Automated Stroke Prediction Using Machine Learning Approach

Saber Hossain
Dept. of ECE
North South University
Dhaka, Bangladesh
saber.hossain@northsouth.edu

Al Sabri Bhuiyan
Dept. of ECE
North South University
Dhaka, Bangladesh
sabri.bhuiyan@northsouth.edu

Maher Ali Rusho*
Senior Scientist
Department of Computational Material & Data Analytics
Mr. R Business Corporation (NGO)
Chennai, Tamil Nadu, India
maher.rusho@colorado.edu

Md Ruhan Afride
Dept. of ECE
North South University
Dhaka, Bangladesh
ruhan.afride@northsouth.edu

Rifat Ibna Azad
Dept. of ECE
North South University
Dhaka, Bangladesh
rifat.ibna@northsouth.edu

Abu Mukaddim Rahi
Dept. of ECE
North South University
Dhaka, Bangladesh
abu.rahi@northsouth.edu

Mithila Arman
Dept. of CSE
BRAC University
Dhaka, Bangladesh
mithila.arman@g.bracu.ac.bd

Md. Khurshid Jahan
Dept. of ECE
North South University
Dhaka, Bangladesh
khurshid.jahan@northsouth.edu

***Abstract*—A specific damage to the central nervous system brought on by decreased blood supply to the brain is known as a stroke. Stroke can be defined as when the blood vessels in the brain burst, harming the brain and causing damage to it. Symptoms may appear when the blood flow of the brain and other nutrients is disrupted. Stroke is becoming a major global hazard that can result in early mortality and severe economic effects. The World Health Organization (WHO) claims that stroke is the leading cause of death and disability worldwide. Therefore, it is urgent to predict how various risk factors affect the likelihood of having a stroke, and artificial intelligence (AI) seems to be the right instrument for the job. In this paper, we have dealt with a typical severe class imbalance problem, which is brought on by the fact that the class of stroke patients is significantly smaller than the class of healthy individuals. SMOTE technique has been implemented to manage the class imbalance problem. Several machine learning (ML) models have been implemented to estimate the probability that a stroke would occur. In this study, we have trained three machine learning models, which are K-Nearest Neighbor (KNN), Random Forest (RF) Classification, and Support Vector Machine(SVM), among these three models we got the best result with Random Forest where the accuracy score is 0.96.**

***Index Terms*—stroke, machine learning, symptoms, accuracy**

I. Introduction

Stroke is described as the rapid demise of brain cells due to the absence of oxygen, and it might occasionally be asymptomatic. It occurs when a blood clot or brain bleeding occurs, permanently harming the body. With 450–750 strokes per 155,000 people and 16.5 million new acute strokes each year, stroke is the 2nd most significant cause of death and adult disability globally [1]. According to the World Health Organization (WHO) estimates, 15.3 million people experience strokes annually, with one victim passing away every 4.0 to 5.0 minutes. According to the limited information that is currently available, noncommunicable diseases (NCDs), such as cancer, stroke, diabetes, chronic obstructive pulmonary disease, and heart disease, account for about 0.51 of deaths in Bangladesh [2].

Stroke can be classified into two categories:

1) Ischemic.
2) Hemorrhagic.

An Ischemic stroke can be known as a blocked artery, and a Hemorrhagic stroke is a burst out of a blood vessel. Many people have a short disruption of blood flow to their brain, known as a transient ischemic attack (TIA), that does not cause any lifelong symptoms to the body. A stroke can also be known as a brain attack that can happen in two possible ways: one blocks the arteries, and the other one is the arteries are ruptured, which occurs when blood supplies are blocked to the brain, or a blood vessel in the brain bursts out.

In this way, the brain dies or gets damaged. You may be seeing a stroke if any of these signs appear suddenly:

- Weakness or numbness, generally on just one side, of the face, arm, or leg.
- Difficulty comprehending or speaking the language.
- Vision loss or blurriness in one or both eyes.
- Sudden dizziness or balance issues.
- Intense headache with no apparent cause.

979-8-3315-3559-9/25/$31.00 ©2025 IEEE

A stroke can lead to long-term brain damage and last as a disability or possibly to death. Strokes can be measured by physical examination, and examination of brain scan pictures are typically used to identify strokes. Stroke can be prevented by living a healthy, balanced lifestyle that excludes harmful habits like smoking and drinking, maintaining a healthy body mass index (BMI), average blood glucose levels, and great heart and kidney function. Kokkotis and his team [3] attempted to predict stroke prediction using several ensemble method-based machine learning models. The authors have used an open-source Cerebral Stroke Prediction-Imbalanced Dataset. For forming a balanced dataset Standardization technique and under sampling technique were implemented. The authors have used various machine learning models; among them, Logistic Regression obtained the best results with an accuracy of 0.7352.

Tazin and her colleagues [4] proposed an automatic stroke disease prediction system using a machine learning approach. The authors used a public dataset. The authors have implemented SMOTE technique and Label encoding in the dataset. After using a different classifier, Random Forest gave the authors the best accuracy, approximately 0.96. Peñafiel and his colleague [5] tried to anticipate stroke using the Dempster-Shafer plausibility theory. They used a dataset for the study provided by a hospital in Okayama, Japan; however, it contained some incorrect data. The researchers used AUC as their metric for accuracy. The example The model achieved a score of 0.669 for the area under the receiver operator curve (AUC ROC). In this work to identify strokes, we have used explainable AI and various machine-learning approaches. We used a Kaggle-sourced dataset on strokes [6]. We substituted the mean value of each feature for some of the features with several missing values. The data has been divided using the holdout validation procedure. We used KNN, SVM, and random forest in this study, three machine learning-based classification techniques. Next, the accuracy, recall, and F1 measure of these classifiers' performance have been analyzed.

In this paper, machine learning is used to implement stroke prediction. As previously indicated, the main contribution of this study is the employment of various machine learning models on a freely available dataset. SMOTE approach is used to reduce the problem of class imbalance. In this work, hyperparameter tuning has also been done. Using the LIME library, an explainable AI technique shows how the model predicts the outcome. The LIME technique clarifies what the model is doing and how each attribute affects how a sample is classified. This method aids in interpreting which features are most important for prediction. Most researchers employed a substantial model to predict stroke disease in earlier studies. However, we used three distinct models and contrasted our findings with earlier research. The next part offers a brief discussion of all the results and comparisons.

II. PROPOSED SYSTEM

The suggested automated diabetes prediction system's procedures and application of several machine learning algorithms are described in this part. The dataset was first gathered and preprocessed to deal with imbalanced class issues, replacing null occurrences with mean values, etc. The holdout validation technique was then used to divide the dataset into training and test sets. The optimal classification algorithm for this dataset was then determined by applying various classification algorithms. The accuracy measures of accuracy score, precision score, recall score, and F1 score are used to compare the models created. At the end, we have deployed our machine learning model in a website so that people can check whether they have a stroke or not at their convenient place. Fig. 1 demonstrates the working sequences of the proposed automatic stroke prediction system.

A. Dataset Overview

The study was carried out using the stroke prediction dataset. This dataset has 12 columns and 5110 rows. The output column stroke has a value of either 1 or 0 [7]. The value 0 denotes no stroke risk was found, but the value 1 indicates a stroke risk. We have identified several significant features in our dataset, including hypertension, heart disease, smoking status, BMI, average glucose level, etc. Data preprocessing is used to balance the data to increase accuracy.

TABLE I. FEATURES OF THE DATASET

Variable	**Type**
Number of records (public dataset)	5110
Number of Attributes	12
age	Fluctuates from 8 to 82 years
hypertension	1(yes) and 0(no)
avg_glucose_level	Ranges from 55 to 271
hypertension	1(yes) and(no)
BMI	Ranges from 0 to 97
heart_disease	1(yes) and 0(no)
stroke	1 (stroke) or 0 (no stroke)

B. Dataset Preprocessing

Data preprocessing is necessary before creating a model to exclude unnecessary noise and outliers from the dataset that can cause the model to deviate from its intended training. All that stops the model from working more effectively is addressed in this step. The required dataset must first be gathered, and then the data needs to be cleaned and ready for model building [8]. The dataset has twelve features, as it was previously mentioned. Firstly, the column id is left out because it does not affect how the model is built. The dataset is next checked for null values, and any identified are filled in. In this instance, the mean of the data column is used to fill in the null values in the BMI column. Due to the different scales in the dataset, it had to be normalized. In this study, a min-max normalizer scalar, which is defined as

$$X_{\text{new}} = \frac{X - \min(X)}{\max(X) - \min(X)}$$

One-hot encoding transforms categorical values into numerical values. The strings must be converted to numerical values since the computer is typically educated on numbers [9].

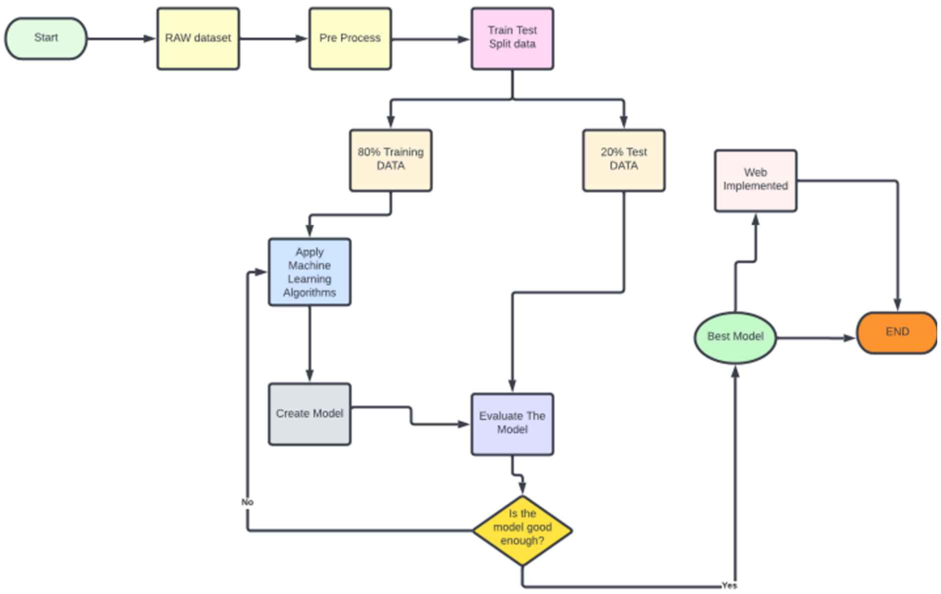


Fig. 1. Proposed Methodology for Calories Burn Prediction

Five columns of the collected dataset are of the string data type. During one-hot encoding, every text is encoded, turning the entire dataset into a set of integers. The dataset used to predict strokes could be more balanced. There are 5110 entries in the dataset; 4.9% indicate that there may have been a stroke, and 95.1% of which indicate there was no stroke found. While accuracy may be achieved by using such data to train a machine-level model, other accuracy metrics like precision and recall are Ineffective, and insufficient forecasting and erroneous discoveries will result from improper handling of such uneven data. Therefore, it is necessary to address this unbalanced data first to build an effective model. The SMOTE technique was used to achieve this.

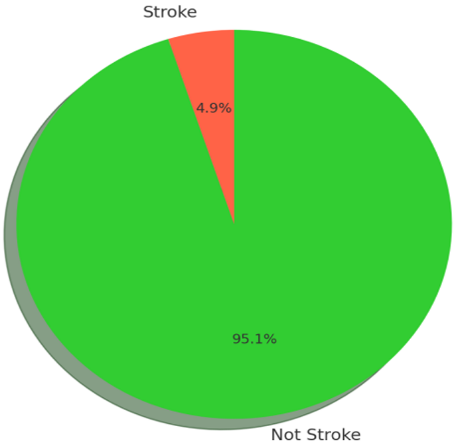


Fig. 2. Before SMOTE technique was applied to the dataset

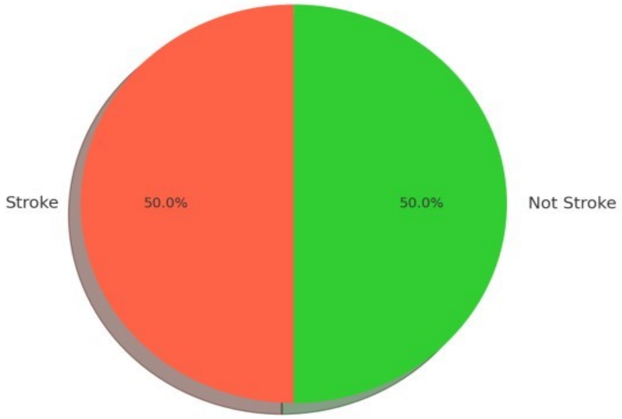


Fig. 3. After SMOTE technique was applied to the dataset

C. *Machine Learning Models*

1) SVM: Support Vector Machine or also known as SVM, can be known as one of the most well-known supervised learning algorithms used to solve Classification and Regression tasks [10]. The SVM method aims to construct the best decision boundary or line that can divide n-dimensional space into classes so that we can quickly classify newer data points in the future. The hyperplane is created using SVM by selecting the extreme points and vectors. Support vectors control the decision boundary line; for this reason, it is known as Support Vector Machine.

2) KNN: K-nearest neighbours (KNN) is a supervised learning algorithm [11]. It is used for both regression and classification models. KNN determines the distance between the test data and all of the training points, and KNN tries to predict the proper class for the test data. Then choose the K points that are closest to the test data.

3) Random forest: Random Forest is known as a machine learning algorithm [12]. It is a part of the supervised learning methodology. Random Forest is used for both Classification and Regression problems in machine learning. It is predicated on the idea of ensemble learning, which is the technique of integrating various classifiers to address a complicated issue and enhance the model's performance.

III. RESULTS AND DISCUSSION

The results and explanation of the suggested automated stroke prediction system are presented in this section. The effectiveness of several machine learning approaches is first addressed. The developed website foundation is then seen in action. We employed precision, recall and F1 scores to assess different ML models. These metrics' equations are represented as

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

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{F1 score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

TP stands for True Positive, TN stands for True Negative, FP stands for False Positive, and FN stands for False Negative. These values were derived from the confusion matrix.

TABLE II. Model Hyperparameter Ranges

Model	Hyperparameter Value Range
SVM	C: [0.1, 1, 10, 100, 1000], gamma: [1, 0.1, 0.01, 0.001, 0.0001], kernel: [rbf]
Random Forest	n_estimators: [200,400,600,800,1000,1200,1400,1600,1800,2000], max_features: [auto, sqrt, log2], max_depth: [10,120,230,340,450,560,670,780,890,1000], criterion: [Gini, entropy]
KNN	n_neighbors: [1], scoring=accuracy, verbose=1

Table II demonstrates the hyperparameter values' ranges for all the ML models. The GridSearchCV technique's associated improved hyperparameters are presented in Table II. The optimized hyperparameters for Random Forest are n_estimator=1600, max_features=log2, max_depth=890 and criterion=gini.

TABLE III. PERFORMANCE METRICS OF VARIOUS CLASSIFIERS FOR DEFAULT PARAMETERS AND SMOTE

Classifier	Precision	Recall	F1 Score	Accuracy
Random Forest	95%	100%	97%	96%
SVM	91%	100%	95%	95%
KNN	95%	97%	96%	95%

Table III demonstrates the performance metrics of several classifiers using the default settings and SMOTE. With 96% accuracy and a 97% F1 score, Random Forest exceeds all other machine learning models, as seen in the table.

TABLE IV. PERFORMANCE METRICS OF VARIOUS CLASSIFIERS WITH OPTIMIZED HYPERPARAMETERS AND SMOTE

Classifier	Precision	Recall	F1 Score	Accuracy
Random Forest	95%	100%	97%	97%
SVM	93%	99%	96%	95%
KNN	95%	94%	95%	95%

Table IV demonstrates the performance metrics of several classifiers using GridSearchCV hyperparameter tuning and SMOTE. It claims that with 97% accuracy and a 97% F1 coefficient, the Random Forest model performed the best.

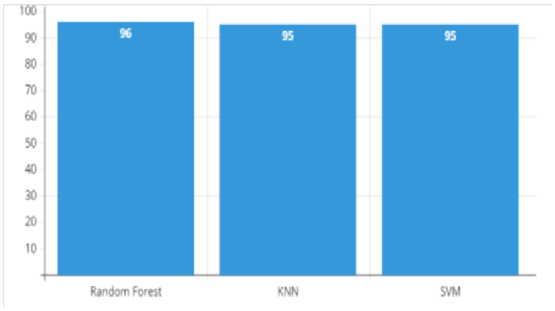


Fig. 4. Demonstration of Bar plot before applying GridSearchCV

Using default settings and the SMOTE approach, a bar graph is used to display the accuracy of the ML models. It shows that Random forest has gained the highest accuracy for the dataset, at 96%. GridSearchCV and SMOTE techniques are used to create a bar chart in Fig. 5 that shows the ML models' accuracy. This statistic shows an improvement in the accuracy of the Random Forest model. The prediction interpretation of the Random Forest model utilizing the LIME explainable AI framework is shown in Fig. 6. The results of the LIME prediction were evaluated using the Random forest model with improved hyperparameters and the SMOTE technique since it performed the best.

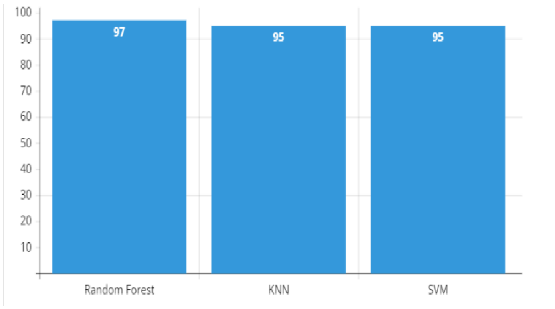
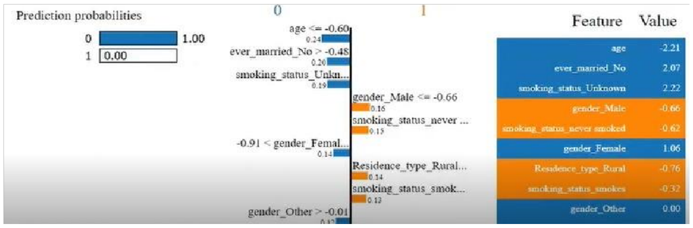


Fig. 5. Demonstration of Bar plot after applying GridSearchCV



Feature	Value
age	-2.21
ever_married_No	2.07
smoking_status_Unknown	2.22
gender_Male	-0.66
smoking_status_never smoked	-0.62
gender_Female	1.06
Residence_type_Rural	-0.76
smoking_status_smokes	-0.32
gender_Other	0.00

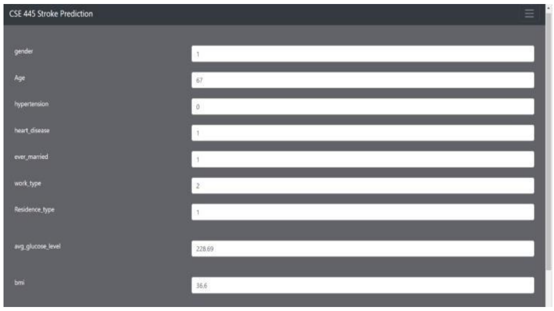
Fig. 6. Evaluation of LIME Explainable AI of Random Forest model.

TABLE V. COMPARISON OF THE PROPOSED SYSTEM WITH EXISTING WORKS

References	ML Model	Accuracy	F1 Score
[1]	Linear Regression	73.52%	N/A
[2]	Random Forest	95%	96%
[3]	SGD	85.4%	45.1%
This work	Random Forest	96%	97%

The suggested automated stroke prediction system is evaluated with past works in Table V. This table demonstrates that, compared to other papers, our suggested model has been highly accurate.

IV. WEB IMPLEMENTATION


CSE 445 Stroke Prediction

gender	1
Age	67
hypertension	0
heart_disease	1
ever_married	1
work_type	2
Residence_type	1
avg_glucose_level	228.69
bmi	36.6

Fig. 7. Deployment of the best model on the website.

The implementation process was done for the people to check their condition from time to time. The deployment was done by (in order):

1) Front-end implementation using flask for numerical input
2) Adding appropriate features
3) Implementing predict button
4) Deploying in localhost

V. CONCLUSIONS

A stroke is a potentially fatal medical condition that must be treated immediately to prevent future consequences. Implementing a machine learning (ML) model might help with early stroke diagnosis and the subsequent reduction of its massive consequences. This study examines the usefulness of several ML algorithms in correctly predicting stroke based on various physiological factors. With a classification accuracy of 0.97, random forest classification exceeds the other techniques. According to the study, the random forest technique performs better than other methods for forecasting strokes using the cross-validation technique. The future purpose of this report is to improve the framework models utilizing a larger dataset and machine learning techniques, including Decision tree, AdaBoost and Bagging. Advanced machine learning algorithms are a great resource for developing more accurate and effective prediction tools that doctors may use.

REFERENCES

[1] C. Kokkotis, G. Giarmatzis, E. Giannakou, S. Moustakidis, T. Tsatalas, D. Tsiptsios, K. Vadikolias, and N. Aggelousis, “An explainable machine learning pipeline for stroke prediction on imbalanced data,” *Diagnostics*, vol. 12, no. 10, p. 2392, 2022.

[2] S. Penafiel, N. Baloian, H. Sanson, and J. A. Pino, “Predicting stroke risk with an interpretable classifier,” *IEEE Access*, vol. 9, pp. 1154–1166, 2020.

[3] T. Tazin, M. N. Alam, N. N. Dola, M. S. Bari, S. Bourouis, and M. Monirujjaman Khan, “Stroke disease detection and prediction using robust learning approaches,” *Journal of healthcare engineering*, vol. 2021, no. 1, p. 7633381, 2021.

[4] A. Géron, *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems.* ” O’Reilly Media, Inc.”, 2022.

[5] L. Chan, H. Li, P. Chan, and C. Wen, “A machine learning-based approach to decipher multi-etiology of knee osteoarthritis onset and deterioration,” *Osteoarthritis and Cartilage Open*, vol. 3, no. 1, p. 100135, 2021.

[6] S. Cloherty, N. Lovell, S. Dokos, and B. Celler, “A 2d monodomain model of rabbit sinoatrial node,” in *2001 Conference Proceedings of the 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, vol. 1, pp. 44–47, IEEE, 2001.

[7] S. Strilciuc, D. A. Grad, C. Radu, D. Chira, A. Stan, M. Ungureanu, A. Gheorghe, and F.-D. Muresanu, “The economic burden of stroke: a systematic review of cost of illness studies,” *Journal of medicine and life*, vol. 14, no. 5, p. 606, 2021.

[8] S. Zhang, J. Wang, L. Pei, K. Liu, Y. Gao, H. Fang, R. Zhang, L. Zhao, S. Sun, J. Wu, *et al.*, “Interpretability analysis of one-year mortality prediction for stroke patients based on deep neural network,” *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 4, pp. 1903–1910, 2021.

[9] F. Di Martino and F. Delmastro, “Explainable ai for clinical and remote health applications: a survey on tabular and time series data,” *Artificial Intelligence Review*, vol. 56, no. 6, pp. 5261–5315, 2023.

[10] A. Géron, *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems.* ” O’Reilly Media, Inc.”, 2022.

[11] P. N. Srinivasu, N. Sandhya, R. H. Jhaveri, and R. Raut, “From blackbox to explainable ai in healthcare: existing tools and case studies,” *Mobile Information Systems*, vol. 2022, no. 1, p. 8167821, 2022.

[12] Q. Teng, Z. Liu, Y. Song, K. Han, and Y. Lu, “A survey on the interpretability of deep learning in medical diagnosis,” *Multimedia Systems*, vol. 28, no. 6, pp. 2335–2355, 2022.