DOI: 10.24996/ijs.2023.64.12.39

ISSN: 0067-2904

Machine Learning Prediction of Brain Stroke at an Early Stage

Mohammad Abood Kadhim, Abdulkareem Merhej Radhi
Department of Computer Science, Al-Nahrain University, Baghdad, Iraq

Received: 15/11/2022 Accepted: 6/3/2023 Published: 30/12/2023

ABSTRACT

The healthcare sector has traditionally been an early adopter of technological progress, gaining significant advantages, particularly in machine learning applications such as disease prediction. One of the most important diseases is stroke. Early detection of a brain stroke is exceptionally critical to saving human lives. A brain stroke is a condition that happens when the blood flow to the brain is disturbed or reduced, leading brain cells to die and resulting in impairment or death. Furthermore, the World Health Organization (WHO) classifies brain stroke as the world's second-deadliest disease. Brain stroke is still an essential factor in the healthcare sector. Controlling the risk of a brain stroke is important for the survival of patients. In this context, machine learning is used in various health-related fields, especially "brain stroke." To that end, an automated model for recognizing and providing helpful information for brain stroke prediction was created. It can predict brain strokes with high accuracy in the early stages. The proposed model aims to examine the patient for effective decision-making. This research study employed a freely accessible dataset and a mix of machine learning methods such as random forest, logistic regression, and decision trees. Furthermore, the Synthetic Minority Over Sampling Technique (SMOTE) was implemented to handle unbalanced data. The result shows a high accuracy of 99% in predicting a brain stroke.

Keywords: Machine learning, Decision Tree, Logistic Regression, Random Forest, SMOTE**.**

التنبؤ المبكر بالجلطة الدماغية باستعمال التعلم الآلي

محمد عبود كاظم, عبد الكريم مرهج راضي
قسم علوم الحاسوب، جامعة النهرين، بغداد، العراق

الخلاصة

لطالما كان قطاع الرعاية الصحية يستعمل في وقت مبكر التكنولوجيا وقد حقق مزايا كبيرة، لا سيما في مجال التعلم الآلي مثل التنبؤ بالأمراض. تعتبر السكتة الدماغية من أهم الأمراض. يعد الاكتشاف المبكر لسكتة دماغية أمرًا بالغ الأهمية لإنقاذ حياة البشر. السكتة الدماغية هي حالة تحدث عند اضطراب أو انخفاض تدفق الدم إلى الدماغ، مما يؤدي إلى موت خلايا الدماغ وتؤدي إلى ضعف أو موت. علاوة على ذلك، صنفت منظمة الصحة العالمية السكتة الدماغية على أنها ثاني أكثر الأمراض فتكًا في العالم. لا تزال السكتة الدماغية عاملاً أساسياً في قطاع الرعاية الصحية. السيطرة على مخاطر الإصابة بسكتة دماغية مهمة لبقاء المرضى على قيد الحياة. في هذا السياق، يتم استعمال التعلم الآلي في مجموعة متنوعة من المجالات المتعلقة بالصحة وخاصة "السكتة الدماغية". تحقيقا لهذه الغاية، تم تصميم نموذج آلي للتعرف وتقديم معلومات مفيدة للتنبؤ بسكتة دماغية.

*Email: mohammed.abood.cs2020@ced.nahrainuniv.edu.iq

يمكنه التنبؤ بسكتة دماغية بدقة عالية في المراحل المبكرة. يهدف النموذج المقترح إلى فحص المريض لاتخاذ قرارات فعالة. استعملت هذه الدراسة البحثية مجموعة بيانات يمكن الوصول إليها مجانًا ومزيجًا من أساليب التعلم الآلي مثل الانحدار اللوجستي، وشجرة القرار ، خوارزمية الغابة العشوائية. علاوة على ذلك، تم تنفيذ تقنية الأقلية الاصطناعية فوق أخذ العينات للتعامل مع البيانات غير المتوازنة. تظهر النتيجة دقة عالية تصل إلى 99٪ للتنبؤ بالسكتة الدماغية.

1. INTRODUCTION

The lives of people all over the world have dramatically improved because of tremendous healthcare technology progress. The development of technologies has greatly enhanced healthcare-related resources. The quality of life has significantly increased thanks to technology, which also provides patients with an effective diagnosis of diseases.

Health data requirements, as well as the data collection process, which may be observed worldwide, are changing. Many people face difficulties finding health information online on ailments, diagnoses, and drugs, which takes time and money [1]. A stroke, sometimes referred to as a "brain circulatory disease," is a disease that happens when the brain's blood supply is stopped or diminished. And if the blood flow to that brain area is inadequate, the cell will not receive appropriate oxygen or nutrients, resulting in death [2]. Stroke is the second most common cause of mortality worldwide. According to the World Health Organization, 15 million people worldwide experience a stroke every year, with low- and middle-income countries accounting for 87% of stroke fatalities [3].

Stroke patients frequently worry about an unplanned attack. It is essential to know the more straightforward way to make predictions since all computations are performed using a computer, which causes these sudden attacks. Due to these characteristics, machine learning techniques are helpful for the medical field since they can more accurately estimate a patient's likelihood of a stroke [4]. ML has been shown to increase optimization and categorization in creating intelligent systems. Machine learning techniques may be used to diagnose and predict stroke in patients directly, and they can work automatically using patient records. Whenever a stroke is predicted, the person can receive healthcare therapy faster without as much harm. Early stroke detection can reduce healthcare expenditures since it costs less and saves the costs of many treatments [5]. Prediction can be used to create a model to predict data that has never been categorized before so that it may be used to classify new data. Many algorithms, such as DT, RF, and LR, can be used in machine learning classification [6].

This study shows the proposed model that aims to create and use an automated model to accurately predict brain stroke in its early stages. This study employed a combination of ML algorithms (random forest, decision tree, and logistic regression). The brain stroke datasets were obtained from the Kaggle data source. In addition, SMOTE technology is used to handle unbalanced data. Most trials had a 99% accuracy rate, which was deemed pretty good. The algorithms employed in this study are substantially greater and more accurate than those used in previous research, suggesting that the algorithms utilized in this study are more trustworthy. The research study has the following contributions:

1.Our assessments of the rebalancing SMOTE techniques and our discovery of the most accurate classifier for predicting the risk of diseases.
2.The model can give quick and accurate diagnoses to people who need help immediately.
3.The proposed model produced excellent accuracy when compared to earlier studies.

In the remainder of the paper, an overview of the literature is presented in Section 2. sections

3 and 4 hold methodology and research results, respectively. Discussions are presented in Section 5. Finally, based on the total of the experiments, a conclusion is given in Section 6.

2. RELATED WORK

Based on the literature, several options for predicting and monitoring healthcare facilities utilize open data sources.

Researchers in [7] gathered data for 507 patients in research to identify stroke disease using a text mining combination and a machine learning classifier. They researched multiple machine learning algorithms using ANN for training purposes, and the SGD algorithm provided the most significant value (95%).

The authors of [8], Decision Tree, Random Forest, and Multi-Layer Perceptron, with just a few small changes, are used to predict the stroke. The accuracy achieved for the three techniques mentioned above was quite close. Random Forest and Decision Tree had computed accuracy of 74.53% and 74.31%, respectively, and the multi-layer perceptron had a calculated accuracy of 75.02%. According to this research, the multi-layer perceptron is more accurate than the other two approaches. To calculate the accuracy score, the confusion matrix was the only parameter used.

In [9], the authors investigated various data mining categorization approaches to predict stroke. The data was gathered from Saudi Arabia's hospitals. Different classification and prediction techniques, including C4.5, JRip, and multi-layer perceptrons, are used. By using these techniques, the best accuracy achieved was about 95%. The learning and prediction times are longer, even after the paper asserts a 95% accuracy rate, as the scientists used a variety of complex techniques.

In [10], his research shows the possibility of having a stroke may be predicted using three different machine learning techniques. These ML techniques include neural networks, decision trees, and Naive Bayes. In the mentioned research, the decision tree performed with the highest accuracy (around 75%) compared to other algorithms. Consequently, depending on the confusion matrix's findings, this model could not be used in real-world situations.

In [11], a study is being done to predict stroke risk. Because stroke is the second-leading cause of death globally, finding it early on would assist the patient in having a better and less detrimental outcome. The most fabulous accuracy rating is 94.26% for a random forest, 92.917% for a decision tree, and 89.4% for a Naive Bayes model. These three models have very excellent performances.

In [12], they implemented a technique known as Synthetic Minority Over-Sampling Technique (SMOTE) to solve the class imbalance issue. Many machine learning algorithms have been used to predict stroke, including Naive Bayes, Decision Trees, Ensemble Learning, Random Forest, and Artificial Neural Network Algorithms. A random forest and an artificial neural network have the highest accuracy ratings of 94.22% and an accuracy score of 0.84%, respectively.

In [2] for early stroke prediction, different classification algorithms, including stochastic multi-layer perceptrons, gradient descent (GD), LR, AdaBoost, Gaussian, quadratic discriminant analysis, KNN, decision trees, gradient boosting, and XGBoost, have been trained using the high feature attributes. The results of the fundamental classifiers are then blended using the weighted voting approach to achieve the highest degree of accuracy. Additionally, the suggested study has a 97 percent prediction rate.

In [13], the authors review patient records for efficient decision-making. This literature review used a patient's publicly available data set and a variety of ML algorithms, including a k-nearest neighbor algorithm, support vector machines, decision trees, and logistic regression. The simulation results show that the decision tree technique has the highest accuracy of 89%.

In [14], the authors utilize six machine learning classification methods: logistic regression, random forest, KNN, support vector machine, decision tree, and Gaussian Naive Bayes. In addition, a thorough comparison of machine learning algorithms was conducted. The researchers claim that machine learning approaches combined with data balancing techniques are helpful tools for stroke prediction when the data is skewed. As a result, the model experienced SMOTE. As a result, it is expected that Random Forest will outperform the competition with a stroke prediction accuracy of 96%.

In [15], the severity of the stroke can be lowered by being aware of as many stroke warning signs as possible. Many machine learning (ML) models have been developed to predict the probability of a brain stroke. This study exploits a variety of physiological indicators and, using four different machine learning techniques (LR, DT, RF, and Voting Classifier), trains to provide appropriate predictions. The best algorithm for this study was Random Forest, which had an accuracy of almost 96%.

In [16], only a few studies have examined the risk of a brain stroke, while most studies have concentrated on predicting the likelihood of a heart attack. In light of this, several machine learning models are being created to predict the possibility of a stroke. In the study mentioned above, they used a variety of physiological factors and five different models for accurate prediction with the help of machine learning techniques, i.e., LR, DT, RF, KNN, SVM, and Naive Bayes. The algorithm that did the best in this challenge was Naive Bayes, which had an accuracy of around 82%.

In [17], Apache Spark, a big data platform, is used to carry out this study. Apache Spark is one of the most widely used extensive data systems, and it comes with the MLlib package. Machine learning algorithms are provided through MLlib, an API linked with Spark. The stroke prediction model was built using four machine learning classification algorithms: decision tree, support vector machine, random forest classifier, and logistic regression. Hyperparameter tweaking and cross-validation were used with machine learning techniques to improve outcomes. The findings revealed that the Random Forest Classifier had the highest accuracy, with a score of 90%.

3. Methodology

This study aims to create a model that can predict brain strokes. There is a portion in this section that includes data collection, dataset description, data pre-processing, feature engineering, and applicable machine learning algorithms. It also consists of a block diagram, evaluation matrices, the study's procedure, and the methodology, describing our approach shown in Fig. 1. Several models are utilized to evaluate the performance and validation of the created model using the brain stroke training data set with its 11 unique features. Confusion matrix measurements are researched to estimate outcomes, and several methodologies are used to assess accuracy. We utilized a train-test split validation approach to split the dataset into two sections, utilizing 70–80% for training machine learning algorithms and a 30–20% assessment ratio. We completed 80% of the training data and 20% of the testing and evaluation due to splitting the dataset.

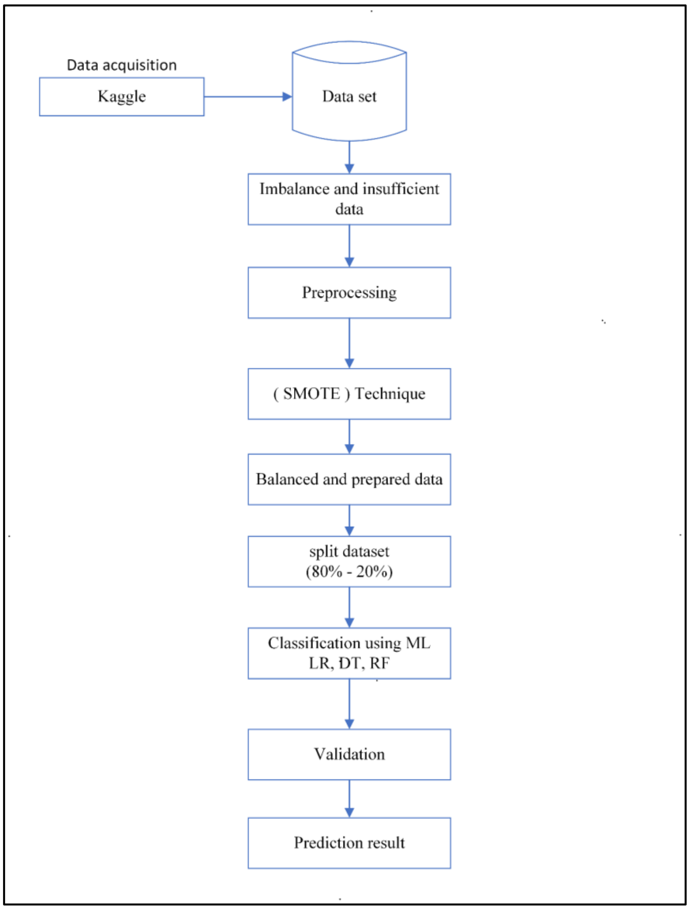

Data acquisition
Kaggle
Data set
Imbalance and insufficient data
Preprocessing
(SMOTE) Technique
Balanced and prepared data
split dataset (80% - 20%)
Classification using ML LR, DT, RF
Validation
Prediction result

Figure 1: The proposed model diagram.

3.1. Dataset

The predictive analysis in this study is performed utilizing the Healthcare Dataset Stroke from Kaggle [18]. The dataset includes 5110 health samples with 12 characteristics listed with their descriptions in Table 1. As demonstrated in Figure 2, the dataset is significantly unbalanced, with 249 out of 5110 individuals experiencing a stroke. The dataset contains 201 missing BMI values, and the smoking status of over 30 percent of the people is unknown, making the model much more complex to fit into the research.

Table 1: Nominal Attributes Explanation

Parameter Names	Values Type	Definition
Age	Integer	Patient Age
gender	String literal (Male, Female, Other)	Tells the patient's gender
hypertension	Integer (0, 1)	Determines if a patient has high blood pressure or not.
id	Integer (number)	A numeric value that cannot be repeated
work_type	String literal (Never_worked, children, Private, Self-employed, Govt_job)	provides several job types.
avg_glucose_level	Float point number	Provides the value of the blood's average glucose level.
ever_married	String literal (No, Yes)	give the patient's state whether he is married or not
heart_disease	Integer (0,1)	identifies the presence or absence of cardiac disease in the patient

Residence_type	String literal (Rural, Urban)	Provides the type of patient's residence
smoking_status	String literal (never smoked, smokes, unknown formerly smoked)	It gives the patient a smoking state
BMI	Float point number	It gives the Body Mass Index value of the patient's
stroke	Integer(0,1)	the outcome of the Stroke gives the presence or absence of disease in the patient

3.2. Data pre-processing

Data preprocessing is crucial to obtaining the best accuracy. One of the most important steps is to prepare the data for analysis by removing unwanted noise and outliers from the sample in the dataset that may cause a divergence from effective training. All issues that hinder the model from operating as effectively as feasible are resolved at this stage. After collecting the appropriate dataset [15], the data must be cleaned and reviewed to ensure it is prepared for model construction. The dataset utilized includes 12 features, as shown in Table 1. The variable "id" is first omitted because its inclusion in model creation has no impact. The dataset is then checked for null values, and any found are filled in. In this instance, the mean of the sample in this column is used to fill in the null values in the BMI column [19]. Over 30% of the population's smoking status is unknown. Depending on the samples' diagnostic categories, mean imputation is utilized to fill in the unknown data [20].

3.2.1. Label Encoding

Label encoding converts the dataset's string literals into integer values that the computer can interpret [21]. Since computers are often programmed using numbers, the strings must be converted into integers. Five columns in the acquired dataset have strings as their data type. All the strings are encoded when you do label encoding, which turns the entire dataset into a collection of integers.

3.3. Handling Data Imbalance

There are several strategies for handling unbalanced data, which may be divided into data and level algorithms. The way the analyst uses contact with data as a preprocessor to edit the datasets, rebalance the imbalanced data, and reduce noise between two classes is the favored among those strategies [22]. The outcomes will be erroneous, and the predictions will be worthless, if such uneven data is not appropriately handled. As a result, the imbalanced data must be dealt with first to produce an efficient model. This was accomplished using the SMOTE approach, as shown in Fig. 2 [23].

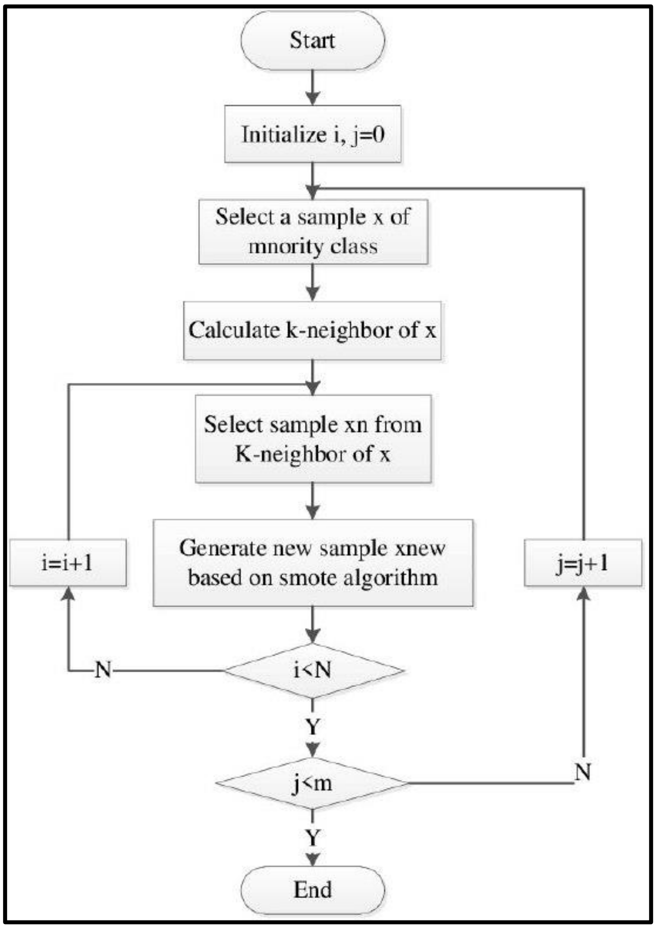

Start
Initialize i, j=0
Select a sample x of mnority class
Calculate k-neighbor of x
Select sample xn from K-neighbor of x
Generate new sample xnew based on smote algorithm
i=i+1
j=j+1
i<N
N
Y
j<m
N
Y
End

Figure 2: SMOTE approach

The dataset used to detect strokes is highly unbalanced. There are 5110 rows in the dataset, with 249 samples showing the chance of having a stroke and 4861 samples showing the absence of disease. Fig. (3a) shows the imbalanced dataset and how using such information to train a machine learning algorithm can lead to less accuracy. Still, various accuracy metrics like recall and precision are insufficient. The balance column output of the data set is depicted in Fig. (3b). There are 9722 rows in the dataset after using SMOTE technology, including 4861 rows that show the absence of a stroke and 4861 rows that indicate the presence of a stroke.

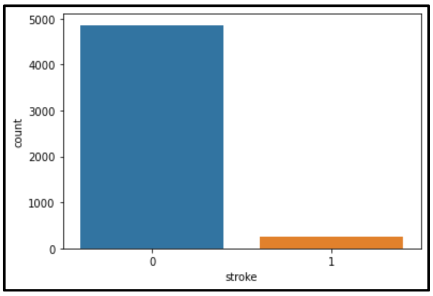


Figure (3a): imbalanced data

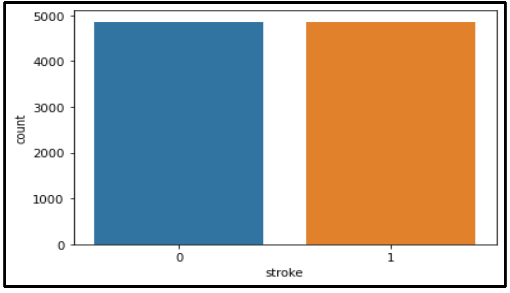


Figure (3b): balanced data

3.4. Partitioning of Data

Splitting is dividing the dataset into two groups, the first of which will be used as training data and the second as testing data. A model is created and demonstrated using training data and then evaluated using testing data [17]. Based on this study, the researcher divided the data into an 80:20 ratio, with 80% used as training data and 20% used as testing data.

3.5. Predictive Analysis

The algorithm proposed for the second most common disease in the healthcare field is stroke, and its prevalence is rising year over year. The following three commonly employed machine learning algorithms for predicting brain stroke were examined in this research: The methods helpful in identifying binary dependent variables were logistic regression, random forest, and decision trees.

1. ***Logistic Regression***: It is an effective and well-known technique in the fields of statistics and healthcare. A classification result and an explanatory variable are compared using LR. When the target variable has two categories, such as active or inactive and healthy or unhealthy, it is possible to apply this prediction model [24]. The goal of our logistic regression technique is to determine the best fit for describing the connection between our target variable and the predictive factors that is diagnostically plausible [25]. Logistic regression steps are shown in the algorithm (1).

```
algorithm (1): Logistic Regression
Input: training data
Output: A class label for feature vector
1.    For i ← 1 to k
2.    For each training data instance di
3.    Set the target value for regression to
          $Zi = \frac{yi - P(1|dj)}{[P(1|dj).(1 - P(1|dj))]}$
4.    Initialize the weight of instance dj to P(1|dj). (1-p(1|dj))
5.    Finalize an f(j) to the data with class value (zj) and weight (wj)
Classification label decision
Assign (class label :0) if p(1|dj) > 0.5, otherwise (class label :1)
```

2. ***Decision Tree***: DT is a supervised algorithm that may be used for both classification and regression, and it is almost always employed for medical applications [26]. A decision tree is divided into two parts: leaf nodes that are labeled, and internal nodes that do not have child nodes that help make decisions. Decision trees accommodate various data formats in categorizing occurrences [27]. The decision tree steps are shown in the algorithm (2).

```
algorithm  (2): Decision Tree
Input training desserts
output desired tree
Process
step1: compute the entropy for the data set
step 2 for every attribute feature
    1.    calculates entropy for all category value
    2.    take average information entropy for the current attribute
    3.    calculate IG for the present attribute
step 3 pick the highest gain attribute
step 4 repeat until we get the tree we desired
End.
```

3. ***Random Forest:*** Multiple decision trees (DT) used independently on a randomized sample of data will make up Random Forests. These trees are created throughout the training process,

and the output for the decision trees is compiled. A voting technique decides the algorithm's final prediction [15]. Each DT in this model should choose between the two output classes of indications of stroke signs or stroke signs that were absent. The eventual prediction is made using the Random Forest method, which selects the class that receives the most votes. Random Forest steps shown in the algorithm (3)

```
algorithm (3): Random Forest
Input: Feature vector S
Output: A class label for feature vector
Begin
Step 1: Choose N features randomly from the overall M features from vector S
Step 2: Between the selected N features, compute the node d utilizing the best point for splitting
Step 3: Divide the node d into smaller nodes within the best split.
Step 4: Repeat the steps from 1 to 3 until the count of nodes is reached to Z number of nodes
Step 5: Establish the forest tree by repeating steps 1 to 4 for a C count of times to obtain the C count of trees.
Step 6: Get the testing group of features, utilize the rules belonging to every decision tree that was created randomly to forecast the output, and save the predicted outcome (goal)
Step 7: Compute the votes for every forecasted goal
Step 8: Consider the high-voted forecasted goal as the final forecasting from the random forest algorithm.
Step 9: Return the class label.
End.
```

3.6. Performance measurement

The confusion matrix, also known as an assessment matrix, is shown in Figure 4. The performance of machine learning classification algorithms can be evaluated using the confusion matrix. The efficiency of each proposed model was assessed using the confusion matrix, which shows how much the accuracy and inaccuracy rate at estimating our models are [28]. We can use the confusion matrix to calculate sensitivity, specificity, accuracy, and precision measurements. Specificity determines the aspect of exactly stated negatives, whereas sensitivity indicates the number of true positives that are precisely defined [6]. True positive and false positive (TP and FP), true negative and false negative (TN and FN), are four values that define the algorithm's accuracy or precision [28].

Accuracy: refers to how near a measured value is to the actual value, and it may be represented mathematically as:

$$\text{Accuracy} = \frac{\text{FN} + \text{TP}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \qquad \text{Eq (1).}$$

The True positive rate, recall, or sensitivity is the ratio of a perceived positive occurrence to all positive cases, such as total projected absence to complete absence, and it will be calculated as:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \qquad \text{Eq (2).}$$

True negative rate or specificity, the ratio of perceived negative samples compared to all negative samples is defined as the mean ratio of expected presence compared to the total sample with the presence of cardiovascular disease & will be calculated as:

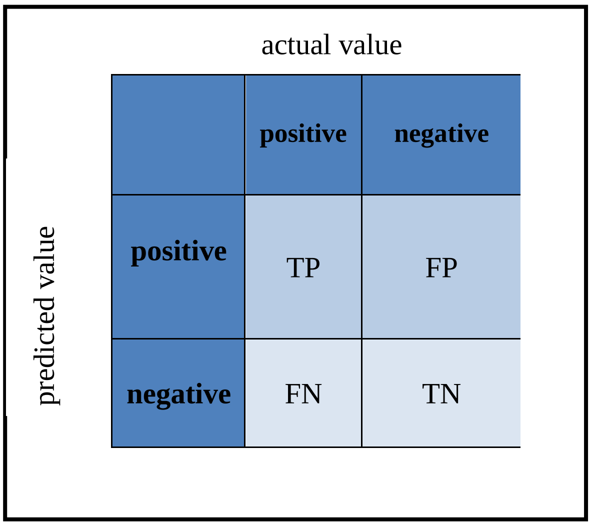
$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \qquad \text{Eq (3).}$$

Positive predictive value or precision is defined as the ratio of genuine positives (absence recognized as absence) to all positive cases (absence determined in all samples).
& will be calculated as:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \qquad \text{Eq (4).}$$

FP, TP, FN, and TN have numerical values that are defined as:

- FP = outcome where the number of people who have had a heart attack that the model incorrectly predicts have never had a heart attack.
- TP = the total number of persons who have a heart attack.
- FN = the number of persons who have never had a heart attack, and the model incorrectly predicts persons who have had a heart attack.
- TN = Total number of persons with heart disease and no heart disease.



actual value

predicted value

	positive	negative
positive	TP	FP
negative	FN	TN

Figure 4: confusion matrix diagram

4. Result

The performance and validity of the developed model are evaluated using various ML techniques using the stroke disease training dataset with one unique feature. Confusion matrix measurements were used to estimate results, and different methods were used to evaluate the accuracy rate. A train-test split validation technique was used to split the dataset into a ratio of 80% for training and 20% for testing and assessment. The models were constructed using the default parameters, and the performance was evaluated based on the unseen test. This section employs a two-phase result. The imbalanced dataset is shown in the first phase, and the balanced dataset uses the SMOTE method. Table (2) summarizes the results of ML algorithms regarding accuracy score and performance measurement (precision, recall, and f1-score). Results of the selected classification techniques, including logistic regression, decision tree, and random forest

Table 2: Ml Algorithms' Performance for imbalance data

algorithms	Accuracy	precision	recall	Specificity	f1_score
LR	0.95	0.0	0.0	1.0	0.0
DT	0.91	0.13	0.16	0.94	0.19
RF	0.95	0.0	0.0	0.99	0.0

Table (3) shows the outcomes of different algorithms applied to the balanced dataset. The best accuracy of LR came out at 0.76; the quality of a positive prediction made by the model (precision) is 0.74; the recall is 0.79; the specificity is 0.72; and the f1_score is 0.77. The second algorithm DT showed that the best accuracy of 0.97 was obtained, precision was 0.96, recall was 1.0, specificity was 0.95, and f1_score was 0.96. The latest algorithm, RF, shows that the best accuracy, 0.99, was obtained, the precision was 0.98, the recall was 1.0, the specificity was 0.98, and the f1_score was 0.99. The RF has the highest classification accuracy compared to other algorithms, followed by the DT. The evaluation of the model is performed with the confusion matrix. The confusion matrix of algorithms in Figures 5 and 6 shows the confusion matrix of the logistic regression and decision tree. In contrast, Fig. 7 shows the confusion matrix of the random forest.

Table 3: Ml Algorithms' Performance with smote

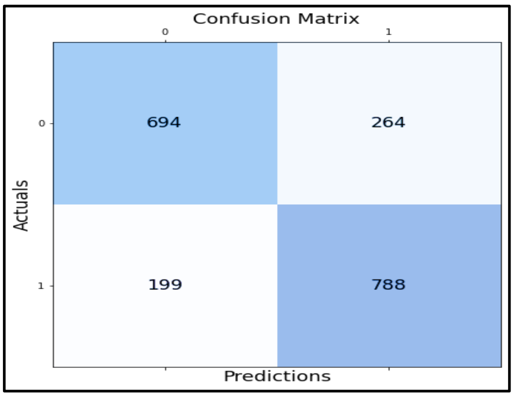
algorithms	Accuracy	precision	recall	Specificity	f1_score
LR	0.76	0.74	0.79	0.72	0.77
DT	0.97	0.96	1.0	0.95	0.96
RF	0.99	0.98	1.0	0.98	0.99

Researchers have compared this work and the results obtained with the rest of the research papers, according to Table 4. The researcher found that the designed model obtained better accuracy results than the rest of the research papers mentioned in the related works. The RF algorithm shows that the best accuracy of 0.99 was obtained, representing better accuracy.

Table 4: comparison with the previous work

Studies	Classifier	best accuracy
[2]	Weighted Voting	0.97
[11]	Random Forest	0.94
[12]	SVM	0.94
[13]	Decision Tree	0.89
[14]	Random Forest	0.96
[15]	Random Forest	0.96
[16]	Naïve Bayes	0.82
[17]	Random Forest	0.90
This work	**Random Forest**	**0.99**

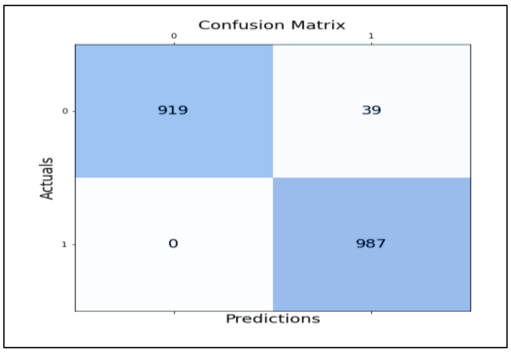
As shown in Fig. 5, the confusion matrix of the logistical algorithm, the output of the matrix is clear: the correct predictions are 1482 out of a total of 1945, as it was found that the people who predict the model have the disease. The truth is that 264 people are not infected with the disease, while the number of people predicted by the model with no disease but who have the disease is 199, as this type of prediction is the most dangerous because the patient suffers from the disease. Still, according to this algorithm, it does not carry the disease but can lead to death. Based on these results, this algorithm is considered unsuitable for this disease.


Confusion Matrix
Predictions
Actuals

	0	1
0	694	264
1	199	788

Figure 5: LR confusion matrix

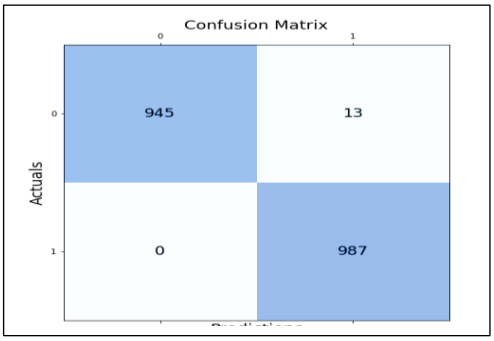
Fig. 6 shows the confusion matrix of the DT algorithm. The output of the matrix clearly shows that the correct predictions are 1906 out of a total of 1945, as it was found that the people who predicted the model had the disease. The truth is that those not infected with the disease are 39. At the same time, the number of people predicted by the model with no disease but who have the disease is 0. According to these results, this algorithm is considered suitable for this disease.


Confusion Matrix
Predictions
Actuals

	0	1
0	919	39
1	0	987

Figure 6: DT confusion matrix

As shown in Fig. 7, the confusion matrix of the RF algorithm, the output of the matrix is clear: the correct predictions are 1932 out of a total of 1945, as it was found that the people who predict the model have the disease. The truth is that those not infected with the disease are 13 At the same time, the number of people predicted by the model with no disease but who have the disease is 0. According to these results, this algorithm is considered the best for this disease.


Confusion Matrix
Actuals

	0	1
0	945	13
1	0	987

Figure 7: RF confusion matrix

5. DISCUSSION

Detecting a brain stroke at an early stage is extremely important to saving human lives. In this case, a fast and reliable judgment is essential. As stated at the outset, machine learning is a helpful assistance tool that provides the necessary data and classification to evaluate the infection based on the input data. The Random Forests technique used in this research has demonstrated that it can be used as an instrument to achieve the paper's primary goal: supporting experts (or medical professionals in general) in recognizing brain stroke and providing a helpful categorization system, regardless of whether a person is incapacitated or not. The Random Forest method, with an accuracy of 0.99, outperformed the other algorithms, followed by the Decision Tree algorithm, which had an accuracy of 0.97. Logistic regression has the lowest accuracy of all the algorithms.

6. CONCLUSION

The brain is one of the most vital organs in the human body. Any brain disease can affect the patient's life. A brain stroke is one of the most dangerous diseases that can lead to death. Stroke is the second-leading cause of death in the world, behind heart attack, among diseases that can lead to mortality. This paper presents a proposed model to design and implement an automated model to accurately predict brain stroke in its early stages. Synthetic Minority Over Sampling Technique (SMOTE) is used to handle imbalanced datasets and affect the ML performance results, and when using SMOTE, the result is better than the imbalanced dataset. Machine learning algorithms like Random Forest, Logistic Regression, and Decision Tree were used to achieve the highest desired efficiency of the model. The RF algorithm shows that the best accuracy of 0.99 was obtained, representing better accuracy. The early detection of stroke allows patients to receive better, less harmful treatment. Medical professionals and specialists in this sector can consistently deliver the finest care. Furthermore, our model might predict the probability of stroke based on comprehensive data, reducing expenditures. This model can also aid decision-making by reducing overcrowding, improving resource use, and reducing wasteful labor expenditures.

REFERENCES

[1] K. D. 1K. Venkatesh1, M. Prathyusha3, C.H. Naveen Teja4, "Identification of Disease Prediction Based on Symptoms Using Machine Learning," *JAC : A Journal Of Composition Theory*, vol. XIV, no. VI, June 2021. [Online]. Available: http://www.jctjournal.com/gallery/10-june2021.pdf.

[2] M. U. Emon, M. S. Keya, T. I. Meghla, M. M. Rahman, M. S. Al Mamun, and M. S. Kaiser, "Performance analysis of machine learning approaches in stroke prediction," *presented at the 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, pp. 1464-1469, Nov 2020. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9297525.

[3] M. S. Phipps and C. A. Cronin, "Management of acute ischemic stroke," *BMJ*, vol. 368, pp. l6983, Feb 2020, doi: 10.1136/bmj.l6983.

[4] J. Heo, J. G. Yoon, H. Park, Y. D. Kim, H. S. Nam, and J. H. Heo, "Machine Learning-Based Model for Prediction of Outcomes in Acute Stroke," *Stroke*, vol. 50, no. 5, pp. 1263-1265, May 2019, doi: 10.1161/STROKEAHA.118.024293.

[5] L. von Rueden *et al.*, "Informed Machine Learning - A Taxonomy and Survey of Integrating Prior Knowledge into Learning Systems," *IEEE Transactions on Knowledge and Data Engineering*, vol 35, no 1, pp. 614-644, 2021, doi: 10.1109/tkde.2021.3079836.

[6] S. S. P. Shimpi, M. Shroff and A. Godbole, "A Machine Learning Approach for the classification of Cardiac Arrhythmia," *presented at the 2017 International Conference on Computing Methodologies and Communication (ICCMC)*, pp. 603-607, Jul 2017. [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8282537&isnumber=8282515.

[7] 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, 2019, doi: 10.1007/s00521-019-04041-y.

[8] S. D. Chidozie Shamrock Nwosu, Peru Bhardwaj, and a. D. J. Bharadwaj Veeravalli, "Predicting stroke from electronic health records," *presented at the 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp.arXiv--1904, 2019, Proc 2019. [Online]. Available: https://www.arxiv-anity.com/papers/1904.11280/.

[9] O. Almadani and R. Alshammari, "Prediction of Stroke using Data Mining Classification Techniques," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 1, pp. 457-460, 2018, doi: 10.14569/ijacsa.2018.090163.

[10] S. T. Teerapat Kansadub and S. Kiattisin, "Stroke risk prediction model based on demographic data," *The 2015 Biomedical Engineering International Conference (BMEiCON-2015)*, pp.1-3, Nov 2015.

[11] N. S. Adi, R. Farhany, R. Ghina, and H. Napitupulu, "Stroke Risk Prediction Model Using Machine Learning," *presented at the 2021 International Conference on Artificial Intelligence and Big Data Analytics*, pp. 56-60, Oct.2021.

[12] C. Rana, N. Chitre, B. Poyekar, and P. Bide, "Stroke Prediction Using Smote-Tomek and Neural Network," *presented at the 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pp.1-5, Jul 2021.

[13] Q. Khan and S. A. Shah, "Using Machine Learning to Predict Patient's Admission Trends in Hospital," in *2021 International Conference on Computing, Electronic and Electrical Engineering (ICE Cube)*, pp. 1-5, Oct 2021: IEEE, doi: 10.1109/ICECube53880.2021.9628249. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9628249

[14] S. Akter, M. Amina, and N. Mansoor, "Early Diagnosis and Comparative Analysis of Different Machine Learning Algorithms for Myocardial Infarction Prediction," in *2021 IEEE 9th Region 10 Humanitarian Technology Conference (R10-HTC)*, pp. 01-06, 2021: IEEE, doi: 10.1109/r10-htc53172.2021.9641080. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9641080

[15] 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," *J Healthc Eng*, vol. 2021, pp. 7633381,Nov. 2021, doi: 10.1155/2021/7633381.

[16] G. L. A. K. Gangavarapu Sailasya, "Analyzing the Performance of Stroke Prediction using ML Classification Algorithms," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 6, pp. 539-545, 2021. [Online]. Available: https://pdfs.semanticscholar.org/df5c/7d1bd7a59009dc51b9db903aa7f144241879.pdf.

[17] A. A. Ali, "Stroke Prediction using Distributed Machine Learning Based on Apache Spark," *Stroke*, vol. 28, no. 15, pp. 89-97, 2019. [Online]. Available: https://www.researchgate.net/profile/Nahla-Omran-2/publication/338458550_Stroke_Prediction_using_Distributed_Machine_Learning_Based_on_Apache_Spark/links/5e1619404585159aa4be6a2e/Stroke-Prediction-using-Distributed-Machine-Learning-Based-on-Apache-Spark.pdf.

[18] Fedesoriano. *Stroke Prediction Dataset*. [Online]. Available: https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

[19] P. Madley-Dowd, R. Hughes, K. Tilling, and J. Heron, "The proportion of missing data should not be used to guide decisions on multiple imputation," *J Clin Epidemiol*, vol. 110, pp. 63-73, Jun 2019, doi: 10.1016/j.jclinepi.2019.02.016.

[20] J. Qin, L. Chen, Y. Liu, C. Liu, C. Feng, and B. Chen, "A Machine Learning Methodology for Diagnosing Chronic Kidney Disease," *IEEE Access*, vol. 8, pp. 20991-21002, 2020, doi: 10.1109/access.2019.2963053.

[21] X. Zeng, J. Huang, and C. Ding, "Soft-Ranking Label Encoding for Robust Facial Age Estimation," *IEEE Access*, vol. 8, pp. 134209-134218, 2020, doi: 10.1109/access.2020.3010815.

[22] C. M. Tae and P. D. Hung, "Comparing ML Algorithms on Financial Fraud Detection," *presented at the Proceedings of the 2019 2nd International Conference on Data Science and Information Technology*, 2019.

[23] D. Dablain, B. Krawczyk, and N. V. Chawla, "DeepSMOTE: Fusing Deep Learning and SMOTE for Imbalanced Data," *IEEE Trans Neural Netw Learn Syst*, PP 1-13, Jan 2022, doi: 10.1109/TNNLS.2021.3136503.

[24] M. M. Mijwil, "Implementation of Machine Learning Techniques for the Classification of Lung X-

Ray Images Used to Detect COVID-19 in Humans," *Iraqi Journal of Science,* vol. 62, no. 6, pp. 2099-2109, 2021, doi: 10.24996/ijs.2021.62.6.35.

[25] S. Ray, "A quick review of machine learning algorithms," in *2019 International conference on machine learning, big data, cloud and parallel computing (COMITCon)*, pp. 35-39,

[26] 2019: IEEE.

[27] F. K. Nasser and S. F. Behadili, "Breast Cancer Detection using Decision Tree and K-Nearest Neighbour Classifiers," *Iraqi Journal of Science,* vol. 63, no. 11, pp. 4987-5003, 2022, doi: 10.24996/ijs.2022.63.11.34.

[28] S. Krishnan and S. Geetha, "Prediction of Heart Disease Using Machine Learning Algorithms," in *2019 1st international conference on innovations in information and communication technology (ICIICT)*, pp. 1-5, 2019: IEEE.

[29] A. U. Haq, J. P. Li, M. H. Memon, S. Nazir, and R. Sun, "A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms," *Mobile Information Systems,* vol. 2018, pp. 1-21, 2018, doi: 10.1155/2018/3860146.