Computer Systems
Science and Engineering

Tech Science Press

DOI: 10.32604/csse.2023.037127
Article

An Ensemble Machine Learning Technique for Stroke Prognosis

Mesfer Al Duhayyim[1],[*], Sidra Abbas[2],[*], Abdullah Al Hejaili[3], Natalia Kryvinska[4], Ahmad Almadhor[5] and Uzma Ghulam Mohammad[6]

[1]Department of Computer Science, College of Computer Engineering and Sciences, Prince Sattam bin Abdulaziz University Al-Kharj, 16273, Saudi Arabia
[2]Department of Computer Science, COMSATS University, Islamabad, 53000, Pakistan
[3]Computer Science Department, Faculty of Computers & Information Technology, University of Tabuk, Tabuk, 71491, Saudi Arabia
[4]Information Systems Department, Faculty of Management, Comenius University in Bratislava, Odbojárov, Bratislava, 440, Slovakia
[5]Department of Computer Engineering and Networks, College of Computer and Information Sciences, Jouf University, Sakaka, 72388, Saudi Arabia
[6]Department of Computer Science and Software Engineering, International Islamic University, Islamabad, 44000, Pakistan
*Corresponding Authors: Mesfer Al Duhayyim. Email: m.alduhayyim@psau.edu.sa; Sidra Abbas. Email: sidra.abbas708@gmail.com
Received: 24 October 2022; Accepted: 28 December 2022; Published: 26 May 2023

Abstract: Stroke is a life-threatening disease usually due to blockage of blood or insufficient blood flow to the brain. It has a tremendous impact on every aspect of life since it is the leading global factor of disability and morbidity. Strokes can range from minor to severe (extensive). Thus, early stroke assessment and treatment can enhance survival rates. Manual prediction is extremely time and resource intensive. Automated prediction methods such as Modern Information and Communication Technologies (ICTs), particularly those in Machine Learning (ML) area, are crucial for the early diagnosis and prognosis of stroke. Therefore, this research proposed an ensemble voting model based on three Machine Learning (ML) algorithms: Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Light Gradient Boosting Machine (LGBM). We apply data preprocessing to manage the outliers and useless instances in the dataset. Furthermore, to address the problem of imbalanced data, we enhance the minority class's representation using the Synthetic Minority Over-Sampling Technique (SMOTE), allowing it to engage in the learning process actively. Results reveal that the suggested model outperforms existing studies and other classifiers with 0.96% accuracy, 0.97% precision, 0.97% recall, and 0.96% F1-score. The experiment demonstrates that the proposed ensemble voting model outperforms state-of-the-art and other traditional approaches.

Keywords: Stroke prediction; machine learning; ensemble model; data analysis; Synthetic Minority Over-Sampling

1 Introduction

Human existence is built on different body components and functions. The heart is the most critical organ since it pumps blood to all other organs. A threatening condition that kills human lives is stroke. After the age of 65, this condition is frequently discovered. As a heart attack affects the functioning of the heart, stroke impacts the brain [1]. The World Stroke Organization (WSO) [2] estimates that 13 million individuals worldwide experience a stroke yearly, with 5.5 million dying.

It tremendously impacts every aspect of life because it is the main mortality element worldwide. Stroke affects the sufferer as well as their social circle, family, and area of employment. Contrary to common belief, it is a condition that can affect any individual of any age, irrespective of gender or health appearance [3]. Strokes are brought on by either a limitation in the blood flow to the brain or the rupture and bleeding of brain blood vessels, which results in one of these two diseases. An obstruction stops oxygen and blood from extending to the brain's tissues [4]. A stroke consists of two types, which are known as ischemic and hemorrhagic. Ischemic stroke is the first type of stroke that happens when blood clots or plaque buildup clog an artery, preventing blood flow to the brain. It might range from minor to severe, regardless of permanent or temporary injury. Blood arteries can rupture during hemorrhages, which are infrequent but can result in brain hemorrhage [5,6].

The presence of myocardial infarction increases the likelihood of having a stroke [7] and other heart diseases like heart attack and cardiac arrhythmia, and age (even though stroke can happen to anybody, even children, people over 55 are more likely to experience it.), alcohol usage, smoking, high cholesterol levels, diabetes, overweight, poor diet, cerebral tightness from plaques, estrogen therapy, use of euphoric drugs, and carotid stenosis [8,9]. Stroke also occurs predominantly and manifests a spectrum of symptoms. In some cases, symptoms develop gradually; in others, they emerge suddenly. Even the possibility exists for someone to experience symptoms even while asleep. After having a stroke, several symptoms may arise unexpectedly. The most frequent ones are stiffness in the arms or legs, sporadically on a face, and rigidity of the arms or legs, usually solely on a single side of the body. Speech problems or moving, disorientation, impaired vision, headache, diarrhea, and a shift in the angle of the mouth. Last, a patient suffering from a severe stroke goes unconscious and enters a coma [1,10].

A Computed Tomography (CT) scan can diagnose a stroke right once it has occurred in the patient. Magnetic Resonance Imaging (MRI) is helpful in the diagnosis of ischemic stroke. Carotid triplex and cardiac triplex are additional auxiliary diagnostic procedures. Strokes can range from minor to severe (extensive). In most situations, the very first 24 h are crucial. The diagnosis will emphasize the form of therapy, which is frequently medical and occasionally surgical and is typically pharmaceutical. The intensive care unit requires intubation and mechanical ventilation when a patient enters a coma [11,12]. However, while some stroke victims recover, the majority continue to experience issues, such as remembrance, adsorbent dose, and behavioral difficulties; speech difficulties; psychological issues like anxiety; lost coordination or mobility; losing sensation solely on a single of the body's sides and having trouble digesting meals, based on the stroke's intensity [13,14]. After a stroke, recovery aids in regaining lost function. With the help of speech therapy, kinesiotherapy and neurologists, the right strategy is made so that the patient quickly regains their psychological and social well-being [15,16]. The stroke risk can be decreased by periodically checking pulse rate, exercising frequently, weight maintenance, avoiding smoking and alcohol consumption alcohol, and following a diet lower in cholesterol and sodium [17,18].

Motivation: The mainstay of the modern day, ML, is employed to anticipate numerous challenges at an early stage. As a severe disease that may be treated if anticipated in the preliminary stages, stroke is one of many that can be prevented if predicted early. In the healthcare sector, ML is crucial for diagnosing and prognosis of diseases. Powerful data analysis tools are needed for enormous amounts of medical data. The quantity of information that hospitals store in patients' medical records keeps growing. ICTs, particularly those in the areas of AI and ML, are crucial for the early identification and prognosis of several illnesses, like hypertension [19], cholesterol [20], and hepatitis C [21].

Numerous research studies have used ML models for this particular condition [1,3–5]. This research uses an ensemble voting model based on ML algorithms such as RF, XGBoost, and LGBM. The other ML algorithms, K-Nearest Neighbor (KNN), Support Vector Classifier (SVC), Logistic Regression (LR), and BernoulliNB, are applied to evaluate the results. The data analysis method (bi-variate analysis, categorical distribution, label encoding pairwise analysis, and standard scaler) examines the dataset's outliers and abnormalities. The dataset is balanced using the SMOTE because class balancing is essential for designing practical algorithms in stroke prediction.

The research's primary contribution is detailed in the list below.

- This study proposed an ensemble voting model built on ML algorithms such as RF, XGB, and LGBM. The other ML algorithms, LR, SVC, KNN, and BernoulliNB, are also applied to evaluate the results.
- The stroke healthcare dataset is collected from Kaggle. Data analysis techniques such as bi-variate, categorical distribution, label encoding, pairwise analysis, and standard scaler are applied to the dataset. Further preprocessing technique SMOTE is applied to represent the minority classes in the dataset better.
- The experiment demonstrates that the suggested ensemble voting model surpasses prior findings and other classifiers with an accuracy of 0.96%, and diagnosing strokes earlier benefits medical professionals and patients.

The following sections make up the organization of the paper: The literature overview of the ML and deep learning methods for recognizing stroke is covered in Section 2. The study methodology for the proposed work using the healthcare dataset for stroke and ML classifiers are explained in Section 3. The outcomes are explained and examined in Section 4. The work's conclusion and suggestions for additional research are presented in Section 5. The list of acronyms is given in Table 1.

Table 1: Acronyms list

No.	Acronyms	Description
1	ICTs	Modern Information and Communication Technologies
2	AI	Artificial intelligence
3	ML	Machine Learning
4	LR	Logistic regression
5	KNN	K-Nearest Neighbor
6	DT	Decision tree
7	DTC	Decision tree classifier
8	RF	Random Forest
9	BernoulliNB	Bernoulli Naïve Bayes
10	NB	Naïve Bayes
11	SVC	Support vector classifier

(Continued)

Table 1: Continued

No.	Acronyms	Description
12	SVM	Support vector machine
13	XGB	Extreme Gradient Boosting
14	LGBM	Light Gradient Boosting Machine
15	SMOTE	Synthetic Minority Over-Sampling Technique
16	CT	Computed Tomography
17	MRI	Magnetic Resonance Imaging
18	SGD	Stochastic Gradient Descent
19	QDA	Quadratic Discriminant Analysis
20	PCA	Principal Component Analysis
21	DNN	Deep neural network
22	AUC	Area Under the Curve
23	CNN	Convolutional neural network
24	LSTM	Long Short-Term Memory
25	GOSS	Gradient-based One-Side Sampling

2 Literature Review

The literature review section provides the background of stroke prediction using different techniques. Two sections comprise the literature: ML-based and deep learning techniques helpful for stroke prediction.

2.1 Machine Learning Techniques

Machine learning and artificial intelligence are essential for the earlier identification and prognosis of several illnesses, including hypertension [9,19], cholesterol [20], hepatitis C [21], malignant mesothelioma [22,23] and others [24–26]. For the specific stroke situation, many research investigations have applied machine learning models [3–8,10,14,18]. In the current analysis, several risk factors for stroke are considered. The authors first examine the traits of people more prone than others to experience a stroke. Multiple classification techniques such as NB, RF, DT, multilayer perceptron, and JRip algorithm are applied to the dataset, which was gathered from a freely accessible source, to forecast the impending occurrence of a stroke. It has been feasible to achieve an accuracy of 98.94% by using the random forest method. Finally, some preventative measures, including smoking, abstaining from alcohol, and other things, are advised to lower stroke risk [1].

The study [4] uses ML to build several models that are then evaluated to lay a strong foundation for the protracted risk stratification of incidence rates. The study makes use of the Kaggle available stroke dataset. The authors selected participants who were older than 18 from this dataset. There are 3254 participants. This research presents a stacking technique that outperforms and is supported by numerous evaluation metrics (EM). The research's findings demonstrated that the stacking algorithm performs better than the other approaches, with 98.9% AUC, 97.4% F-measure, precision, recall, and 98% accuracy.

The author in [27] trains five different models to precisely assess the performance of stroke detection using different physiological elements and ML algorithms like LR, DT, RF, KNN, SVM, and NB. The dataset used for the study to predict strokes comes from Kaggle. The dataset contains data in which rows are 5110 and 12 columns. The main attributes are id, age, sex, heart disease, hypertension, work type, ever-married, average glucose level, type of residence, smoking status, BMI, and stroke. The output column "stroke" value is either "0" or "1." The number "0" shows that there is no stroke incidence at all, whereas the number "1" indicates that there may be a stroke threat. The NB technique works effectively and has an accuracy rating of 82%.

To determine whether a patient will experience a stroke, [28] states that an ML model is used. The Random Forest classifier outperforms other cutting-edge models, including LR, Decision Tree Classifier (DTC), and KNN. The research uses datasets containing 5110 observations and 12 attributes for its studies. Additionally, the author used feature techniques to balance the datasets and exploratory data analysis to preprocess the data. As a final point, a cloud-centered mobile application aggregates user information to estimate and show the risk of stroke for notifying the person with 96% accuracy and precision, 96% recall, and 96% F1-score. This easy method can protect a human's life because it enables them to swiftly take a life-saving alert from any place as long as they possess a smartphone.

The author of [29] suggests using several ML algorithms to immediately forecast stroke problems based on the prevalence of the illness, body weight, cardiovascular illness, average sugar level, smoking habits, preceding stroke, and gender. Ten different classifiers: LR, Stochastic Gradient Descent (SGD), DTC, AdaBoost, Quadratic Discriminant Analysis (QDA), MLP Classifier, KNN, Gradient Boosting, and XGBoost Classifier, have been trained using these high features attributes to predict strokes. Following the weighted voting approach, the result of the basic classifiers is incorporated to acquire the highest accuracy. The weighted voting predictor surpasses the base classifiers in the suggested study, which has a 97% accuracy rate. The stroke prediction accuracy provided by this model is the highest.

2.2 Deep Learning-Based Techniques

Deep Learning (DL) models can significantly improve disease classification and identification procedures such as ischemic stroke, Alzheimer, and others [30–33]. A revolutionary methodology is proposed in [34] that enables the direct implementation of DL techniques on unprocessed Electroencephalography (EEG) data. Real-time EEG sensor data has been utilized to generate and train the proposed deep learning-based model for predicting stroke disease. The author implements and compares DL models (LSTM, Bidirectional LSTM, CNN-LSTM, and CNN-Bidirectional LSTM) focused on data categorization and prognosis. The experiment outcomes demonstrate that the CNN-bidirectional LSTM model, when applied to the unprocessed EEG data, effectively predicts stroke with 94.0% accuracy and low FPR (6.0%) and FNR (5.7%), demonstrating great trust in the proposed approach. The result of this study shows the possibility of quasi-methods, which can rapidly collect neural activity from forecasting and analyzing stroke problems in real time while going about daily activities. Compared to previous testing approaches, these findings are anticipated to improve early stroke identification with lower costs and pain significantly.

The author [35] describes DL-based prognostic methodological approaches for stroke utilizing a heart illness dataset. Cardiac arrhythmia symptoms in people with heart disease are a significant cause of stroke and also have characteristics that can forecast brain hemorrhage. The study's conclusions are more accurate than currently used physician grading methods for alerting cardiac sufferers in danger of stroke. This study employed a DL approach to a cardiac illness. There are 899 records with 76 attributes each. The deep learning model effectively predicts the stroke with a 36.73% average value and standard deviation of 0.084.

The author of [36], using data on medical facility use and health behavior, used a DNN to prognosis stroke; researchers found 15,099 patients with the disease. From medical data, pertinent background features are extracted using Principal Component Analysis (PCA) with quantile scaling, and they are then utilized to predict stroke. The author proposes the scaled PCA/DNN strategy against five ML techniques. Since the proposed method's Area Under the Curve (AUC) value is 83.48%, patients and medical professionals can use it to check for potential strokes.

The author of [37] proposes a data-driven classifier called a Dense Convolutional Neural Network (Dense Net) for stroke detection built on the data of 12-lead ECG. The study obtained a training accuracy of 99.99% and a prediction accuracy of 85.62% with the proposed fine-tuned model. The findings show that ECG is an important complementary tool for diagnosing stroke. According to the author, it is the first paper that is aware of using deep learning to examine the relationship between a stroke and an ECG. Some other studies [38,39] have presented security and privacy concerns regarding using these approaches.

In short, several restrictions are being addressed using ensemble learning techniques, ML strategies, and a variety of datasets utilized to assess the comprehensiveness of techniques such as stroke healthcare prediction. The literature review summary is given in Table 2.

Table 2: Summary of literature review

References	Techniques	Results	Limitations
[27]	NB, RF, DT, LR, SVM	82% accuracy	Low performance
[34]	CNN-bidirectional LSTM	94% accuracy	Model provides better results with raw data
[36]	DNN	83.48% AUC, 84.03% accuracy	Low performance
[37]	Dense-Net	85.62% accuracy	Low performance

3 Proposed Methodology

This section explains the proposed methodology step, such as experimental dataset, data analysis, balance data with SMOTE, and training model. Fig. 1 demonstrates the proposed methodology. First, collect the stroke dataset from Kaggle, then performs the data analysis using bi-variate, categorical distribution, and pairwise analysis. The next step is to preprocess data using SMOTE to balance the dataset. The used dataset is split into training and testing data in the next step, with the proportion keeping 80% and 20% of training and testing data. A voting ensemble model is created for stroke prediction using ML algorithms such as RF, XGBoost, and LGBM. Then different ML algorithms LR, SVC, KNN, and BernoulliNB, are utilized to assess the model's performance. Finally, the model makes a stroke prediction for each patient. After applying the ML models, the best training and testing results are obtained.

3.1 Experimental Dataset

This research collects the stroke prediction dataset from Kaggle for an experiment. This dataset comprises 5110 records (41% male and 59% female) and 12 features. The dataset includes 11 features such as id (unique number given to patients), sex of the patient (male, female, or others), age of the patient in years (age), hypertension (no hypertension = 0 and hypertension = 1), heart illness (heart disease = 1 and no heart disease = 0), ever-married (either Yes or No), work's kind (private, self-employed, government job and children), residence type (Urban or rural) level of glucose (avgglucoselevel), Mass index of the body (BMI), smokingstatus (smoke, formally smoked, never smoked, no-information and unknown). The target feature includes two categories: "1" is regarded as a stroke, and 0 represents no stroke.

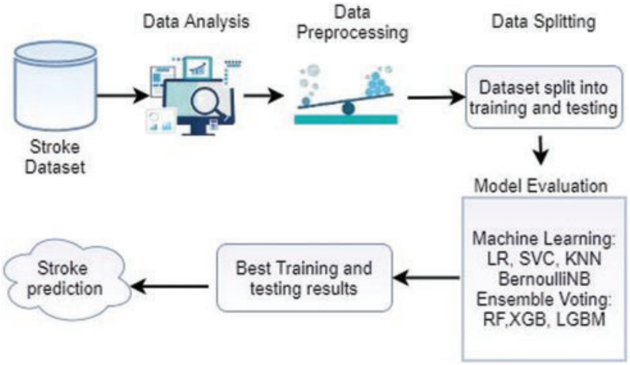
Data Analysis
Data Preprocessing
Data Splitting
Stroke Dataset
Dataset split into training and testing
Model Evaluation
Machine Learning: LR, SVC, KNN BernoulliNB Ensemble Voting: RF,XGB, LGBM
Best Training and testing results
Stroke prediction

Figure 1: Proposed model overview

3.2 Data Analysis

After data preparation, the dataset is examined to determine whether or not data cleaning is necessary. This step also includes an analysis of significant features. Identifying outliers and anomalies in the data allows customizing the testing of the hypothesis in a specific way, which is the fundamental purpose of data analysis. The data is effectively viewed, plotted, and updated without making assumptions to assess the data quality and develop models.

By exploring the dataset, it has both numerical and categorical type variables. The dataset has "5" categorical and "7" numerical features, and the feature BMI has a 4% missing value. Since the model requires numbers as input, the research must first encode the categorical characteristics before supplying them to the model. This research used bi-variate analysis for numerical features, distribution for categorical features, and some pairwise analysis to visualize all features. Some exploring techniques depend on data types, and objective analysis is used.

In Bi-variate analysis, this research analyzes how stroke condition is affected by two features according to the target feature. Age, avgglucoselevel, and BMI features are used for Bi-variate analysis. People aged 65–85 have a high chance of getting a stroke, and BMI cannot distinguish stroke patterns and has 4% missing values; hence we drop this feature. A discrete probability distribution known as a categorical distribution defines the likelihood that a random variable will have a value that falls into one of the K categories, where each category has a corresponding probability. In categorical variable distribution, this study has used gender, evermarried, worktype, residencetype, smokingstatus, hypertension, and heart disease features to analyze either stroke effect or not in these conditions. This research analyzes that there are higher samples of no stroke (stroke = 0) in every feature compared to the other class, and other categories in' gender' can be ignored.

In pairwise analysis, analyze different pair analyses such as residencetype and smokingstatus. The study observes that the Type of Residence, either Urban or Rural, does not affect having a stroke, hence can also be dropped. In worktype and gender pairwise analysis, it observed that only "2" children have strokes, and both female and older females in govtjobs have a higher risk of stroke. In worktype and smokingstatus analysis, it has been observed that most people who had a stroke worked in the' Private' sector.

3.3 Dataset Preprocessing

In this stage, data processing is carried out to standardize and divide the categorical data into integer values. The data preprocessing phase is central since it supports accomplishing extra correct features and increases the model's performance. In this paper, encoding categorical data, standard scalar techniques are used.

3.3.1 Encoding Categorical Variables

Categorical variables are complex to handle for some machine learning algorithms. It is necessary to transform the categorical variable into numeric data. It is a crucial step for the effective functioning of the implemented algorithms. The way categorical variables are codified affects how well various algorithms perform. One or more labels in word or numeric format can be included in a feature's dataset. That makes it simpler for people to interpret the data, but it needs to be more comprehensible to computers. Therefore, this study utilizes encoding to make these labels understandable to computers. There are different encoding techniques, such as one-hot encoding, hash encoding, etc. [40]. In this study label, an encoding technique is used for encoding categorical variables.

3.3.2 Label Encoder

Label encoding enables numerical label input into an ML model. It is a crucial stage in the preprocessing of data for supervised learning techniques. Label Encoder merely assigns a number value to each label to replace each different label's value in the dataset. When the labels have varied priorities, they can be used effectively. This strategy typically replaces numbers from 0 to N – 1 for each value in a categorical column [41]. This paper uses a label encoder to assign the value from 0 to 1 for each categorical feature.

3.3.3 Data Scalar

One of the most widely used methods for standardizing information is standardizing. A standard scaler is an essential tool used primarily as a preprocessing phase preceding several machine learning models to normalize the input dataset's operating range. It scales a feature to unit variance after subtracting the mean to normalize it. The Standard scalar scales each component so that the distribution is centered on 0 and has a standard deviation (SD) of 1 since it anticipates that information is typically appropriated within each component [42]. This research uses the standard scaler technique to normalize the functional range of the stroke input dataset.

3.3.4 Data Splitting

The process of building models comes next, after data pre-processing and handling the unbalanced. The resampled dataset is split into training and testing, keeping the proportion at 80% training phase and 20% test data to increase accuracy and effectiveness for this operation. Following dividing, several classifiers are used to build the model. The classification methods employed in this situation are LR, DT, RF, KNN, SVC, BernoulliNB Classification, XGBoost, and LGBM.

3.4 Balance Dataset with SMOTE

Real-world data sets typically have a high ratio of "normal" classes and a low ratio of "abnormal" ones. If the distribution of the classifying classes is inconsistent across the dataset, it is considered imbalanced data. Two techniques are used mainly for making the dataset balanced: the first is oversampling, which is applied to the minority class, and the second is under-sampling, which is applied to the majority class. These techniques accomplish better performance of the classifier. SMOTE is an appropriate solution for the categorization imbalance issue and has shown solid outcomes across several domains [43]. The limited training set combines synthetic data via the SMOTE method to create a balanced and stable dataset [44]. The class imbalance problem relates to the inequality between the categories of datasets utilized to create prediction methodologies, a widespread issue not specific to medical data. Class unbalances issues are addressed by tampering with data, an algorithm, or together to increase the model's performance because classification algorithms have a tendency to approval the majority classes whenever the training set with negative outcomes has a disproportionately smaller number of assumptions than the categories with majority inferences [45].

The method's fundamental step is to randomly under and over-sample for larger and smaller samples. This research dataset has two classes: stroke and no stroke. The dataset is highly imbalanced; therefore, we employed the SMOTE technique for balancing the data. Before applying SMOTE technique, the model performed poorly without training it with enough samples from both classes. After applying SMOTE technique, the model performs well with proper training of machine learning classifiers.

3.5 Classification Models

A classification model aims to derive some meaning from the training set of inputs. It will forecast the categories and class labels for the updated data. Various ML algorithms are utilized in this study, like LR, KNN, SVC, RF, BernoulliNB, XGBoost, and LGBM. The details of these algorithms are given below.

3.5.1 Logistic Regression

A supervised learning algorithm determines the likelihood that goal values will occur. Those goal value natures are shaky; they imply only two conceivable outcomes. The parameter is binary, with data represented as either 1 (for stroke) or 0 to make it easier to understand (for no stroke) [46].

3.5.2 K-Nearest Neighbor

KNN is the least challenging machine learning algorithm that uses supervised learning. It is utilized to create resemblance information by assembling and storing other information. As new information is produced, it can be sorted using this technique into a class that best fits its characteristics. Although, in general, it utilizes for classification and regression, it is usually used to address classification-related issues. It keeps the data during preparation, and when it finds any modifications, it reclassifies the data to make it seem like it just happened. The KNN method performs better when the value of k is chosen optimally [46].

3.5.3 Decision Tree

The decision tree algorithm is the class of algorithms used in supervised AI. Both the classification and regression issues can be solved with a decision tree. This algorithm aims to create a system that predicts the number of target parameters. To accomplish this, this study utilizes a decision tree to approach the issue. The interior nodes of the tree are referred to as attributes, and the leaf nodes act as classes [46].

3.5.4 Random Forest

The researchers can utilize machine learning to handle problems involving regression and classification. A well-known artificial intelligence class that develops through supervised learning is called Random Forest. Decision trees are built on multiple subsets of the supplied dataset to improve prediction accuracy. This classifier has not relied solely on one tree. To locate different types of results, different types of trees are created. The most trustworthy outcome from these trees is chosen in the end. The correct tree is selected; as a result, [46].

3.5.5 BernoulliNB

The BernoulliNB algorithm uses discrete values and interacts with the Bernoulli distribution. The primary characteristics of this classifier are dual attributes such as either true or false, yes or no, success or failure, 0 or 1, etc. [46]. Therefore, the Bernoulli NB algorithm is used when the parameter values are binary.

3.5.6 Support Vector Classifier

In a support vector classifier, information emphases are arranged from an information network to a high-dimensional component space employing a kernel function. The SVDD algorithm finds the smallest sphere in the kernel's subspace, including data representation. This circle creates several forms that enclose the information when it is projected back into the information network. Then, those forms are translated as group constraints, with each shape encapsulating emphases that are associated through SVC to a related group [44].

3.5.7 XGBoost

A good application of the gradient augmentation technique is XGBoost (XGB). It is a gradient gain substitute that may be precisely engineered for precision and optimization. It includes a linear model, and the baby tree may be a technique that tests if a weak beginner would create a reliable beginner using various AI algorithms to improve the model's performance. From the random forest, for instance, and parallel learning (bagging). Data gathering is a technique that can regulate the presentation of an advanced AI version whose precision processing is quicker than improving gradients. These techniques for filling the data gap are built in.

The XGBoost's s fundamental concept is to use variances to create the model. The ensemble learner is then created by linearly weighting each weak learner. The error is steadily reduced over several serial repeats using the result of the best current tree training as the input. The best solution is the optimal quadratic solution, which XGBoost achieves by approximating the objective function using a second-order Taylor expansion. Additionally, a greedy method based on information gain is used to select the optimal splitting point for training the XGBoost tree. Furthermore, a standard term is included to control the complexity of the spanning tree, reducing the model overfitting risk [46].

3.5.8 Light Gradient Boosting Machine

Light Gradient Boosting Machine (LGBM) is another gradient-boosting technique and ensemble learning application. The primary purpose of LGBM is to produce huge gradients that increase the gain ratio. LGBM builds trees upwards using a leaf-wise method; a branch that minimizes the damage is selected for splitting [47].

One leaf that decreases loss the utmost is selected to divide and develop the tree. The best splitting candidates are found by LGBM using a histogram-based approach. To improve training, LGBM employs the Gradient-based One-Side Sampling (GOSS) method to determine the significance of the data samples. Its core objective is to ignore data trials with smaller gradients and concentrate on those with more significant gradients. The underlying presumption is that data with modest slopes would have less error and be more trained. GOSS advised rejecting fewer datasets and utilizing the entire dataset to quantify the acquired knowledge when finding the proper divides. However, this will modify the primary data distribution and introduce bias favoring the sample with more significant gradients. GOSS resolves this issue by keeping all trials with large gradients and periodically choosing the data with smaller gradients. Because the sample is biased toward the data with large gradients, GOSS increases the weights of the samples with lesser gradients when calculating the information gain (IG).

To address dataset sparsity, LGBM uses the Exclusive Feature Bundling method. Since it aggregates independently exclusive features in a nearly lossless way, the number of features is decreased while the most informative ones are kept [48].

3.5.9 Voting Classifier

The ML classifier that builds many base models or classifiers is a voting algorithm. It combines their outcomes to produce predictions. Combining voting with the accumulating parameters for each classifier result is possible. The base classifiers RF, XGBoost, and LGBM, are employed in this study to create a voting predictor.

Algorithm 1: Pseudo code of Proposed Model

```
Input: Stroke healthcare dataset $D_s$
Output: Model Performance
Data Analysis ($DA$)
    Bi-variate Analysis
    categorical Variables Distribution
    Pairwise Analysis
Data Preprocessing ($D_p$)
$x, y$ {Dataset $D_s$}
$x_{train}, x_{test}, y_{train}, y_{test}$ {data split into train and test set}
$SM$ = smote () {smote formula}
Ensemble {Create Ensemble Models}
$RF \leftarrow$ RandomForest ()
$XGB \leftarrow$ XGBoost ()
$LGBM \leftarrow$ LGBmodel ()
$E_m \leftarrow$ Accuracy, Precision, Recall, F1-Measure {Evaluation metrics}
Return $\leftarrow$ Best Training Results
Ensemble {Recreate Model for test results}
Return $\leftarrow$ Result
```

Algorithm 1 explains the overall working of the proposed ensemble voting model. The ensemble voting model *Ensemble* is created using three machine learning algorithms (*RF, XGB, LBGM*). The stroke dataset D_S is taken as input, and output is the performance of the proposed ensemble voting model. Data analysis is utilized for the visualization of all the features. Bi-variate, categorical variables distribution, and pairwise analysis techniques are applied. Data is distributed into training and tests with sizes of 0.8 and 0.2 and preprocessed using SMOTE (SM) to balance the data. Evaluation metrics E_m accuracy, precision, recall, and F1-score are used. The best results are obtained on the training dataset. Then recreate, the model and outcomes on the test dataset are obtained.

4 Experimental Analysis and Results

This section analyzes the performance of the suggested model. The suggested model uses different machine learning classifiers on a stroke healthcare dataset. It is assessed using various criteria, including accuracy, recall, precision, and f1-score. These criteria are used to assess the proposed model's performance compared to the current methods and their suitability for early stroke detection.

4.1 Evaluation Metrics

The suggested model's categorization outcomes are identified using various evaluation indicators. This research used accuracy, precision, recall, and F1-score performance metrics [49]. Determine the proportion of all affirmative values in the data to find precision. The proportion of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) are utilized to compute the accuracy. Manipulative the true positive by the actual negative and false negative yields the recall, also known as sensitivity. Taking the mean of the precision and recall values yields the F1-score.

The total effectiveness of ML algorithms on the stroke dataset is shown in Table 3. The LR algorithm achieved 0.78% accuracy, 0.76% precision, 0.82% recall, and 0.79% f1-score. K-NN algorithm obtained 0.89% accuracy, 0.84% precision, 0.96% recall, and 0.90% f1-score. DT obtained 0.91% accuracy, precision, recall, and f1-score.

Table 3: Proposed model results

Model	Accuracy	Precision	Recall	F1-score
Logistic regression	0.78%	0.76%	0.82%	0.79%
K-Nearest neighbor	0.89%	0.84%	0.96%	0.90%
Decision tree	0.91%	0.91%	0.91%	0.91%
Random forest	0.95%	0.95%	0.95%	0.95%
BernoulliNB	0.72%	0.65%	0.92%	0.77%
Support vector classifier	0.82%	0.78%	0.89%	0.83%
XGB	0.96%	0.98%	0.95%	0.95%
LGBM	0.93%	0.96%	0.90%	0.93%
Ensemble voting	0.96%	0.97%	0.97%	0.96%

RF achieved 0.95% accuracy, 0.95% precision, 0.95% recall, and 0.95% f1-score. BernoulliNB achieved 0.72% accuracy, 0.65% precision, 0.92% recall, and 0.77% f1-score. The SVC algorithm achieved an accuracy of 0.82%, 0.78% precision, 0.89% recall, and 0.96% f1-score. XGBoost algorithm attained 0.96% accuracy, a precision of 0.98%, a recall of 0.95%, and an f1-score of 0.95%. The LGBM algorithm achieved an accuracy of 0.93%, precision of 0.96%, recall of 0.90%, and 0.93% f1-score.

The ensemble voting model achieved the best outcomes of 0.96% accuracy, precision 0.97%, recall 0.97%, and F1-score 0.96%. Also, showing the comparison of the proposed model with a simple base classifier. The assessment displays that the proposed ensemble voting model performed better with the base classifier. Fig. 2 graphically represents XGBoost, RF, and LGBM's confusion matrix. It provides an overview of how a classification algorithm operates. Because it has more continuous, better true positive (TP) and negative values and fewer false negative and false positive values, the proposed method performs better.

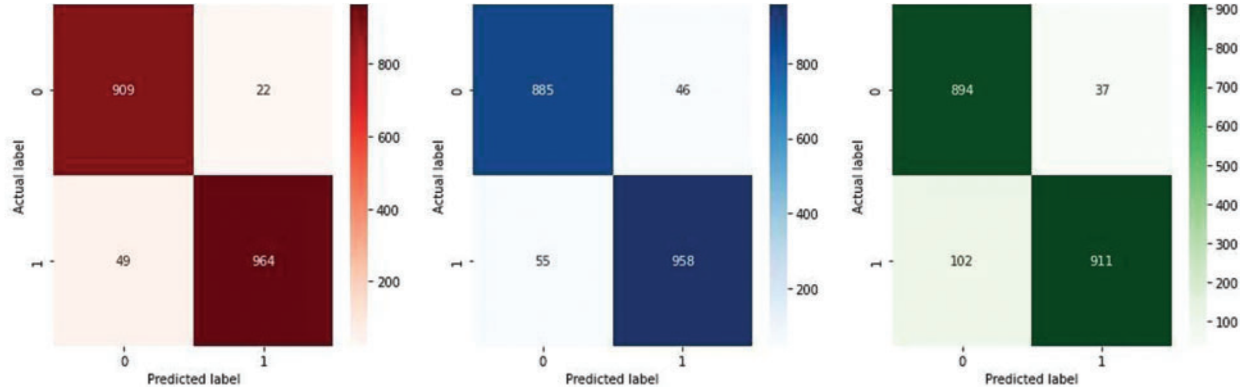

Figure 2: Confusion matrix of XGB, RF, and LGBM

Fig. 3 indicates the Receiver Operating Characteristic (ROC) curve of the RF and XGBoost classification algorithm. The green line in the curve represents the XGBoost, and the yellow line indicates RF. The AUC of RF and XGBoost is 0.986% and 0.989%, respectively.

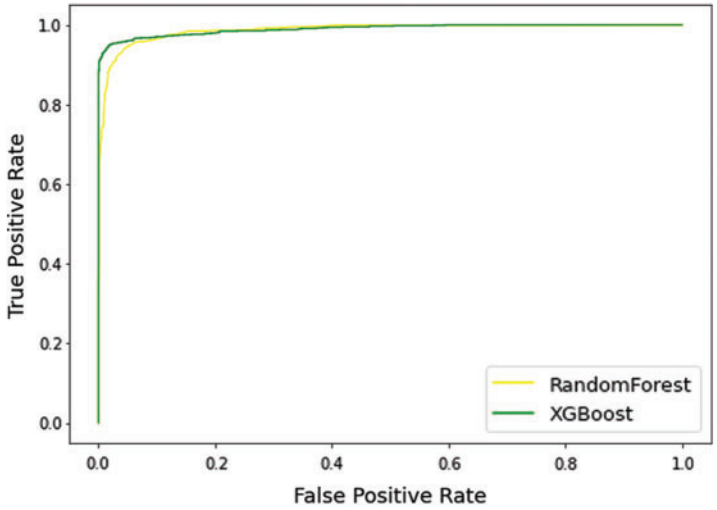

Figure 3: Receiver operating characteristic curve

Fig. 4 graphically represents the confusion matrix (CM) of the proposed model. It diagnoses the stroke in two categories: stroke or no stroke. Because it has more continuous, better true positive and negative results and fewer false positive and negative values, the proposed technique performs better.

The comparative analysis of the suggested approach with existing techniques [27,30,36,37] is provided in Table 4. The comparison is provided regarding the accuracy, precision, recall, and F1-score. The proposed approach outperforms as associated to the existing techniques.

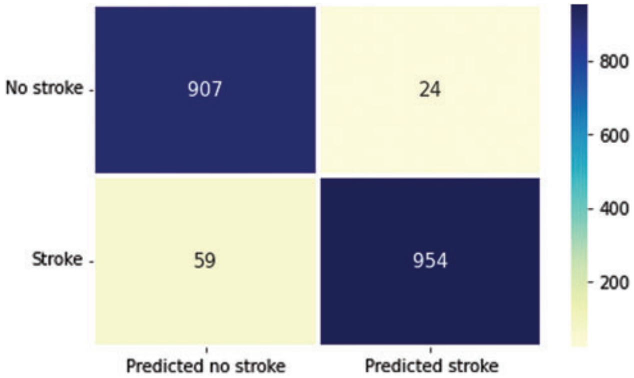

Figure 4: Confusion matrix of proposed model

Table 4: Comparison with existing techniques

References	Techniques	Accuracy	Precision	Recall	F1-score
[27]	NB	82%	79.2%	85.7%	82.3%
[30]	CNN-bidirectional LSTM	94%	94.6%	NA	94.1%
[36]	DNN	84.03%	NA	NA	NA
[37]	Dense-Net	85.62%	NA	NA	NA
Proposed approach	Ensemble voting	96%	97%	97%	96%

5 Conclusion

Human existence is based on different body components and their functions. As the organ that pumps blood to all other organs, the heart is regarded as the most important. Stroke is a severe disease that kills lives. Every aspect of life is substantially impacted because it is the foremost global reason for mortality and disability. In this research, three machine learning algorithms, RF, XGBoost, and LGBM, are used to propose an ensemble voting model. Data analysis is applied using bi-variate, categorical distribution, and pairwise analysis. It visualizes the feature individually to check the anomalies and outliers in the dataset. The stroke healthcare dataset with 11 important features is gathered from the Kaggle. The ensemble voting model performed well on the training dataset with 0.99% accuracy and 0.96% F-score on the test dataset. In the future, deep learning techniques (CNN, VGG16, and inception) and multiple datasets will be utilized to check the generalizability of the proposed model.

Funding Statement: The authors received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] C. Sharma, S. Sharma, M. Kumar and A. Sodhi, "Early stroke prediction using machine learning," in *2022 Int. Conf. on Decision Aid Sciences and Applications (DASA)*, Chiangrai, Thailand, IEEE, pp. 890–894, 2022.

[2] World Stroke Organization, 2022. [Online] Available: https://www.world-stroke.org/world-stroke-day-campaign/why-stroke-matters/learn-about-stroke

[3] 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.

[4] E. Dritsas and M. Trigka, "Stroke risk prediction with machine learning techniques," *Sensors*, vol. 22, no. 13, pp. 4670, 2022.

[5] M. Katan and A. Luft, "Global burden of stroke," in *Seminars in Neurology*, vol. 38, no. 2, United States: Thieme Medical Publishers, pp. 208–211, 2018.

[6] A. Bustamante, A. Penalba, C. Orset, L. Azurmendi, V. Llombart *et al.,* "Blood biomarkers to differentiate ischemic and hemorrhagic strokes," *Neurology*, vol. 96, no. 15, pp. e1928–e1939, 2021.

[7] Y. Guo, H. Wang, Y. Tian, Y. Wang and G. Y. Lip, "Multiple risk factors and ischemic stroke in the elderly Asian population with and without atrial fibrillation," *Thrombosis and Haemostasis*, vol. 115, no. 1, pp. 184–192, 2016.

[8] X. Xia, W. Yue, B. Chao, M. Li, L. Cao *et al.,* "Prevalence and risk factors of stroke in the elderly in northern China: Data from the national stroke screening survey," *Journal of Neurology*, vol. 266, no. 6, pp. 1449–1458, 2019.

[9] A. Alloubani, A. Saleh and I. Abdelhafiz, "Hypertension and diabetes mellitus as a predictive risk factors for stroke," *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, vol. 12, no. 4, pp. 577–584, 2018.

[10] J. Lecouturier, M. J. Murtagh, R. G. Thomson, G. A. Ford, M. White *et al.,* "Response to symptoms of stroke in the uk: A systematic review," *BMC Health Services Research*, vol. 10, no. 1, pp. 1–9, 2010.

[11] L. Gibson and W. Whiteley, "The differential diagnosis of suspected stroke: A systematic review," *The Journal of the Royal College of Physicians of Edinburgh*, vol. 43, no. 2, pp. 114–118, 2013.

[12] M. Rudd, D. Buck, G. A. Ford and C. I. Price, "A systematic review of stroke recognition instruments in hospital and prehospital settings," *Emergency Medicine Journal*, vol. 33, no. 11, pp. 818–822, 2016.

[13] B. Delpont, C. Blanc, G. Osseby, M. Hervieu-Bègue, M. Giroud *et al.,* "Pain after stroke: A review," *Revue Neurologique*, vol. 174, no. 10, pp. 671–674, 2018.

[14] S. A. Chohan, P. K. Venkatesh and C. H. How, "Long-term complications of stroke and secondary prevention: An overview for primary care physicians," *Singapore Medical Journal*, vol. 60, no. 12, pp. 616, 2019.

[15] M. J. M. Ramos-Lima, I. d. C. Brasileiro, T. L. d. Lima and P. Braga-Neto, "Quality of life after stroke: Impact of clinical and sociodemographic factors," *Clinics*, vol. 73, 2018.

[16] M. Gittler and A. M. Davis, "Guidelines for adult stroke rehabilitation and recovery," *JAMA Clinical Guidelines Synopsis*, vol. 319, no. 8, pp. 820–821, 2018.

[17] J. D. Pandian, S. L. Gall, M. P. Kate, G. S. Silva, R. O. Akinyemi *et al.,* "Prevention of stroke: A global perspective," *The Lancet*, vol. 392, no. 10154, pp. 1269–1278, 2018.

[18] V. L. Feigin, B. Norrving, M. G. George, J. L. Foltz, G. A. Roth *et al.,* "Prevention of stroke: A strategic global imperative," *Nature Reviews Neurology*, vol. 12, no. 9, pp. 501–512, 2016.

[19] E. Dritsas, N. Fazakis, O. Kocsis, N. Fakotakis and K. Moustakas, "Long-term hypertension risk prediction with ml techniques in elsa database," in *Int. Conf. on Learning and Intelligent Optimization*, Switzerland AG, Springer, pp. 113–120, 2021.

[20] N. Fazakis, E. Dritsas, O. Kocsis, N. Fakotakis and K. Moustakas, "Long-term cholesterol risk prediction using machine learning techniques in elsa database," in *Proc. of the 13th Int. Joint Conf. on Computational Intelligence*, pp. 445–450, 2021.

[21] M. A. Konerman, L. A. Beste, T. Van, B. Liu, X. Zhang *et al.,* "Machine learning models to predict disease progression among veterans with hepatitis c virus," *PLoS One*, vol. 14, no. 1, pp. e0208141, 2019.

[22] T. M. Alam, K. Shaukat, I. A. Hameed, W. A. Khan, M. U. Sarwar *et al.,* "A novel framework for prognostic factors identification of malignant mesothelioma through association rule mining," *Biomedical Signal Processing and Control*, vol. 68, pp. 102726, 2021.

[23] M. Z. Latif, K. Shaukat, S. Luo, I. A. Hameed, F. Iqbal *et al.,* "Risk factors identification of malignant mesothelioma: A data mining-based approach," in *Int. Conf. on Electrical, Communication, and Computer Engineering (ICECCE)*, Istanbul, Turkey, IEEE, pp. 1–6, 2020.

[24] T. M. Alam, K. Shaukat, H. Mahboob, M. U. Sarwar, F. Iqbal *et al.,* "A machine learning approach for identification of malignant mesothelioma etiological factors in an imbalanced dataset," *The Computer Journal*, vol. 65, no. 7, pp. 1740–1751, 2022.

[25] K. Shaukat, F. Iqbal, T. M. Alam, G. K. Aujla, L. Devnath *et al.,* "The impact of artificial intelligence and robotics on the future employment opportunities," *Trends in Computer Science and Information Technology*, vol. 5, no. 1, pp. 050–054, 2020.

[26] M. Khushi, K. Shaukat, T. M. Alam, I. A. Hameed, S. Uddin *et al.,* "A comparative performance analysis of data resampling methods on imbalance medical data," *IEEE Access*, vol. 9, pp. 109960–109975, 2021.

[27] 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.

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

[29] M. U. Emon, M. S. Keya, T. I. Meghla, M. M. Rahman, M. S. Al Mamun *et al.,* "Performance analysis of machine learning approaches in stroke prediction," in *4th Int. Conf. on Electronics, Communication and Aerospace Technology (ICECA)*, Coimbatore, India, IEEE, pp. 1464–1469, 2020.

[30] G. Fang, Z. Huang and Z. Wang, "Predicting ischemic stroke outcome using deep learning approaches," *Frontiers in Genetics*, vol. 12, pp. 827522, 2021.

[31] R. C. Suganthe, M. Geetha, G. R. Sreekanth, K. Gowtham, S. Deepakkumar *et al.,* "Multiclass classification of Alzheimer's disease using hybrid deep convolutional neural network," *Nveo-Natural Volatiles & Essential Oils Journal Nveo*, pp. 145–153, 2021.

[32] S. Tiwari, A. Jain, V. Sapra, D. Koundal, F. Alenezi *et al.,* "A smart decision support system to diagnose arrhythmia using ensembled convnet and convnet-lstm model," *Expert Systems with Applications*, vol. 213, pp. 18933, 2023.

[33] S. Tiwari, "An ensemble deep neural network model for onion-routed traffic detection to boost cloud security," *International Journal of Grid and High Performance Computing (IJGHPC)*, vol. 13, no. 1, pp. 1–17, 2021.

[34] Y. -A. Choi, S. -J. Park, J. -A. Jun, C. -S. Pyo, K. -H. Cho *et al.,* "Deep learning-based stroke disease prediction system using real-time bio signals," *Sensors*, vol. 21, no. 13, pp. 4269, 2021.

[35] P. Chantamit-O-Pas and M. Goyal, "Prediction of stroke using deep learning model," in *Int. Conf. on Neural Information Processing*, Sydney, Australia, Springer, pp. 774–78, 2017.

[36] S. Cheon, J. Kim and J. Lim, "The use of deep learning to predict stroke patient mortality," *International Journal of Environmental Research and Public Health*, vol. 16, no. 11, pp. 1876, 2019.

[37] Y. Xie, H. Yang, X. Yuan, Q. He, R. Zhang *et al.,* "Stroke prediction from electrocardiograms by deep neural network," *Multimedia Tools and Applications*, vol. 80, no. 11, pp. 17291–17297, 2021.

[38] K. Shaukat, S. Luo, V. Varadharajan, I. A. Hameed and M. Xu, "A survey on machine learning techniques for cyber security in the last decade," *IEEE Access*, vol. 8, pp. 222310–222354, 2020.

[39] K. Shaukat, S. Luo, V. Varadharajan, I. A. Hameed, S. Chen *et al.,* "Performance comparison and current challenges of using machine learning techniques in cybersecurity," *Energies*, vol. 13, pp. 2509, 2020.

[40] N. Sharma, H. V. Bhandari, N. S. Yadav and H. Shroff, "Optimization of ids using filter-based feature selection and machine learning algorithms," *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, vol. 10, no. 2, pp. 96–102, 2020.

[41] V. G. Raju, K. P. Lakshmi, V. M. Jain, A. Kalidindi and V. Padma, "Study the influence of normalization/-transformation process on the accuracy of supervised classification," in *2020 Third Int. Conf. on Smart Systems and Inventive Technology (ICSSIT)*, Tirunelveli, India, IEEE, pp. 729–735, 2020.

[42] N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer, "Smote: Synthetic minority oversampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[43] B. S. Raghuwanshi and S. Shukla, "Classifying imbalanced data using smote based class-specific kernelized elm," *International Journal of Machine Learning and Cybernetics*, vol. 12, no. 5, pp. 1255–1280, 2021.

[44] D. Mpanya, T. Celik, E. Klug and H. Ntsinjana, "Machine learning and statistical methods for predicting mortality in heart failure," *Heart Failure Reviews*, vol. 26, no. 3, pp. 545–552, 2021.

[45] S. Hossain, P. Biswas, P. Ahmed, M. R. Sourov, M. Keya *et al.,* "Prognostic the risk of stroke using integrated supervised machine learning techniques," in *2021 12th Int. Conf. on Computing Communication and Networking Technologies (ICCCNT)*, Kharagpur, India, IEEE, pp. 1–5, 2020.

[46] J. Yang and J. Guan, "A heart disease prediction model based on feature optimization and smote-xgboost algorithm," *Information*, vol. 13, no. 10, pp. 475, 2022.

[47] S. -Y. Lin, K. -M. Law, Y. -C. Yeh, K. -C. Wu, J. -H. Lai *et al.,* "Applying machine learning to carotid sonographic features for recurrent stroke in patients with acute stroke," *Frontiers in Cardiovascular Medicine*, vol. 9, 2022.

[48] F. Alzamzami, M. Hoda and A. El Saddik, "Light gradient boosting machine for general sentiment classification on short texts: A comparative evaluation," *IEEE Access*, vol. 8, pp. 101840–101858, 2020.

[49] K. Shaukat, S. Luo and V. Varadharajan, "A novel method for improving the robustness of deep learning-based malware detectors against adversarial attacks," *Engineering Applications of Artificial Intelligence*, vol. 116, pp. 105461, 2022.