Asian Research Journal of Current Science

Volume 6, Issue 1, Page 122-131, 2024; Article no.ARJOCS.1353

Stroke Prediction with Random Forest Machine Learning Model

Okpe Anthony Okwori [a*], Moses Adah Agana [b], Ofem Ajah Ofem [b] and Obono I. Ofem [b]

[a] Department of Computer Science, Federal University Wukari, Nigeria.
[b] Department of Computer Science, University of Calabar, Nigeria.

Authors' contributions

This work was carried out in collaboration among all authors. All authors read and approved the final manuscript.

Article Information

Open Peer Review History:
This journal follows the Advanced Open Peer Review policy. Identity of the Reviewers, Editor(s) and additional Reviewers, peer review comments, different versions of the manuscript, comments of the editors, etc are available here: https://prh.globalpresshub.com/review-history/1353

Original Research Article

Received: 24/06/2023
Accepted: 30/08/2023
Published: 17/06/2024

ABSTRACT

Stroke is a medical condition associated with either blockage or rupture of blood vessels which prevents the free flow of blood to the brain cells causing the brain cells to die. The dead brain cells cause malfunctions of the part of the body that it controls leading to stroke that can further result in permanent disability. Both ischemic and hemorrhagic stroke though occurring suddenly, are associated with some stroke risk factors such as age, hypertension, and body mass index among others. These two types of stroke are very dangerous to human health and are a threat to life, ischemic stroke occurs more frequently than haemorrhagic stroke. In an attempt to reduce stroke occurrence, medical doctors use stroke biomarkers to predict stroke occurrence and confirm suspected stroke cases using several diagnostic tests. This technique of stroke prediction and diagnosis is highly time consuming, especially at an early stage when decision making is most important and no individual candidate or multimarker panel has proven to have adequate performance for use in an acute clinical setting hence a need for more efficient stroke prediction technique such as machine learning models. Machine learning is one of the modern areas in

**Corresponding author: Email: okwori@fuwukari.edu.ng;*

Cite as: *Okwori, Okpe Anthony, Moses Adah Agana, Ofem Ajah Ofem, and Obono I. Ofem. 2024. "Stroke Prediction With Random Forest Machine Learning Model". Asian Research Journal of Current Science 6 (1):122-31. https://jofscience.com/index.php/ARJOCS/article/view/111.*

artificial intelligence that deals with the ability of a machine to imitate intelligent human behavior. This field is widely applied in healthcare services due to the ever-evolving patient dataset that can be used to train machine learning algorithms for pattern detection that enable medical professionals to recognize new diseases, predict treatment outcomes as well as make medical decisions about the risk of developing disease or medical condition like stroke**.** this paper aims to predict the stroke vulnerability status of patients using a random forest (RF) machine learning model. The model was built on Python programming language using healthcare_dataset_stroke data obtained from the Kaggle machine learning dataset repository. The dataset was properly cleaned and the clean dataset was used to train the random forest machine learning model for efficient prediction of stroke. the results obtained from the random forest model were evaluated using a confusion matrix and it was found that random forest is a very good choice of algorithm for predicting stroke vulnerability as evidenced in its prediction accuracy of 93%.

Keywords: Stroke; biomarkers; prediction; machine; learning; random; forest.

1. INTRODUCTION

A stroke could be explained as a medical condition that occurs when there is an absence or insufficient flow of blood to the brain cells causing the cells to die. There are basically two types of strokes caused by two different factors namely: Ischemic Stroke caused by blockage of arteries and hemorrhagic stroke caused by leaking of blood vessels [1]. In either case, the brain cells are starved of oxygen and nutrients necessary for its growth and proper functionalities consequently causing the brain cells to die. Stroke is a chronic disease that reduces the quality of life as it usually results to some complications such as difficulties in moving the body muscles, talking, swallowing food and other substances, loss of memory, depression, loss of self-care ability and sometimes results in permanent physical disability [2]. To reduce the menace associated with stroke occurrence, there is an urgent need for stroke prediction, diagnosis and control. Stroke prediction and diagnosis prior to the age of artificial intelligence and in particular machine learning techniques were done by medical doctors using some stroke prediction biomarkers (biomarkers refer to indicators measured by chemical or biological tests using blood or urine that predict physiologic or disease states, or increased disease risk.) such as high-sensitivity C reactive protein (hsCRP), interleukin-6 (IL-6), N-terminal pro-brain natriuretic peptide (NT-proBNP) [3] among others and conclude that such patient is likely to have a stroke if those biomarkers are present in abnormal quantity. To confirm stroke occurrence and the particular type of stroke that a given patient is suffering from, doctors use several diagnostic tests such as [4]:

i) **Physical Examination:** A doctor will ask about the person's symptoms and medical history. They will check muscle strength, reflexes, sensation, vision, and coordination. They may also check blood pressure, listen to the carotid arteries in the neck, and examine the blood vessels at the back of the eyes.
ii) **Blood Tests:** A doctor may perform blood tests to determine if there is a high risk of bleeding or blood clots, measuring levels of particular substances in the blood, including clotting factors, and checking whether or not an infection is present.
iii) **Computerized Tomography (CT) scan:** A series of X-rays can show hemorrhages, strokes, tumors, and other conditions within the brain.
iv) **Magnetic Resonance Imaging (MRI) scan:** These use radio waves and magnets to create an image of the brain, which a doctor can use to detect damaged brain tissue.
v) **Carotid Ultrasound:** A doctor may carry out an ultrasound scan to check blood flow in the carotid arteries and to see if there is any narrowing or plaque present.
vi) **Cerebral Angiogram:** A doctor may inject a dye into the brain's blood vessels to make them visible under X-ray or MRI. This provides a detailed view of the blood vessels in the brain and neck.
vii) **Echocardiogram:** This creates a detailed image of the heart, which doctors can use to check for any sources of clots that could have traveled to the brain.

Although these methods of stroke prediction and diagnosis exist, they are highly time consuming, especially at an early stage when decision making is most important and no individual candidate or multimarker panel has proven to have adequate performance for use in an acute clinical setting where decisions about an

individual patient are being made [5] hence there is a need for more efficient techniques such as machine learning approach. Machine Learning (ML) is a computer application in the field of artificial intelligence that deals with the study of computer algorithms that improve automatically through experience using data. Machine learning algorithms build a model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to do so [6]. The application of machine learning models is very versatile as it is used in almost all life sectors such as in system security which is needed by every user [7] and in healthcare for improving services to humanity. The healthcare sector has specifically benefited enormously from machine learning applications as it is used for disease prediction and biomedical studies to increase care quality through efficient medical decisions. In recent times, there has been an increase in the application of machine learning in healthcare resulting in improved diagnostic accuracy and enhanced personalized healthcare services. Research from relevant literature has provided enormous evidence in favour of some machine learning classification algorithms such as Random Forest [8] in terms of binary classification efficiency as it has been successfully used in a wide range of biomedical applications, such as the automatic detection of the pulse during electrocardiogram-based cardiopulmonary resuscitation, breast cancer diagnosis using mammography images, cardiovascular disease detection, prediction of obesity, prediction of diabetes, stroke outcome prediction among others.

1.1 Random Forest Algorithm

A random forest is simply an assemblage of tree-based models that are trained using random subsets of the training dataset which is obtained by randomly choosing x number of features (columns) and y number of instances (rows) from the available dataset consisting of n features and m instances. It is an ensembled learning model of decision trees where every tree is trained on a random subset of the entire training dataset thereby making the ensemble less likely to overfit [9] the training dataset, very simple and highly stable. It can be viewed as a bag containing a finite number (n) of decision trees (DT) having different sets of hyper-parameters and are trained on different subsets of the training dataset which brings about the variation in the predicted result of the trees as shown in Fig. 1.

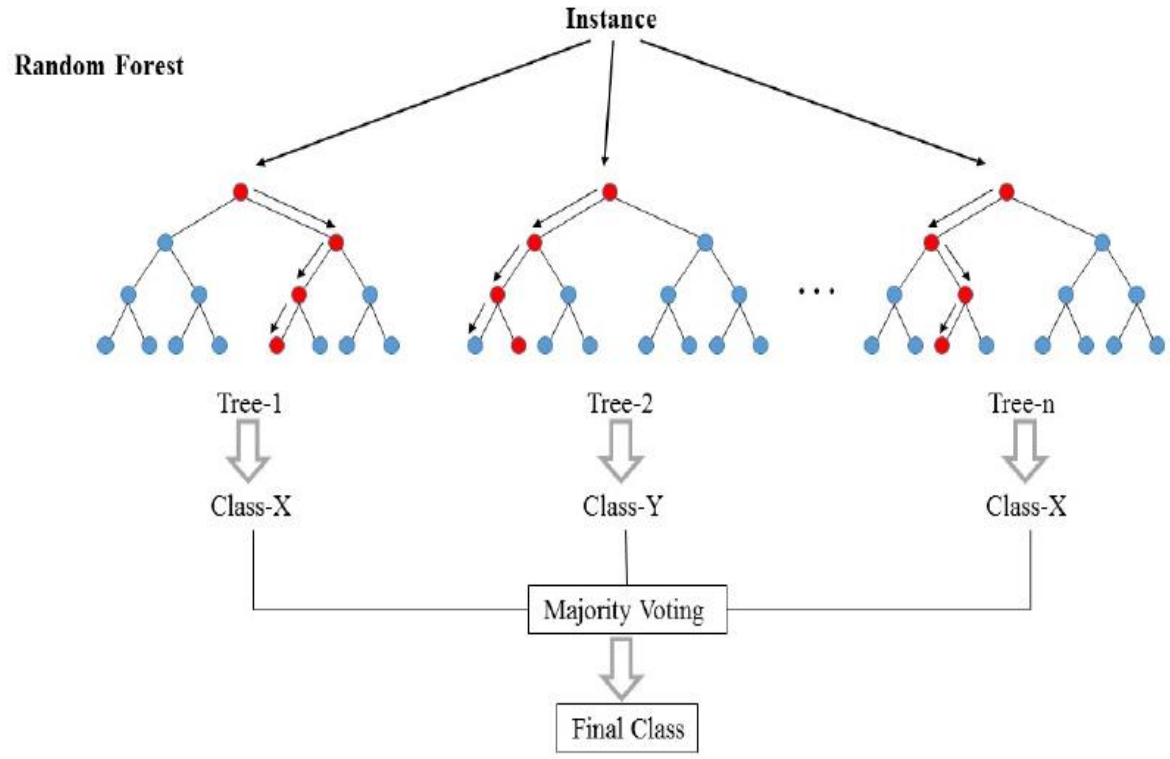

Fig. 1. Prediction process with random forest

As seen in Fig. 1, the predictive capability of random forest is derived from the collective perception of the individual trees that constitute the entire forest under consideration. In the domain of machine learning, random forest can be used to provide solutions to both classification and regression problems [10]. In this research, the random forest algorithm has been used to predict stroke vulnerabilities among individuals using the hearthcare-dataset-stroke-data with a total of 5110 instances and contains 11 independent features with two target classes. As a classification problem, the standard random forest classification model was adopted and used to identify and assign every instance of the dataset to one of the target classes depending on the feature values. The random forest algorithm is shown in Algorithm 1 below.

Algorithm 1: Random Forest Algorithm

Step 1: Take n number of random records from the data set.
Step 2: Construct decision trees for each sample.
Step 3: predict stroke vulnerability with each decision tree.
Step 4: Make final prediction by considering the Majority Vote

2. RELATED WORKS

This section reviews the efficacy of the random forest machine learning model in binary classification to show that random forest is one of the best algorithms for binary disease classification and diagnosis and hence suitable for predicting the stroke vulnerability status of an individual. The use of machine learning models for disease classification to support clinical decisions has been shown in various studies and random forest continue to demonstrate quality performance as evident in the work of [11] where a random forest algorithm was used to predict the survival rate of people with breast cancer based on the stage of cancer detection, adequacy of prognosis, the kind of treatment the patient receives as well as the patient's characteristics. The researchers generated their dataset from the medical records of 60 patients with breast cancer and reduced its dimension using the relief feature selection technique to improve the performance of the basic random forest algorithm. The result shows that the random forest algorithm performs very well in classifying breast cancer survivability among cancer patients using the dataset obtained from the patient's medical records however it was recommended that a larger number of patient records be collected over a long period of time to generate a fairly large dataset for more accurate predictions. In like manner, [12] developed a random forest-based machine learning model for the prediction of breast cancer using the cancer microarray dataset. The researchers trained the model using various percentages of the dataset as a training set and found that random forest is the best model for cancer classification when the training sample is high. Random forest machine learning models are generally good in binary classification and specifically very accurate in disease prediction for adequate medical decision making as its viability is not only in breast cancer but adequately predicts lung cancer according to [13] who developed a machine learning model for the prediction of lung cancer Survival rate using random Forest based decision tree algorithms. The model developed in the study shows outstanding performance in the prediction of lung cancer survivability rate as evident in its area under the receiver operating characteristic (ROC) curve and high accuracy of about 85%. The efficiency of the prediction capability of random forest was evident not only in cancer classification but other diseases such as diabetes. In a bid to reduce the damage caused by untimely detection of diabetes [14] developed a diabetes prediction model using a random forest machine learning algorithm on a diabetes dataset obtained from Sawanpracharak Regional Hospital within the period of 2009-2013. The researchers improved the performance of the basic random forest algorithm through adequate feature selection using the Gain Ratio Feature Selection technique. The study shows that a random forest with appropriate feature selection produces a very good result in diabetes classification as the feature selection improves the model by eliminating some features that are less relevant to the prediction of the target class thereby reducing the dataset dimension to improve the overall prediction time and efficiency. The researchers recommended the use of random forest with appropriate feature selection techniques in diabetes classification as well as other related binary disease classification problems such as obesity, stroke, and cancer among others. [15] used random forest (RF) algorithm to predict the prognoses of COVID-19 patients and identify the optimal diagnostic predictors for patients' clinical prognoses. The RF model was implemented using the dataset obtained from 126 COVID-19 patients from Wuhan Fourth Hospital. After resampling of the

dataset using synthetic minority oversampling techniques (SMOTE) and subsequent application of other dataset preprocessing techniques, the random forest achieved a very high prediction accuracy. The proficiency of random forest in predicting stroke-related problems was demonstrated by [4] where a random forest machine learning model was used in predicting stroke outcomes in terms of mortality and morbidity within 3 months after admission. The researchers obtained the dataset from ischemic stroke and non-traumatic intracellular hemorrhagic stroke patients who were admitted to the Stroke Unit of European Tertiary Hospital and were prospectively registered to develop and evaluate the random forest model. It was shown that random forest performs very well in classifying stroke outcomes and hence can be used for efficient clinical decision making regarding long-term mortality or morbidity outcomes of stroke patients and other related stroke conditions.

3. METHODOLOGY

The general methodology used is the quantitative approach to scientific research which focuses on both analytical and empirical methods. Ensemble learning using Random Forest approaches was used to create and train our model. Confusion matrix, using the True and False Positives and True and False Negatives was used to evaluate the performance of the model. This system was implemented using Python programming language as Python is a general-purpose dynamic programming language that provides high-level readability, Simplicity, consistency, flexibility, platform independence, and fewer Codes as it provides access to great libraries and frameworks. Some of the Python programming language libraries and frameworks used in this research for efficient implementation include pandas for reading the stroke.csv file from Microsoft Excel into the Python programming environment for further processes, numpy for manipulating the dataset as multidimensional data structure, matplotlib.pyplot for graphical representation of outputs, LabelEncoder for transforming the categorical variables in the stroke risk factors to numeric once, StandardScaler for scaling the dataset, synthetic minority oversampling technique (SMOTE) for balancing the dataset, RandomForestClassifier for developing the random forest model, classification_report for producing the evaluation result of the classification models using the sklearn evaluation metrics.

3.1 Design of the Random Forest-based Stroke Prediction Models

To design the random forest based stroke prediction model, general research information was obtained from relevant literatures such as journal articles, books, papers among others about stroke prediction using various machine learning models including their strengths and limitations which enabled us to make a good choice of algorithms for optimal prediction performance. The model was implemented using the healthcare-dataset-stroke-data obtained from the Kaggle machine learning dataset repository as structured data. The dataset was downloaded and renamed as stroke and converted to comma-separated value (CSV) file “stroke.csv” using Microsoft Excel for easy data preprocessing using the Pandas machine learning library. It originally consisted of 12 columns and 5110 rows with 95.13% NoStroke samples and 4.8% Stroke samples. The dataset has 11 independent variables features: id, gender, age, hypertension, heart_disease, ever_married, work_type, residence_type, avg_glucose_level, bmi, and smoking_status with one dependent variable class label “stroke”.

The dataset is described in Table 1 as shown.

The top view of healthcare-dataset-stroke-data used in the system is shown in Table 2.

To optimize the performance of the original random forest model, the healthcare dataset stroke data was adequately preprocessed to remove data inconsistencies. Some of the preprocessing techniques used in this paper to improve the prediction performance include: the selection of relevant dataset features using the correlation feature selection technique, feature encoding to convert categorical data into numerical data, filling of missing values in the dataset, balancing the target class which makes the NoStroke sample 50% and Stroke sample 50%, removing of outliers, scaling of the dataset features as well as turning of the various hyperparameters. The cleaned healthcare-dataset-stroke-data was used for training, testing, and validation of the random forest based stroke prediction models for optimal performance.

Table 1. Dataset description

Feature No.	Feature Name	Feature Description
1	Id	Unique identification number for each data point in the dataset
2	Gender	Male or Female
3	Age	Number of years of a patient
4	Hypertension	Presence or absence of hypertension
5	Heart_disease	Presence or absence of heart disease
6	ever_married	Married or not married
7	work_type	Children, Private, Never worked, Govt. job or Self employed
8	Residence_type	Urban or rural residence
9	avg_glucose_level	The average quantity of glucose in the patient body
10	Bmi	The body mass index
11	smoking_status	Never smoked, formally smoked, or smokes
12	Stroke	Presence or absence of stroke

Table 2. Top view of healthcare-dataset-stroke-data

id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
9046	Male	67	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
51676	Female	61	0	0	Yes	Self-employed	Rural	202.21	N/A	never smoked	1
31112	Male	80	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
60182	Female	49	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
1665	Female	79	1	0	Yes	Self-employed	Rural	174.12	24	never smoked	1
56669	Male	81	0	0	Yes	Private	Urban	186.21	29	formerly smoked	1
53882	Male	74	1	1	Yes	Private	Rural	70.09	27.4	never smoked	1
10434	Female	69	0	0	No	Private	Urban	94.39	22.8	never smoked	1
27419	Female	59	0	0	Yes	Private	Rural	76.15	N/A	Unknown	1
60491	Female	78	0	0	Yes	Private	Urban	58.57	24.2	Unknown	1
12109	Female	81	1	0	Yes	Private	Rural	80.43	29.7	never smoked	1
12095	Female	61	0	1	Yes	Govt_job	Rural	120.46	36.8	smokes	1
12175	Female	54	0	0	Yes	Private	Urban	104.51	27.3	smokes	1
8213	Male	78	0	1	Yes	Private	Urban	219.84	N/A	Unknown	1

3.2 Split Dataset into Training, Validation, and Testing Sets

After the entire data preprocessing exercise, the preprocessed stroke.csv had 9722 rows with 11 features that were split into a training dataset, validation dataset, and testing dataset respectively using sklearn python library. Out of the 9722 data records, 80% (7776) data items were used for training, 10% (973) data items were used for validation and 10% (973) data items were used for testing of the models respectively.

3.3 Stroke Prediction with Random Forest Classifier

The term Random Forest classifier is used to describe a supervised machine learning algorithm that performs classification by generating a very large quantity of decision trees from a random subset of the dataset in a homogeneous ensemble manner. It is the combination of decision trees that classifies a new object based on an attribute by enabling each decision tree to generate its own classification (this process is called voting). After the classification, the forest observes all the votes and chooses the classification with the highest votes among all the trees. Unlike predictions of a decision tree that is very sensitive to noise in the training dataset, the voting technique uses the average prediction of many uncorrelated decision trees hence not sensitive to noise from the training set. The uncorrelated decision trees in the random forest algorithm are achieved by training many trees on different training datasets which is why this random forest algorithm was trained using the bootstrap aggregation or simply bagging of decision tree base learners. The bagging approach uses a training dataset X, Y where X is the set of independent variables in the dataset and is given by x1, x2, …, xn while Y is the set of responses and is given by y1, y2, …, yn to repeatedly draws a random subset of the training dataset with replacement and fit decision trees to the selected sample. To classify unseen sample data, the process adopts the majority vote of the constituent decision tree-based learners. The bagging process decreases the random forest model variance with no increase in the bias term thereby increasing the overall model performance. To predict stroke with random forest, we consider an instance $X \in R^N$, a vector consisting of N features given by: X = [x1, x2, x3, …, xN], there exists a tuple (xi, yi) collection of labeled observations which contains the instance vectors xi and the actual target variable yi that constitute the training dataset given by: L = {(x1, y1), (x2, y2), (x3, y3), … (xL, yL)}. The predicted result of a random forest is the dominant class as predicted by individual trees in the forest. Therefore, if the random forest is established with T number of trees, then the number of votes given to a target class m is given by:

$$v_m = \sum_{t=1}^{T} I(\hat{y}_t == m) \quad (1)$$

Where;

$\hat{y}_t$ = predicted result of the t^{th} tree on a given instance of the dataset

$I(\hat{y}_t == m)$ = Indicator function that assumes the value 1 when the test condition evaluates to true and 0 otherwise

From equation (1), the class with the highest vote can be determined, hence the predicted result of random forest is given by the formula:

$$\hat{y} = \underset{m \in \{1,\ldots,M\}}{\arg\max}\ v_m \quad (2)$$

To generate the random forest model, a Random Forest Classifier was imported from the Python sklearn library and the various hyperparameters were tuned. To train the model, the training dataset was fit into the Random Forest Classifier model and evaluated using both stratified 10-fold cross-validation method and confusion matrix.

4. RESULTS AND DISCUSSION

The AUC and confusion matrix together with its associated matrixes generated by the random forest model after successful testing are shown in Fig. 2 and Table 3 and Table 4 respectively.

From Table 4 above it can be seen that random forest has a prediction accuracy of 93%, precision of 92%, recall of 94%, F1 score of 93%, sensitivity of 94%, specificity of 92%, and AUC of 98%. This shows that the algorithm performs very well in stroke disease prediction using the healthcare-dataset-stroke-data. Its eminent performance is evident in the value of its area under receiver operating characteristic AUC-ROC as shown in Fig. 2.

4.1 Comparative Analysis of this Model and Existing Work

After the successful evaluation of this random forest based stroke prediction model using the confusion matrix and its related matrices, the results generated from this empirical research were compared with the results from other related work reviewed in this paper as shown in Table 5.

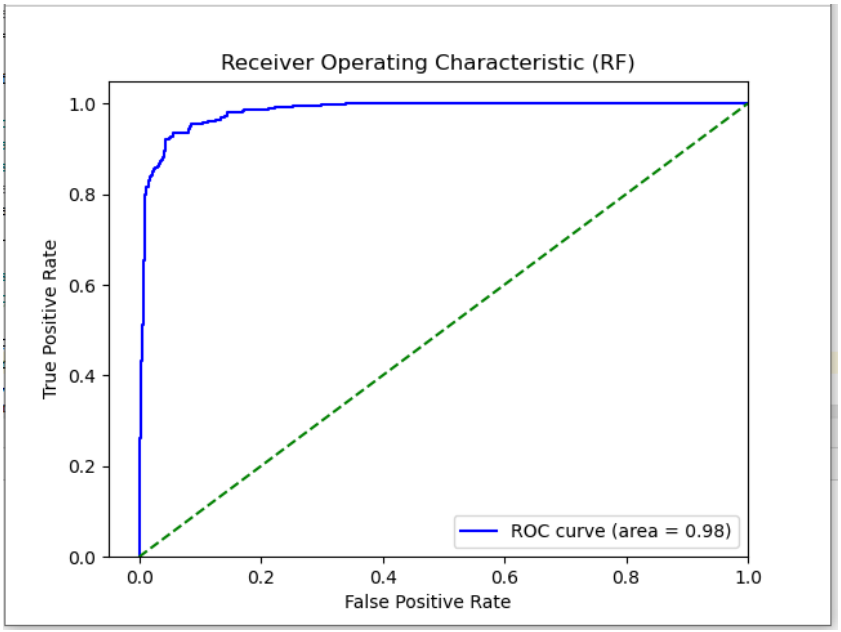

Fig. 2. Area under receiver operating characteristic (ROC) cure (AUC) for Random Forest

Table 3. Confusion matrix for random forest

N = 973	**Actual Values**		
Predicted Values		**Positive (Yes)**	**Negative (No)**
	Positive (Yes)	TP = **448**	FP = **39**
	Negative (No)	FN = **31**	TN = **455**

Table 4. Evaluation results of random forest

Matrix	**Accuracy**	**Precision**	**Recall**	**F1 score**	**Sensitivity**	**Specificity**	**AUC**
Value	0.93	0.92	0.94	0.93	0.94	0.92	0.98

Table 5. Comparative analysis of the proposed work with other related works

S/NO	**Author(S)**	**Title**	**Result**
1	Octaviani and Rustama [12]	Random Forest for Breast Cancer Prediction	Accuracy above 90%
2	Queiroz et al. [10]	Prediction of survival in breast cancer patients using Random Forest classifier and ReliefF feature selection method	Accuracy of 93.33%
3	Sittidech and Nai-arun [13]	Random forest analysis on diabetes complication data	Accuracy of 94.743%
4	Hashi, [12]	Lung Cancer Survival Prediction Using Random Forest-Based Decision Tree Algorithms	Accuracy of 85%
5	Fernandez-Lozano et al. [7]	Random forest-based prediction of stroke outcome	AUC of 90.0%
6	Proposed Work (2023)	Stroke prediction with random forest machine learning model	Accuracy of 93% AUC of 98%

From Table 5 above, it can be seen that the random forest machine learning model is a good algorithm for binary event classification as demonstrated in its high prediction accuracy and AUC values. The model has a prediction accuracy of 85% and above as well as an AUC of 90% and above in all the classification instances in this paper. In particular, the random forest produces 93% accuracy and 98% AUC in the proposed work and this result is within the acceptable range when compared to all the related works in this paper.

5. CONCLUSIONS

The application of machine learning in the healthcare sector for disease prediction promotes proactive medical intervention that may result in total prevention or effective control of chronic diseases such as stroke. This paper investigated the efficiency of the random forest machine learning model in the prediction of stroke vulnerability. It is empirical research that uses the healthcare_dataset_stroke_data obtained from the Kaggle machine learning dataset repository to build the random forest stroke prediction model in the Python programming language. The prediction results obtained were evaluated using a confusion matrix and its associated metrics such as accuracy, precision, and AUC, among others and it was found that the random forest algorithm performs very well in classifying stroke vulnerability status of patients using this dataset. The high performance of the model is evident in its accuracy score of 93% and area under the curve (AUC) of 98% respectively.

DISCLAIMER (ARTIFICIAL INTELLIGENCE)

Author(s) hereby declare that NO generative AI technologies such as Large Language Models (ChatGPT, COPILOT, etc) and text-to-image generators have been used during writing or editing of manuscripts.

COMPETING INTERESTS

The authors have declared that no competing interests exist.

REFERENCES

1. Parmar P. Stroke: classification and diagnosis, Journal of the Royal Pharmaceutical Society; 2018.
   Available:https://pharmaceutical-journal.com/article/ld/stroke-classification-and-diagnosis
   Accessed on 26th August, 2023.
2. Baye M, Hintze A, Gordon-Murer C, Tatiana Mariscal T, Belay GJ, Gebremariam AA, Hughes CML. Stroke Characteristics and Outcomes of Adult Patients in Northwest Ethiopia, Frontiers in Neurology; 2020.
   Available:https://www.frontiersin.org/articles/10.3389/fneur.2020.00428/full
   Accessed on 26th August, 2023.
3. Sara R. Biomarkers for Prediction of Stroke; 2020.
   Available:https://www.news-medical.net/health/Biomarkers-for-Prediction-of-Stroke.aspx
   Accessed on 28th December, 2022.
4. James M. Everything you need to know about stroke; 2020.
   Available:https://www.medicalnewstoday.com/articles/7624
   Accessed on 28th December, 2022.
5. Marie D, Geoffrey AD, Stephen MD, Helen MD, David WH. Acute Stroke Biomarkers: Are We There Yet?, Front Neurol; 2021.
   Available:https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7902038/
   Accessed on 30th June, 2023.
6. Hamdi B. Machine Learning: For Beginners; 2020.
   Available:https://bouzouitina-hamdi.medium.com/machine-learning-for-beginners-b552ec0067a
   Accessed on 26th August, 2023.
7. Ogbu HN, Agana MA. Intranet Security using a LAN Packet Sniffer to Monitor Traffic. In Natarajan M.(Eds). 2019;9(8):57-68: CCSIT, NCWMC, DaKM
8. Fernandez-Lozano C, Hervella P, Mato-Abad V, Rodríguez-Yáñez M, Suárez-Garaboa S, López-Dequidt I, et al. Random forest-based prediction of stroke outcome, Journal of Scientifc Reports; 2021.
   Available:https://www.nature.com/articles/s41598-021-89434-7
   Accessed on 16th February, 2023.
9. Ellis C. Random Forest overfitting; 2023.
   Available:https://crunchingthedata.com/random-forest-overfitting/
   Accessed on 26th August, 2023.
10. Logunova I. Random Forest Classifier: Basic Principles and Applications; 2022.
    Available:https://serokell.io/blog/random-forest-classification

Accessed on 26th August, 2023.

11. Queiroz DA, Assunção GSA, Ferreira KAS, Moura VV, Lima VPB, Dias FA, et al. Prediction of survival in breast cancer patients using Random Forest classifier and ReliefF feature selection method, International Journal of Computer Science and Information Security (IJCSIS). 2021;19(5):41-47.
12. Octaviani TL, Rustama Z. Random Forest for Breast Cancer Prediction, Proceedings of the 4th International Symposium on Current Progress in Mathematics and Sciences (ISCPMS 2018). 2018;1-6.
13. Hashi Z. Lung cancer survival prediction using random forest based decision tree algorithms, proceedings of the international conference on industrial engineering and Operations Management Washington, DC, USA; 2018.
Available:https://www.researchgate.net/publication/328772631_Lung_Cancer_Survival_Prediction_Using_Random_Forest_Based_Decision_Tree_Algorithms
Accessed on 11th March, 2023.
14. Sittidech P, Nai-arun N. Random forest analysis on diabetes complication data, Proceedings of the IASTED International Conference Biomedical Engineering (BioMed) Zurich, Switzerland. 2014;315-320.
15. Wang J, Heping Y, Hua Q, Jing S, Liu Z, Peng X, Cao C, Luo Y. A descriptive study of random forest algorithm for predicting COVID-19 patients outcome, Journal of Biomedical and Life Sciences; 2020.
Available:https://www.ncbi.nlm.nih.gov/pmc/
Accessed on 26th August, 2023.

Peer-review history:
The peer review history for this paper can be accessed here:
https://prh.globalpresshub.com/review-history/1353