International Journal of Electronics and Communications System
Volume 1, Issue 2, 57-62.
ISSN: 2798-2610
http://ejournal.radenintan.ac.id/index.php/IJECS/index
DOI: *10.24042/ijecs.v1i2.10393*

Stroke prediction analysis using machine learning classifiers and feature technique

Md. Monirul Islam *
Uttara University, Dhaka-1230, BANGLADESH

Sharmin Akter
Atish Dipankar University of Science & Technology, Dhaka-1230, BANGLADESH

Md. Rokunojjaman
Chongqing University of Technology, Chongqing 400054, CHINA

Jahid Hasan Rony
Dhaka University of Engineering and Technology, Gazipur, Gazipur-1700, BANGLADESH

Al Amin
Chongqing University of Technology, Chongqing 400054, CHINA

Susmita Kar
Dhaka University of Engineering and Technology, Gazipur, Gazipur-1700, BANGLADESH

Article Info

Article history:

Received: October 4, 2021
Revised: December 8, 2021
Accepted: December 15 2021

Keywords:

Feature Technique,
Random Forest Classifier,
Stroke disease

Abstract

Stroke is one of the fatal brain diseases that cause death in 3 to 10 hours. However, most stroke mortality can be prevented by identifying the nature of the stroke and reacting to it promptly through smart health systems. In this paper, a machine learning model is approached for predicting the existence of stroke of a patient where the Random forest classifier outperforms the state-of-the-art models, including Logistic Regression, Decision Tree Classifier (DTC), K-NN. We conduct the experiments on datasets which has 5110 observations with 12 attributes. We also applied EDA for preprocessing and feature techniques for balancing the datasets. Finally, a cloud-based mobile app collects user data to analyze and provide the possibility of stroke for alerting the person with the accuracy of precision 96%, recall 96%, and F1-score 96%. This user-friendly system can be a lifesaver as the person gets an essential warning very easily by providing very little information from anywhere with a mobile device.

To cite this article: M. M. Islam, S. Akter, M. Rokunojjaman, J. H. Rony, A. Al Amin, and S. Kar, "Stroke Prediction Analysis using Machine Learning Classifiers and Feature Technique," Int. J. Electron. Commun. Syst., vol. 1, no. 2, 57-62, 2021.

INTRODUCTION

A stroke happens to interrupt blood flow to a portion of your brain [1]. A loss of blood circulation to some brain areas causes a stroke, which is also known as a brain attack [2]. Furthermore, clot blocking is the major cause of stroke in the brain (thrombosis). The blood vessel delivers the brain portion and is subsequently run down of blood and oxygen. The brain cells expire as an outcome of the lack of blood and O2, and the part of the body it regulates ceases working [3]. Death and disability happen for stroke in the United States badly. Ischemic embolic and hemorrhagic strokes cause the majority of strokes. An ischemic embolic stroke happens when a blood clot exits the patient's brain, travels through the circulatory system, and becomes lodged in smaller brain arteries. Another type is hemorrhagic stroke, which occurs when leaks or ruptures a blood vessel in the brain. [4]. The use of various predictive indicators to predict the outcome of a stroke could help doctors identify high-risk patients and reduce morbidity. Overweight, physical inactivity, diabetics, and other parameters such as age, sex, race can be used to predict the possibility of stroke. On the other hand, machine learning offers an option, particularly for large-scale multi-institutional data that

• **Corresponding author:**
Md. Monirul Islam, Uttara University, Dhaka-1230, BANGLADESH. ✉ monirul@uttarauniversity.edu.bd

may be readily included in a forecast [5] based on freshly available data.

Smartphones can play a vital role in establishing a between the healthcare system and the global population. A mobile app is very user-friendly and popular in the current world. According to statistics, there are more than 3.2 billion smartphone users. As a result, a mobile app could be one of the most popular and effective mediums. In case of stroke, a disease avoidable through awareness, a smartphone could be the easier way to reach people.

Machine learners have various applications expanding within the study of bioinformatics, a subfield of artificial intelligence which includes improving calculations to discover how projections are dependent on information. Bioinformatics manages computational and numerical approaches for comprehension and manipulating natural data. Six natural environments have been subjected to machine learning. To assist in the analysis of stroke, machine learning algorithms for examining neuroimaging data are used. The diagnosis and treatment of stroke disease in underdeveloped countries is extremely difficult due to a lack of diagnostic technologies and a scarcity of doctors and other resources that impede the accurate prediction and treatment of heart patients. Recently, computer technology and machine learning approaches have been developed with this goal in mind to improve the system's ability to assist doctors in the initial phases of disease decision-making [6]. Our motivation is to benefit stroke prediction to prevent casualty and ensure accessibility for everyone.

Among various studies in this area, in [7], stroke prediction directions were designed as risk assessment and web-based cooperative Java applets. These Java applets enable risk calculations and can be run interactively with any web browser that supports Java 1.1. With this method, patient data can simply be entered into a computer that uses complex statistical models to produce instant calculations of risk scores. Authors in [8] examined the utility of the echo planer magnetic perfusion imaging and diffusion-weighted imaging in predicting stock with a critical hemispheric infraction. In [9], type 2 diabetes patients have an increased risk of stroke. In this approach, they examined the stroke predictors and effects of atorvastatin on certain stroke subtypes in type 2 diabetes in the collaborative atorvastatin diabetes study, which used Cox regression models to evaluate atorva's impact statins on stroke, and assess the risks associated with stroke and underlying stroke. The authors determined how many self-measures of blood pressure they took home compared to their predictive value for the risk of a stroke. In [10], they have designed and compared several methods of learning machines, which can predict the result of endovascular intervention in the previous histosa circulation. The authors developed a CPS to detect the appearance of the patients who are at high risk or survived a stroke before [11]. CPS developed send data registered by the doctor and warned to find a stroke.

Furthermore, the proposed system works in data purchased by the patient's brain electroencephalography sensor. The authors have developed a model learning model (ML) calculated by threshold (ML) to predict the tracking infarction in patients with acute ischemic stroke [12]. The author determined the optimal number of self-measurements of blood pressure at home based on its predictive value for stroke risk. Therefore, the Cox proportional hazard regression model [13] was used to investigate the prognostic significance of blood pressure for the risk of stroke, which was adjusted for possible confounding factors.

The author has developed a very accurate and highly interpretable predictive model. These predictive models will be provided in the form of sparse decision lists [14], which are derived from a series of if . . .Then . . . Statements where the if statement defines a set of feature partitions and the then statement corresponds to the predicted result of interest. In [15], the authors predicted stroke occurrence using a large population-based EMC database and also compared DNN with three other ML methods. The authors compared the Cox proportional hazard model with an automatic stroke prediction approach based on a cardiovascular health study (CHS) dataset [16]. The author developed a hybrid machine learning method to predict stroke based on incomplete and unbalanced physiological data for clinical diagnosis [17]. Using this method, the whole process involves

two steps. First, use random forest regression to estimate missing values before classification. Secondly, automatic hyperparameter optimization (AutoHPO) based on deep neural networks (DNN) predicts stroke on unbalanced data sets. The author applies machine learning principles to existing large data sets to effectively predict strokes based on potentially changeable risk factors [18] to develop applications that provide personalized warnings and related information based on each user's stroke risk level Lifestyle of stroke risk factors. The authors raised the hypothesis that the degree of stenosis, the irregularity of the plaque's surface, eColity, and consistency, complicated in a total score of risk (TPR), are predictors of the ischemic blow [19]. Three classification algorithms that include the decision-making tree, naive bayes, and neuronal network are used to predict the stretch based on models higher than general statistics and obtained an adequate model for identification [20].

This paper proposed Stroke prediction analysis using a machine learning algorithm using a healthcare dataset, including various kinds of risk factors.

The rest of the paper is organized as tracks: the methodology is stated in the next section. Study outcome and discussion are in the results and discussion section. Finally, the paper concludes with future scope.

METHOD

Figure 1 shows the detailed block diagram of the proposed methodology.

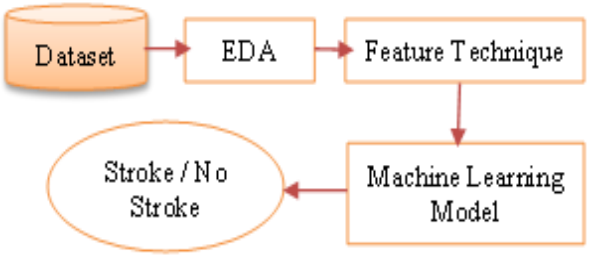

Figure 1. Block Diagram of the Proposed Methodology

Dataset Description

The utilized dataset [21] contains 5110 observations with 12 attributes. The attributes are gender, age, hypertension, heart_disease, ever_married, work_type, Residence type, average glucose_level, BMI, smoking_status, and stroke. Stroke is a dependent variable, and others are independent variables.

Exploratory Data Analysis (EDA)

EDA often uses data visualization approaches to analyze and examine data sets and summarize their key characteristics. It can help determine how best data sources can be handled to get the needed answers, facilitating the finding of patterns, spot anomalies, hypotheses, or assume checks for data scientists. In this part, we defined the missing values, data counts, dropped the id column, exploring each variable.

Feature Techniques

Feature engineering means transforming raw data into features that better signify the predictive models' underlying problem and improve model accuracy in unsightly data. Many techniques can be employed, including NearMiss, SMOTE, Tomak Links, etc. This paper utilized the synthetic minority over-sampling technique (SMOTE) after preprocessing the datasets in the EDA step. The target variable has 201 stroke occurrences and 4908 non-occurrence patients.

Machine Learning Analysis

This paper utilized various machine learning (ML) models containing Naïve Bayes, Random Forest, Ada Boost Algorithm. Among them, the Random Forest model outperforms the best accuracy. So Random forest model is described here.

Random Forest (RF)

RF is a supervised learning algorithm. It creates a "forest" from a series of decision trees that are usually trained using a "bagging" process. The basic premise of the bagging method is that combining different learning models can improve the overall result. The advantage of RF is that it can solve classification and regression problems that make up most of the existing machine learning systems. Decision trees or bagging classifiers have almost the same hyperparameters as random forests. Fortunately, you can use random forest classifiers instead of combining decision trees and bagging classifiers. You can use the algorithm's suppressor to handle the regression task of the random forest. The RF adds additional unpredictability to the model

as the tree develops. Instead of the most relevant feature, splitting a node looks for the optimal function in a random selection of features. Hence, many types lead to better models. Therefore, the algorithm for splitting nodes in the random forest only considers a random subset of features. Instead of looking for the best possible threshold, you can make the tree more random by using random thresholds for each function [22].

The random forest training algorithm uses the general aggregation bootstrap technique, or bags, for train trees students. Figure 2 demonstrates the concept of a random forest model where Tree 1 and Tree 2 associate Class X. So, the majority vote/predicted output is Class X.

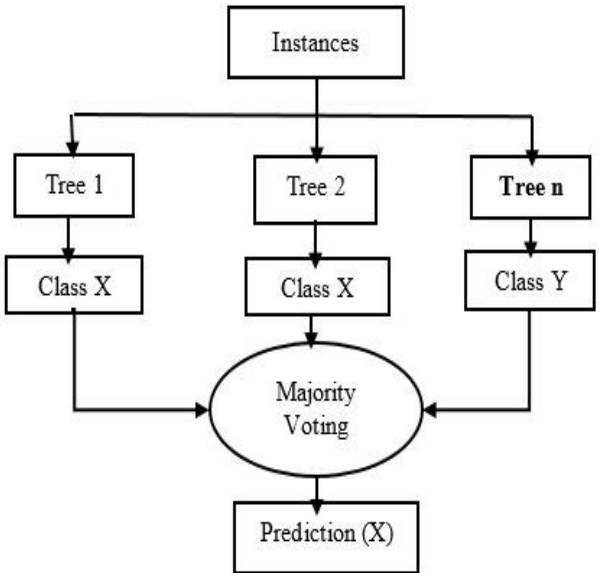

Figure 2. Random Forest

Predictions for unseen samples I can be produced after training by summing the predictions from all of the separate regression trees on i':

$$\hat{f} = \frac{1}{D}\sum_{n=1}^{D} f_b(i')$$

or by taking the majority vote in the case of classification trees [24].

User Interface

User data are collected through mobile apps. Users input gender, age, work_type, heart_disease, hypertension, ever_married, Residence_type, BMI, avg glucose level, smoking_status through the mobile app. In Figure 3, the mobile app interface is shown. User data are stored in the cloud Firestore database. After the processing, the result is stored in the Firestore and shown on the user end.

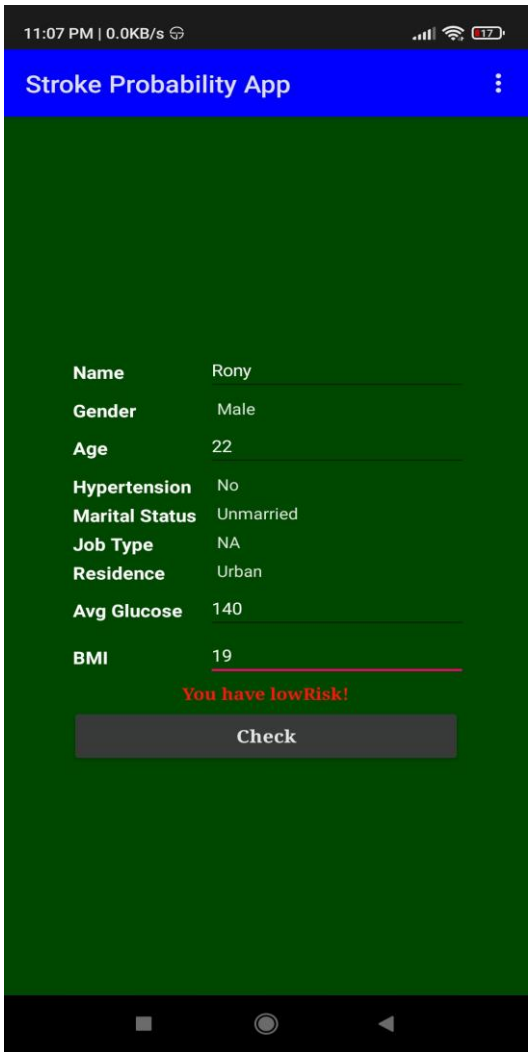

Figure 3. Mobile App

RESULTS AND DISCUSSION

Python programming language is used to classify the proposed model and describe other models for data analysis. The instrument is very useful for analysis and includes different methods. For each model species, we have used 20% of the values for testing and 80% for training. We take precision, recall, and f1-score as performance metrics.

Precision (P): P is the ratio of the positive cases correctly predicted to the positive cases. The low false positive rate refers to high accuracy. It is a measure of a classifier's accuracy. In equation 1, it is defined mathematically.

$$P = \frac{TP}{TP + FP} \qquad (1)$$

Recall (R): R refers to the ratio of positive cases correctly predicted to all positive classification cases. It is a measure of a classification's completeness. In Equation 2, R is defined mathematically.

$$R = \frac{TP}{TP + FN} \quad (2)$$

F1-Score: is an average weighted accuracy and recall. F1, if there is an inconsistent class distribution in the data set, is usually more useful than precision. It is displayed in equation 3 mathematically, and the result of accuracies can be seen in table 1.

$$F1 - score = \frac{2 \times (P \times R)}{P + R} \quad (3)$$

Table 1. Result Accuracies

ML Model	Accuracies (%)		
	Precision	***Recall***	***F1-Score***
Logistic Regression [23]	87	87	87
DTC [24]	93	93	93
K-NN [25]	90	91	90
Random Forest (proposed)	96	96	96

Table 1 describes the result of accuracies. The random forest model gives the highest accuracies in all performance metrics as 96%. K-NN achieves 3rd place as holding 90% performance metrics, DTC stays 2nd position as 93% accuracy, and logistic regression receives 87% accuracy.

CONCLUSION

This paper presented a machine learning approach to the stroke dataset. The Random Forest models showed the best accuracy as precision 96%, recall 96%, and F1-score 96%, outperforming the state-of-art models including logistic regression, decision tree classifier, and K-NN. The utilized dataset is imbalanced, therefore, SMOTE feature engineering is used to process the data. In the future, we will plan to analyze the dataset using deep learning methods and try to enhance the accuracy.

REFERENCES

[1] A. Point, "Stroke (Causes, Symptoms, and Complications) - Assignment Point." https://www.assignmentpoint.com/science/medical/stroke-causes-symptoms-and (accessed May 16, 2021).

[2] M. Clinic, "Stroke - Symptoms and Causes," Mayo Clinic, Nov. 06, 2020. https://www.mayoclinic.org/diseases-conditions/stroke/symptoms-causes/syc-20350113.

[3] B. Wedro, "Stroke Warning Signs, Symptoms, Treatment, Types & Causes," MedicineNet, 2019. https://www.medicinenet.com/stroke_symptoms_and_treatment/article.htm

[4] H. Rodgers, "Stroke," Neurological Rehabilitation, pp. 427–433, 2013, doi: 10.1016/b978-0-444-52901-5.00036-8

[5] H. Asadi, R. Dowling, B. Yan, and P. Mitchell, "Machine Learning for Outcome Prediction of Acute Ischemic Stroke Post Intra-Arterial Therapy," PLoS ONE, vol. 9, no. 2, p. e88225, Feb. 2014, doi: 10.1371/journal.pone.0088225.

[6] P. Govindarajan, R. K. Soundarapandian, A. H. Gandomi, R. Patan, P. Jayaraman, and R. Manikandan, "Classification of stroke disease using machine learning algorithms," Neural Computing and Applications, vol. 32, no. 3, pp. 817–828, Jan. 2019, doi: 10.1007/s00521-019-04041-y.

[7] T. Lumley, R. A. Kronmal, M. Cushman, T. A. Manolio, and S. Goldstein, "A stroke prediction score in the elderly," Journal of Clinical Epidemiology, vol. 55, no. 2, pp. 129–136, Feb. 2002, doi: 10.1016/s0895-4356(01)00434-6.

[8] P. A. Barber et al., "Prediction of stroke outcome with echoplanar perfusion- and diffusion-weighted MRI," Neurology, vol. 51, no. 2, pp. 418–426, Aug. 1998, doi: 10.1212/wnl.51.2.418.

[9] G. A. Hitman et al., "Stroke prediction and stroke prevention with atorvastatin in the Collaborative Atorvastatin Diabetes Study (CARDS)," Diabetic Medicine, vol. 24, no. 12, pp. 1313–1321, Dec. 2007, doi: 10.1111/j.1464-5491.2007.02268.x.

[10] H. Asadi, R. Dowling, B. Yan, and P. Mitchell, "Machine Learning for Outcome Prediction of Acute Ischemic Stroke Post Intra-Arterial Therapy," PLoS ONE, vol. 9, no. 2, p. e88225,

Feb. 2014, doi: 10.1371/journal.pone.0088225.

[11] A. Laghari, Z. A. Memon, S. Ullah and I. Hussain, "Cyber Physical System for Stroke Detection," in IEEE Access, vol. 6, pp. 37444-37453, 2018, doi: 10.1109/ACCESS.2018.2851540.

[12] H. Kuang et al., "Computed Tomography Perfusion–Based Machine Learning Model Better Predicts Follow-Up Infarction in Patients With Acute Ischemic Stroke," stroke, vol. 52, no. 1, pp. 223–231, Jan. 2021, doi: 10.1161/strokeaha.120.030092.

[13] T. Ohkubo et al., "How many times should blood pressure be measured at home for better prediction of stroke risk? Ten-year follow-up results from the Ohasama study," Journal of Hypertension, vol. 22, no. 6, pp. 1099–1104, Jun. 2004, doi: 10.1097/00004872-200406000-00009.

[14] B. Letham, C. Rudin, T. H. McCormick, and D. Madigan, "Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model," Annals of Applied Statistics, vol. 9, no. 3, pp. 1350–1371, Sep. 2015, doi: 10.1214/15-AOAS848.

[15] C. Hung, W. Chen, P. Lai, C. Lin and C. Lee, "Comparing deep neural network and other machine learning algorithms for stroke prediction in a large-scale population-based electronic medical claims database," 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2017, pp. 3110-3113, doi: 10.1109/EMBC.2017.8037515.

[16] A. Khosla, Y. Cao, C. C.-Y. Lin, H.-K. Chiu, J. Hu, and H. Lee, "An integrated machine learning approach to stroke prediction," Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '10, 2010, doi: 10.1145/1835804.1835830.

[17] T. Liu, W. Fan, and C. Wu, "A hybrid machine learning approach to cerebral stroke prediction based on imbalanced medical dataset," Artificial Intelligence in Medicine, vol. 101, p. 101723, Nov. 2019, doi: 10.1016/j.artmed.2019.101723.

[18] M. Monteiro et al., "Using Machine Learning to Improve the Prediction of Functional Outcome in Ischemic Stroke Patients," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 15, no. 6, pp. 1953-1959, 1 Nov.-Dec. 2018, doi: 10.1109/TCBB.2018.2811471.

[19] P. Prati et al., "Carotid Plaque Morphology Improves Stroke Risk Prediction: Usefulness of a New Ultrasonographic Score," Cerebrovascular Diseases, vol. 31, no. 3, pp. 300–304, 2011, doi: 10.1159/000320852.

[20] T. Kansadub, S. Thammaboosadee, S. Kiattisin and C. Jalayondeja, "Stroke risk prediction model based on demographic data," 2015 8th Biomedical Engineering International Conference (BMEiCON), 2015, pp. 1-3, doi: 10.1109/BMEiCON.2015.7399556.

[21] "Stroke Prediction Dataset," kaggle.com.https://www.kaggle.com/fedesoriano/stroke-prediction-dataset.

[22] MM. Islam, MA. Kashem, J. Uddin, "Fish survival prediction in an aquatic environment using random forest model," IAES International Journal of Artificial Intelligence (IJ-AI), vol. 10, no. 3, pp. 614–622, 2021, doi: 10.11591/ijai.v10.i3.pp614-622.

[23] Md. M. Islam, J. Uddin, M. A. Kashem, F. Rabbi, and Md. W. Hasnat, "Design and Implementation of an IoT System for Predicting Aqua Fisheries Using Arduino and KNN," Intelligent Human Computer Interaction, pp. 108–118, 2021, doi: 10.1007/978-3-030-68452-5_11.

[24] A. Esmael, M. Elsherief, and K. Eltoukhy, "Predictive Value of the Alberta Stroke Program Early CT Score (ASPECTS) in the Outcome of the Acute Ischemic Stroke and Its Correlation with Stroke Subtypes, NIHSS, and Cognitive Impairment," Stroke Research and Treatment, vol. 2021, pp. 1–10, Jan. 2021, doi: 10.1155/2021/5935170.

[25] C. Y. Baek, W. N. Chang, B. Y. Park, K. B. Lee, K. Y. Kang, and M. R. Choi, "Effects of dual-task gait treadmill training on gait ability, dual-task interference, and fall efficacy in people with stroke: A Randomized Controlled Trial," Physical Therapy, Feb. 2021, doi: 10.1093/ptj/pzab067.