

# Stroke Detection

<sup>1</sup>Md Sajid Mansoori , <sup>2</sup>Naveen Kumar

<sup>1,2,3,4</sup>Master of Computer Application DS & AI, Assistant Professor, Dept. Of Computer Science and Engineering, Babu Banarasi, University Uttar Pradesh, India.

**ABSTRACT:** Stroke is a condition in which a blood vessel in the brain bursts, causing brain damage. Symptoms can occur when blood and other nutrients to the brain are disrupted. According to the World Health Organization (WHO), stroke is the leading cause of death and disability worldwide. Early recognition of the many warning signs of a stroke can help reduce the risk of stroke. Different machine learning (ML) models have been developed to predict the probability of stroke in the brain. This study uses various physical and machine learning algorithms such as logistic regression (LR), decision tree (DT) classification, random forest (RF) classification, and surveys to train four different models with reliable predictions. Random Forest is the best algorithm for this task with around 96% accuracy. The data used in the development of this method is an open access prediction dataset. The fact that the accuracy of the model used in this survey is higher than in previous studies indicates that the model used in this survey is more reliable. Comparison of general models has shown that they are robust and the process can be inferred from clinical trials.

## 1 INTRODUCTION

A stroke occurs when blood flow to different parts of the brain is cut off or reduced, causing cells in those areas of the brain to lose the nutrients and oxygen they need and die. Stroke is a medical emergency that requires immediate treatment. Early diagnosis and appropriate management are necessary to prevent further damage to the affected area of the brain and other problems elsewhere in the body. The World Health Organization (WHO) estimates that 15 million people worldwide suffer a stroke each year, and one person dies every four to five minutes among those affected. According to the Centers for Disease Control and Prevention (CDC), stroke is the sixth leading cause of death in the United States [1]. Stroke is a non-communicable disease that kills about 11% of the population. Approximately 795,000 people are affected by stroke in the United States [2]. It is the fourth leading cause of death in India. Stroke is divided into ischemic stroke and hemorrhagic stroke. In stroke medicine, a clot blocks fluid; In hemorrhagic stroke, a fragile blood vessel ruptures and causes bleeding in the brain. Stroke can be prevented through a healthy diet and lifestyle, including quitting unhealthy habits such as smoking and alcohol, controlling body weight (BMI) and moderate blood sugar, and maintaining good heart and kidney function. Predicting stroke is important and prompt treatment is essential to avoid damage or death. With the development of medical technology, it is now possible to use machine learning techniques to predict the onset of stroke [3] used text mining and machine learning classifiers to classify stroke in 507 people images taken in the system's image processing. CNN techniques achieved 90% accuracy. Song et al.[12] conducted a study to establish a stroke index. They collected data on 3577 patients with ischemic stroke. However, we used four different models and compared the results with previous studies. All results and comparisons are briefly discussed in the next section. The remainder of the article is as follows: experiments and methods are described in Chapter 2; an analysis of the results is given in Chapter 3; and the results are discussed in Chapter 4.

## 2 PROCEDURE AND EXPERIMENTAL METHODOLOGY

This section includes a description of the dataset, a block diagram, a flow diagram, and evaluation matrices, as well as the process and methodology used in the study.

### 2.1. Proposed System

The data is processed and prepared for design. The building model requires preliminary data and machine learning techniques. LR, DT classification, RF classification and voting are some of the methods used. After you create the four proxy models, compare them using the correct metrics (eg Horse Count, Precision Score, Recall Score, and F1 Score). The design block diagram is shown in Figure 1.

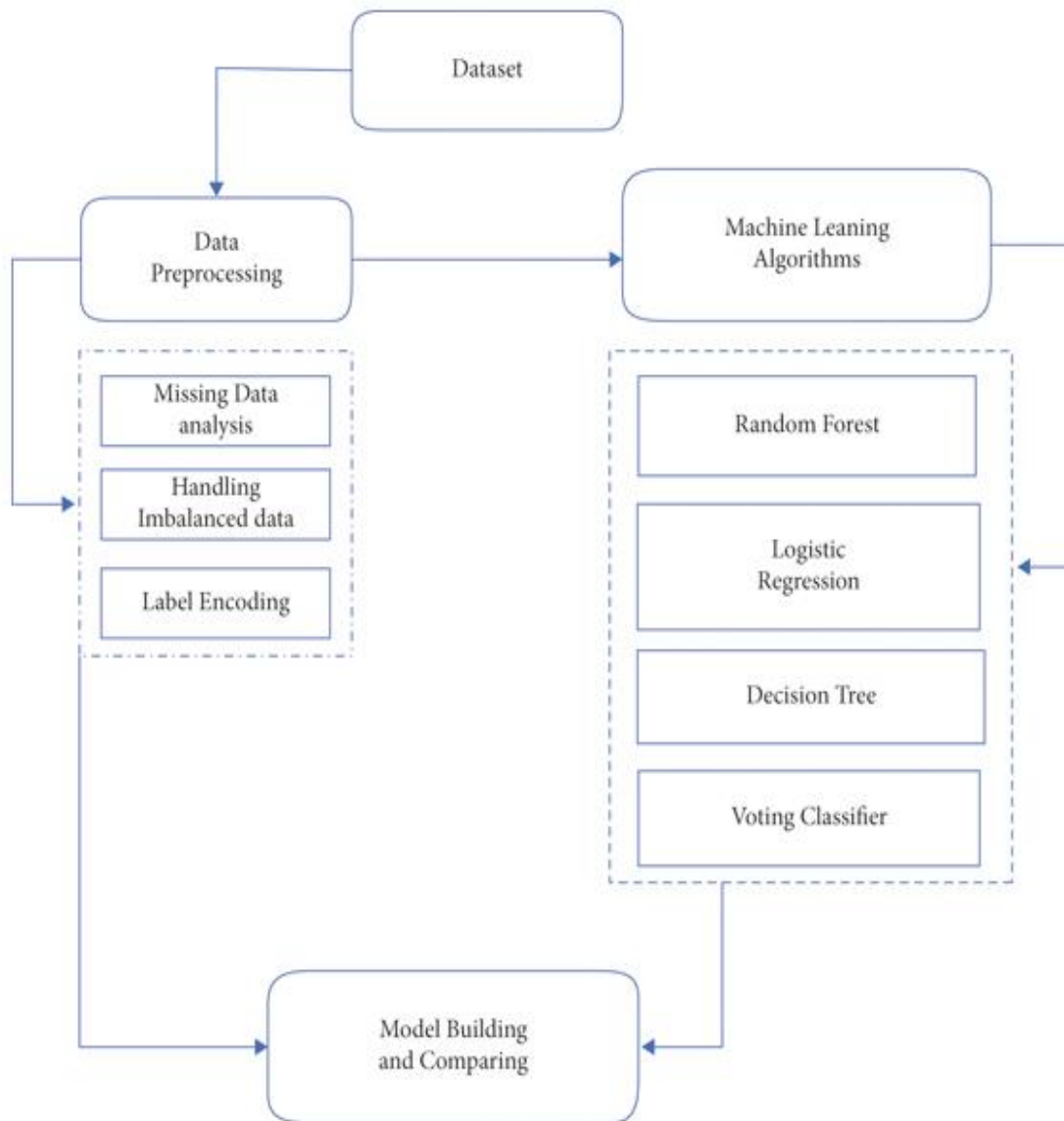


Figure 1 : Proposed System's block diagram

All the components of the block diagram have been discussed in the following subsections.

## 2.2. Dataset

The stroke predictive database [16] was used to conduct this study. This file contains 5110 rows and 12 columns. The stroke indicator has a value of 1 or 0. A number of 0 means that the risk of stroke has not been determined, while a value of 1 means the risk of stroke. The occurrence of 0 in the output column (stroke) exceeds the occurrence of 1 in the same column in this document. The stroke line alone has 249 lines of 1 and 4861 lines of 0. To increase accuracy, data is balanced with previous data. Figure 2 shows all beat and non-hit data in the first column.

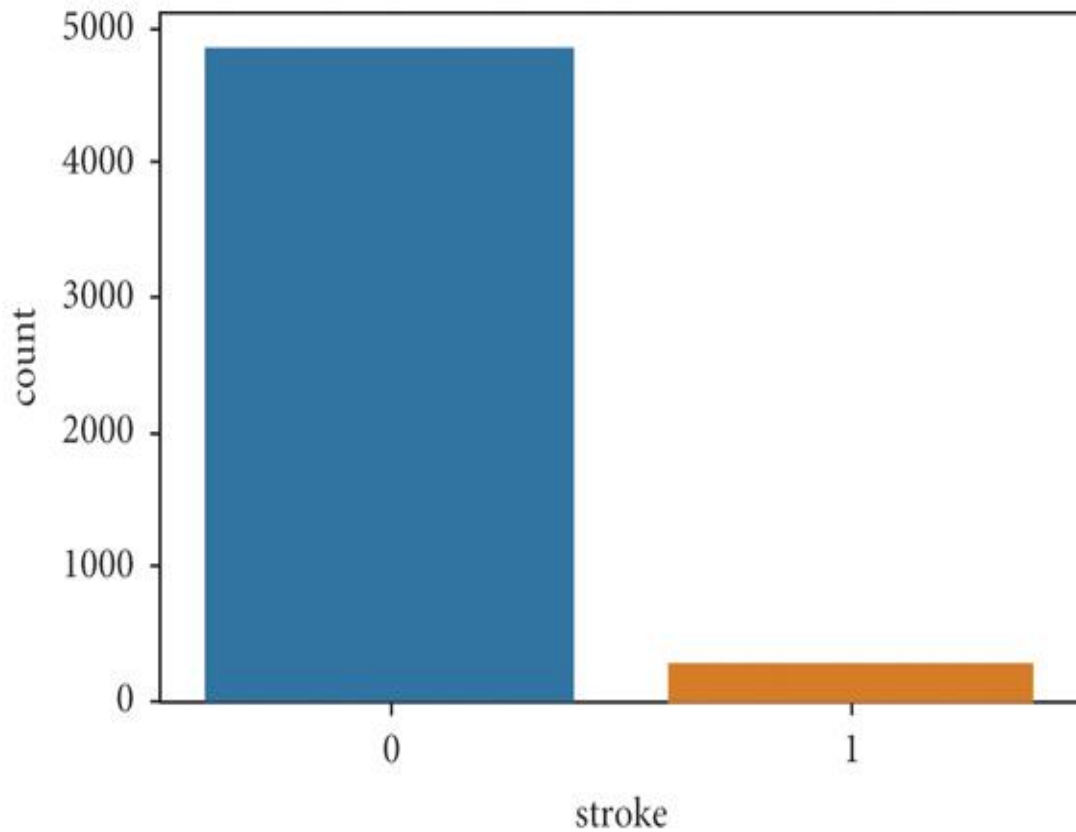


Figure 2 : Total number of stroke and normal data (16)

From Figure 2, it is clear that this dataset is an imbalanced dataset. The SMOTE technique has been used to balance this dataset.

### 2.3. Preprocessing

Before building a model, preliminary data should be free of unwanted noise and bad data that could cause the model to deviate from its training. This phase addresses any issues that prevent the model from working better. After collecting the necessary data, the data should be cleaned and prepared for modelling. As mentioned earlier, the reference material has twelve elements. First, the id column is removed as its presence does not affect the design. The dataset is then checked for nulls and collected if detected. In this example, the values in the BMI column are populated with the average value of the column. The tag encoding converts the string literals of a dataset into machine-understandable integer values. Since most computers are trained with numbers, strings need to be converted to numbers. The data collection consists of five rows of data type columns. During label encoding, all strings are encoded and the entire dataset is converted into a collection of numbers. Data for predicting stroke are inconsistent. The data includes a total of 5110 rows, of which 249 represent stroke probability and 4861 confirmed strokes. While training machine-level models with such data can improve accuracy, other accuracy measures such as precision and recall are insufficient. If these data imbalances are not corrected, the results will be inaccurate and the estimate will be invalid. Therefore, to obtain a good model, unbalanced data must be dealt with first.

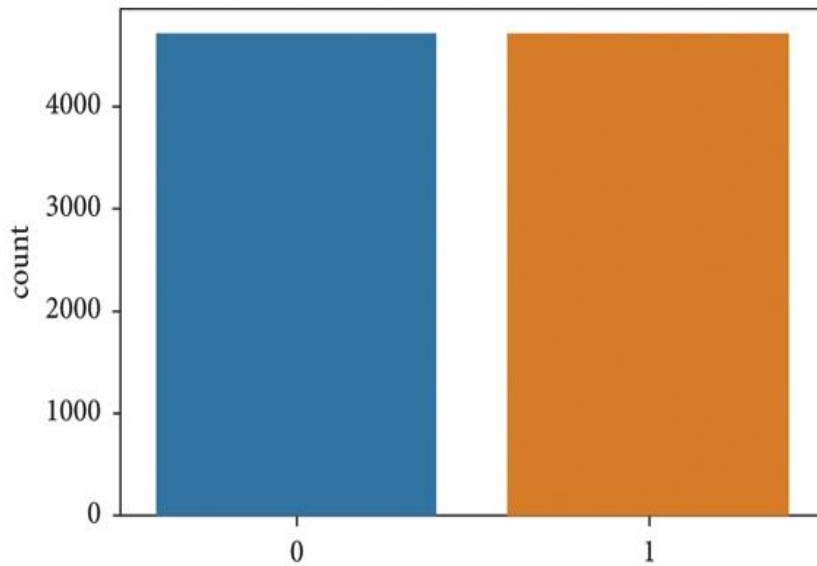


Figure 3 : Output column after preprocessing

After completing the data preparation and managing the unbalanced data, it is time to create the model. To increase the accuracy and efficiency of this study, the data is divided into training data and test data; training data corresponds to 80%, Data evaluation data corresponds to 20%. After segmentation, the model is trained using various classifiers. In this study, classification algorithms such as random forest, decision tree classification, voting and logistic regression were used.

**2.4. Proposed Algorithms**

The most common disease in treatment is stroke, the incidence of which increases every year. Using publicly available stroke data, this study evaluated four machine learning methods to predict stroke recurrence: (i) Random Forest (ii) Decision Tree (iii) Voting Classifier (iv) Logistic Regression

**2.4.1. Random Forest**

The preferred classification algorithm is RF classification [17]. RF consists of several independent decision trees that are individually trained on random data. These trees are created during training and record the results of the decision tree. A process called voting is used to determine the final prediction made by the algorithm. Each CE in this method must vote for one of the two exit groups (in this case, beat or no hit).The final estimate is determined by the RF method, which selects the class with the most votes. A block diagram of a random forest distribution is shown in Figure 4.

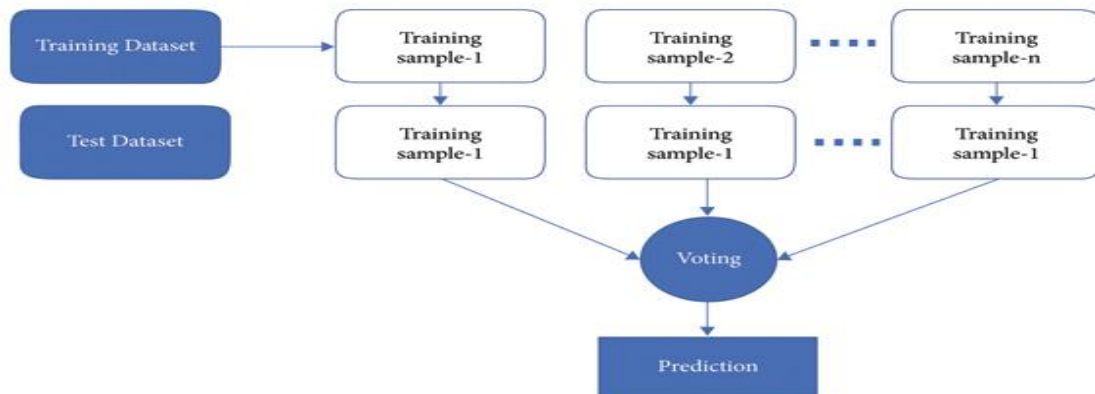


Figure 4 : Block diagram of Random Forest Classifier (17)

The simplicity of random forests is one of its most appealing features. It can be used for recurring calls and group work, and the overall weight given to the information is obvious. It's also a good method, as it uses default hyperparameters that usually provide clear expectations. Understanding hyperparameters is important because there are some of them in the first place. Overfitting is a well-known problem in machine learning, although it rarely occurs in arbitrary random forest classifiers. If there are enough trees in the forest, the classifier will not fit the model.

#### 2.4.2. Decision Tree

Both regression and classification problems can be solved using DT classification [18]. In addition, these models are maintenance models because different ideas are already associated with different products. It is like a tree. In this way, data is continuously segmented according to certain parameters. Order nodes and leaves are two parts of the order tree. While the first node is partitioning the data, the next node is the node causing the result. The basic structure of the DT classifier is shown in Figure 5.

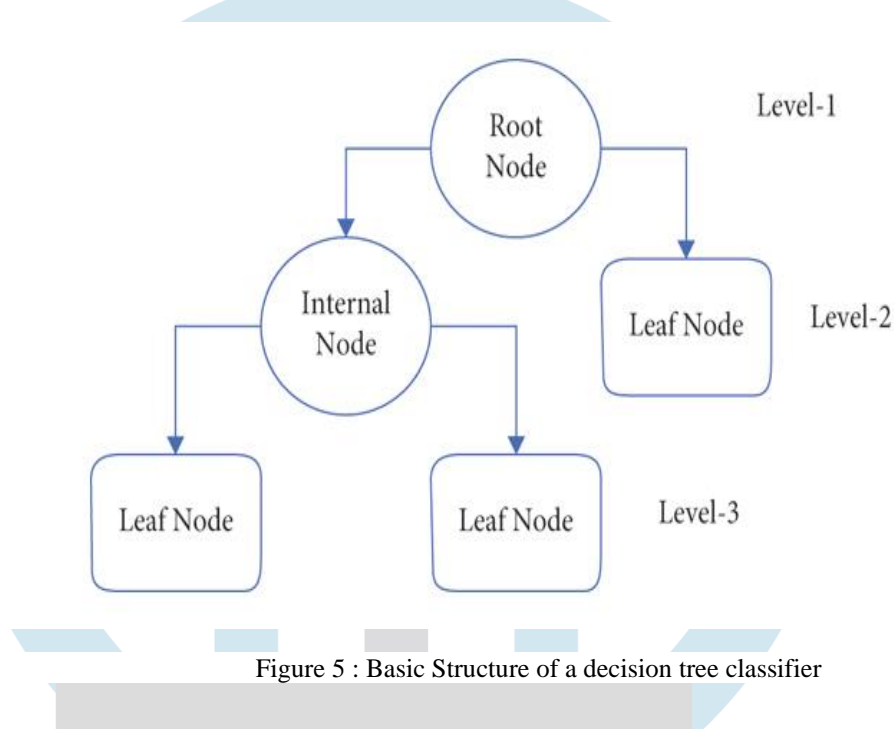


Figure 5 : Basic Structure of a decision tree classifier

DT is easy to understand because it repeats the individual level until a decision is made in the real world. This can be very helpful for decision making. Consider all possible solutions to the problem. There is no need to clear data like other methods.

#### 2.4.3. Voting Classifier

Voting is a classification model that is trained from the inputs of various models and predicts the outcome (class) based on the class most likely to be selected as the output [19]. It is used to predict voting results. A map of the voting pattern is shown in Figure 6.

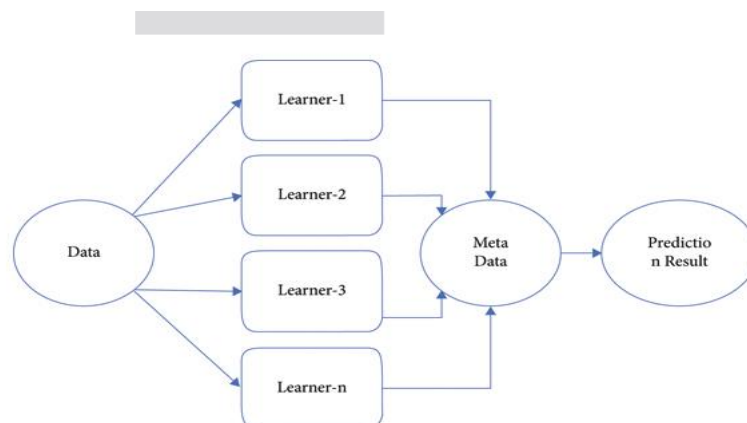


Figure 6 : Flowchart of a voting classifier

The survey shows the process we will use to compare the various models studied. There are two voting methods as follows: (i) Soft voting: In this phase, the estimated result of each model is summed up and averaged. The category with the highest value is considered the winner and its content is output. While this may seem like a fair and reasonable idea, it is only recommended if individual classes are properly evaluated. This is similar to calculating the weighted average of a set of numbers, except that each sample contributes the same amount to the final output vector. (ii) Difficulty voting: This stage combines the classification of all the different models and gives the final output value based on the value type of the output. This method is similar to calculating the arithmetic sum of numbers, as certain values associated with each sample are ignored. Evaluate the output of each model separately.

#### 2.4.4. Logistic Regression

The flowchart of the logistic regression model is shown in Figure 7. Among supervised learning, LR is one of the most widely used ML algorithms [20]. It is an estimation method that uses a series of degrees of freedom to estimate the variance of a variable.

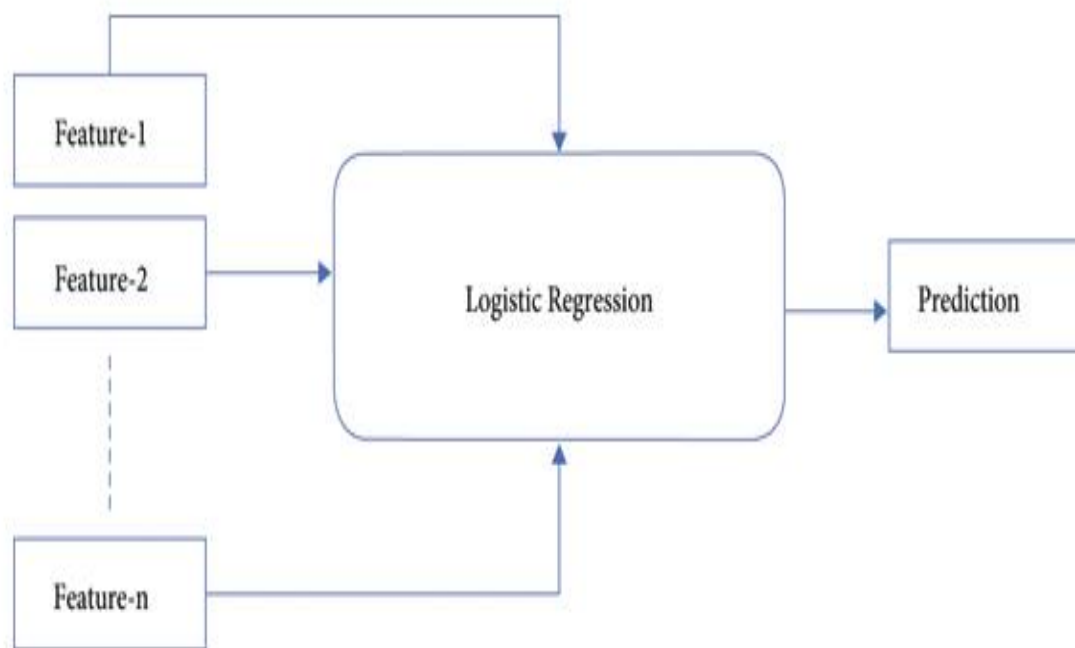


Figure 7 : Structure of a logistic regression classifier

Estimate the output of the categorical dependent variable using logistic regression. Therefore, the output must be discrete or categorical in nature. True or false, 0 or 1, true or false, etc. but a value between 0 and 1 is given. Logistic regression and linear regression are used in similar ways. The classification problem is solved with LR, and the regression problem is solved with linear regression. Instead of linear regression, we use the sigmoid logistic function to estimate two maximums (0 or 1).

#### 2.5. Evaluation Matrix

Figure 8 shows the confusion matrix or evaluation matrix. A confusion matrix is a tool for evaluating the performance of machine learning classification algorithms. A confusion matrix was used to evaluate the performance of each design. The confusion matrix shows how well our model was predicted and how wrongly predicted. False positives and false negatives are given for incorrectly estimating values, while true positives and true negatives are given to correct the need. After inserting all the predictions into the matrix, evaluate the model's performance using its accuracy, true return trade-off, and AUC.

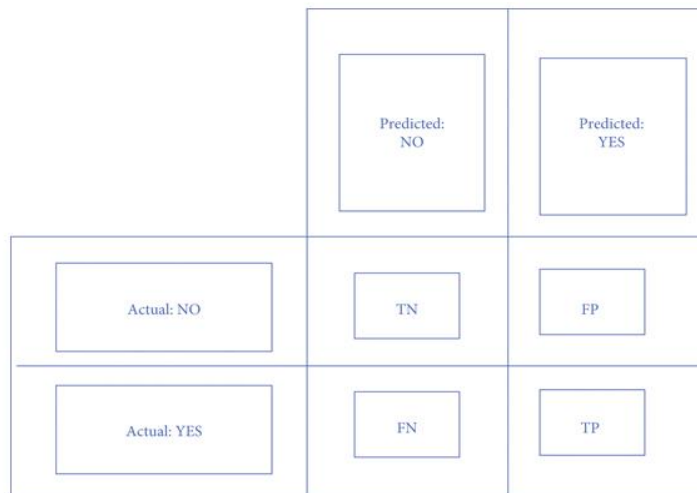


Figure 8 : Block diagram of confusion matrix

### 3 RESULT ANALYSIS

This section examines the scope of the model, model predictions, findings and final results.

#### 3.1. Data Visualization

The histogram shows the reproducible distribution for infinite classes. It is based on a square area with a class boundary proportional to the frequency of the class. The entire square shape is connected, and the bottom is the middle of the class boundary. The squareness of the neck is proportional to the relative class frequency and the velocity of the different classes. Figure 9 shows the distribution of data for sex, age, blood pressure, heart disease, marital status, mean blood glucose, and weight. For the gender attribute, 0 means male, 1 female. More women than men were tested during this operation. However, when we look at the age distribution, it is clear that the average age of the sample is around 40 and it can go up to 60. For high blood pressure, 0 means no and 1 means yes. IT. All healthy individuals without a history of cardiovascular disease were included in this database. Regarding BMI and mean blood glucose, Figure 10 shows the relationship between positive behavior and goals.

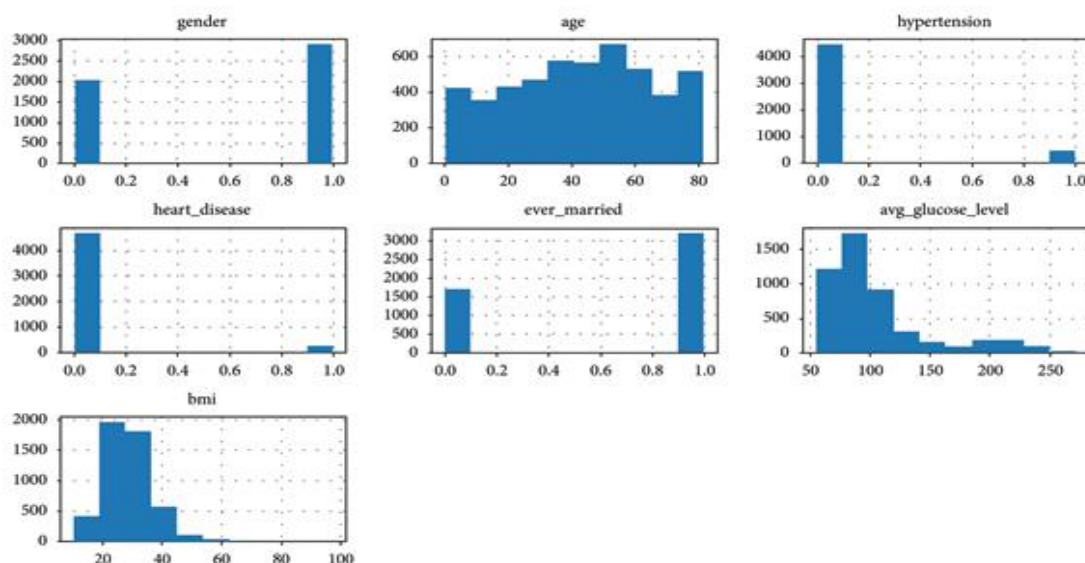


Figure 9 : Histogram of sum important features of the dataset (16)

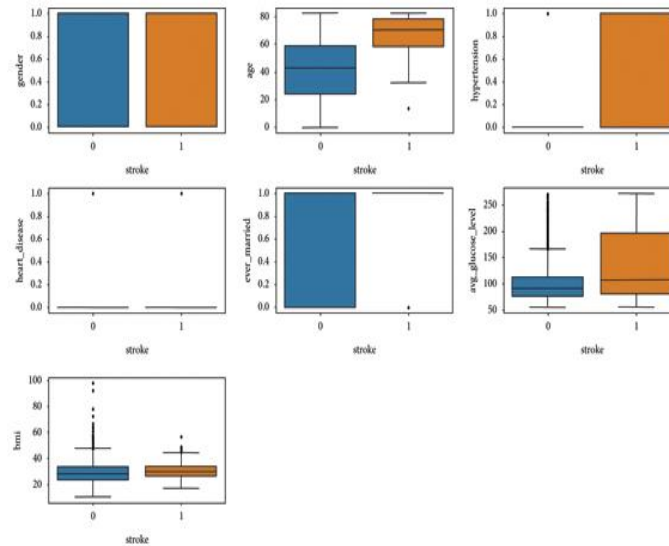


Figure 10 : Relationship between some important features with the target feature (16)

Figure 10 shows the relationship between gender and stroke, age and stroke, high blood pressure and stroke, heart disease and stroke, never married and stroke, average diabetes stage and stroke, and BMI and stroke.

### 3.2. Visualization of Feature Selection

The feature selection process is shown in Figure 11. Feature selection helps to understand how features are related to each other.

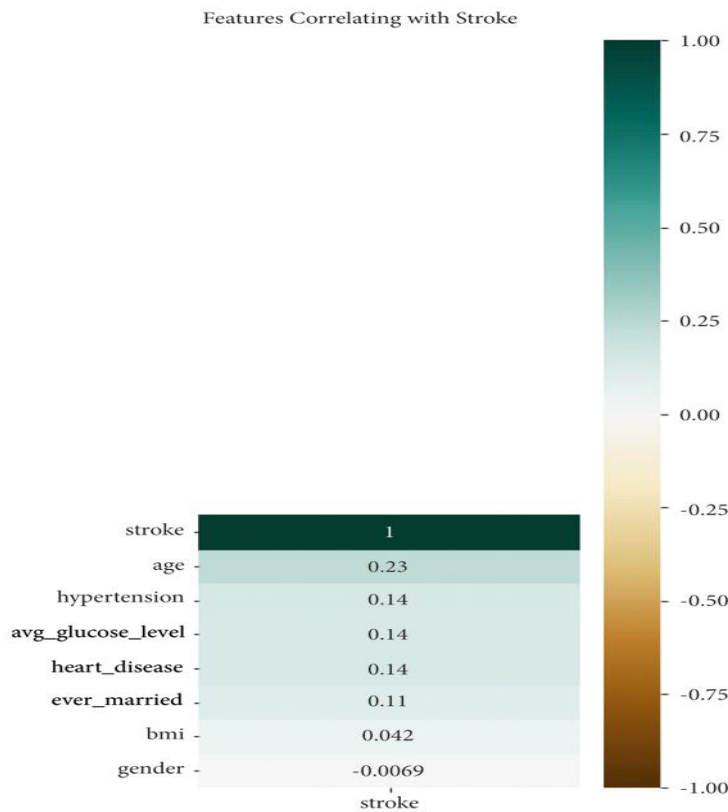


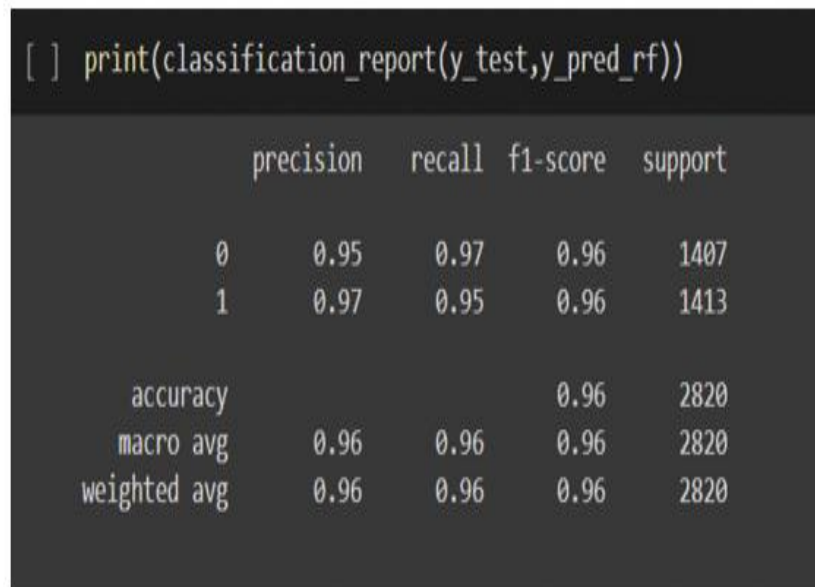
Figure 11 : Features correlation with stroke (16)

Figure 11 shows a positive association of age, hypertension, avg\_glucose\_level, heart disease, so far, and BMI with target. However, gender was associated with stroke.

### 3.3. Evaluation of the Model

#### 3.3.1. Random Forest (RF)

Figure 12 shows the distribution map for the RF model.

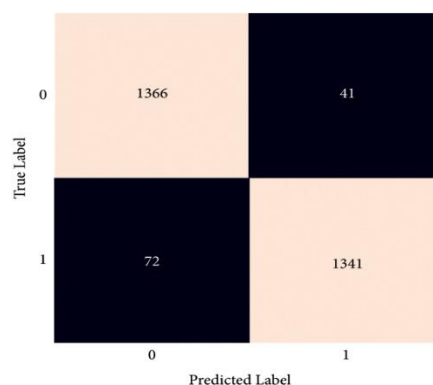


```
[ ] print(classification_report(y_test,y_pred_rf))
```

	precision	recall	f1-score	support
0	0.95	0.97	0.96	1407
1	0.97	0.95	0.96	1413
accuracy			0.96	2820
macro avg	0.96	0.96	0.96	2820
weighted avg	0.96	0.96	0.96	2820

Figure 12 : Classification report of random forest (17)

In this case, the total F1 gain is 96%. The F1 score of the healthy person was 96%, and that of the paralyzed patient was 96%. The model achieves the most accurate results after optimization. Before correction, the model is 92% accurate. Figure 13 shows the prediction of the random forest model. The prediction results and computational performance of the model are shown in the confusion matrix. 2707 correct guesses and 113 wrong guesses.



True Label	Predicted Label	
	0	1
0	1366	41
1	72	1341

Figure 13 : Confusion matrix of random forest (17)

#### 3.3.2. Decision Tree

The distribution map for the decision tree distribution is shown in Figure 14.

```
[ ] y_pred_Clf = Clf.predict(X_test_s)
    print(classification_report(y_test, y_pred_Clf))
```

	precision	recall	f1-score	support
0	0.95	0.94	0.94	1407
1	0.94	0.95	0.95	1413
accuracy			0.94	2820
macro avg	0.94	0.94	0.94	2820
weighted avg	0.94	0.94	0.94	2820

Figure 14 : Classification report of decision tree (18)

The final F1 score in this example is 94%. The F1 score was 94% in healthy subjects and 95% in paralyzed subjects. Additionally, decision and recovery are shown in Figure 14. A good decision tree model is also used. But after the fix the value is not good. Figure 15 shows the estimation of the DT model. 2664 correct guesses and 156 wrong guesses.

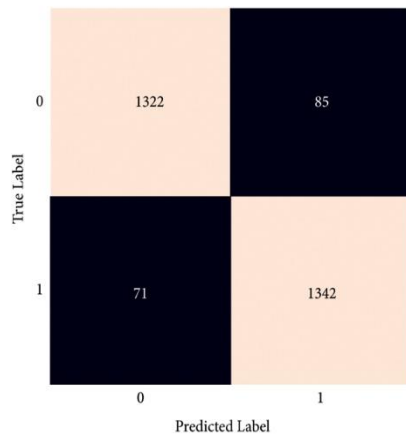


Figure 15 : Confusion matrix of a decision tree (18)

### 3.3.3. Voting Classifier

Distribution map for voters is shown in figure 16.

```
y_pred_VC = VC.predict(X_test_s)
print(classification_report(y_test, y_pred_VC))
```

	precision	recall	f1-score	support
0	0.91	0.91	0.91	1407
1	0.91	0.91	0.91	1413
accuracy			0.91	2820
macro avg	0.91	0.91	0.91	2820
weighted avg	0.91	0.91	0.91	2820

Figure 16 : Classification report of a voting classifier (19)

F1's overall score in this regard is 91%. Individual F1 scores were 91% for healthy subjects and 91% for paralyzed subjects. Also, the decision and the response are shown in Figure 16. The model achieves 91% accuracy without any optimization. Estimates made by interviewers are shown in figure 17. The total number of correct guesses is 2565, and the total number of wrong guesses is 255.

True Label	0	1280	127
	1	128	1285
		0	1
		Predicted Label	

Figure 17 : Confusion matrix of a voting classifier (19)

## 4 CONCLUSION

Stroke is a life-threatening condition that must be treated as soon as possible to prevent further complications. The development of machine learning models can help detect stroke early and reduce its later serious consequences. This study explores the effectiveness of various machine learning algorithms in predicting stroke risk based on a large number of physical variables. Arbitrary Woodland Classification outflanks other exploratory strategies with 96% classification exactness. The model can be powered by larger data and machine learning models such as AdaBoost, SVM, and Bagging. This will increase the reliability of the framework and framing. In exchange for providing some basic information, machine learning architecture can help citizens determine the likelihood of an adult having a stroke. Ideally, it will help patients get early treatment for stroke and rebuild their lives after a stroke.

### Data Availability

The data used to support the findings of this study are available online: <https://www.kaggle.com/fedesoriano/Stroke-prediction-dataset>

### REFERENCES

1. "Healthline Stroke Concept" [Online]. Muaj: <https://www.cdc.gov/Stroke/index.Um>.
2. "Stroke Statistics from Centers for Disease Control and Prevention" [Online]. Muaj: <https://www.cdc.gov/stroke/facts.htm>. Tshooj
3. S. Govindarajan, R.K. Soundarapandian, A.H. Gandomi, R. Patan, Note: Jayaraman thiab R. Manikandan, "Stroke Disease Classification Using Machine Learning Algorithms", Neural Computing and Applications, vol. 32. Tsis muaj. 3, issue 817–828, 2020.
4. L. Amini, R. Azapa Zhou, M.T. Farzadfar et al, "Stroke Decision and Management through Data Mining", International Journal of Preventive Medicine, Vol. 4. No. 2 H. S245–S249, 2013.
5. SM Riza, M.M. Rahman and S. A. Mamun, "A new approach to communication - collaboration between tool xml and big data", Proceedings of the International Electronics and ICT Conference, Vol. 1-5, Mirpur, Dhaka, April 2014.
6. ib.W. Chiu, "Using Artificial Neural Networks to Predict Outcomes in Ischemic Stroke Patients After Intravenous Thrombolysis," Health Technology and Informatics Research, vol. 202, p. 115-118, 2014.
7. p.S. Cheon, J. Kim and J. Lim, "Using deep learning to predict mortality in stroke patients," International Journal of Environmental Research and Public Health, vol. 16. No. November 2019.
8. M.S. Zulfikar, N. Kabir, A. A. Biswas, P. Chakraborty ve M. M. Rahman, "Siv Machine Learning Methods for Predicting Student Performance at Private Universities in Bangladesh", International Journal of Advanced Computer Science and Applications, vol. 11. Tsis muaj. Peb Hlis 2020
9. P. Rahman, T. Sharma, S. Reza, M. Rahman, M. Kaiser, thiab Nurjahan, "Pso-nf-based vertical migration decisions for ubiquitous heterogeneous wireless networks (uhwn)", International Computing Intelligence 2016 Symposium Proceedings (IWCI), p. ua. Choudhary, "Stroke Prediction Using Artificial Intelligence," in Proceedings of the 2017

10. 8th Annual Industrial Automation and Mechatronics Engineering Conference (IEMECON), p. 158–161, Bangkok, Thailand, Lub Yim Hli 2017.
11. C.-L."Automated Early Ischemic Stroke Detection System Using CNN Deep Learning Algorithm", 2017 IEEE 8th International Conference on Conscious Science and Technology (iCAST), p. 368-372, Taichung, Taiwan, November 2017.
12. ib. S.-F.Song by C.-Y. Thank you Y.-H. Gao Yang et al."Development of a stroke index from controlled data is possible using data mining techniques", Journal of Clinical Epidemiology, vol. 68, no. 11, p. 1292–1300, 2015. 4 444
13. M. Monteiro, A.C. Fonseca, A.T.Freitas et al, "Using Machine Learning to Improve Prediction of Functional Outcomes in Stroke Patients," IEEE/ACM Proceedings of Applied Bioinformatics and Bioinformatics, vol. 15. No. 6, p. 1953-1959, 2018.
14. T. Kansadub, S. Thammaboosadee, S. Kiattisin ve C.Jalayondeja, "Demographic Data-Based Stroke Risk Estimation Model", 2015 8th International Conference on Biomedical Engineering (BMEiCON), p. 1-3, Pattaya, Thailand, Kaum Ib Hlis 2015.
15. S. Y.Adam, A. Yousif and M. B. Bashir, "Classification of Ischemic Stroke Using a Machine Learning Algorithm", International Journal of Computer Applications, vol. 149, no.pm 10. 26-31, 2016.
16. "Stroke Prediction Dataset", [Online]. siv tau:.
17. "Scikit-learn Documentation for Random Forest Classification", org.
18. "Scikit-learn Documentation for Decision Tree Classification", org.. [Online
19. "Logistic Regression hauv Machine Learning", [online].available: Section
20. G. Celasia and G.L.A. Kumari, "Exploring the Performance of Stroke Prediction Using ML Classification Algorithms", International Journal of Advanced Computer Science and Applications, vol. 12. No. 6, issue 539-545, 2021.

