

When blood that flows to the brain reduces, there is a lack of nutrients in the cells, quickly leading to cell dysfunction. The symptoms of stroke appear when any part of the brain fails. For example, a core area in a stroke is where blood is almost completely blocked and the cells die within five minutes [4]. There are many reasons for a stroke occurring in a person. These include age, hypertension, diabetes, heart failure, ethnicity, heredity, physical inactivity and peripheral artery disease [5]-[6]. A stroke generally increases with age, but can occur at any period. In 2014, 38% of people hospitalized for a stroke were under 65 and 30% of patients aged 85 and above died from stroke [7].

Stroke is a curable condition that can be considerably reduced in severity if diagnosed or anticipated early. In various investigations and clinical trials, several risk factors for stroke have been found [8]. Proper management and controlled trials, such as preventing high blood pressure, avoiding smoking and

[1] https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

T. Ahammad is with Department of Computer Science & Engineering, Jagannath University, Bangladesh. ORCID: 0000-0002-0218-1365
Email: tanvir@cse.jnu.ac.bd

alcohol, controlling diabetes, lowering cholesterol, surgery for carotid stenosis and maintaining height and weight adjustment, can reduce the risk of stroke [6], [9]-[11]. Moreover, other diets and mobile technology effectively prevent initial stroke in combination with salt restriction. On the other hand, health agencies can build secondary preventive measures for stroke [12]. Therefore, providing insightful information about stroke prognosis through research from patients' medical history as a tertiary action with personal, medical and secondary management is essential in today's world context.

The inclusion of Artificial Intelligence (AI), especially Machine Learning Algorithm (MLA), has changed the traditional healthcare paradigm into an intelligent health service system. MLAs find hidden patterns from a health data repository and establish models to predict disease for making data-driven healthcare decisions [13]. Predicting the sign of stroke using an MLA is a promising task. Two potential procedures, such as CT scan/MRI and risk factor analysis, can easily monitor the incidence of stroke. Brain imaging can detect real-time stroke on bio-signal data more accurately than risk analysis, as shown in [14]. However, the main drawback of this CT/MRI approach is not anticipating the probability of other diseases (e.g. cardiovascular disease or diabetes). In addition, this approach cannot identify the correlations between the risk factors or the most influential feature importance. Therefore, predictive analysis of risk factors is a prominent approach to observing the likelihood of stroke symptoms.

In recent years, numerous studies have identified predictive analysis of stroke disease using the MLA approach based on the publicly available stroke datasets. In 2019, H. Ahmed et al. [15] examined the presence of stroke with 90% accuracy. Then, P. Govindarajan et al. [16] demonstrated the stroke prognosis for only 507 patients with an accuracy of 96% in 2020. Afterward, in 2021, A. Kumar [17] and T. Tazin et al. [18] showed how to detect stroke using different MLAs with 82% and 95% accuracy, respectively. Finally, in 2022, S. Dev et al. [19] proposed an approach for predictive analysis of stroke risk factors and found four attributes that showed the best accuracy rate, around 80% only. However, the research gap in these studies includes choosing the combination of various pre-processing tasks to improve the quality of classifiers significantly. Moreover, these studies should identify the key risk factors responsible for stroke disease and predict the likelihood of stroke with high-performance MLA models.

This article presents an enhanced approach for identifying possible risk parameters of stroke and predicting its presence in publicly available stroke datasets. First, this approach collects and loads the clinical data containing patients' diagnoses with stroke disease. Next, the dataset is pre-processed and transformed into a standard format to improve the performance of the approach. Then, the best-fit features are identified to find the key risk factors of a stroke. Afterward, ten classification models are used to predict the presence of stroke. Finally, the performance of the classifiers is recorded and compared in terms of accuracy, F1-score, precision, recall and auc_roc to find the best-performing algorithms. The experimental results revealed that Extreme Gradient Boosting (XGB), Gradient Boosting Machine (LGBM), Category Boosting Classifier (CBC) and Adaptive Boosting Classifier (ABC) showed the highest accuracy (97%) of stroke prediction with all feature classifications. In addition, patients' age, cardiovascular disease, diabetes, hypertension and marital status are the most significant risk factors. Overall, the proposed approach demonstrated a higher accuracy of 97% compared to the Machine Learning (ML) models used in existing research [15]-[19] on the same publicly available stroke dataset.

The main contribution of this research is to present an enhanced method that identifies the critical risk factors of stroke and then predicts the possibility of receiving a stroke. In sum, the contributions of this article are as follows:

- Choosing a combination of various pre-processing tasks to improve the quality of classifiers significantly;
- Identifying the best-fit features (risk factors) of the stroke dataset to feed into ML models;
- Ranking key risk factors that are responsible for stroke;
- Achieving the highest percentage of accuracy using advanced gradient boosting-based classifiers that can be the most appropriate ones for physicians to prognose stroke based on the patients' medical history in the real world.

The rest of the paper is structured as follows. First, Section 2 discusses the background of the research. Then, Section 3 represents the description of the dataset, materials and proposed methodology used in this study. Next, Section 4, entitled results and discussion shows the discussion and analysis of the

experimental results. Finally, the paper concludes by suggesting future directions in Section 5, entitled conclusions.

2. RELATED WORKS

Stroke is one of the highest reasons for death globally and causes mental and functioning concerns. So, extensive research is required to find ways to monitor, prevent and treat stroke. The benefits of artificial neural networks (ANNs) and other MLAs have been noticed in the literature to diagnose or predict the occurrence of stroke in a patient [20]. For example, D. Shanthi et al. [21] used an ANN to predict Thromboembolic strokes caused by a thrombus (blood clot) that forms in the arteries delivering blood to the brain. They used stroke data from healthcare datasets with eight attributes of patients. Their investigation improved accuracy to 89%. Although their approach emphasizes prediction accuracy, it is challenging to identify risk factors with a higher performance.

A significant number of research studies have been conducted in the literature to anticipate the possibility of stroke in the human brain using machine-learning (ML) models. First, H. Ahmed et al. [15] used MLAs to identify the presence (90% accuracy) of stroke on the Apache Spark, an open-source distributed processing system used for Big Data workload. Then, G. Sailasya and G. L A. Kumari [17] examined a similar type of study. They compared traditional ML methods and obtained 82% accuracy using the Naive Bayes classifier. Finally, T. Tazin et al. [18] examined how to detect the probability of stroke with a higher accuracy (95%) than in previous studies. However, their methodologies require normalization before feature selection and rank physiological factors to detect strokes more accurately.

Medical imaging and bio-signal analysis are promising research methods to monitor stroke as early as possible. For example, J. Yu et al. [14] developed an AI-based real-time stroke-prediction system on patients' EMG (electromyography, measuring muscle response or activity) bio-signals. They collected and measured real-time left and right biceps femoris (thigh muscle located in the posterior portion or back) and gastrocnemius muscles (large back muscles or back part of the lower leg of humans) from health monitoring devices at 1500 Hz. Their experimental results revealed that the proposed approach could be an alternative to stroke detection with a low-cost diagnosis. However, though their system effectively detects early stroke, it overlooks the risk factors in predicting pre-stroke conditions, because risk-factor analysis shows which parameters are responsible for stroke in advance.

Anticipating the likelihood of a similar type of stroke is a robust approach. This investigation was conducted by L. Amini et al. [22]. They collected 50 different attributes of healthy and sick patients in two hospitals from 2010 to 2011. They used data-mining techniques to classify high-risk groups of patients' history of cardiovascular disease, hyperlipidemia, diabetes, smoking and alcohol consumption. In continuation of predictive stroke analysis, C. Colaka, E. Karaman and M. G. Turtay [23] proposed knowledge discovery from data (KDD) methods on nine attributes. They used 297 data samples (130 sick and 167 healthy persons) and showed the highest accuracy, approximately 93%, by using an ANN. Similarly, L. I. Santos et al. [24] used a decision tree-based ML model to predict the stroke outcomes for the imbalance dataset. They obtained 70% and 78% accuracy to show the significance of their study with the state-of-the-art approach. However, these investigations incorporated small and limited data samples, resulting in poor approximation. In addition, the most significant risk issues of stroke were unrevealed in these studies.

Predictive analysis of risk factors is a promising research approach for stroke disease. S. Dev et al. [19] introduced a method that analyzes and identifies potential physiological attributes related to stroke disease. Using a perceptron neural network, they found four critical risk factors that exhibited the best performance, about 80% accuracy rate. Although they examined the significant risk factors, the accuracy of their approach could improve by choosing a combination of different pre-processing tasks. Besides, they reduced many critical attributes that could give a better predictable rate. D. Paikaray and A. K. Mehta [25] examined a similar approach to predicting stroke before its occurrence. They used nine different ML models in their experiment. They achieved a promising result with an accuracy of 95.10%. Although their experimental result was better than that of S. Dev et al. [19], they could not discover the possible risk factors that may cause a patient's stroke.

Analyzing the effects of risk factors on stroke monitoring is an emerging research trend. For example,

P. Songram and C. Jareanpon [26] showed that people could prevent stroke by predicting its risk factors. They identified seven health issues for a stroke. Using the decision-tree approach, they achieved 74.29% of accuracy in the F1-score. Likewise, R S Jeena and A Sukeshkumar [27] developed a stroke risk-assessment model by detecting relevant predictors. They categorized the risk factors into low-risk, medium-risk and high-risk factors. In addition, Fang et al. [28] used an integrated ML methodology to select the essential features for stroke prognosis. They chose twenty-three parts to predict the acute stroke with an accuracy of 69% only. While these researches have shown the potential of feature selection related to stroke, they have demonstrated a lower accuracy than that obtained in our proposed approach. Moreover, they have identified many attributes that can be difficult to correlate with the probable stroke signs in patients. Therefore, an enhanced approach for identifying ranking-based stroke risk factors and predicting stroke incidence is essential as an alternative to the existing methodologies.

3. MATERIALS AND METHODS

This section represents the description of the stroke dataset, the methodology and the analysis of results from the ten classifiers used in this research.

3.1 Dataset

The dataset used in this research was related to stroke disease. The dataset indicates whether a patient is likely to suffer a stroke based on different parameters, such as gender, age, various diseases and physical conditions. The dataset was publically accessible on the Kaggle[2] online community platform. It contains 12 different attributes and around 5,110 records or rows of data. Each row comprises relevant patient medical history information, as shown in Table 1. The dataset has 201 missing values in BMI attribute and 1,544 in smoking_status attribute. Moreover, it is a binary classification with a strongly imbalanced dataset involving 4,861class label 0 and 249 class label 1.

Table 1. Summary of stroke dataset.

Attribute Name	Attribute Description
id	Unique identifier of the patient
gender	"Male", "Female" or "Other"
age	Age of the patient
hypertension	0: if the patient doesn't have hypertension; 1: if the patient has hypertension
heart_disease	0: if the patient doesn't have any heart disease; 1: if the patient has a heart disease
ever_married	"No" or "Yes"
work_type	"children", "Gov. job", "Never worked", "Private" or "Self-employed"
Residence_type	"Rural" or "Urban"
avg_glucose_level	Average glucose level (mg/dL) in blood after meal
BMI	Body Mass Index
smoking_status	"formerly smoked", "never smoked", "smokes" or "Unknown". Unknown status indicates that the information is unavailable for this patient
stroke	1: if the patient had a stroke; 0: if the patient had no stroke. (It is the class label attribute)

3.2 Machine-learning Classification Models

This study focuses on identifying risk factors for the binary classification of stroke disease. We employed ten different classification models from various fields of machine learning [29], as shown in Table 2. The models consist of three-tree based methods, including Random Forest (RF) [30], XGB [31] and Decision Tree (DT) [32]; three ensemble boosting approaches, such as LGBM [33], CBC [34] and ABC [35]; one Support Vector Machine (SVM) [36] and neural network-based Multilayer Perceptron (MLP) [37]; one K-Nearest Neighbor (KNN) [38] and linear statistical-based approach Logistic Regression (LR) [39]. The classifiers are evaluated independently using different performance metrics and the outcomes are recorded for further analysis.

[2] https://www.kaggle.com/fedesoriano/stroke-prediction-dataset

Table 2. Review of different classification models.

Classifiers	Description	Strengths	Weaknesses
RF	It performs random selection of features to build different decision trees and applies voting policy to obtain the final result.	It is efficient for classification problems with numerical and categorical features.	Making predictions is quite slow once they are trained.
XGB	It is an ensemble method that supports various functions, such as classification, regression and ranking.	It is computationally efficient and predicts the result with high accuracy.	It is slow for a large number of classes.
DT	DT is a popular classification approach. It constructs tree data structure, where an internal node denotes the test on an attribute and a leaf node determines the class	It is simple and fast and has good accuracy depending on dataset.	It takes a long time when training the dataset and deals with memory unavailability with respect to large data.
LightGBM	It is a gradient boosting algorithm for classification problems.	It has a faster training speed, higher efficiency and a lower memory utilization. It can also handle large-scale data.	It is prone to overfitting; it can easily overfit small data.
CatBoost	It is a gradient boosting algorithm that predicts with a less amount of time for unseen data.	It is very useful in categorical data without explicit pre- processing.	It needs to construct deep decision trees in order to get better accuracy.
AdaBoost	It is an ensemble boosting classifier by the combination of multiple classifier models to increase accuracy.	It provides high-accuracy outcomes.	It does not perform well with noisy data and outliers.
SVM	It performs classification by setting the hyper-plane that distinguishes between two class labels.	It works very well with a strong margin of segregation for high-dimensional spaces.	It is slow with large datasets.
MLP	It is a feedforward neural network-based classifier, which learns on the non-linear functions for complex data.	It is very powerful and works with high accuracy for both small and large datasets.	The training process is time-consuming to determine the exact parameters for obtaining expected performance.
KNN	It solves classification and regression problems by setting the K-neighbors.	It is a non-parametric algorithm, which implies that certain assumptions must be met in order for it to work.	KNN requires to find tune K-value that may be challenging for large dataset.
LR	It is a statistical model that solves classification and regression problems.	It is easier to extend additional classes and a probabilistic view of class predictions.	The assumption of linearity between the dependent and independent variables is a key constraint of LR.

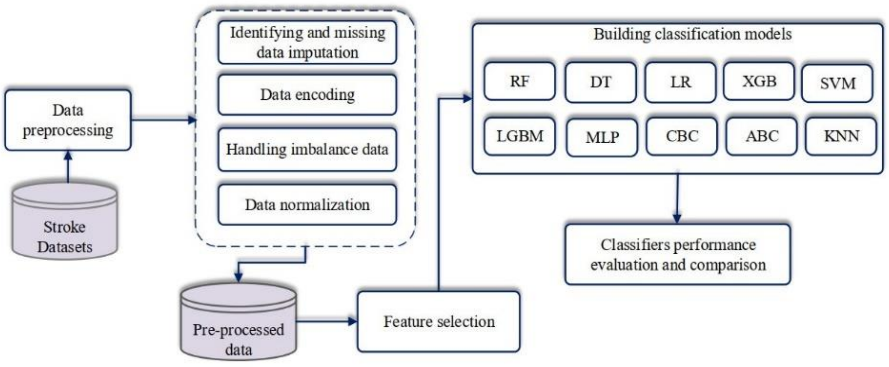


Figure 1. Methodology for the identification of risk factors and prediction of stroke disease.

3.3 Methodology

The proposed approach of this article is a refinement of several methodologies, such as [15]-[18], [40] in the context of stroke-disease analysis. It incorporates six-stage processing phases for identifying and predicting the main risk factors of stroke disease, as illustrated in Figure 1. The stages are stated as follows:

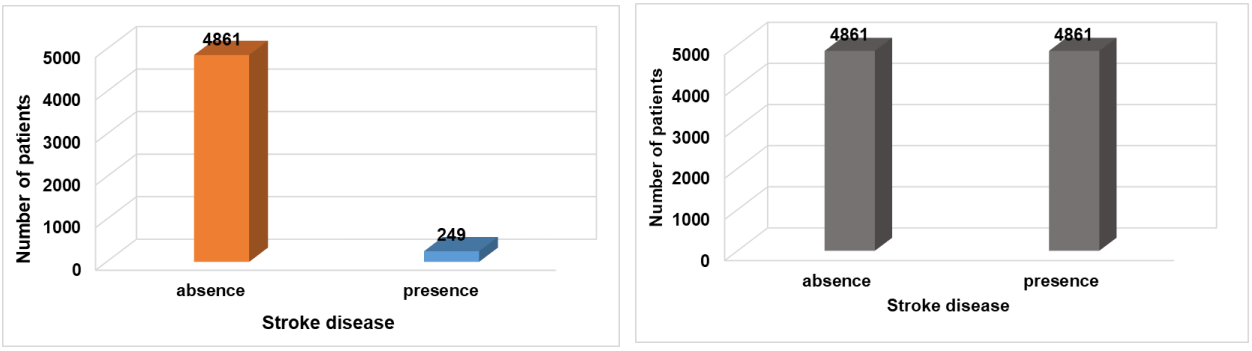
1) **Collecting and loading dataset**. Select and load the target dataset from the health data archive containing patients' medical records related to stroke disease. Since this paper focuses on a publicly accessible stroke dataset, the dataset is first loaded into the program for analysis.

2) **Data pre-processing**. Before feeding target data into the classifiers, this step involves analyzing the datasets to find any inconsistency (e.g. missing values, noise or extreme values). Moreover, this stage transforms the data into a well-formed format to enhance the performance of classifiers. As stated earlier, the stroke dataset has twelve attributes, where *bmi* and *smoking_status* contain missing values. So, these missing values are predicted and replaced by certain values analogous to non-missing data, called missing-data imputation. In continuation of the pre-processing data phases, the columns or attributes containing the categorical or text data are encoded to numeric values so that the ML models can process them properly.

Furthermore, the stroke dataset is rigorously checked to determine the class-label imbalances. As there are a total of 5,110 data records where 249 of them indicate the incidence of a stroke and 4,861 rows indicate the absence of a stroke (Figure 2a), these disparities (imbalance ratio 20:1) may lead many ML models to low predictive accuracy (e.g. metrics like precision and recall) with infrequent class. Consequently, the unbalanced data must be dealt with first to obtain an efficient model. Improved Synthetic Minority Over-sampling Technique (ISMOTE) [41] is possibly a novel approach that selects new samples nearest to the minority-class neighbors; it then balances the minority-class with the majority-class instances. Figure 2 shows the ratio of data samples in the class distribution used in this study. In the final pre-processing stage, the dataset is changed to a standard scale using z-score normalization[3], as shown in Equation 1.

$$z_{ij} = \frac{f_{ij} - m_i}{sd_i} \tag{1}$$

where, z_{ij}: normalized score j^{th} value of i^{th} feature, f_{ij}: j^{th} value of i^{th} feature, m_i: mean value of i^{th} feature and sd_i: standard deviation of i^{th} feature.



(a) Imbalanced distribution (b) Balanced distribution

Figure 2. Proportion of samples in the number of stroke absence to the number of stroke incidence.

3) **Archiving pre-processed data**. Different preprocessing methods convert the raw data into various understandable formats. For example, various encoding schemes or data-normalization techniques generate distinct data values that may affect the performance of ML models. So, storing all of these formats in a data archive or data files is necessary. In other words, archived data allows ML algorithms to get comprehensible dataset features during the training or learning.

4) **Feature selection**. This stage is essential for deciding the best-fit features for the classifiers' best performance. Firstly, use all the attributes in the target stroke dataset to build and measure the

[3] https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.zscore.html

accuracy of the classification models. Next, calculate feature-importance scores using either tree- based classifiers or correlation coefficient; select the top-most n − 1, n − 2, n − 3, ..., 1 features and find classifiers' accuracy, respectively. Finally, a voting procedure is applied to get the best accuracy among all the choices of feature selection. In other words, all combinations of attributes in the dataset are used in the ML models and then recorded as the best-fit feature selection classification models.

5) **Model building**. As stated earlier, ten classification algorithms are used in this study to show the performance of the proposed approach. Therefore, the ratio of data samples used for training and testing purposes is 80:20. Since many ML models have different parameters/variables that control the model's performance, the parameters can not directly predict (e.g. KNN, MLP) from data to obtain the desired accuracy. So, we need to tune the parameters. However, we train all ML models by setting different parameters, grid searching or random searching of model hyperparameters to be learned from data for the best accuracy.

6) **Applying evaluation metrics and performance comparison**. It is the final stage of the proposed methodology. After building the classification models, analyze them using five metrics: accuracy, F1-score, precision, recall and roc_auc (compute area under the receiver operating characteristic curve from prediction scores). Then, the performances of the classifiers are compared based on these criteria. The metrics are defined as follows:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{2}$$

$$Precision = \frac{TP}{TP+FP} \tag{3}$$

$$Sensitivity = Recall = \frac{TP}{TP+FN} \tag{4}$$

$$F1 = \frac{2*Precision*Recall}{Precision+Recall} = \frac{2*TP}{2*TP+FP+FN} \tag{5}$$

$$Specificity = \frac{TN}{FP+TN} \tag{6}$$

$$ROC_{AUC} = Sensitivity - (1 - Specificity) \tag{7}$$

where: $TP = TruePositive$: ML model correctly predicts that a patient has stroke disease.
$TN = TrueNegative$: ML model correctly predicts that a patient has no stroke disease.
$FP = FalsePositive$: ML model incorrectly predicts that a patient has stroke disease.
$FN = FalseNegative$: ML model incorrectly predicts that a patient has no stroke disease.

4. RESULTS AND DISCUSSION

4.1 Exploratory Analysis of Dataset

Exploratory data analysis is necessary to analyze the presence of stroke disease. It is the process of discovering patterns and irregularities and checking premises with the help of summary statistics and visual representations before applying ML models. The stroke dataset used in this study comprises 11 feature attributes and one attribute containing two class labels, shown in Table 1. We did not consider the attribute, *id*, in our analysis, because it does not influence the performance of the classifiers. Since most features are categorical, it is easy to find patterns in the medical history responsible for a patient's stroke.

Figure 3 depicts various distributions of categorical features concerning stroke. For example, Figure 3a illustrates three attributes the value of which is in binary type. Looking closer at this figure, we can see that 13.25% of patients who suffer from hypertension have stroke disease, while the number is below 4% for stroke patients with no hypertension. On the other hand, the number of stroke patients who got married is three times more than those not married. Besides, the ratio of people having a stroke with heart disease is 17% which is more than four times higher than the patients with no heart disease (4.18%). Therefore, heart-disease patients are likely to have a higher risk of stroke than stroke patients with hypertension and those ever married, as shown in Figure 3a. Turning to the attribute work type (Figure 3b), we observe that self-employees (8%) suffer a slightly higher percentage of

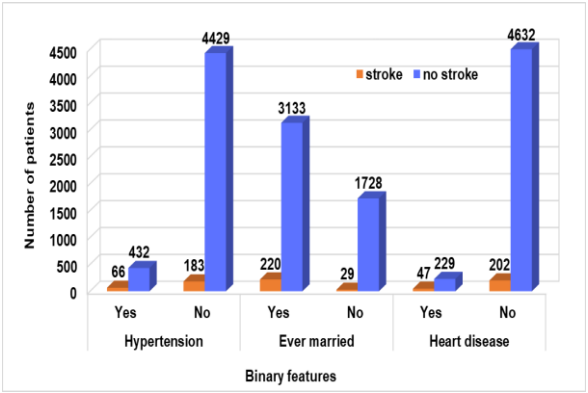
strokes than government (5%) and private (5.09%) workers. Moreover, the stroke patient rate trend seems to be almost similar in residence type and gender groups.

In the stroke dataset, numerical data with missing values can be detected in some entities, as demonstrated in Figure 4. However, to visualize the stroke disease trend, first, numerical features are converted into categorical ones based on predefined rule-based approaches. For instance, the attribute, age, is grouped into four clusters based on reference [42], as stated below:

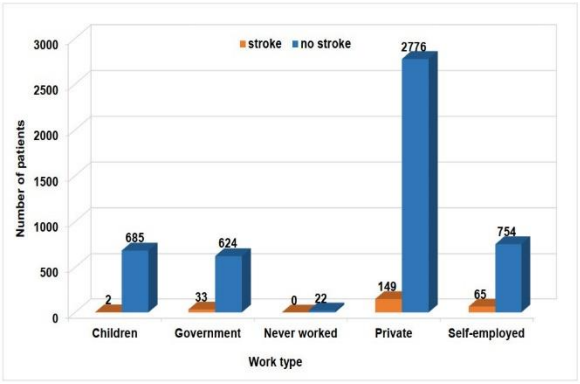
- Child: 0-12 years;
- Adolescent: 13-18 years;
- Adult: 19-59 years;
- Senior Adult: 60 years and above.

Then, the feature bmi containing null/missing values is organized into several categories following the Centers for Disease Control and Prevention (CDC) interpretation[4], defined as follows:

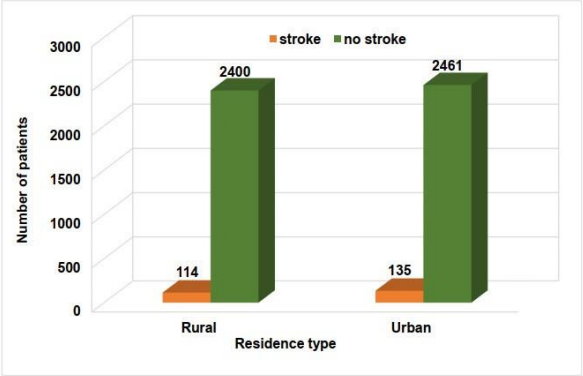
- below 18.5: underweight;
- between 18.5 and 24.9: healthy weight;
- between 25 and 29.9: overweight;
- between 30 and 39.9: obese;
- missing values: null values.



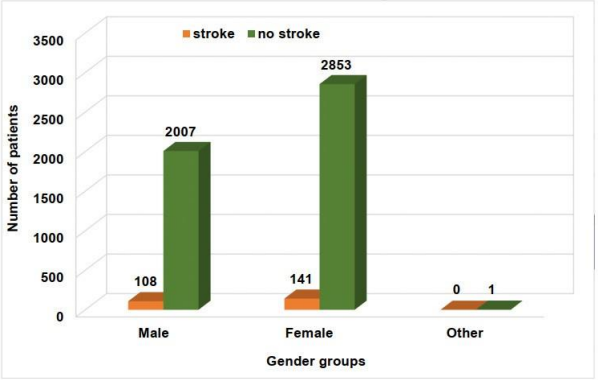
(a) Binary features



(b) Work type status



(c) Residence type



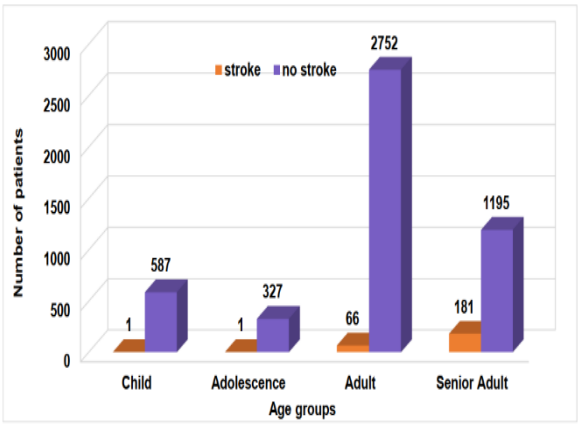
(d) Gender

Figure 3. Representation of categorical attributes.

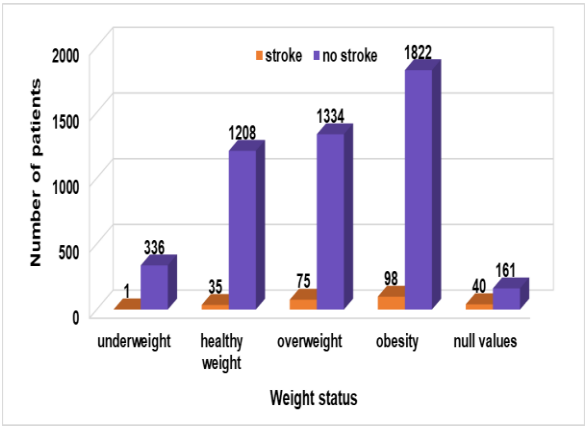
Next, the average glucose level measured after the meal is grouped into three types using CDC report[5], as indicated below:

- Diabetes: 200 mg/dL or above;
- Pre-diabetes: 140 to 199 mg/dL;
- Normal: 140 mg/dL or less.

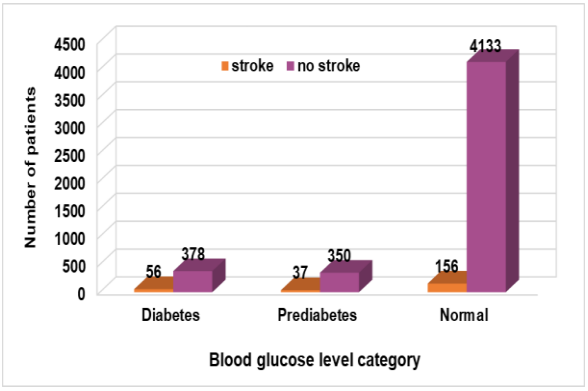
[4] https://www.cdc.gov/healthyweight/assessing/bmi/adult_bmi/
[5] https://www.cdc.gov/diabetes/basics/getting-tested.html



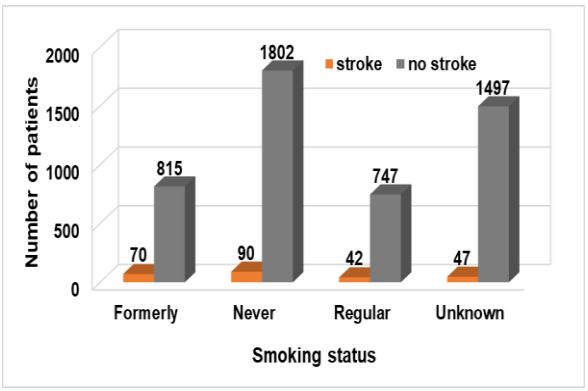
a. Age



b. BMI



c. Average glucose level



d. Smoking status

Figure 4. Representation of numerical features as categorical.

Finally, three categorical values were specified in the original dataset in the smoking status attribute, except for missing values that were later grouped into "Unknown" status, as indicated in Figure 4d. However, if we look at Figure 4, we can notice that senior adults are more likely to suffer from stroke (13%) than other age groups (Figure 4a). Likewise, stroke patients who are overweight and obese have higher numbers than others. In addition, nearly 20% of patients were found to be more likely to suffer from stroke in the missing BMI values, as shown in Figure 4b. Most importantly, the average glucose level in blood is another feature that reveals a noticeable portion, nine and a half percent, of diabetic patients who suffer from stroke (Figure 4c).

Identifying what kind of risk factor can predict a stroke is an important step. In other words, before applying ML models, feature correlation is a practical approach to determine the closeness between features and the target class. This method groups the related health information (e.g. age, bmi or smoking status) to reduce personal data processing, eliminate less essential data and improve the performance of ML models. Figure 5 illustrates the relationship of features with the stroke attribute. It shows that most features are positively correlated with the target variable other than gender and smoking status.

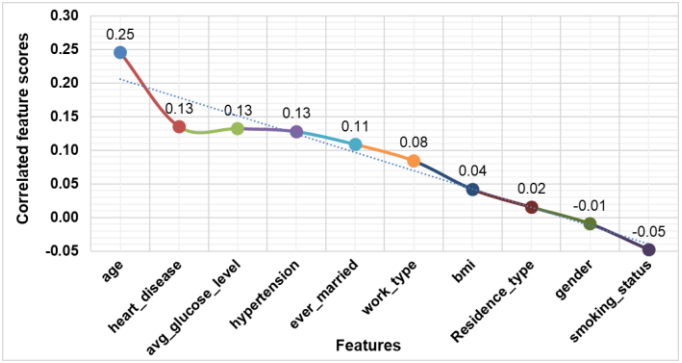
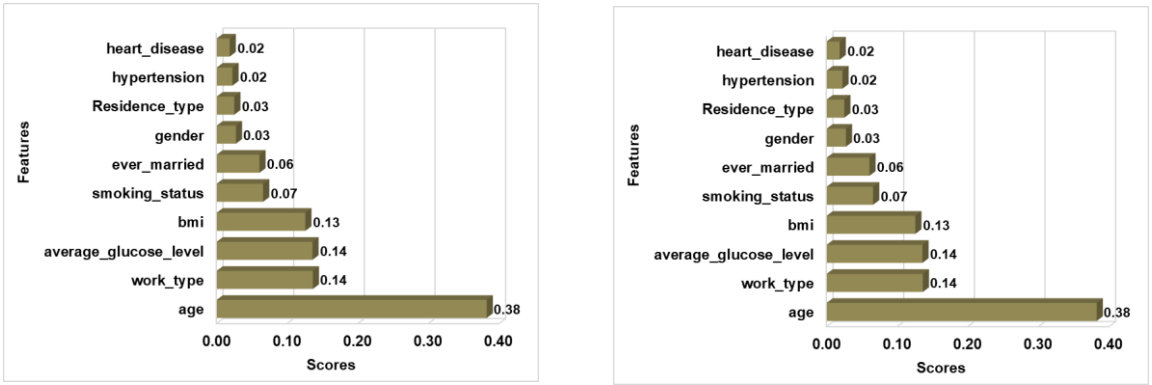


Figure 5. Feature correlation with the *stroke* attribute.

4.2 Performance Evaluation of Classifiers

The performance of the ML classification models is measured using all features and top-most main attributes in the dataset, as mentioned in the section containing the research methodology. The top-most important attributes are selected and ranked according to feature-importance scores by RF and XGB classifiers (Figure 6). A higher score implies that the specific feature will significantly impact the classification model. However, if we observe closely Figure 6a and Figure 6b, we can see that the two approaches, RF and XGB, generated different feature rankings despite a similar tree-based paradigm. In addition, three are the most common features in the top five scores. So, we applied these top-most features in classification. We also considered listing key attributes from correlated feature scores.

As said previously, first, we tested and evaluated the performance of ten classifiers on all features from the stroke dataset. Table 3 shows the results from the analysis in the experiment. We can see that gradient boosting-based classifiers, including XGB, LGBM, CBC and ABC, showed the best accuracy (97%). On the other hand, the LR model gave the lowest result in terms of accuracy (80%), F1-score (86%), precision (81%), recall (86%) and roc_auc (86%), respectively. Besides, RF and MLP showed the second-highest accuracy, whereas SVM and K-NN performed similarly. However, Figure 7 depicts the performance analysis of all classifiers used in this study on the stroke dataset. It reveals that the performance rate of most classifiers slightly fluctuates between 94% and 97% except for the LR model. Overall, in all feature selections, gradient and ensemble boosting-tree-based ML models exhibit higher performance (97%) for stroke-disease detection compared to the ML models used in the existing research studies [15]-[19].



a. RF for feature importance b. XGB for feature importance

Figure 6. Ranking of features for stroke-disease prediction.

Table 3. Train-test performance evaluation of classifiers on all feature sets.

	Performance-evaluation metrics				
Classifier	**Accuracy (%)**	**F1-score (%)**	**Precision (%)**	**Recall (%)**	**ROC_AUC (%)**
RF	96.4	96	96	96	96
XGB	**97**	**97**	**97**	**97**	**97**
SVM	95	95	95	95	95
DT	94	94	94	94	94
LGBM	**97**	**97**	**97**	**97**	**97**
CBC	**97**	**97**	**97**	**97**	**97**
ABC	**97**	**97**	**97**	**97**	**97**
MLP	96	96	96	96	96
K-NN	95	95	95	95	95
LR	80	81	81	86	86

Top-most attribute selection is another benchmark for predicting stroke disease. So, we selected different subsets (except null and all feature sets) of attributes on the target dataset. Based on the order indicated in Figure 5, we found the seven most essential feature classification presented the best accuracy. Table 4 summarizes the obtained results on the performance of ten classifiers. It shows that XGB, LGBM and CBC have achieved the highest performance, similar to the results in [18] but better than the results in works [15]-[17]. Figure 8 illustrates the visual representation of stroke prediction accomplishment on the top-most seven-feature dataset. We can see that the performance rate starts with a rising trend from RF to XGB.

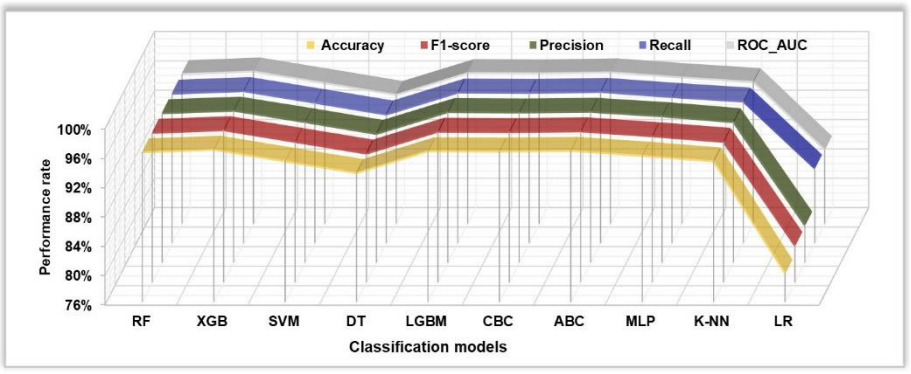


Figure 7. Train-test performance analysis on stroke dataset for all features.

Table 4. Train-test performance evaluation of classifiers on top-most feature sets.

	Performance evaluation metrics				
Classifier	**Accuracy (%)**	**F1-score (%)**	**Precision (%)**	**Recall (%)**	**ROC_AUC (%)**
RF	94	94	94	94	94
XGB	**96**	**96**	**96**	**96**	**96**
SVM	91	91	91	91	91
DT	92	92	92	92	92
LGBM	**96**	**96**	**96**	**96**	**96**
CBC	**96**	**96**	**96**	**96**	**96**
ABC	94	94	94	94	94
MLP	92	92	93	92	92
K-NN	92	92	92	92	92
LR	78	79	81	83	85

Then, it falls and remains constant for SVM and DT; it increases again sharply and reaches the highest peak at 96% for LGBM and CBC; afterward, it presents a downward trend and reaches the lowest point (78%). Therefore, LR was the lowest-performing model, whereas XGB, LGBM and CBC were the best-performing models on the top-most seven features.

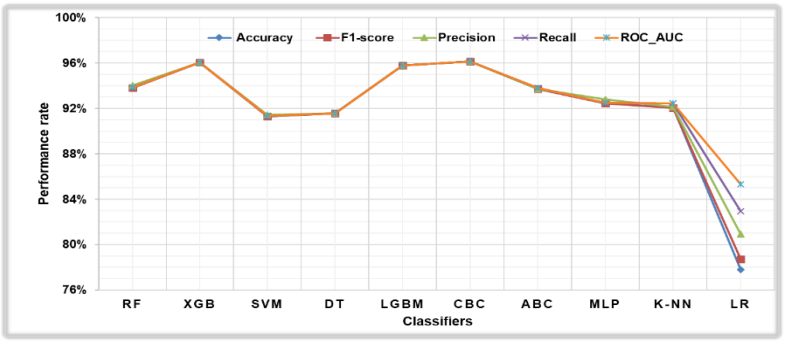


Figure 8. Performance from the top-most seven features.

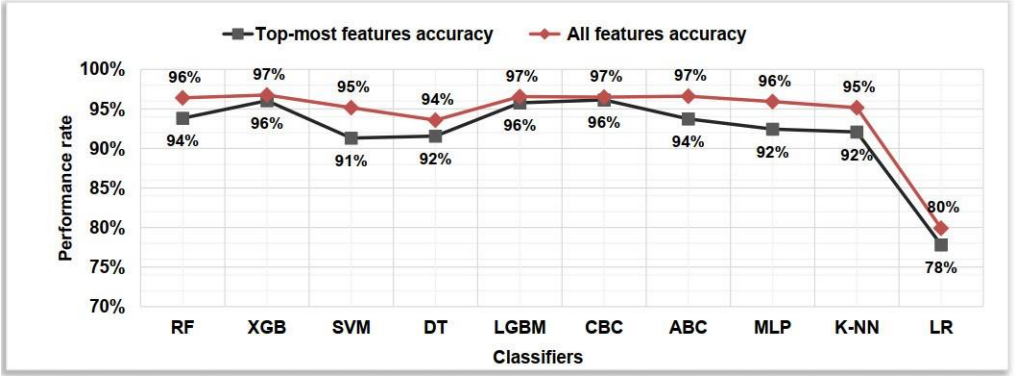


Figure 9. Comparison of classifiers' performance on different feature selections.

The implication of this research lies in finding the best performance from the ten ML classifiers to determine the subset of attributes for stroke-disease detection and prevention. We identified two test cases, including all features and the top-most seven attributes. We also separately compared the results obtained from ML classifiers on these two types of attribute selection. As a result, the attribute/feature selections show different patterns, as depicted in Figure 9. Looking at the graph, we can observe that the results obtained from the ten ML models demonstrate an average of 2.1% better accuracy for all features than the top-most seven features. In addition, SVM, ABC, MLP and K-NN models revealed the most asymmetry differences (3%-4% accuracy deviations) in these classifications. Therefore, considering our analysis, we conclude that the ML classification models performed sufficiently on all attributes of the stroke dataset. In other words, the classifiers do not perform considerably well by selecting different subsets of attributes rather than all (ten features in the stroke dataset).

As stated above, this article presents an enhanced method that performs well for all feature classifications. So, we compare the best-performing results of this article with those of other approaches. Table 5 represents a summary of the classification accuracy of different methods in several studies, including this article. The table shows that RF and DT classifiers were common in all papers for stroke prediction. We can also see from the table that every classification model used in this article exhibits a better accuracy rate than others, despite the similarity in [18] for the DT classifier. One significant point is that none of the existing research studies used gradient and ensemble boosting-tree-based (except [16]) classifiers. In other words, gradient and ensemble boosting-tree-based ML models showed the highest percentage (97%) of stroke prognosis on the same stroke dataset used in previous studies. In conclusion, this article outperformed previous works [15]-[19] using the methodology of six-stage processing phases.

Table 5. Performance comparison of classifiers in different studies.

ML models	Ref. [15]	Ref. [16]	Ref. [17]	Ref. [18]	Ref. [19]	This article
RF	90%	90.9%	72%	96%	75%	96.4%
XGB						**97%**
SVM	77%	91.5%	78.6%		68%	95%
DT	79%	90.7%	77.5%	94%	74%	94%
LGBM						**97%**
CBC						**97%**
ABC		91.5%				**97%**
MLP		95%			80%	96%
K-NN			77.4%			95%
LR	77%	90.6%	77.5%	79%		80%

We found in our analysis that all the medical records presented in the stroke dataset are essential for the stroke disease of the patients. We also ranked the patients' health history in terms of feature importance. Age, heart disease, diabetes, high blood pressure and marital status are considered the most critical factors for a patient's stroke. Besides, the type of workplace and the ratio of height and people's weight are also significant factors. Although patients' residential environment, gender type or smoking habits demonstrated less importance in the analysis, we can not ignore them to detect and prevent stroke.

5. CONCLUSIONS

Stroke is one of the deadly diseases at the global level. Healthcare providers should identify its causes and take preventive measures as early as possible to avoid complications. However, it is a critically challenging problem for healthcare professionals and researchers. This research focuses on an enhanced approach for identifying the risk factors and detecting the presence of stroke for clinical stroke datasets using ML models to solve the issues. First, we analyzed the dataset to find any inconsistencies and discover hidden patterns. Next, we selected different subsets of features to identify and rank stroke risk factors for classification. Then, we relied on ten ML classification models to predict the presence of stroke using the train-test splitting technique. Finally, we evaluated the performance of classifiers using five metrics, including accuracy, precision, F1-score, recall and roc_auc.

We compared the results of the ten ML models in all features and the top-most seven feature

classifications. We observed that the classifiers performed differently (2.1% divergence) on these two feature selections. XGB, LGBM, CBC and ABC models showed the highest accuracy rates in all feature (ten attributes) classification, whereas XGB, LGBM and CBC were the most accurate ones for the top-most seven attributes. We also showed that every classification model used in this article exhibits a higher accuracy rate than other studies in most of the cases. Overall, we obtained a higher accuracy of 97% than the existing approaches on the stroke dataset using gradient and ensemble boosting-tree-based classifiers. Therefore, healthcare providers can use these classifiers that are the most suited for predicting stroke based on the medical history of a patient in the real world.

Furthermore, our experimental results revealed that age, heart disease, glucose level, hypertension and marital status are significant risk factors. Other attributes, such as employment variety, bmi, residential status, gender and smoking status, are essential in predicting stroke to achieve the best accuracy. However, in the future, an intelligent stroke -diagnosis and- monitoring system will be proposed to capture real-time health status (e.g. blood pressure, pulse rate/ECG and glucose level) and then predict the probability of stroke. Moreover, the system will apply to analyzing other diseases (e.g. heart disease and kidney disease).

ACKNOWLEDGMENTS

I want to express my gratitude to the Department of Computer Science and Engineering, Jagannath University, Dhaka-100, Bangladesh, for allowing me to use the lab to conduct the work.

CONFLICTS OF INTEREST

There are no conflicts of interest regarding the publication of this paper.

REFERENCES

[1] C. O. Johnson, M. Nguyen, G. A. Roth et al., "Global, Regional and National Burden of Stroke, 1990–2016: A Systematic Analysis for the Global Burden of Disease Study 2016," The Lancet Neurology, vol. 18, no. 5, pp. 439–458, 2019.

[2] B. C. Campbell, D. A. De Silva, M. R. Macleod, S. B. Coutts, L. H. Schwamm, S. M. Davis and G. A. Donnan, "Ischaemic Stroke," Nature Reviews Disease Primers, vol. 5, no. 1, pp. 1–22, 2019.

[3] S. S. Virani, A. Alonso, H. J. Aparicio et al., "Heart Disease and Stroke Statistics—2021 Update: A Report from the American Heart Association," Circulation, vol. 143, no. 8, pp. e254–e743, 2021.

[4] A. Subudhi, M. Dash and S. Sabut, "Automated Segmentation and Classification of Brain Stroke Using Expectation-maximization and Random Forest Classifier," Biocybernetics and Biomedical Engineering, vol. 40, no. 1, pp. 277–289, 2020.

[5] J. J. Noubiap, V. F. Feteh, M. E. Middeldorp, J. L. Fitzgerald, G. Thomas, T. Kleinig, D. H. Lau and P. Sanders, "A Meta-analysis of Clinical Risk Factors for Stroke in Anticoagulant-Naïve Patients with Atrial Fibrillation," EP Europace, vol. 23, no. 10, pp. 1528–1538, 2021.

[6] M. S. Elkind and R. L. Sacco, "Stroke Risk Factors and Stroke Prevention," Seminars in Neurology, vol. 18, no. 04, pp. 429–440, Thieme Medical Publishers, Inc., 1998,.

[7] G. Jackson and K. Chari, "National Hospital Care Survey Demonstration Projects: Stroke Inpatient Hospitalizations," Natl Health Stat Report, vol. 132, pp. 1-11, National Library of Medicine, 2019.

[8] V. Malik, A. N. Ganesan, J. B. Selvanayagam, D. P. Chew and A. D. McGavigan, "Is Atrial Fibrillation a Stroke Risk Factor or Risk Marker? An Appraisal Using the Bradford Hill Framework for Causality," Heart, Lung and Circulation, vol. 29, no. 1, pp. 86–93, 2020.

[9] H.-J. Lin, J.-H. Yeh, M.-T. Hsieh and C.-Y. Hsu, "Continuous Positive Airway Pressure with Good Adherence Can Reduce Risk of Stroke in Patients with Moderate to Severe Obstructive Sleep Apnea: An Updated Systematic Review and Meta-analysis," Sleep Medicine Reviews, vol. 54, p. 101354, 2020.

[10] K. Furie, "Epidemiology and Primary Prevention of Stroke," CONTINUUM: Lifelong Learning in Neurology, vol. 26, no. 2, pp. 260–267, 2020.

[11] C. English, L. MacDonald-Wicks, A. Patterson, J. Attia and G. J. Hankey, "The Role of Diet in Secondary Stroke Prevention," The Lancet Neurology, vol. 20, no. 2, pp. 150–160, 2021.

[12] J. D. Pandian, S. L. Gall, M. P. Kate et al., "Prevention of Stroke: A Global Perspective," The Lancet, vol. 392, no. 10154, pp. 1269–1278, 2018.

[13] K. Shailaja, B. Seetharamulu and M. Jabbar, "Machine Learning in Healthcare: A Review," Proc. of the 2nd IEEE International Conference on Electronics, Communication and Aerospace Technology (ICECA), pp. 910–914, Coimbatore, India, 2018.

[14] J. Yu, S. Park, S.-H. Kwon, C. M. B. Ho, C.-S. Pyo and H. Lee, "Ai-based Stroke Disease Prediction

System Using Real-time Electromyography Signals," Applied Sciences, vol. 10, no. 19, p. 6791, 2020.

[15] A. A. Ali, "Stroke Prediction Using Distributed Machine Learning Based on Apache Spark," Stroke, vol. 28, no. 15, pp. 89–97, 2019.

[16] P. Govindarajan, R. K. Soundarapandian, A. H. Gandomi et al., "Classification of Stroke Disease Using Machine Learning Algorithms," Neural Computing and Applications, vol. 32, no. 3, pp. 817–828, 2020.

[17] 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, pp. 539–545, 2021.

[18] T. Tazin, M. N. Alam, N. N. Dola, M. S. Bari, S. Bourouis and M. Monirujjaman Khan, "Stroke Disease Detection and Prediction Using Robust Learning Approaches," Journal of Healthcare Engineering, vol. 2021, Article ID 7633381, 2021.

[19] 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.

[20] N. Kasabov, V. Feigin, Z.-G. Hou, Y. Chen, L. Liang, R. Krishnamurthi, M. Othman and P. Parmar, "Evolving Spiking Neural Networks for Personalized Modeling, Classification and Prediction of Spatio-temporal Patterns with a Case Study on Stroke," Neurocomputing, vol. 134, pp. 269–279, 2014.

[21] D. Shanthi, G. Sahoo and N. Saravanan, "Designing an Artificial Neural Network Model for the Prediction of Thrombo-embolic Stroke," International Journal of Biometrics and Bioinformatics (IJBB), vol. 3, no. 1, pp. 10–18, 2009.

[22] L. Amini, R. Azarpazhouh, M. T. Farzadfar, S. A. Mousavi, F. Jazaieri, F. Khorvash, R. Norouzi and N. Toghianfar, "Prediction and Control of Stroke by Data Mining," International Journal of Preventive Medicine, vol. 4, no. Suppl. 2, p. S245, 2013.

[23] 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. 181–185, 2015.

[24] L. I. Santos, M. O. Camargos, M. F. S. V. D'Angelo et al., "Decision Tree and Artificial Immune Systems for Stroke Prediction in Imbalanced Data," Expert Systems with Applications, vol. 191, p. 116221, 2022.

[25] D. Paikaray and A. K. Mehta, "An Extensive Approach towards Heart Stroke Prediction Using Machine Learning with Ensemble Classifier," Proc. of the International Conference on Paradigms of Communication, Computing and Data Sciences, pp. 767–777, Springer, 2022.

[26] P. Songram and C. Jareanpon, "A Study of Features Affecting on Stroke Prediction Using Machine Learning," Proc. of the International Conference on Multi-disciplinary Trends in Artificial Intelligence, pp. 216–225, Springer, 2019.

[27] R. S. Jeena and A. Sukeshkumar, "Development of a Stroke Risk Assessment Model for a Small Population in South Kerala Using Logistic Regression," Proc. of TENCON 2019-2019 IEEE Region 10 Conference (TENCON), pp. 350–355, Kochi, India, 2019.

[28] G. Fang, W. Liu and L. Wang, "A Machine Learning Approach to Select Features Important to Stroke Prognosis," Computational Biology and Chemistry, vol. 88, p. 107316, 2020.

[29] L. R. Guarneros-Nolasco, N. A. Cruz-Ramos, G. Alor-Hernández, L. Rodríguez-Mazahua and J. L. Sánchez-Cervantes, "Identifying the Main Risk Factors for Cardiovascular Diseases Prediction Using Machine Learning Algorithms," Mathematics, vol. 9, no. 20, p. 2537, 2021.

[30] A. Parmar, R. Katariya and V. Patel, "A Review on Random Forest: An Ensemble Classifier," Proc. of the International Conference on Intelligent Data Communication Technologies and Internet of Things, pp. 758–763, Springer, 2018.

[31] T. Chen, T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho et al., "Xgboost: Extreme Gradient Boosting," R Package Version 0.4-2, vol. 1, no. 4, pp. 1–4, 2015.

[32] M. Salman Saeed, M. W. Mustafa, U. U. Sheikh, T. A. Jumani, I. Khan, S. Atawneh and N. N. Hamadneh, "An Efficient Boosted C5. 0 Decision-tree-based Classification Approach for Detecting Nontechnical Losses in Power Utilities," Energies, vol. 13, no. 12, p. 3242, 2020.

[33] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye and T.-Y. Liu, "Lightgbm: A Highly Efficient Gradient Boosting Decision Tree," Advances in Neural Information Processing Systems, vol. 30, pp. 3146–3154, 2017.

[34] J. T. Hancock and T. M. Khoshgoftaar, "Catboost for Big Data: An Interdisciplinary Review," Journal of Big Data, vol. 7, no. 1, pp. 1–45, 2020.

[35] W. Wang and D. Sun, "The Improved Adaboost Algorithms for Imbalanced Data Classification," Information Sciences, vol. 563, pp. 358–374, 2021.

[36] S. Suthaharan, "Support Vector Machine," Proc. of Machine Learning Models and Algorithms for Big Data Classification, pp. 207–235, Springer, 2016.

[37] 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.

[38] H. Saadatfar, S. Khosravi, J. H. Joloudari, A. Mosavi and S. Shamshirband, "A New K-nearest Neighbors

Classifier for Big Data Based on Efficient Data Pruning," Mathematics, vol. 8, no. 2, p. 286, 2020.
[39] K. Shah, H. Patel, D. Sanghvi and M. Shah, "A Comparative Analysis of Logistic Regression, Random Forest and KNN Models for the Text Classification," Augmented Human Research, vol. 5, no. 1, pp. 1–16, 2020.
[40] J. Yu, S. Park, S.-H. Kwon, C. M. B. Ho, C.-S. Pyo and H. Lee, "Ai-based Stroke Disease Prediction System Using Real-time Electromyography Signals," Applied Sciences, vol. 10, no. 19, p. 6791, 2020.
[41] S. Wang, Y. Dai, J. Shen and J. Xuan, "Research on Expansion and Classification of Imbalanced Data Based on Smote Algorithm," Scientific Reports, vol. 11, no. 1, pp. 1–11, 2021.
[42] J. Nithyashri and G. Kulanthaivel, "Classification of Human Age Based on Neural Network Using FG-net Aging Database and Wavelets," Proc. of the 4th IEEE International Conference on Advanced Computing (ICoAC), pp. 1–5, Chennai, India, 2012.

ملخص البحث:

النّوبة القلبية حالة تهدّد الحياة، وتعدّ ثاني أسباب الوفاة عالمياً. وهي مشكلة تنطوي على تحدٍّ في ميدان الصّحة العامة في القرن 21 للعاملين والباحثين في حقلالرعاية الصحية. لذا فإنّ الرّصد الملائم للنّوبة القلبية يمكن أن يقود الى منعها أو تخفيف شدّتها. هناك العديد من الطّرق التي ركزت على توقّع النّوبات القلبية لدا المرضى. والكثير من تلك الطّرق حقّقت معدّل دقّةٍ عالياً، وصل الى ما يقرب من 90% على مجموعات البيانات المتاحة للعموم. ويمكن لجمْع مهامّ تدريبٍ قبْلي متنوعة أن يحسّن على نحوٍ ملموس جودة المصنِّفات التي تشكل مجالاً يحتاج الى البحث. مع تحديد الباحثين عوامل الخطورة الرئيسية للنّوبة القلبية واستخدام مصنِّفات متقدّمة لتوقّع احتمال الإصابة بالنّوبة القلبية.

تقدم هذه الورقة طريقة محسَّنة لتحديد عوامل الخطورة المحتملة وتوقّع الإصابة بالنّوبة القلبية، وذلك باستخدام إحدى مجموعات البيانات المتاحة للعموم. وتسعى الطّريقة المقترحة الى ردم فجواتٍ في الأدبيّات السّابقة المتعلّقة بالموضوع. وتستخدم الدراسة الحالية عشرة نماذج تصنيف تشمل مصنِّفات معزّزة متقدّمة للكشف عن الإصابة بالنّوبة القلبية. وقد تمّ تحليل أداء المصنِّفات على جميع مجموعات البيانات الفرعية الممكنة المتعلّقة باختيارات الخصائص/السِّمات، بأخذ خمسة مقاييس بعين الاعتبار، لتحديد الخوارزميات الأفضل أداءً. وبينت النّتائج التجريبية أنّ الطريقة المقترحة حقّقت الدّقة الأعلى على تصنيفات السِّمات كافّةً. وكان الإنجاز الحقيقي لهذه الدراسة تحقيق نسبة دقّة أعلى (97% باستخدام مصنِّفات معزّزة) مقارنة بالطّرق الأخرى. وعليه يمكن للأطباء الاستفادة من الطريقة المقترحة في توقّع الإصابة النّوبة القلبية في العالم الحقيقي. وأوضحت النتائج أنّ أبرز عوامل الخطورة المرتبطة بالنّوبة القلبية هي: العمر، وأمراض القلب، ومستوى الغلوكوز، وإرتفاع ضغط الدّم، والحالة الاجتماعية. وفي الوقت ذاته، فإنّ السِّمات الأخرى أساسية للحصول على الأداء الأفضل.