Healthcare Analytics 2 (2022) 100116

Contents lists available at ScienceDirect

Healthcare Analytics

journal homepage: www.elsevier.com/locate/health

A comparative analysis of machine learning classifiers for stroke prediction: A predictive analytics approach

Nitish Biswas [a], Khandaker Mohammad Mohi Uddin [a,*], Sarreha Tasmin Rikta [a], Samrat Kumar Dey [b]

[a] Department of Computer Science and Engineering, Dhaka International University, Dhaka 1205, Bangladesh
[b] School of Science and Technology, Bangladesh Open University, Gazipur 1705, Bangladesh

ARTICLE INFO

Keywords:
Machine learning
Stroke
Support vector machine
Random forest
Random over-sampling
Hyperparameter tuning
Cross-validation

ABSTRACT

Stroke is the third leading cause of death in the world. It is a dangerous health disorder caused by the interruption of the blood flow to the brain, resulting in severe illness, disability, or death. An accurate prediction of stroke is necessary for the early stage of treatment and overcoming the mortality rate. This study proposes a machine learning approach to diagnose stroke with imbalanced data more accurately. Random Over Sampling (ROS) technique has been used in this work to balance the data. Eleven classifiers, including Support Vector Machine, Random Forest, K-nearest Neighbor, Decision Tree, Naïve Bayes, Voting Classifier, AdaBoost, Gradient Boosting, Multi-Layer Perception, and Nearest Centroid, are analyzed in this study. Ten classifiers show more than 90% accurate results before balancing the data and four classifiers display more than 96% accurate results after data-balancing using the oversampling method. The Hyperparameter tuning and cross-validation are performed in each model to enhance the results. Moreover, Accuracy, F1-Measure, Precision, and Recall are used to measure the performance of machine learning models. The results show the Support Vector Machine has the highest accuracy of 99.99%, with recall values of 99.99%, precision values of 99.99%, and F1-measure of 99.99%. Random Forest achieves the second-highest accuracy of 99.87%, with a 0.001% error. In addition, a user-friendly web app and a user-friendly mobile app are built based on the most accurate model.

1. Introduction

Stroke is one of the leading causes of death and a globally serious threat to public health; is known as a cerebrovascular accident (CVA). According to World Health Organization (WHO), strokes are defined as an acute, global disturbance or dysfunctions of the blood vessels supplying the causes of limb paralysis, severe morbidity, and unconsciousness [1]. The symptoms of stoke last more than 24 h and can cause death 3 to 10 h [2]. This is found in a WHO survey that fifteen million people are undergoing stroke and every 4 to 5 min, individuals passed away [3]. Strokes can mainly be divided into two types — ischemic and hemorrhagic [4]. When a blood vessel supplying to the brain is obstructed or blocked because of a blood clot called an ischemic stroke which is accounting for 87% of all strokes according to the American Heart Association (AHA) [5].

On the contrary, Hemorrhagic stroke occurs when a weakened blood vessel bursts or leaks blood, 15% of strokes account for hemorrhagic [5]. This disease is rapidly increasing in developing countries such as China, with the highest stroke burdens [6], and the United States is undergoing chronic disability because of stroke; the total number of people who died of strokes is ten times greater in those countries than in the past five decades. It is a piece alarming news that the WHO proposed mortality rate due to stroke is number 84 position in the world [7]. About 700 thousand people are affected by this disease every year. Researchers have studied different methodologies such as prospective cohort studies, case-control studies, and case series in the past several decades and identified nonmodifiable risk markers (genetic factors, male gender, older age) and modifiable risk factors (Hypertension, Cigarette smoking, Diabetes Mellitus) [8]. Fig. 1 represents risk factors for ischemic stroke.

The mortality rate and the number of affected people by this disease are expected to grow with the population of the world. But this mortality rate can prevent by early treatment and early prediction. There are some tools like the cox proportional hazard model that are used to predict stroke. But it is unable to effectively prognosis to stroke with high dimensional data because of being a traditional method. In that case, machine learning can play a vital role in predicting stroke

* Corresponding author.
E-mail addresses: nitishbiswas.cse@gmail.com (N. Biswas), jilanicsejnu@gmail.com (K.M.M. Uddin), tasmin.sarreha@gmail.com (S.T. Rikta), samrat.sst@bou.ac.bd (S.K. Dey).

https://doi.org/10.1016/j.health.2022.100116
Received 2 July 2022; Received in revised form 30 August 2022; Accepted 7 October 2022

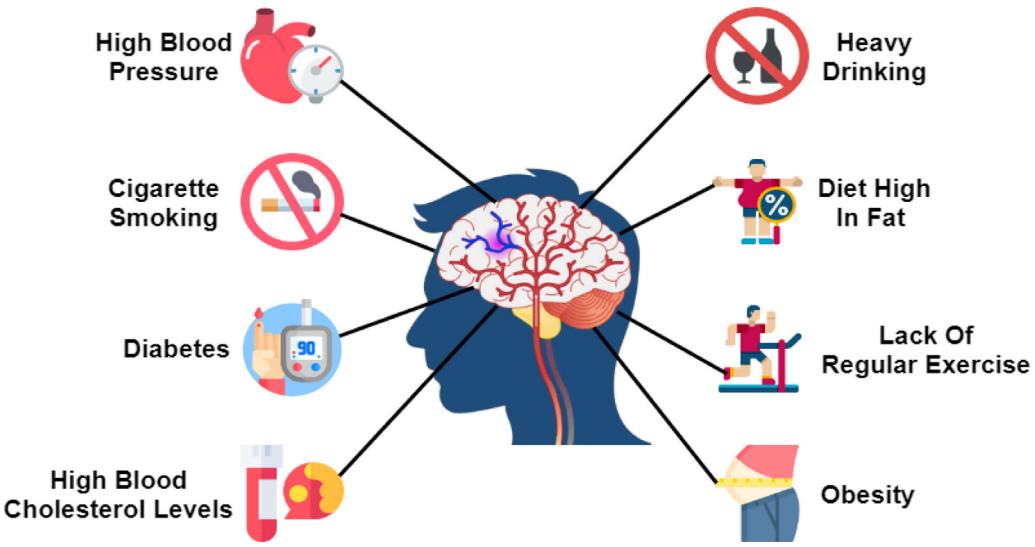

Fig. 1. Risk factors for ischemic stroke.

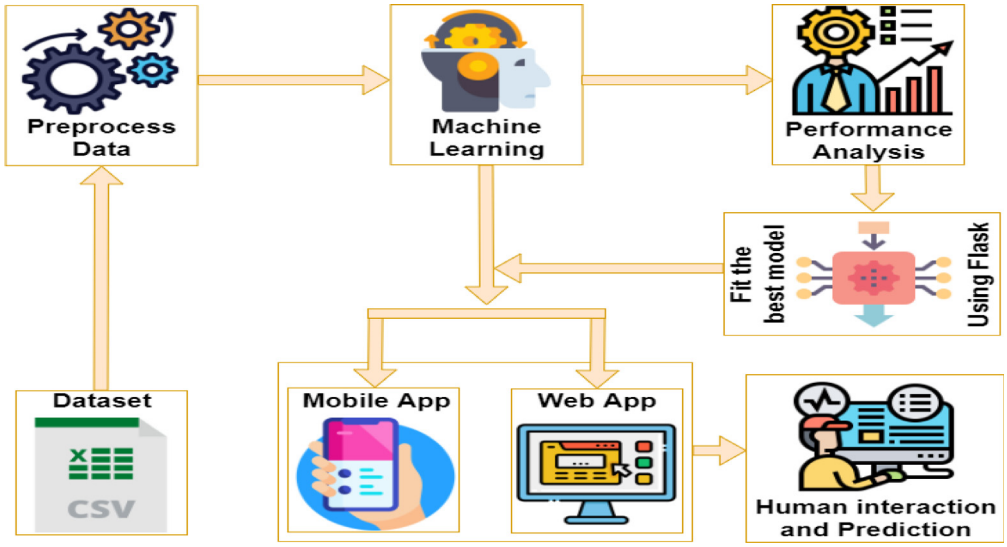

Fig. 2. Overview of the proposed system.

effectively and efficiently with a lower cost. Different machine learning classifiers such as Random Forest, Support Vector Machine, Decision Tree, and Logistic Regression are used in medical fields for many years making correct analyses and predicting accurate results relying on patterns with the big imbalanced dataset. This study shows the highest result for stroke prediction using data balancing techniques, machine learning algorithms with various kinds of risk factors, and an imbalanced dataset.

The overview of the proposed system is illustrated in Fig. 2. The dataset is pre-processed in the first phase. The pre-processed dataset is then input into several machine learning algorithms in the second phase. The output of the models is then examined using various metrics in the third phase. In the subsequent phase, the model with the best accuracy is used to identify any individual's stroke and is connected with a web-based and mobile application.

Our contribution to this proposed research work-

1. This work achieves a higher accuracy with 99.99% than the previous accuracy on this specific topic performed by other researchers.
2. Eleven classifiers and different machine learning techniques including oversampling, hyperparameter tuning, and cross-validation are employed in this research work to reach the best result.
3. A web page, as well as a mobile app, are developed based on this research work that can calculate the result accurately using real-time inputs.
4. Among the eleven classifiers, SVM and RF show the maximum accuracy respectively 99.99% and 99.87%.

The rest of the paper is organized into four sections where Section 2 represents related work. The proposed materials and method of this paper for predicting stroke disease are illustrated in Section 3. Section 3 is divided into nine subsets: dataset description, data pre-processing, label encoding, Imbalanced data handling, Hyperparameter tuning, data visualization, machine learning classifiers, implementation of a web application, and implementation of the mobile app. The result and analysis are discussed in detail in Section 4. This section is also organized into two subsections — environment setup, confusion matrix, and result analysis. Finally, Section 5 concludes the entire work of this paper.

2. Related works

The computational dependability of artificial intelligence has been a blessing for medical diagnostics and clinical data processing in the modern era (AI). Where human eyesight is restricted, AI has extended. Currently, a number of applications using various machine learning algorithms are being employed in the field of medical research for data processing and innovation. The utilization of machine learning techniques has been observed in a number of recent healthcare studies, including the detection of COVID-19 using X-rays [9,10], the detection of tumors using MRIs [11,12], the prediction of heart diseases [13, 14], the detection of dengue diseases [15,16] and the diagnosis of cancer [17,18], and [19]. Many studies have previously utilized machine learning to predict stroke prognosis. This section presents the accomplishments of other researchers on this field.

Minhaz et al. [20] also conducted research on stroke where they obtained the data from many hospitals in Bangladesh. After data preprocessing, ten algorithms are employed to train. Then, the weighted voting classifier is applied to improve the performance of all classifiers. Then all model is optimized and find out the best model uses a weighted voting classifier. This research concludes that the weighted voting classifier has a maximum accuracy of 97%.

Yoon-A et al. [21] has done a study on stroke where EEG (Electroencephalography) biometrics signals during walking can detect the stroke. This paper claim that Random Forest can predict stroke with the biometric signal's methods.

In another study, Priya et al. [22] predict the stroke using text mining tools as well as machine learning algorithms. The authors use 14 classification methods such as a simple tree, medium tree, complex tree, logistic regression, Linear SVM, Quadratic SVM, and ANN. From this experiment, ANN gives a higher accuracy of 95.3% compared to others.

Sailasya et al. [23] considered different datasets from Kaggle and they operated data preprocessing including missing value handling, label encoding, and imbalanced data handling. After data preprocessing, six machine learning algorithms are applied to this dataset. After comparing these algorithms' accuracy, the Naïve Bayes classification gives the maximum accuracy of 82%. In addition, they developed an HTML page from which the user can get the result whether he or she has a stroke or not giving some parameters.

Hager et al. [24] used four types of classifiers Logistic regression, Random Forest, Decision Tree, and Support Vector Machine to predict stroke. They applied hyperparameter tuning and cross-validation in the machine learning algorithms for getting the result. After that, they evaluated the performance of these four models and attain the highest 90% accuracy for Random Forest among these four models.

Wu et al. [25] proposed a stroke prediction model with imbalanced data. In this study, they collected the imbalanced data from a Chinese longitudinal Healthy longevity study where they processed the balanced data using ROS, RUS, and SMOTE techniques. In this work, the authors used regularized logistic regression, a support vector machine, random forest models were used to predict the stroke for the imbalanced dataset as well as balanced dataset where the best result was compared with the individual dataset. From this comparison, they showed SVM and RLR give the highest accuracy of 95% for the imbalanced dataset but extremely low sensitivity.

Badriyah et al. [26] have collected the CT scan data from the Hajj Hospital in Surabaya, Indonesia of those patients who is suffering from a stroke. During pre-processing, image processing techniques such as data conversion, cropping, scaling, grayscale, and data augmentation are performed here to improve the image quality. Moreover, feature extraction is also applied to image data. After that, eight algorithms' accuracy like Decision Tree, Logistic Regression, Naïve Bayes, and Random Forest are compared. In this experiment, Random Forest achieves the highest accuracy of 95.97% compared to other classifiers.

Jaehak et al. [27] also conducted research on stroke where they predict the stroke using real-time bio-signals with AI. Random Forest (Machine learning) and Long Short-Term Memory (Deep learning) algorithms are used in this system where LSTM attains the best result (98.58%).

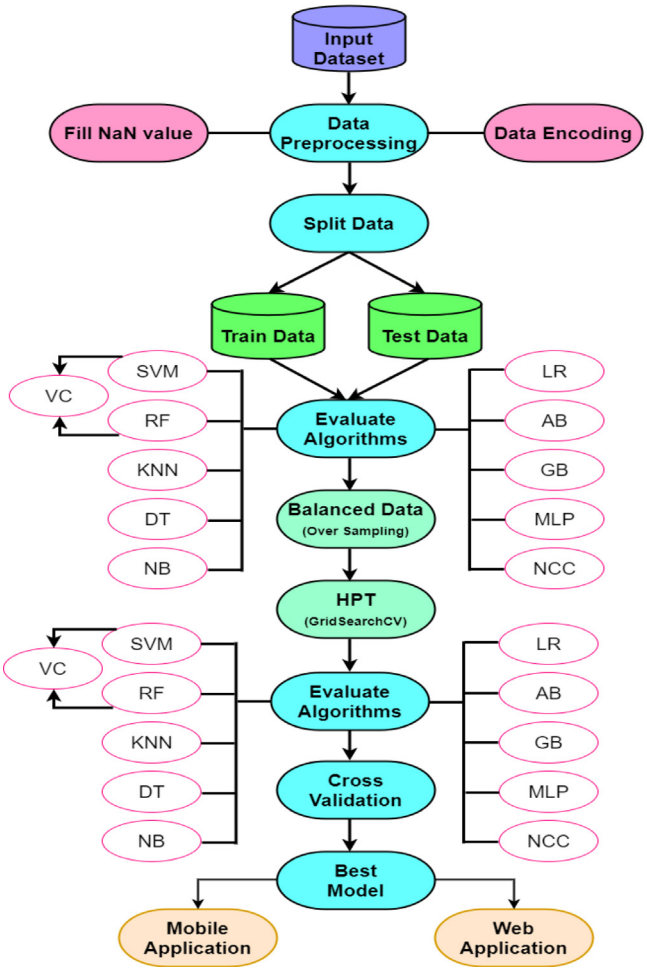

Fig. 3. Working procedure of proposed model.

Yoon-A et al. [28] publish another report using EEG data with deep learning models where LSTM shows the maximum result with 94% accuracy.

In this section, it is observed that different machine learning approaches, bio-metric signals, and text mining tools are used here to obtain the best accuracy. But all these approaches and results show different results which are actually baffling. That is why an easy machine learning approach is implemented here that has been able to achieve the highest accuracy compared to the previous work.

3. Materials and methods

The methodology of the proposed system is described in this section. First of all, data has been pre-processed from the datasets where null values or missing values are filled up and data encoding is performed to transform categorical variables into numerical values. Then, the pre-processed data is divided into 2 parts-train data and test data. After that, the training data is fed into the different machine learning algorithms to predict the outcome of stroke. Afterward, the Random oversampling method is used for balancing the data, and GridSearchCv is applied to evaluate the machine learning models for a range of hyperparameter values. Thereafter, machine learning classifiers are evaluated again, and by performing cross-validation the performance of machine learning models are estimated. Fig. 3 depicts the proposed methodology for stroke prediction.

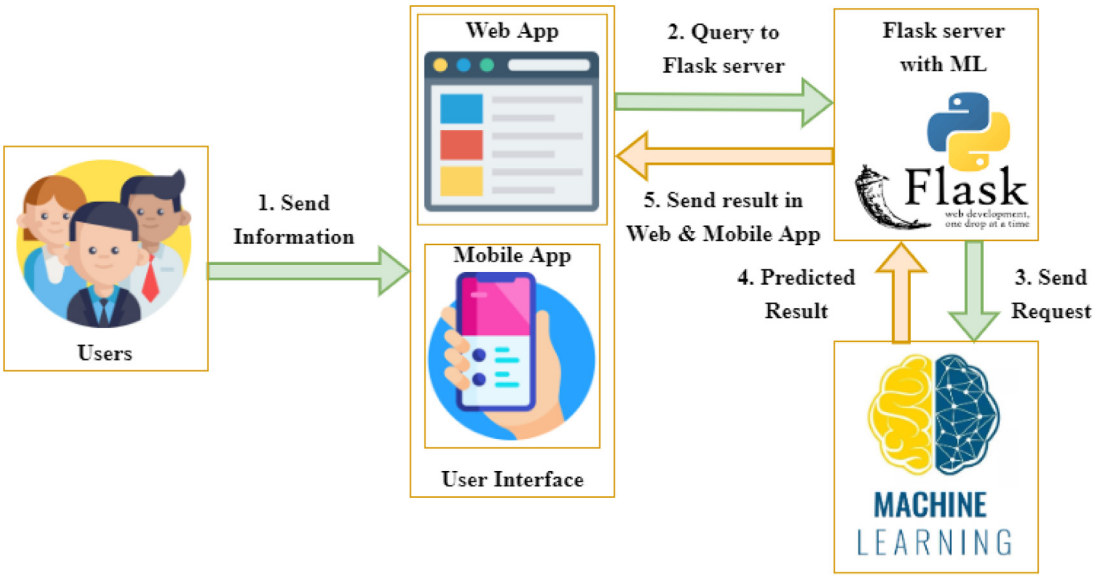

Fig. 4. Work flow of the web and mobile application.

After the training of all models and the calculation of accuracy, a web application and a mobile app are created from which a user may forecast whether or not they will have a stroke by entering the input variables. Algorithm 1 shows the working procedure of the proposed framework. Work flow of the web and mobile application of the proposed system is depicted in Fig. 4.

Algorithm 1: Working Procedure of Stroke Prediction

Input: *Kaggle Stroke Dataset*

Output: *Predicted value of Stroke (Positive or Negative)*

Begin
    data ← load dataset
    if *data.value* ***is*** *equal NaN or empty*
        replace NaN or missing_value
    pre-processing:
        if *data.dtypes* ***is*** *equal object*
            encoding the data
    x ← data.drop[stroke]
    y ← data.stroke
    x1, x2, y1, y2 ← split_data of x and y
    model ← train_model using x1and y1
    predict ← testing_model using x2 and y2
    b_x ← balanced data of x
    for *i in* ***range(len****(models)):*
        checking for HTP
        apply to all models
    classifier ← train_model using b_x
    model ← apply_voting_classifier using classifier
    predict ← cross_validation with model
    computes performance evaluation metrics
End

3.1. Dataset description

The dataset in this research work were obtained from the Kaggle [29] containing 43 400 instances and 12 features. The columns consist of 'age', 'hypertension', 'gender', 'heart_disease', 'Residence_type', 'avg_glucose_level', 'bmi', 'ever_marri ed', 'work_type', 'smoking_status' and 'stroke'. The stroke column represents the outcome where 0 indicates no stroke and 1 indicates stroke detected. This dataset contains 42 617 non-stroke and 31 962 strokes after balanced the data; 42 617 non-stroke detection and 783 strokes detected before balanced the data. Table 1 shows the predictive variables, known as an independent variable which is used to predict the dependent variables. The priority ratio of these independent features refers that how much they have a strong relationship with the dependent features or the outcome.

Table 1
The importance of predictive variables in Stroke.

Features	Priority ratio (%)
age	0.015723
work_type	0.007660
ever_married	0.006517
bmi	0.003886
heart_disease	0.003712
avg_glucose_level	0.002885
smoking_status	0.002875
Residence_type	0.001731
gender	0.001556
hypertension	0.001409

3.2. Data pre-processing

Data pre-processing is needful before feeding these datasets into machine learning models to reach maximum accuracy. The pre-processing techniques are used to remove unwanted noise, missing values, outliers, label encoding, and so on. After data cleaning machine learning models are applied to the dataset. This dataset consists of 12 attributes from which the 'id' column is eliminated as it does not much affect the outcome. The column 'bmi', and 'smoking_status' contains the null value that has been filled. Then label encoding is employed in the datasets.

After that normalization technique is used to prepare data which transforms the value of the numeric column to use a common scale that lies between 0 and 1 and the data standardization rescales a dataset using a mean of 0 and standardization deviation of 1 for scaling of model.

The process of normalization is frequently used to prepare data for machine learning. The purpose of normalization is to convert the values of the dataset's numeric columns to a standard scale without losing information or distorting the ranges of values. The Table 2 represents

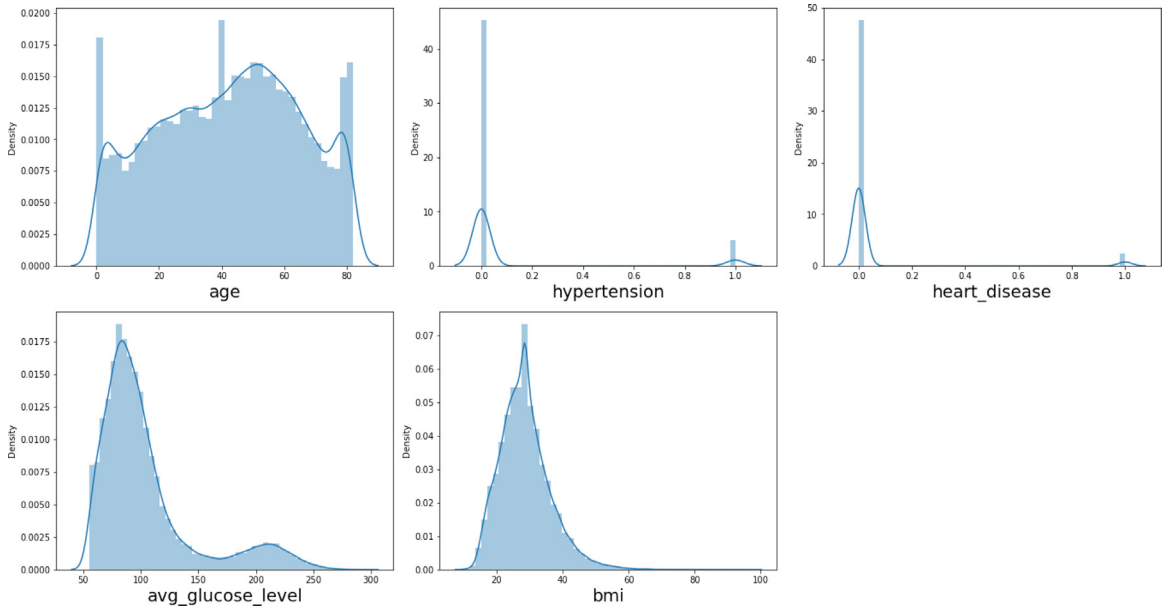

Fig. 5. Numerical column representation.

Table 2
Statistical analysis of the dataset of numerical features.

Features	Mean	Std	Max	Min
age	42.217894	22.519649	82.00000	0.080000
Bmi	28.605038	7.638023	97.60000	10.10000
heart_disease	0.047512	0.212733	1.000000	0.000000
avg_glucose_level	104.482750	43.111751	291.0500	55.00000
hypertension	0.093571	0.291235	1.000000	0.000000

Table 3
Statical analysis of the dataset of categorical features.

Features	Label	Count
gender	0	25 665
	1	17 724
	2	11
ever_married	0	15 462
	1	27 938
work_type	0	5440
	1	177
	2	24 834
	3	6793
	4	6156
Residence_type	0	21 644
	1	21 756
smoking_type	0	7493
	1	16 053
	2	6562
	3	13 292

distinct features of this dataset and Fig. 5 represents the numerical column representation of age, bmi, heart_disease, avg_glucose_level and hypertension.

3.3. Label encoding

Label encoder refers to encoding the categorical value with a numerical value to fit in the machine learning model smoothly [23]. There are five columns named 'gender', 'ever_married', 'work_type', 'residence_type', and 'smoking_status' that comprise string values. All these string values are converted into a combination of numerical values. The encoded categorical features are shown in Table 3.

3.4. Imbalanced data handling

The used dataset in this study for stroke prediction is highly asymmetry which influences the result. Fig. 6 shows the graphical representation of the imbalanced data as well as balanced data. Training a machine learning model with an imbalanced dataset gives poor performance and inaccurate results. So, for achieving the promising accuracy with the highest result, the imbalance dataset is handled first. Therefore, the data balancing technique named Random Oversampling, known as non-heuristics algorithms is applied to the dataset. The main purpose of this method is to randomly duplicate the examples from the minority class [30]. By repeating the original samples, the Random Over Sampler expands the dataset. The key aspect is that the random over sampler does not generate new samples, and the sample diversity remains constant [31].

After oversampling, the dataset contains 42 617 rows with a value of 0 which means no stroke detected, and 31 962 rows with a value of 1 means stroke detected.

3.5. Hyperparameter tuning

Hyperparameters are the parameters whose value is defined before beginning the machine learning algorithms processes [32]. It controls the machine learning models' behaviors. Without using it, more errors may occur in the model. Therefore, hyperparameter tuning is applied with gridsearchcv in this study. In GridSearchCv techniques, Grid search along with cross-validation has been implemented in this work. GridSearchCv checks all combinations of the values passed in the dictionary and evaluate the model for each combination. Thus, the best accuracy is achieved for every combination of hyperparameters and selected the best model.

3.6. Data visualization

Data visualization is a machine learning tool that provides a graphical representation, analyzes, and observes the datasets which helps to understand the human brain easily. Fig. 7 depicts the correlation of this dataset where blue color refers negative and green color refers positive. A correlation matrix measures the relationship between two variables. The stronger the green color, the stronger relationship. It is observed from the above Fig. 7 that age is highly correlated with ever_married, they have a 69% positive relationship between them which is strong enough.

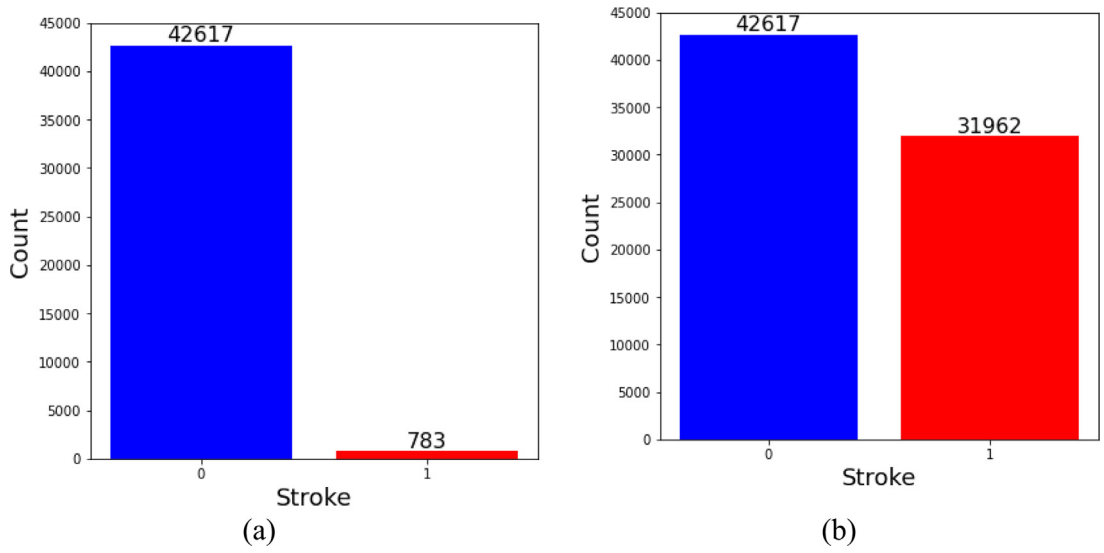

Fig. 6. Class distribution of stroke (a) before balanced and (b) after balanced using oversampling.

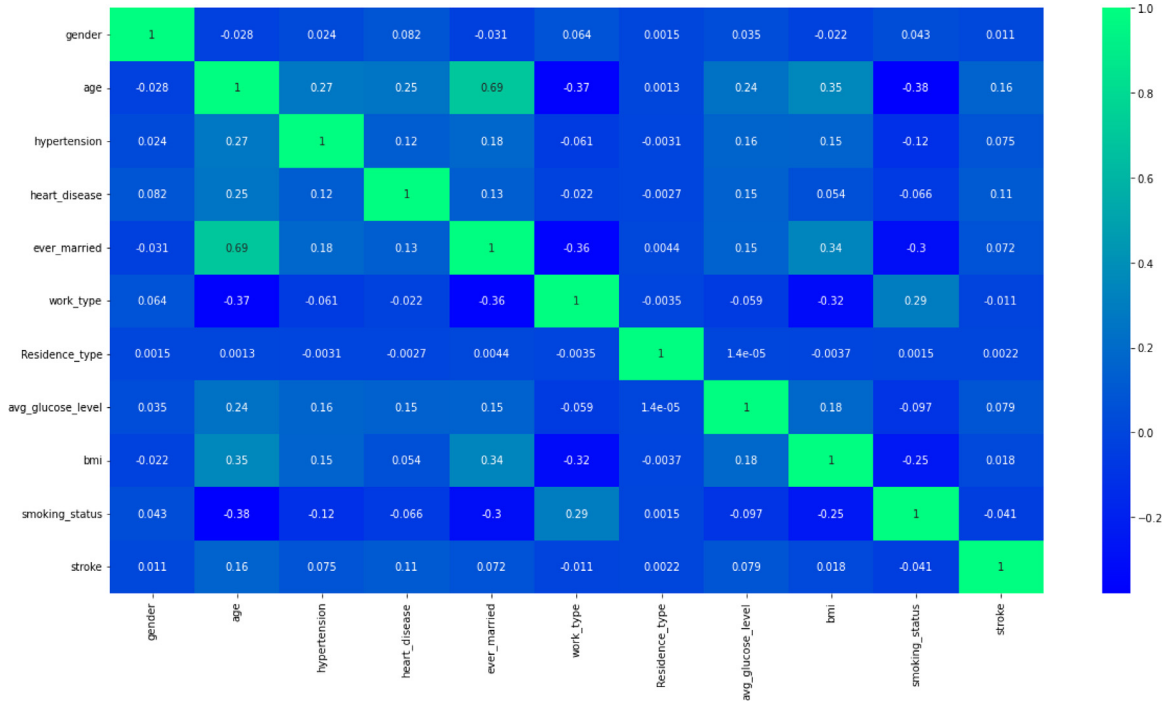

Fig. 7. Correlation with each feature. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3.7. Machine learning classifiers

The performance is measured using 11 classifiers namely Support Vector Machine, Random Forest, K-Nearest Neighbor, Decision Tree, Naïve Bayes, Logistic Regression, AdaBoost, Gradient Boosting, Nearest Centroid, voting classifier, and Multilayer perceptron which are discussed in this section. After that, the best model is evaluated with the highest accuracy among all these classifiers.

3.7.1. Support vector machine

SVM is a popular supervised learning algorithm that is used in medical fields for many years for predicting the result and used for classification and regression problems [33,34]. This classification performs the most suitable hyperplane distinguishing the dataset between two classes [35]. Eq. (1) refers to a linear combination linking kernel with SVM where S is the SVM, p_j denotes the patterns of the training set and $q_j \ \epsilon\{+1, -1\}$ is the respective class labels.

$$\mathrm{f(p)} = \sum_{p_j \epsilon S} \alpha_j q_j K(p_j, p) + b \tag{1}$$

3.7.2. Random forest

Random Forest (RF) is the collection of decision trees on different samples that takes the average to improve the accuracy of the given dataset. It can solve classification problems along with regression problems [36]. Eq. (2) is the equation of basic decision tree in RF [37].

$$\mathrm{m(x)} = \sum_{m=1}^{M} w_m \Pi(x \in R_m) = \sum_{m=1}^{M} w_m \phi(x; \upsilon_m) \tag{2}$$

where $\mathrm{w_m}$ is the average response of the m_{th} region R_m, V_m is the variable that are used to split on, and $\phi(\mathrm{x}; \mathrm{v}_m)$ is the threshold information.

3.7.3. K-nearest neighbor

K-NN is a simple and non-parametric supervised learning algorithm that finds similar features in the training set [38]. It is commonly based on Euclidian, Manhattan, and Minkowski distance algorithm that evaluates the distance from new data to all others data. Eqs. (3), (4), and (5) represents the formula of Euclidian, Manhattan, and Minkowski distance calculation, respectively.

$$Euclidean = \sqrt{\sum_{i=1}^{k} \left(x_i - y_i\right)^2} \tag{3}$$

$$Manhattan = \sum_{i=1}^{k} |\boldsymbol{x_i} - \boldsymbol{y_i}| \tag{4}$$

$$Minkowski = \left(\sum_{i=1}^{k} (|\boldsymbol{x_i} - \boldsymbol{y_i}|)^q\right)^{\frac{1}{q}} \tag{5}$$

3.7.4. Decision tree

DT is also a supervised learning classifier that is non-parametric. A decision tree contains a number of internal nodes that makes the decision and the final outcome is represented using leaf nodes. A node of DT is designed by using the Eq. (6) [39].

$$\mathrm{E}''(\boldsymbol{d_i}) = [\boldsymbol{T_0}(\boldsymbol{d_i}) - \boldsymbol{T}(\boldsymbol{d_i})] + \boldsymbol{K}[\in_{\mathbf{0}} (\boldsymbol{d_i}) - \in (\boldsymbol{d_i})] + \sum_{j=1}^{c_i} \boldsymbol{P_{i+j}} \boldsymbol{E_0}(\boldsymbol{d_{i+j}}) \tag{6}$$

It is the evaluation function for a node where the heuristic search is applied based on $\mathbf{E}''$ that gives the accurate decision.

3.7.5. Naïve Bayes

Naive Bayes is a simple and effective classifier that can predict the result quickly on many complicated problems. It works based on the Bayes theorem and assumes a certain feature independently. Eq. (7) represents the Bayes theorem where p(B|A) means the posterior probability, P(A|B) is the likelihood probability and p(B) is the prior probability [40].

$$\mathrm{p}\,(B|A) = \frac{p(A,\,B)}{p(A)} = \frac{\mathrm{p}\,(A|B)\,\mathrm{p(B)}}{\sum_{y'=1}^{C} p\,(A|B')\,p(B')} \tag{7}$$

3.7.6. Logistic regression

Logistic Regression is a popular supervised learning algorithm that calculate the probability and predicts the outcome of a categorical dependent variable making a relationship between dependent variables and independent variables [41]. The sigmoid function maps the predicted values either 0 or 1 to probabilities. The probability is calculated by the Eq. (8).

$$\hat{y}j = \frac{e^x \hat{b}j}{1 + e^x \hat{b}j} \tag{8}$$

3.7.7. Adaptive boosting

An AdaBoost is called Adaptive boosting which is a boosting technique that is propagated from the boosting algorithm [42]. The aim of this algorithm is to combine multiple weak classifiers into a single strong classifier. The Eq. (9) represents the AdaBoost classifier.

$$\mathrm{F(x)} = \ \mathrm{sign}(\sum_{m=1}^{m} \theta_{\mathrm{m}} \mathrm{f_m}\,(x)) \tag{9}$$

where m denotes the weight and f_m denotes the mth weak classifier.

3.7.8. Gradient boosting

Gradient boosting is a group of machine learning algorithms that combine many weak learning models together to create a strong predictive model. Decision trees are usually used when doing gradient boosting. Gradient boosting models are becoming popular because of their effectiveness at classifying complex datasets. The Eq. (10) represents the gradient decent size [43].

$$\rho_t = argmin_\rho \sum_{i=1}^{N} \psi \left[y_i, \hat{f}_t - 1\left(x_i\right) + \rho h(x_i, \theta_i)\right] \tag{10}$$

3.7.9. Multi-layer perceptron

MLP is a feedforward neural network, consists of three layers including input layer, output layer and hidden layer. It has a linear activation function. Eq. (11) describes the two-activation function where both are sigmoid.

$$\mathrm{y}(\upsilon \mathrm{i}) = \tanh(\upsilon \mathrm{i}) \text{ and } \mathrm{y}(\upsilon \mathrm{i}) = (1 + \mathrm{e} - \upsilon \mathrm{i}) - 1 \tag{11}$$

And Eq. (12) denotes compact notation for a MLP where x represents the input vector, $\mathrm{F^l}$ is the activation function, $\mathrm{W^l}$ represents first column of matrix and $\hat{\mathrm{O}}^{(\mathrm{l}-1)}$ is an operation [44].

$$O^0 = X, O^1 = F^1(W^1 \hat{O}^{(1-1)}) \ for \ 1 = 1, \ldots, \tag{12}$$

3.7.10. Nearest centroid

A machine learning algorithm for linear classification is called nearest centroids. It entails determining which class-based centroid from the training dataset a new example will belong to in order to forecast its class label. According to this approach, test samples are assigned to the class that has the closest centroid after being represented by each class' centroid. Eq. (13) [45] calculates the centroid of per class.

$$\vec{\mu}_l = \frac{1}{|C_l|} \sum_{i \in C_i} \vec{x}_i \tag{13}$$

where C_l is the collection of sample indexes that belong to the class $l_{\in} Y$.

3.7.11. Voting classifier

A voting classifier is a machine learning classifier which trains numerous models and finds the highest probability based on the chosen class category. It has two types-Hard voting and soft voting [46]. Hard voting is used in this proposed work which mathematical formula is represented in Eq. (14).

$$\hat{q} = mode\{C_1\,(p), C_2\,(p), \ldots, C_m\,(p)\} \tag{14}$$

3.8. Implementation of web application

The construction of a web application for stroke prediction is described in this section. The web page is developed using react. This webpage can take the input from a user and predict the stroke. Input form of a user is shown in Fig. 8

For developing this webpage, firstly a pickle file is built after training the model in the Jupyter notebook [47] and an API (Application Programming Interface) is created using flask application. This flask actually python code that works as a bridge between the webpage and machine learning model. When a user enters the input values and click on the 'predict' button, the given input parameters are sent to the flask. After that, the flask sends the entered input to the machine learning model for the stroke prediction. After predicting the stroke, the result is transmitted on the web page through the flask so that the users can see the result on the screen. The output is shown in Fig. 9.

3.9. Implementation of mobile application

This section discusses the development of mobile application. This app is built using react native. The API which was created during web application used in also mobile application. By using this API machine learning model is inserted with android app which shows the result like following Fig. 10.

4. Result and discussion

The proposed system has been tested and trained in 35% and 65% of data respectively. The best model is evaluated using different machine learning classifiers such as NB, DT, RF, SVM, LR, VC, K-NN, GB, MLP and NCC. Among these eleven classifiers, the supreme machine learning model is discovered using performance measures including accuracy, recall, precision, and f1-score. This section covers the environmental setup, confusion matrix for individual classifier and accuracy.

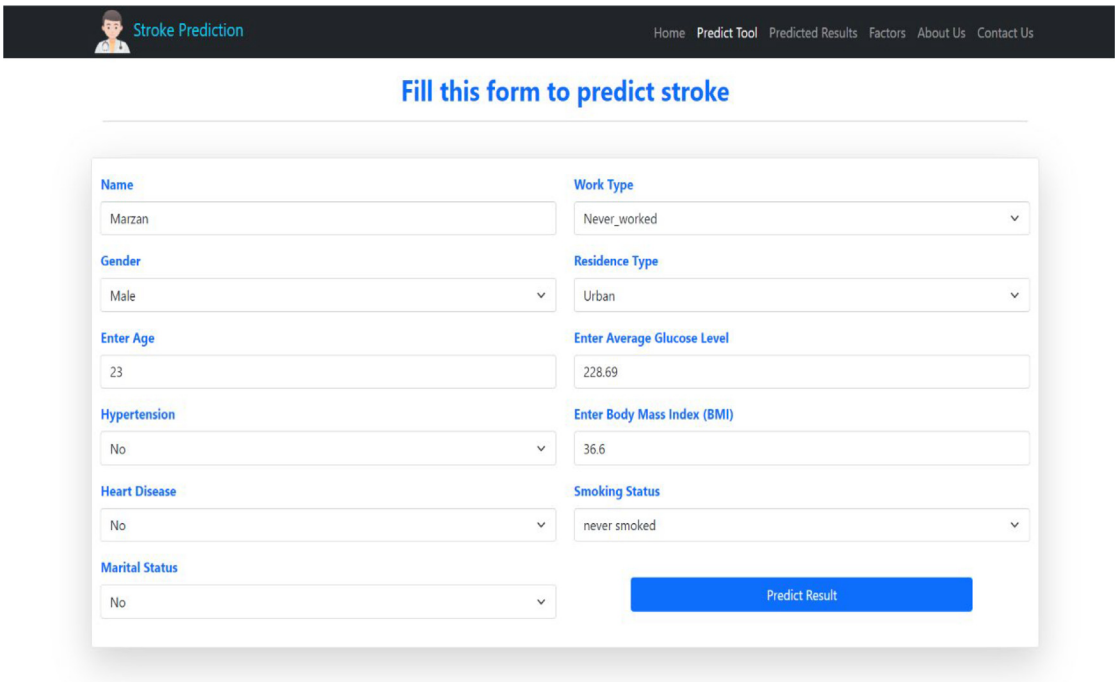

Fig. 8. Input form of a user.

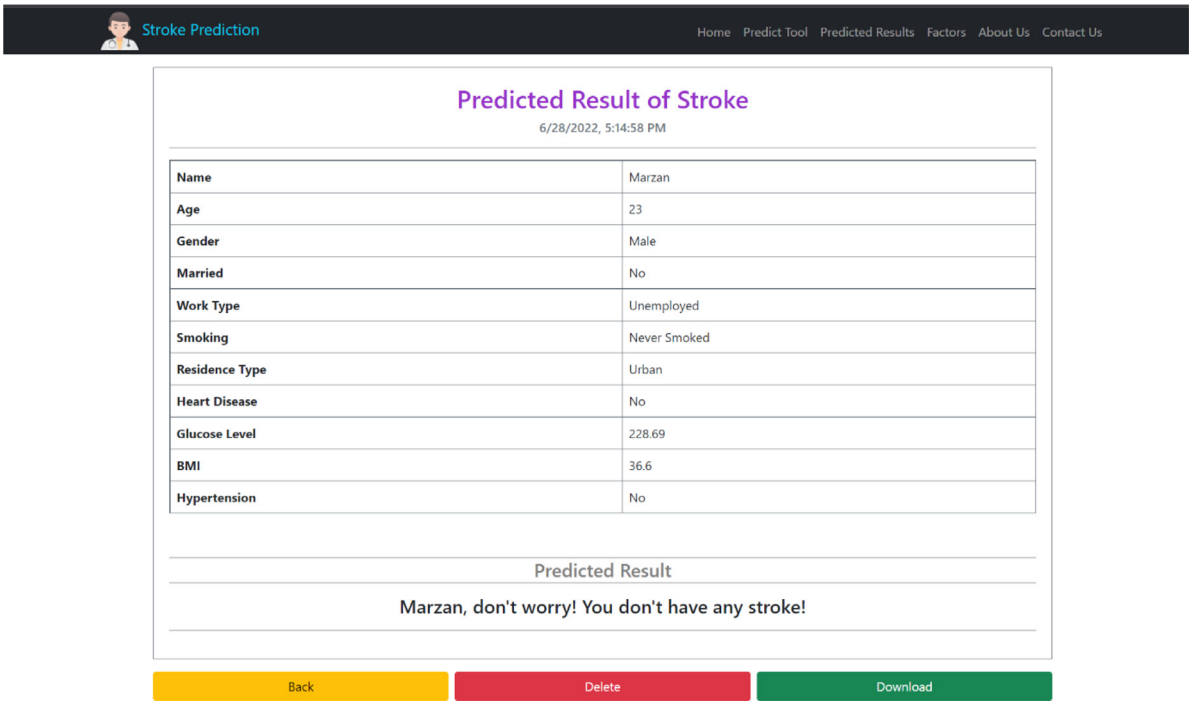

Fig. 9. Predict result through web application.

4.1. Environment setup

For conducting this research some resources were needed. This proposed model is developed using the following resources which is shown in Table 4.

Table 4
Environment setup of the proposed system.

Resource	Details
CPU	Intel® Core™ i-3-1005G1 CPU @ 1.20 GHz
RAM	12 GB
GPU	Intel® UHD Graphics
Experimental tool	Jupyter notebook

4.2. Confusion matrix

A confusion matrix is the $N \times N$ matrix which is used to perform measurement for machine learning classification, where N denotes the

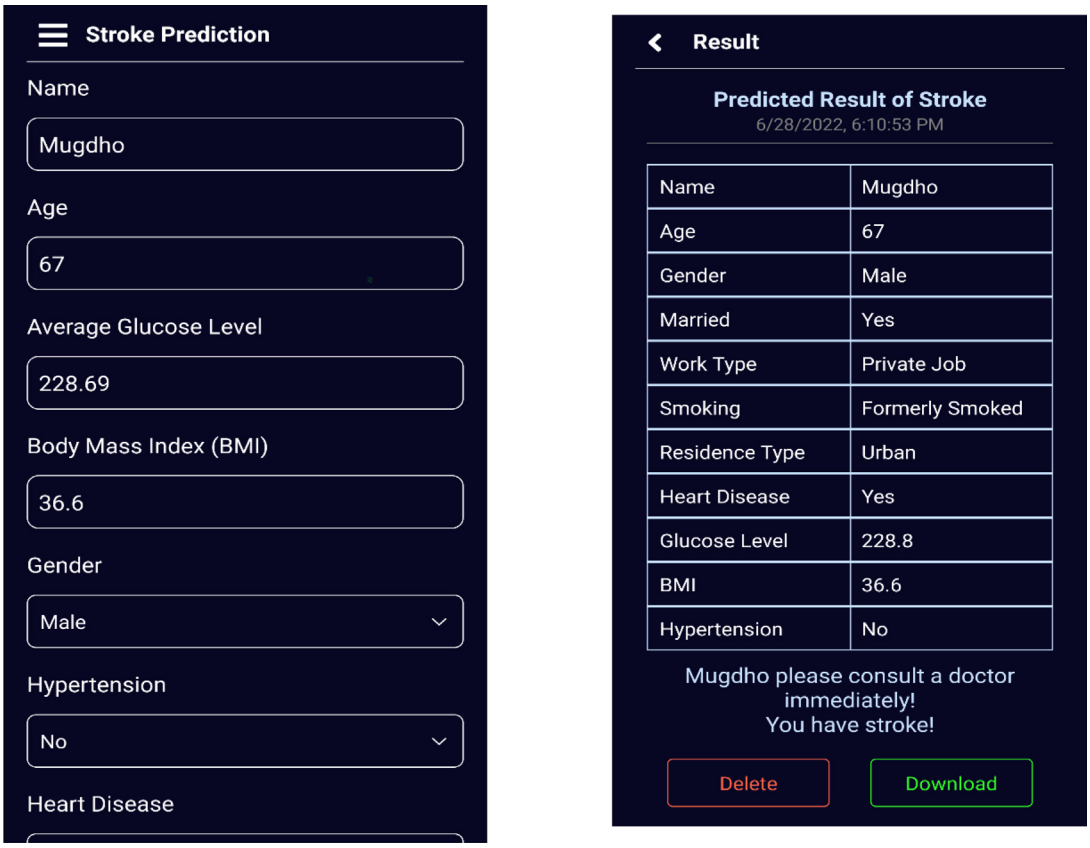

Fig. 10. Predict stroke through mobile app.

number of target classes. Using this method, the number of correct and incorrect prediction is summarized and observed which machine learning classifiers performs well. In this section the used classifiers in this study has been evaluated and compared with classifier. The Fig. 11 shows the performance of each classifier.

The confusion matrix contains four outcomes that measures the performance of each classifier on positive and negative class independently; where two types give correct prediction and two types give incorrect prediction for individual classifier which are including true positive (TP), true negative (TN), false positive (FP), and false-negative (FN) [48]. Next, each classifier is analyzed using the (15)–(19) metrics formula.

$$Accuracy = \frac{(TP + TN)}{(TP + FP + FN + TN)} \tag{15}$$

$$Precision = \frac{TP}{(TP + TN)} \tag{16}$$

$$Recall = \frac{TP}{(TP + FN)} \tag{17}$$

$$\text{F1\_score} = \frac{2 \times \text{ Recall} \times \text{ Precision}}{\text{Recall} + \text{ Precision}} \tag{18}$$

$$\text{Error rate} = \frac{\text{FP} + \text{ FN}}{\text{TP} + \text{ TN} + \text{ FP} + \text{ FN}} \tag{19}$$

Using the above performance metrics, the performance of each classifier is evaluated before optimization and after optimization shown in Tables 5 and 6, respectively. It is obvious from Table 5 that 98.18% is the highest accuracy that belongs to SVM, LR, AB, MLP. After optimization the accuracy of SVM and RF is 99.99% with 0.00001% error and 99.87% with 0.001%, respectively. But the accuracy of MLP, AB, LR becomes poor after optimization. In spite of decreasing accuracy for some classification, a few classifiers show the highest result near to 100%. Therefore, it is said that optimization technique enhances the performance.

Table 5
Evaluation of classification methods before optimization.

Methods	Precision	Recall	F-measure	Accuracy	Error
SVM	49.09%	50.00%	49.54%	98.18%	0.018%
RF	49.09%	49.99%	49.54%	98.17%	0.018%
KNN	49.09%	49.99%	49.53%	98.16%	0.018%
DT	51.58%	51.86%	51.70%	96.25%	0.038%
NB	53.43%	64.60%	54.34%	91.24%	0.088%
LR	49.09%	50.00%	49.54%	98.18%	0.018%
AB	49.09%	50.00%	49.54%	98.18%	0.018%
GB	61.59%	50.34%	50.24%	98.16%	0.018%
MLP	49.09%	50.00%	49.54%	98.18%	0.018%
NCC	51.81%	66.63%	48.55%	79.52%	0.204%
VC (LR+SVM)	49.09%	50.00%	49.54%	98.18%	0.018%

Table 6
Evaluation of classification methods after optimization.

Methods	Precision	Recall	F_Measure	Accuracy	Error
SVM	99.99%	99.99%	99.99%	99.99%	0.00001%
RF	99.85%	99.88%	99.86%	99.87%	0.001%
KNN	98.66%	98.97%	98.81%	98.82%	0.01%
DT	96.63%	97.27%	96.86%	96.90%	0.03%
NB	74.26%	74.39%	74.32%	74.77%	0.25%
LR	77.07%	77.32%	77.17%	77.53%	0.22%
AB	78.11%	78.39%	78.21%	78.55%	0.21%
GB	80.82%	81.29%	80.95%	81.18%	0.19%
MLP	79.58%	80.05%	79.71%	79.94%	0.20%
NCC	67.78%	66.21%	66.36%	68.22%	0.32%
VC (LR+SVM)	92.33%	87.95%	89.00%	89.67%	0.10%

For better understand the graphical representation of classifiers are shown in Figs. 12 and 13, respectively. The used dataset in this study is splinted within k equal size of fold where the value of k has been assumed 10 for this study. This 10-fold cross validation is used after hyperparameter tuning. Cross validation is a resampling technique

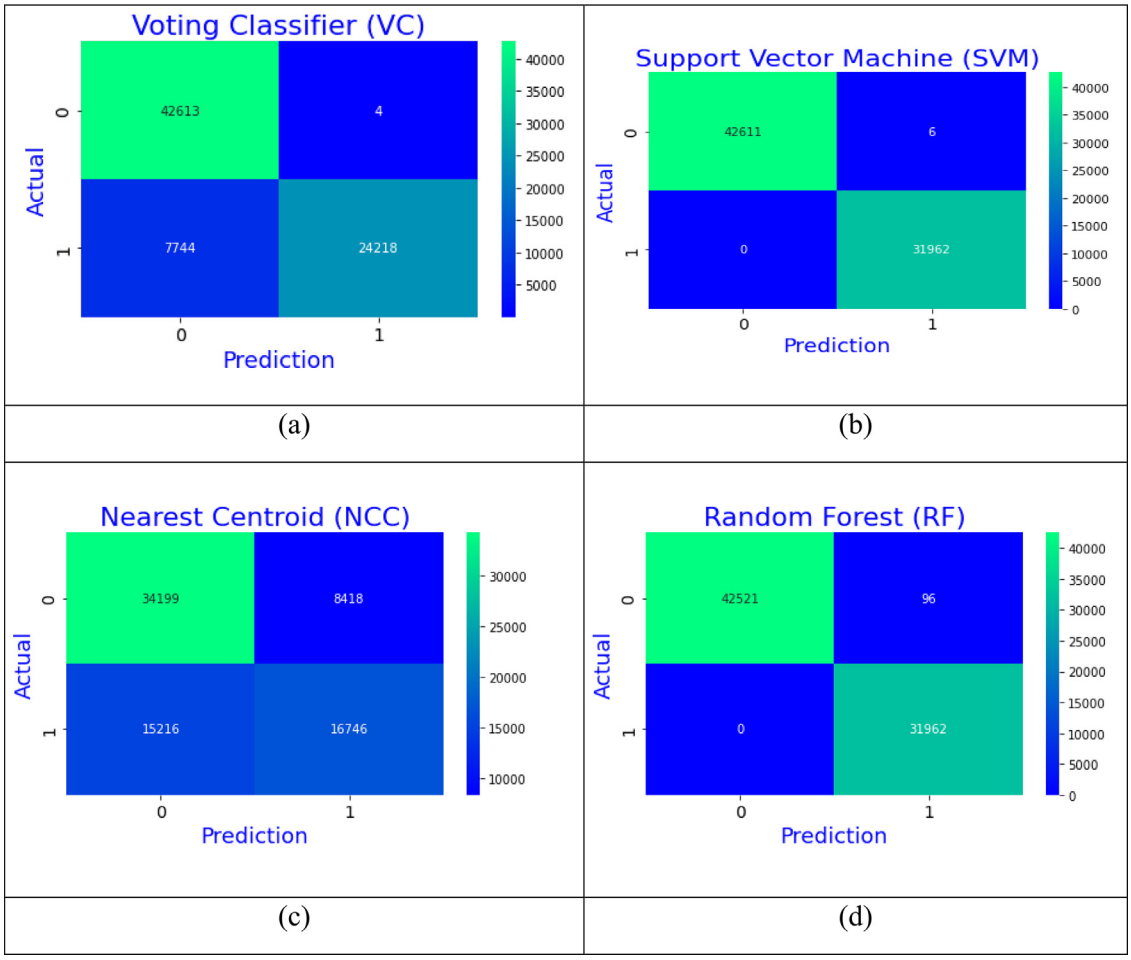

Fig. 11. Confusion matrix for (a) Voting classifier (b) Support Vector Machine (c) Nearest Centroid (d) Random Forest (e) Naïve Bayes (f) Multilayer perceptron (g) K-Nearest Neighbor (h) Logistic Regression (i) Decision Tree (j) Gradient Boosting (k) AdaBoost.

evaluating machine learning algorithm, which checks the performance of each models on unseen data. However, for 10-fold cross validation, 90% data is used for training for the models whereas 10% data is used for testing the models. After using the cross validation, the maximum promising result is measured using the performance metrics. Thus, the best classifier is selected among all these classifiers for this research work.

However, many researchers have conducted experiments on stroke for many years and used various machine learning technique and shown different results.

As stroke is an alarming disease worldwide, it is necessary to detect the result very accurately at the early stage so that people can recover from this awful disease. Therefore, this model has been developed to work with this disease to overcome the limitations. The comparison analysis is shown in Table 7 confirms that our proposed methodology achieves the highest accuracy among their methodology. In future, explainable machine learning approach will be applied to check the explainability of the machine learning algorithms.

Table 7
Comparison of the proposed system with the most related works.

Author	Dataset	Method	Accuracy
Emon et al., 2021 [9]	Medical clinic of Bangladesh, 5110 instances, 11 features	Weighted voting	97%
Choi et al., 2021 [10]	Chungnam national university hospital, total 273 instances, 67 features	RF	92.52%
Govindarajan et al., 2019 [11]	Sugam multispecialty hospital, Tamil Nadu, India, 507 instances, 23 features	ANN	95.3%
Sailasya et al., 2021 [12]	Kaggle, 5110 instances, 12 features	NB	82%
Ahmed et al., 2019 [13]	Kaggle, 5110 instances, 10 features	RF	90%
Wu et al., 2020 [14]	CLHLS	SVM	95%
Badriyah et al., 2020 [15]	Hajj hospital in Surabaya, Indonesia.	RF	95.97%
J. Yu et al. 2020 [16]	Chungnam national university hospital, 287 instances, 29 features	RF	98.95%
Choi et al., 2021 [17]	Chungnam national university Hospital, total 273 instances, 66 features	LSTM	94%
Proposed system	Kaggle, 43 400 instances, 12 features	RF SVM	99.87% 99.99%

5. Conclusion

Stroke is the dangerous threats all over the world. It should be recovered before it worsens. In this critical situation, machine learning model can play a vital role to predict the stroke in the beginning stage. This paper detects and predicts the result of stroke based on various machine learning classification methods. The proposed methodology contains mainly five stages including loading stoke dataset, data

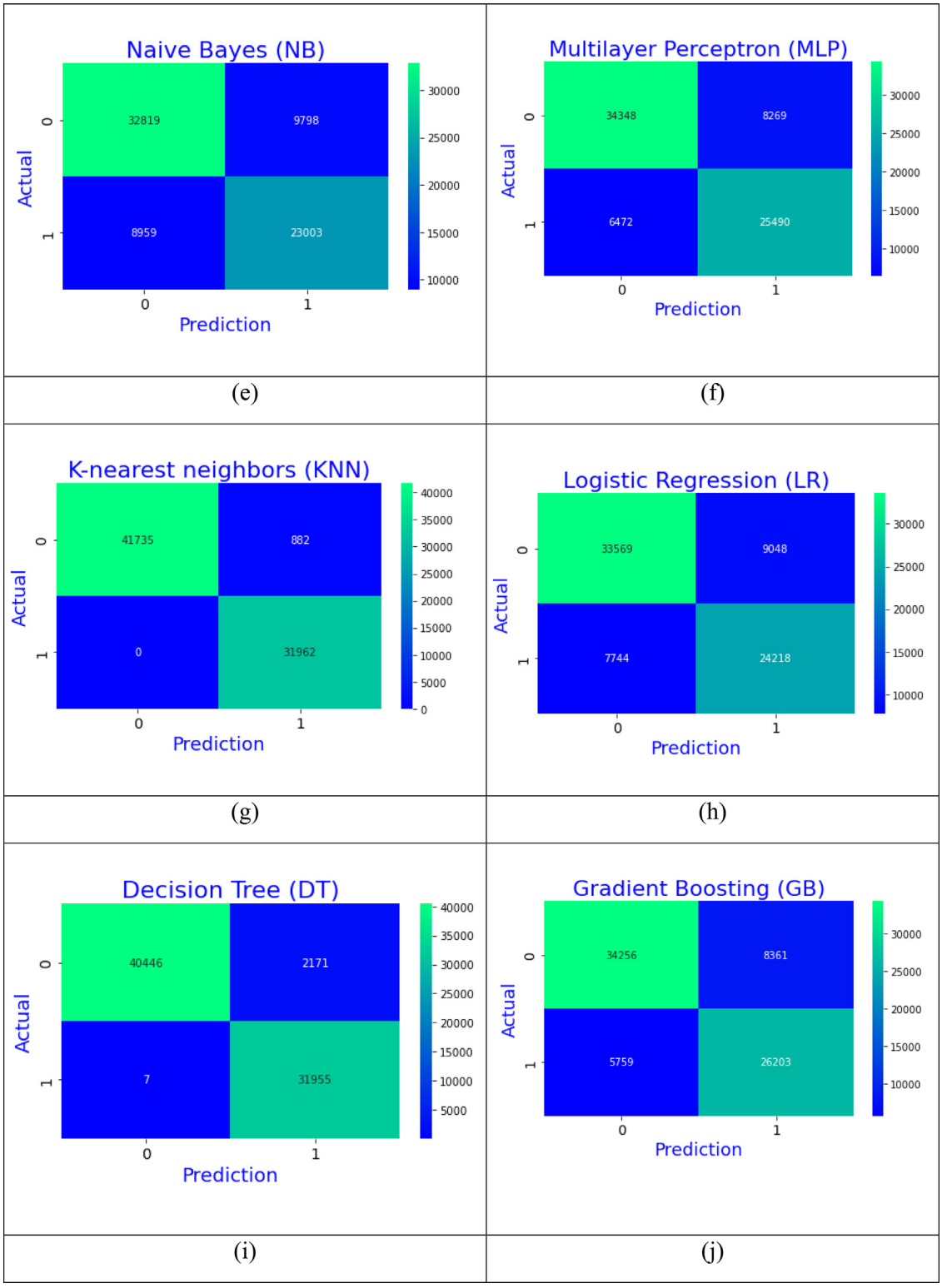

Fig. 11. (*continued*).

pre-processing, hyperparameter tuning, evaluating classifiers and cross-validation. The results show that Support Vector Machine and Random Forest classifier achieves the maximum accuracy result respectively 99.99% and 99.85%. In addition, a user-friendly web page and mobile app are developed for better representation of results. Another research study will be conducted using image processing and deep learning techniques to achieve accurate result in future. We are optimistic that this experiment will improve the treatments of this disease properly.

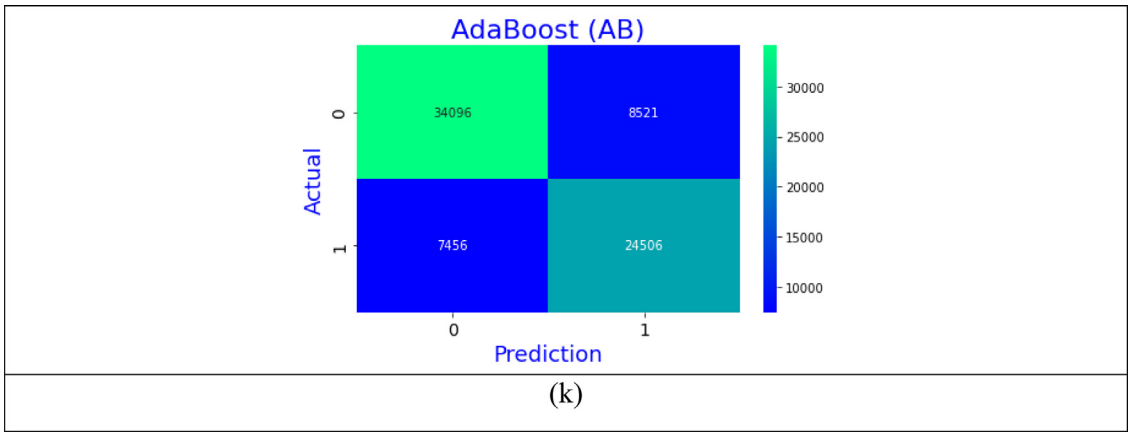

Fig. 11. (*continued*).

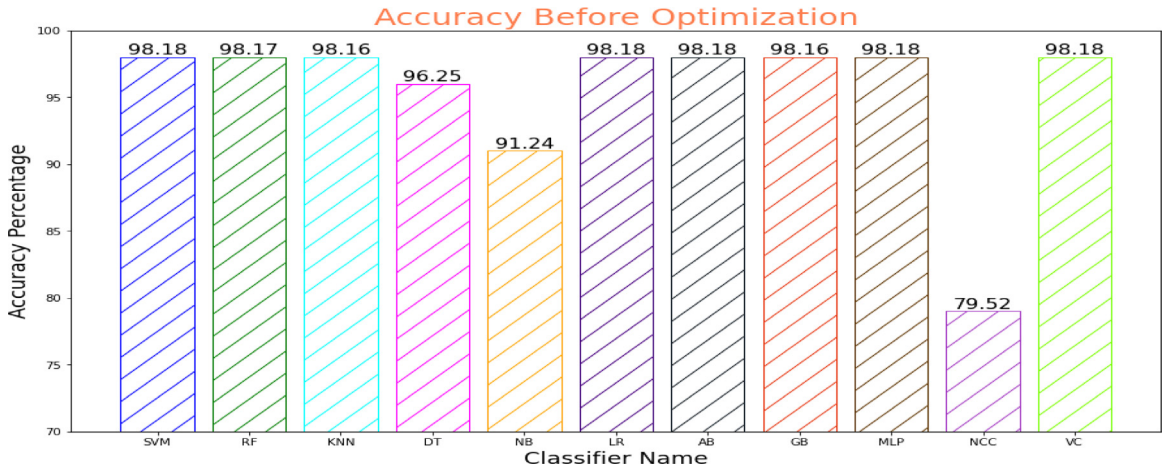

Fig. 12. Accuracy before optimization.

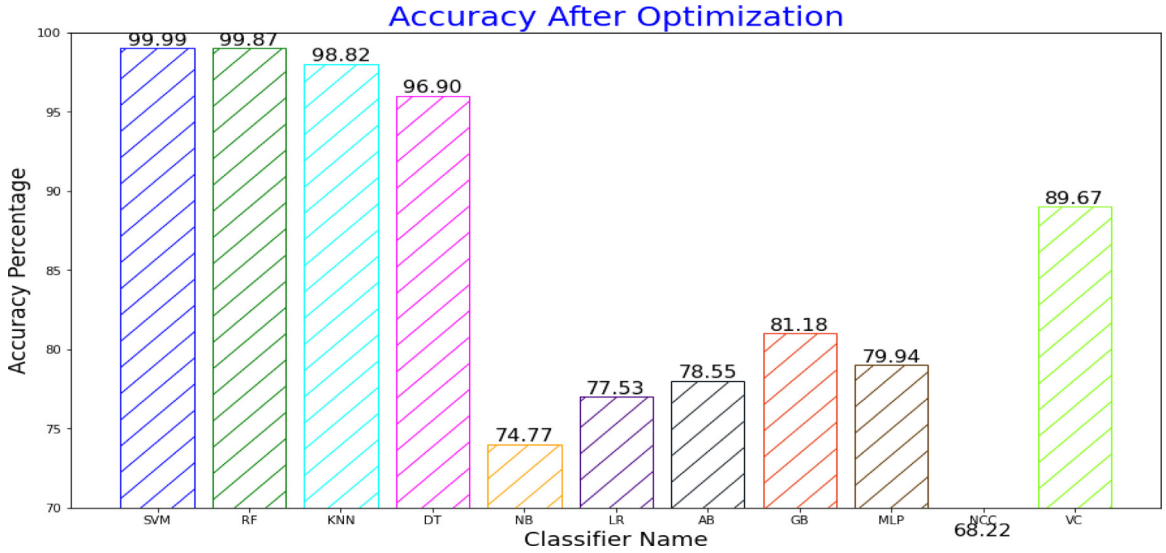

Fig. 13. Accuracy after optimization.

CRediT authorship contribution statement

Nitish Biswas: Conceptualization, Methodology, Software, Formal analysis, Data Curation, Writing – Original Draft, Visualization. **Khandaker Mohammad Mohi Uddin:** Conceptualization, Methodology, Formal analysis, Investigation, Writing – Review & Editing, Supervision, Project administration. **Sarreha Tasmin Rikta:** Methodology, Validation, Resources, Writing – Original Draft. **Samrat Kumar Dey:** Conceptualization, Resources, Writing – Review & Editing, Supervision, Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data supporting this study's findings are available from the corresponding author upon reasonable request.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

[1] J. van Gijn, G.J. Rinkel, Subarachnoid haemorrhage: diagnosis, causes and management, Brain: J. Neurol. 124 (2) (2001) 249–278.
[2] M. Mahmud, et al., A brain-inspired trust management model to assure security in a cloud based iot framework for neuroscience applications, Cogn. Comput. 10 (5) (2018) 864–873.
[3] M.B.T. Noor, N.Z. Zenia, M.S. Kaiser, S. Al Mamun, M. Mahmud, Application of deep learning in detecting neurological disorders from magnetic resonance images: a survey on the detection of alzheimer's disease, parkinson's disease and schizophrenia, Brain Inform. 7 (1) (2020) 1–21.
[4] M. Mahmud, M.S. Kaiser, A. Hussain, Deep learning in mining biological data, 2020, arXiv preprint arXiv:2003.00108.
[5] T.V. Glotzer, A.S. Hellkamp, J. Zimmerman, M.O. Sweeney, R. Yee, et al., Atrial high rate episodes detected by pacemaker diagnostics predict death and stroke: report of the Atrial Diagnostics Ancillary Study of the MOde Selection Trial (MOST), Circulation 107 (12) (2003) 1614–1619.
[6] M. Chun, et al., Stroke risk prediction using machine learning: A prospective cohort study of 0.5 million Chinese adults, J. Am. Med. Informatics Assoc. 28 (8) (2021) 1719–1727, http://dx.doi.org/10.1093/jamia/ocab068.
[7] M.N. Islam, et al., Burden of stroke in Bangladesh, Int. J. Stroke 8 (3) (2013) 211–213, http://dx.doi.org/10.1111/j.1747-4949.2012.00885.x.
[8] T. Omae, Stroke risk factors and stroke prevention, J. Stroke Cerebrovasc. Dis. 2 (1) (1992) 45–46, http://dx.doi.org/10.1016/S1052-3057(10)80035-7.
[9] R. Mostafiz, M.S. Uddin, K.M.M. Uddin, M.M. Rahman, COVID-19 along with other chest infections diagnosis using faster R-CNN and generative adversarial network, ACM Trans. Spatial Algorithms Syst. (2022) http://dx.doi.org/10.1145/3520125, Just Accepted.
[10] R. Hertel, R. Benlamri, A deep learning segmentation-classification pipeline for x-ray-based covid-19 diagnosis, Biomed. Eng. Adv. (2022) 100041.
[11] A. Chattopadhyay, M. Maitra, MRI-based brain tumor image detection using CNN based deep learning method, Neurosci. Inform. (2022) 100060.
[12] S.K. Mamatha, H.K. Krishnappa, N. Shalini, Graph theory based segmentation of magnetic resonance images for brain tumor detection, Pattern Recognit. Image Anal. 32 (1) (2022) 153–161.
[13] G.N. Ahmad, H. Fatima, A.S. Saidi, Efficient medical diagnosis of human heart diseases using machine learning techniques with and without GridSearchCV, IEEE Access (2022).
[14] M.M. Rahman, M.R. Rana, Nur-A-Alam, M.S.I. Khan, K.M.M. Uddin, A web-based heart disease prediction system using machine learning algorithms, Netw. Biol. 12 (2) (2022) 64–81.
[15] S.K. Dey, M.M. Rahman, A. Howlader, U.R. Siddiqi, K.M.M. Uddin, R. Borhan, E.U. Rahman, Prediction of dengue incidents using hospitalized patients, metrological and socio-economic data in Bangladesh: A machine learning approach, PLoS One 17 (7) (2022) e0270933.
[16] D.V.B. Oliveira, J.F. da Silva, T.A. de Sousa Araújo, U.P. Albuquerque, Influence of religiosity and spirituality on the adoption of behaviors of epidemiological relevance in emerging and re-emerging diseases: The case of dengue fever, J. Religion Health 61 (1) (2022) 564–585.
[17] X. Meng, X. Pang, K. Zhang, C. Gong, J. Yang, H. Dong, X. Zhang, Recent advances in near-infrared-II fluorescence imaging for deep-tissue molecular analysis and cancer diagnosis, Small 18 (31) (2022) 2202035.
[18] A. Sharma, K. Dulta, R. Nagraik, K. Dua, S.K. Singh, D.K. Chellappan, D. Kumar, D.S. Shin, Potentialities of aptasensors in cancer diagnosis, Mater. Lett. 308 (2022) 131240.
[19] H. Liao, R. Fang, J.B. Yang, D.L. Xu, A linguistic belief-based evidential reasoning approach and its application in aiding lung cancer diagnosis, Knowl.-Based Syst. (2022) 109559.
[20] M.U. Emon, M.S. Keya, T.I. Meghla, M.M. Rahman, M.S. Al Mamun, M.S. Kaiser, Performance analysis of machine learning approaches in stroke prediction, in: Proc. 4th Int. Conf. Electron. Commun. Aerosp. Technol. ICECA 2020, no. January, 2020, pp. 1464–1469, http://dx.doi.org/10.1109/ICECA49313.2020.9297525.
[21] Y.A. Choi, et al., Machine-learning-based elderly stroke monitoring system using electroencephalography vital signals, Appl. Sci. 11 (4) (2021) 1–18, http://dx.doi.org/10.3390/app11041761.
[22] P. Govindarajan, R.K. Soundarapandian, A.H. Gandomi, R. Patan, P. Jayaraman, R. Manikandan, Classification of stroke disease using machine learning algorithms, Neural Comput. Appl. 32 (3) (2020) 817–828, http://dx.doi.org/10.1007/s00521-019-04041-y.
[23] G. Sailasya, G.L.A. Kumari, Analyzing the performance of stroke prediction using ML classification algorithms, Int. J. Adv. Comput. Sci. Appl. 12 (6) (2021) 539–545, http://dx.doi.org/10.14569/IJACSA.2021.0120662.
[24] H. Ahmed, S.F. Abd-El Ghany, E.M.G. Youn, N.F. Omran, A.A. Ali, Stroke prediction using distributed machine learning based on apache spark, Int. J. Adv. Sci. Technol. 28 (15) (2019) 89–97, http://dx.doi.org/10.13140/RG.2.2.13478.68162.
[25] Y. Wu, Y. Fang, Stroke prediction with machine learning methods among older chinese, Int. J. Environ. Res. Public Health 17 (6) (2020) 1–11, http://dx.doi.org/10.3390/ijerph17061828.
[26] T. Badriyah, N. Sakinah, I. Syarif, D.R. Syarif, Machine learning algorithm for stroke disease classification, in: 2020 International Conference on Electrical, Communication, and Computer Engineering, ICECCE, IEEE, 2020, pp. 1–5.
[27] J. Yu, S. Park, S.H. Kwon, C.M.B. Ho, C.S. Pyo, H. Lee, AI-based stroke disease prediction system using real-time electromyography signals, Appl. Sci. 10 (19) (2020) http://dx.doi.org/10.3390/app10196791.
[28] Y.A. Choi, et al., Deep learning-based stroke disease prediction system using real-time bio signals, Sensors 21 (13) (2021) http://dx.doi.org/10.3390/s21134269.
[29] Stroke Prediction dataset, https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset/metadata?fbclid=IwAR2yNahYaZ9itbIme6TFmSI7_QDBQgAYeeXY_uRLkRgPEl0apySPfUqZkWA.
[30] R. Mohammed, J. Rawashdeh, M. Abdullah, Machine learning with oversampling and undersampling techniques: Overview study and experimental results, in: 2020 11th Int. Conf. Inf. Commun. Syst. ICICS 2020, 2020, pp. 243–248, http://dx.doi.org/10.1109/ICICS49469.2020.239556.
[31] H. Li, J. Li, P.-C. Chang, J. Sun, Parametric prediction on default risk of Chinese listed tourism companies by using random oversampling, isomap, and locally linear embeddings on imbalanced samples, Int. J. Hospital. Manage. 35 (2013) 141–151.
[32] R. Bardenet, M. Brendel, B. Kégl, M. Sebag, Collaborative hyperparameter tuning, in: International Conference on Machine Learning, PMLR, 2013, pp. 199–207.
[33] H. Tan, Machine learning algorithm for classification, J. Phys. Conf. Ser. 1994 (1) (2021) 12–13, http://dx.doi.org/10.1088/1742-6596/1994/1/012016.
[34] S. Ray, A quick review of machine learning algorithms, in: Proc. Int. Conf. Mach. Learn. Big Data, Cloud Parallel Comput. Trends, Prespectives Prospect. Com. 2019, 2019, pp. 35–39, http://dx.doi.org/10.1109/COMITCon.2019.8862451.
[35] S.R. Amendolia, G. Cossu, M.L. Ganadu, B. Golosio, G.L. Masala, G.M. Mura, A comparative study of K-nearest neighbour, support vector machine and multi-layer perceptron for thalassemia screening, Chemom. Intell. Lab. Syst. 69 (1–2) (2003) 13–20, http://dx.doi.org/10.1016/S0169-7439(03)00094-7.
[36] S. Wan, Y. Liang, Y. Zhang, M. Guizani, Deep multi-layer perceptron classifier for behavior analysis to estimate Parkinson's disease severity using smartphones, IEEE Access 6 (2018) 36825–36833, http://dx.doi.org/10.1109/ACCESS.2018.2851382.
[37] M.J. Vowels, Trying to outrun causality with machine learning: Limitations of model explainability techniques for identifying predictive variables, 2022, pp. 1–27, [Online]. Available: https://arxiv.org/abs/2202.09875v4.
[38] L.E. Peterson, K-nearest neighbor, Scholarpedia 4 (2) (2009).
[39] P.H. Swain, H. Hauska, Decision tree classifier: Design and potential, IEEE Trans. Geosci. Electron. GE-15 (3) (1977) 142–147, http://dx.doi.org/10.1109/tge.1977.6498972.
[40] K.P. Murphy, Naive Bayes classifiers generative classifiers, Bernoulli 4701 (October) (2006) 1–8, http://dx.doi.org/10.1007/978-3-540-74958-5_35.
[41] T. Rymarczyk, E. Kozłowski, G. Kłosowski, K. Niderla, Logistic regression for machine learning in process tomography, Sensors (Switzerland) 19 (15) (2019) 1–19, http://dx.doi.org/10.3390/s19153400.
[42] R. Rojas, AdaBoost and the Super Bowl of Classifiers a Tutorial Introduction to Adaptive Boosting, Tech. Rep., Freie University, Berlin, 2009.
[43] A. Natekin, A. Knoll, Gradient boosting machines, a tutorial, Front. Neurorobot. 7 (DEC) (2013) http://dx.doi.org/10.3389/fnbot.2013.00021.
[44] P. Nieminen, Classification and multilayer perceptron neural networks, Training (2010).

[45] I. Levner, Feature selection and nearest centroid classification for protein mass spectrometry, BMC Bioinformatics 6 (1) (2005) 1–14.
[46] D. Ruta, B. Gabrys, Classifier selection for majority voting, Inf. Fusion 6 (1) (2005) 63–81.
[47] Jupyter notebook, 2022, https://jupyter.org/, [Last accessed: 20.06.22].
[48] Gustavo E.A.P.A. Batista, R.C. Prati, M.C. Monard, A study of the behavior of several methods for balancing machine learning training data, ACM SIGKDD Explor. Newsl. 6 (1) (2004) 20–29, http://dx.doi.org/10.1145/1007730.1007735.