Mathematical Modelling of Engineering Problems
Vol. 11, No. 12, December, 2024, pp. 3369-3378
Journal homepage: http://iieta.org/journals/mmep

Efficient Detection of Brain Stroke Using Machine Learning and Artificial Neural Networks

Manal T. Ali, Adham Hadi Saleh*, Haneen S. Aziz

Department of Electronic Engineering, College of Engineering, University of Diyala, Baquba 32001, Iraq

Corresponding Author Email: adham.hadi@yahoo.com

https://doi.org/10.18280/mmep.111215

Received: 29 April 2024
Revised: 21 August 2024
Accepted: 30 August 2024
Available online: 31 December 2024

Keywords:
neural network, machine learning, decision tree classification, k-nearest neighbors, logistic regression, support vector machine, stroke

ABSTRACT

The medical term for the devastating effect of a ruptured blood vessel on the brain is a stroke. A disruption in the delivery of oxygen and nutrients to the brain through the circulatory system is another potential cause. According to a report released by the World Health Organization, the World Health Organization, there are many reasons of death and disability on the globe, but the most common cause is a brain stroke. Although cardiac stroke prediction has received a lot of attention, brain stroke risk has received comparatively little attention. The severity for a stroke can be reduced by detecting it early on. This research aims to use neural network (NN) and machine learning (ML) techniques to assess the probability of a stroke in the brain occurring early on. Reliable stroke prediction data has been obtained from the website of Kaggle in order for testing the algorithm's performance. This data has 11 columns and 4982 rows, with 10 columns representing features and the final column representing stroke prediction. In this research, five separate models were trained to accurately predict based on multiple physiological variables utilizing AI techniques that include logistic regression, artificial neural networks, decision trees categorization, k-nearest neighbors' categorization, and linear support vector models. The top approach to this challenge was k-nearest neighbor classification, which had nearly 100% accuracy.

1. INTRODUCTION

A brain stroke considered one of the most serious medical conditions that caused a death to people over 65 years old, which classified as a one of main three reasons of death in developing nations and America, similar to how a "heart attack" harms the heart. In addition to the high expense of treatment and the possibility of long-term incapacity or even death, a stroke can cause significant financial hardship for those affected. Although a stroke claims the life of someone every four minutes, it is possible to avert as many as 80% of these tragedies by detecting and preventing strokes in their early stages [1].

The strokes in brain tissue occur more frequently in men compared with women, particularly in middle age and old age. Stroke affects around 8% of sickle cell children. Every year, 15 million people in the world suffer from strokes. Five million die, while a different five million are severely disabled, imposing a burden on communities and their families [2].

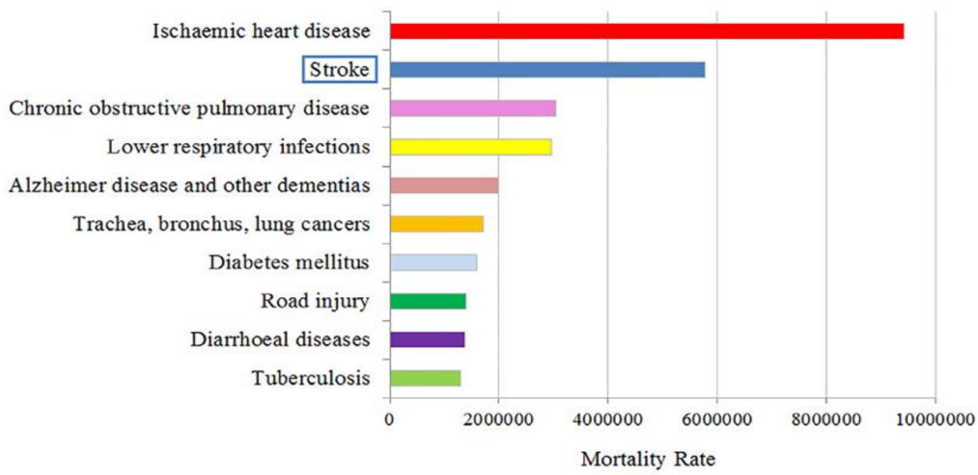

Figure 1. Displays the estimated number of deaths in 2016 [3]

In the past few years, machine learning has developed and grown rapidly in a variety of healthcare organizations. Figure 1 indicates the most recent health worldwide predictions pursuant to cause to the period 2000-2016. It highlights ischemic heart failure and stroke, as a main causing of disability and death [3]. Access for medical treatment increases the probability of a full recovery for stroke victims in the early stages. Any person who has suffered a stroke should not delay seeking medical attention. Loss of life, crippling disability, and brain damage await those who fail to comply. Heart attacks can happen for many different reasons. According to the National Hearts, Pulmonary, and Blood Institution, the leading causes of stroke include sedentary lifestyle, alcohol and tobacco consumption, as well as one's own genetic makeup, prior medical conditions, and other health issues [4].

There are actually three kinds of strokes: hemorrhagic, ischemic, and a temporary ischemic stroke. Ischemic stroke is the most prevalent kind of stroke overall. From what we know, most of the deaths were caused by ischemic strokes. An ischemic stroke mostly happens when blood clots in the brain. The two most common forms are thrombotic and embolic strokes. An embolic stroke usually occurs when a blood clot forms somewhere in the body and moves to the brain, stopping the flow of blood. A thrombotic stroke occurs when a clot impedes an artery carrying blood to the brain [5]. Machine learning (ML) is one of the most important reasons for scientific development in the health field, as machine learning helps specialists in health diagnosis of diseases, making appropriate medical decisions, and predicting the various conditions of the patient. Several research efforts had been performed in the past several years to improve stroke diagnosis accuracy and speed using machine learning [3].

ML is a powerful testing technology that depends upon training as well as testing. Machine learning that is a wide field of learning in which automated processes mimic human skills, is a subset of artificial intelligence, that combination of technological advances is known as machine intelligence. Machine learning structures, on the reverse hand, are trained to process and use data. One of the main principles of machine learning (ML) is comparison with natural phenomena, and in this case, it was based on many factors such as high blood pressure, the number of heart attacks, sex and age of the patient, and others [6].

Data on the patient's mental health, blood, and diagnoses were analyzed using a Kaggle dataset containing a variety of physiological traits [7]. These characteristics will be investigated further and used to make conclusive predictions. To understand the machine learning approach, the dataset has to be cleaned up. Preparing data is a process at this stage. This is achieved by verifying and entering the zero values in the data record. Following preprocessing, and the dataset classified for two main groups: training and testing data.

This study aims toward developing a system that is able to reliably detecting brain strokes. The stages that must be completed are:

• Early detection of stroke helps reduce complications, take the necessary medical precautions to avoid complications and reduce the risk of death.

• A brain stroke could devastate a global system of healthcare.

• This approach can be helpful to the doctor.

• The early identification of a brain stroke can also result in recommendations on how to handle it safely.

• A brain stroke can be used to reduce morality.

• The final outcomes are utilized to assess the efficiency of contrastive models.

The contribution of this work involves is using different algorithms on a freely available dataset (from the Kaggle website), as well as methods for pre-processing the brain stroke dataset. Deep learning algorithms are usually used to detection and diagnostics brain strokes Brain stroke detection and diagnostic algorithms are evaluated using performance matrices to compare and determine the best approach for predicting the onset of stroke. Future research directions for researchers studying brain stroke detection and diagnosis.

In this paper, we only look at a few high-performance machine learning techniques, such support vector machine (SVM), decision tree (DT), logistic regression (LR), k-nearest neighbours (KNN), and neural network (NN), both with as well as without the smoking status feature. Some logical approaches are utilized to pre-process data with the objective to balance the database and attain peak performance. Among the five classification algorithms tested, k-nearest neighbours (KNN) outperformed the others, producing a higher accuracy score; this is an important aspect of the methods and results. The study presents five different machine learning classification algorithms.

There are six sections to the rest of the paper. The works that serve as a foundation are detailed in Section 2. Materials and methods that contents description of dataset, pre-processing, and classification are all covered in Section 3, which also details the technique. Experiment settings and results are discussed in Section 4; Section 5 is conclusion; and finally, Section 6 considers the future scope.

2. RELATED WORKS

Previous research on a range of topics related to stroke prediction has been published in the literature.

Akter et al. [8] suggested an ML algorithm for identifying stroke victims. The data used in their study came from 507 patients at Kumbakonam's Sugam Multispecialty Hospital. They employed boosting, bagging, DT, LR, SVM, and ANN methods. Their method achieved a maximum accuracy of 95.3% using ANN.

Al-Zubaidi et al. [9] implemented a SMOTE method to address the imbalance problem, and it did so use the identical dataset and features selected for this paper. It's worth noting that the "unknown" class for the "smoking status" feature hadn't been set to null. Furthermore, it performed well by merging three classifiers: logistic regression, k-nearest neighbors, and RF. The random forest classifier (RFC) yielded accurate results (96% accuracy, precision, recall, and F1 score).

Govindarajan et al. [10] diagnosed two ANN algorithms ischemic stroke with accuracy rates of 79.2% and 95.1%, respectively.

Tazin et al. [11] suggested utilizing decision trees, random forest approach, voting classifier, and logistic regression categorization approaches to estimate the probability of stroke in its early stages. With a 96% success rate, random forest outperformed all other algorithms tested.

Bandi et al. [12] suggested enhancing the random forest method for predicting strokes following obtaining an accuracy rate of 96.97% with data assembled from the National Institutes of Medicine Stroke Score.

Sakri et al. [13] aimed to improve stroke prediction by conducting a thorough analysis of the various components that comprise electronic health records (EHR). They use statistical and principal component analysis methods to determine the most important stroke prediction factors. The RF, SVM, and DT models were deployed to train on EHR data. It was determined that RF was the best model, with a 95.3 percent accuracy.

Merdas [14] proposed a classification approach to strokes based on whether they occur or do not. This model blends artificial intelligence methods with text-mining tools. When properly trained, algorithms for machine learning are valuable tools in many domains, such as data management and medicine. In order to train the system to prevent strokes in the future, the study extracts the symptoms of patients from case sheets. 507 patients at Kumbakonam, Tamil Nadu, India's Sogam Multispecialty Hospital provided the researchers with information. Next, labeling and optimized entropy techniques were used to extract case papers. From the data, researchers identified significant, crucial, and practical attributes that they utilized to categorize strokes. The following algorithms were employed: "neural networks, support machine algorithms, boosting, bagging, and random forests". With an accuracy of roughly 95% and a standard deviation of 14.69, The model demonstrated that artificial neural networks trained using the stochastic gradient descent technique outperformed other models. It also produced good results.

Patereha and Melnyk [15] produced a stroke prediction model utilizing stacking ensemble classifiers and machine learning classifiers. Furthermore, the study proposes strategies such as selecting features and model parameter tuning to increase classification accuracy. Random forest was chosen for meta-learning among base classifiers such as k-nearest neighbor, random forest, logistic regression, support vector machine, and Naive Bayes in the suggested stacking prediction model. 97% accuracy was accomplished. However, the study has certain limitations that may make it less useful. For example, there is little detail about the data preparation and selection procedure, which may affect the accuracy of the results. Furthermore, the authors did not compare their approach to other machine-learning strategies.

In order to look into stroke early detection, Mridha et al. [16] found that when the brain's blood supply is abruptly cut off, a stroke occurs. It is imperative that these episodes, which may cause impairment or even death, be identified early on. Many machine learning techniques and approaches were tested during this search, and the stacking strategy proved to be the most efficient. A range of models, including RF, NBs, LR, KNN, and organizing, were tested for accuracy, precision, recall, and F-measure on datasets containing 3254 people aged 18 and up. The stacking technique produced results with 98.9% AUC, 97.4% precision, and 80% accuracy.

Purwono et al. [17] utilized algorithmic learning, especially a random forest approach, to avoid stroke by merging carotid ultrasonography image-based phenotypes and their harmonics with conventional risk parameters, achieving an accuracy of 93.15%.

Saleh et al. [18] predicted stroke using a combination of support vector machine algorithm (SVM), stochastic gradient boosting (SGB), penalized logistic regression (PLR) on a dataset collected by Turgut Ozal Medical centres, Inonu University, the Malatya, Turkey. The study's findings indicated that SVM had the best accuracy of 98%.

3. MATERIALS AND METHODS

The process for predicting a brain stroke, as seen in Figure 2, is explained in detail in this section. Acquiring data, preparing data, describing classifiers, and evaluating performance are the four corners of the methodology. This part is described in depth below.

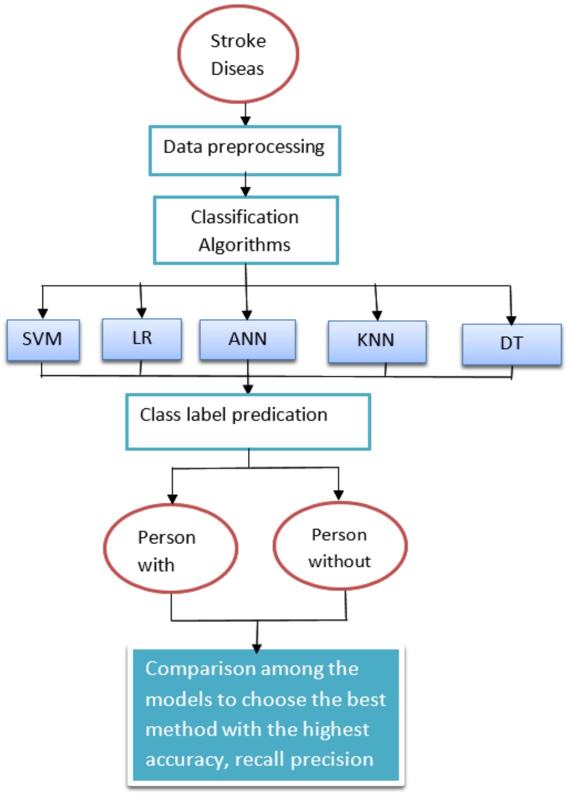

Figure 2. A flow chart of the proposed work

Table 1. Description of the dataset's features

Features	Description	Value Range
Gender	The person's gender [0: Female, 1: Male]	0, 1
Age	The human being's age in years (0.24 - 82)	0.24-82
Hypertension	High blood pressure (0 = No, 1 = Yes)	0, 1
Heart disease	Heart disease (0-There's no, 1 - In fact)	0, 1
Ever_married	Status of Marriage (Yes/No, 0 – 1)	0, 1
Work type	1-Children 2-Govt_Job 3-Never Worked 4-Private 5-Self_Employed	1, 2, 3, 4, 5
Residence type	0 = Rural, 1 = Urban	0, 1
Avg glucose level	Average Glucose Level	55 - 291.0
BMI	Body Mass Index	10.1 97.6
Smoking status	0 = no smoke, 1 = smoke	0, 1
Stroke	Class Attributes: (0-No Stroke Risk, 1-SrokeRisk)	0, 1

3.1 Description of dataset

The machine learning algorithms provide a consistent result, and the data is their main input. therefor in this proposal work "Stroke prediction dataset", the used data is obtained from an independent data source called Kaggle [7]. There are 4982 rows and 11 columns in this dataset. The first ten of those columns contain the features that we will use subsequently to forecast the final column, "target(stroke)", which will indicate whether or not the patient will experience a stroke. The 4982 rows are the 4982 patients' data that we were able to locate in the dataset. Table 1 provides a brief overview of the dataset's features that used in proposal work.

3.2 Pre-processing of data

Before developing a model, the dataset must be preprocessed to exclude outliers and undesirable noise that could deviate from standard training procedures. This step eradicates any obstacle that prevents the model from operating at peak efficiency. Next, after gathering the right dataset, you need to clean it up and make sure it's suitable to use for model creation. Eleven characteristics make up the examined dataset. The missing values of dataset has been detected and filled during dataset examination.

mean and unit variance, normalization is used to scale the Data preprocessing is one of the most important and crucially phase in the data mining process, which require preparing raw data for analysis by cleaning and transforming it. Typical data pre-processing steps include the following:

The processor of finding and fixing mistakes or inconsistencies in the data is a data cleaning, such as duplicates, outliers, and missing values. A variety of methods can be accomplished for data cleaning process, including imputation, removal, and transformation is a process to convert a data into a form that is appropriate for analysis. Some of common methods that used in data transformation such as normalization, standardization, and discretization. To have a zero data to a common range, standardization is used to change the data. Continuous data can be discretized using the discretization process.

The process of dividing continuous data into distinct intervals (categories) is called the data discretization process. We need to use discretization frequently for developing algorithms for data mining and machine learning that need categorical data. To achieve the discretization, some techniques used like clustering, equal width binning, and equal frequency binning.

Data normalization is the process of scaling the data to a standard range, like 0 to 1. In order to manage data with various scales and units, normalization is frequently utilized. Decimal scaling, z-score normalization, and minimum-maximum normalization are examples of common normalization methods. The 'hypertension, heart disease, and stroke' attributes "data type should be changed to string since the attribute's data displays 'yes' and 'no' statements, but they are represented by the numbers '0' and '1'.

3.3 Methodologies

Many machine-learning algorithms are using to carry out the work we have proposed.

3.3.1 Support vector machine (SVM)

It is an efficient and highly straightforward algorithm used for the purpose of classification and pattern identification. The support vector machines algorithm was pioneered by Vladimir Vapnik in 1995.The primary objective of this approach is to derive a function that generates hyperplanes, or boundaries. Hyperplanes are employed to segregate distinct groups of input data points. support vector machines (SVM) employ a binary classification approach [11]. It provides a viable alternative approach for classifying data in supervised machine learning programming. When utilizing machine learning programming in nuclear facilities, it is necessary at times to perform historical data classification in order to assess new input data, such as operator actions and field signals for pressure and temperature. This classification involves categorizing the data into one of two states, such as correct or incorrect, true or false, etc. In this case, support vector machines (SVM) can be employed [19].

3.3.2 Logistic regression (LR)

One of the most well-known ML algorithms at supervised learning is a Logistic Regression (LR). This method of forecasting makes use of several an independent variable to a dependent variables for predictable categorical. Regardless of the application, logistic regression and linear regression are quite similar.

Where linear regression corrects regression problems and logistic regression corrects classification problems. Any multicollinear data can be analyzed using ridge regression, a model tuning technique. This technique carries out L2 regularization [10].

3.3.3 Decision tree (DT)

Parts of a typical tree are its trunk, branches, and leaves. In a decision tree, the same structure is used. It has nodes at the root, branches, and leaves [20]. The process begins on every internal node with an attribute test, continues on every branch, and ends on every leaf node with the class label. A root node is the very first node in a tree, and, as its name implies, it is the parent node of every other node [21]. Every node in a decision tree represents an attribute, every link represents a rule, and every leaf represents the conclusion, which can be a continuous or categorical value. Because decision trees are designed to replicate human reasoning, they make it easy to gather data and draw accurate conclusions. The goal is to organize all of the data in a tree structure and process a single outcome at each leaf [22].

3.3.4 K-nearest neighbors (KNN)

It is a classification technique can be considered a non-parametric, which means it does not make any assumptions about the elementary dataset. The K-nearest-neighbors (KNN) approach fits this description [23]. Both its ease of use and its efficiency have earned it a reputation. Specifically, it is a method for supervised learning. The unlabeled data and a labeled training dataset are providing to predict this class. Within this dataset, the data points are categorized into a variety of classes.

A variety of distinct characteristics are ultimately responsible for determining the category to which the unlabeled data belongs when it comes to the categorization process. When it comes to classification, KNN is utilized the majority of the time. Its objective is to classify data in a particular location based on the training exam. During the classification process, the class to which the unlabeled data belongs is finally determined. The goal of its application is to classify data in a particular location based on the training examples that are located in close proximity to or immediately next to the site in question.

The implementation of this approach is uncomplicated, and the time required to compute it is rather negligible. Making the use of Euclidean distance is the method that is utilized in order to calculate its nearest neighbors for continuous data [24].

3.3.5 Neural network (NN)

Machine learning techniques are categorized to: supervised and unsupervised. With neuron inputs and outputs, perception is the most basic ANN design. Both ramp and step functions are utilized as activation functions. In order to divide data into two categories, perceptions are utilized [25]. Among the many neural network methods for classification and prediction, it sees heavy use. The BP algorithm's output is computed by propagating the results of the hidden layers to the output layer [26]. We compare this output to the desired outcome for the input that was provided. Error path is starting with output layer, passing to hidden layer, and input layer, based on this difference. The relative importance of each neuron is altered when the flow reverses direction. Epochs are defined as this forward-and-backward cycle of input/output and output/input. To train a neural network to produce a specific output, one must first provide it with a collection of input data that is already known [27]. This process is known as training the network. As long as the error (the gap between the network's actual and expected outputs) remains within a predetermined tolerance, the network will continue to go through this cycle. We say the network is trained now. The training procedure determines the relative importance of each neuron in each layer. In order to determine the network's reaction to unknown data, the weights that have been learned from a trained network are utilized [28].

4. EXPERIMENT SETTINGS AND RESULTS

An ML model's efficacy is assessed using performance evaluation metrics. To predict the occurrence of a brain stroke, we employed five different classifiers. To determine which classifiers worked best, we calculated the performance assessment measures.

4.1 Matrix of confusion

The confusion matrix is a simple evaluation method to check performance of classification model, which is depended on compare between positive instances were (correctly/incorrectly) classified and negative instances were (correctly/incorrectly) classified.

Rows in a confusion matrix represented an actual label, and columns represented a predicted label, as explained in Table 2.

Table 2. The matrix of confusion

	Positive Prediction	**Negative Prediction**
Real Positive	TP	FN
Real Negative	FP	TN

True positives (TP): These cases occur when the predicted and actual categories are True (1), i.e., at this state the patient has complications and also the model classifies them as the same.

True negatives (TN): These cases occur when the predicted and actual categories are False (0), i.e., at this state the patient does not have complications and also the model classifies them as the same.

False negatives (FN): These cases happen when there are different results between predicted class and actual class, where predicted class is False (0), and the actual class is True (1), i.e., which mean the model classifies patient have no complications, but they have.

False positives (FP): These cases can occur when there are different results between predicted class and actual class, where predicted class is True (1), and the actual class is False (0), i.e., which mean the model classifies patient have a complication, but they do not have.

4.1.1 Accuracy (ACC)

Accuracy is defined as the rate between the correct predictions and the overall model's predictions. In spite of it is used widely, it is not considered a measure of performance in unbalanced data. The accuracy formula is:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \times 100\% \tag{1}$$

4.1.2 Precision

Precision is rate between the number of patients who actually get complication among a patient who identified they have by the system. The formula of precision is show below:

$$Precision = \frac{TP}{TP + FP} \times 100\% \tag{2}$$

4.1.3 Recall/sensitivity/true positive rate (TPR)

Recall or sensitivity: it is a measure of the number of patients who were expected to have complications and it actually happened to them as shown in formula:

$$Sensitivity = \frac{TP}{TP + FN} \times 100\% \tag{3}$$

4.1.4 Specificity

Specificity is the ability of the algorithm/model's to predict a true negative to each available category. In general, it is known as a true negative rate.

$$Specificity = \frac{TN}{TN + FP} \times 100\% \tag{4}$$

4.1.5 F1 score

F1 score is a harmonic average of recall and precision, which is used for testing accuracy. Here's the formula:

$$F1\,score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{5}$$

4.1.6 False positive rate (FPR)

It is indicated to a percentage or ratio of positive cases that were incorrectly identified or classified as positive in a test, or, in layman's terms, false alarms. To evaluate binary context problems, the false positive rate can be used. These include predictions, disease detection, quality control, and cybersecurity threats. Another application is machine earning (ML), which using to estimate the performance of classification algorithms or models. FPR is frequently expressed as a percentage or ratio.

$$False\ positive\ rate(FPR) = \frac{FP}{TN + FP} \times 100\% \tag{6}$$

4.1.7 False negative rate (FNR)

It refers to the predicted value is negative, and actual value is positive, which mean a model is incorrectly and predicted value is positive class labels to be negative.

$$False\ negative\ rate\ (FNR) = \frac{FN}{TP + FN} \times 100\% \tag{7}$$

4.1.8 AUC-ROC

AUC-ROC is an abbreviation of "Area Under the Receiver Operating Characteristic Curve", which represented a method to check the performance in machine learning to estimate binary classification models. AUC-ROC is using to evaluate the ability of model to recognize between positive and negative classes. This metric represented the probability curve that compare between true positive rate (TPR) and false positive rate (FPR) with different thresholds with rang (0 to 1), where 0 indicated to a poor mode, and 1 indicated to an ideal model (Table 3).

Table 3. Confusion matrix for each classifier

Classification Algorithms	Confusion Matrix	
SVM	TP	FN
	4733	0
	FP	TN
	248	0
LR	TP	FN
	4733	0
	FP	TN
	248	0
DT	TP	FN
	4713	20
	FP	TN
	188	60
KNN	TP	FN
	4733	0
	FP	TN
	0	248
NN	TP	FN
	118	29
	FP	TN
	130	4708

4.2 Experimental results

The performance comparison of stroke-level prediction models and benchmark machine learning models is shown in this section. The comparison is based on specified measures. Normalization was performed on the confusion matrices that were constructed using percentages per row.

In the first model, a linear support vector machine (SVM) classifier was used. Figure 3 depicts the confusion matrix that is associated with this paradigm. It has been observed that the linear support vector machine classifier achieved an accuracy of 95.022%.

The second model was created with logistic regression classification, and Figure 4 shows a linear logistic regression classification achieved 95.02% accuracy.

The third model, which used a neural network algorithm, a neural network was built consisting of five layers, with 10 inputs and 2 outputs, The neural network was trained with backpropagation in MATLAB program, stroke prediction data set was classified to three groups, 80% used for training and used for testing, and 20% for test validity. Figure 5 explains ANN schematic.

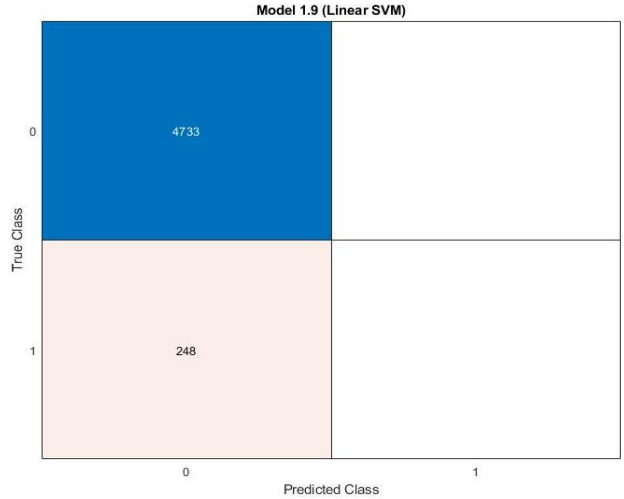

Figure 3. Confusion matrix for a linear support vector machine

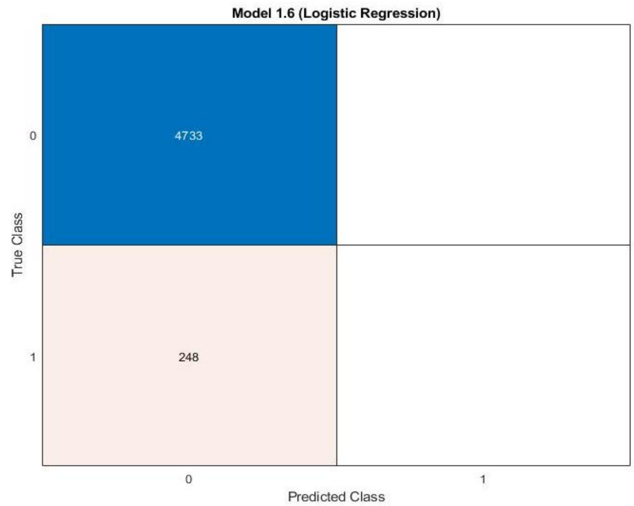

Figure 4. Confusion matrix with logistic regression model

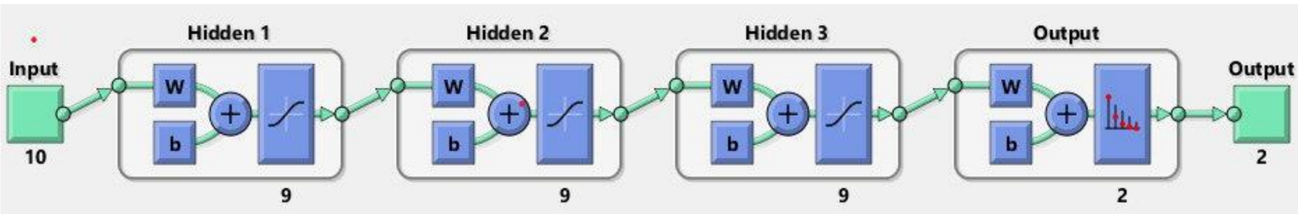

Figure 5. An artificial neural network schematic

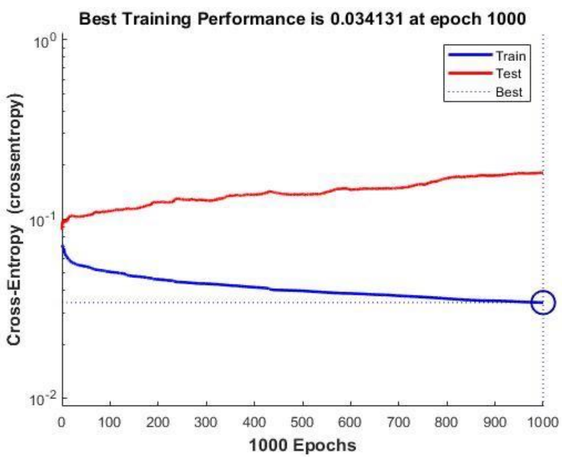

Figure 6. Neural network training

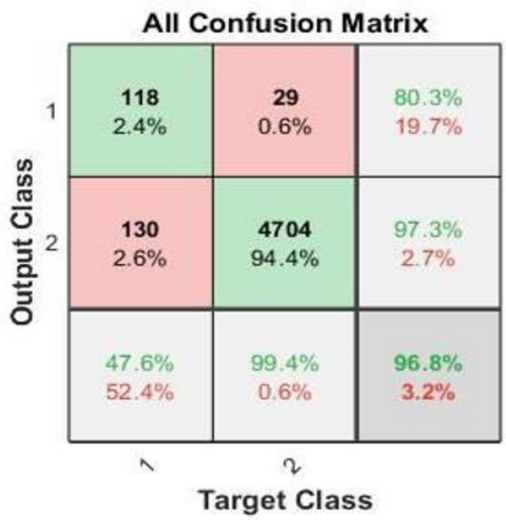

Figure 7. Neural networks with confusion matrix

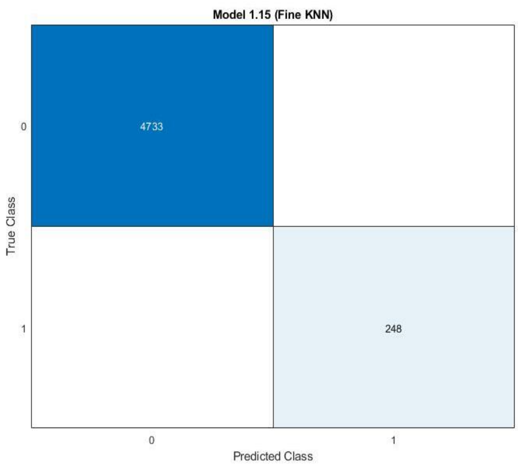

Figure 8. KNN classifier of confusion matrix

Figure 6 shows the training results of the dataset using the neural network algorithm.

An accuracy of 96.8% in this model was obtained from the confusion matrix, as in Figure 7.

By using KNN classifier, fourth model was constructed, and it achieved the most successful outcomes compared to the other five algorithms. One hundred percent accuracy was achieved, Figure 8 displays this matrix model.

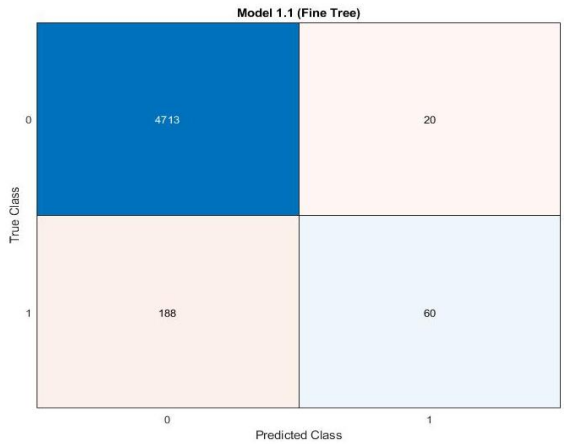

Figure 9. Decision tree's (DT) confusion matrix

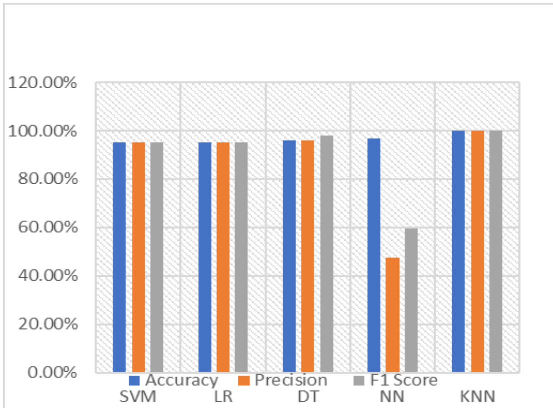

Figure 10. Comparison of the five classifiers' accuracy

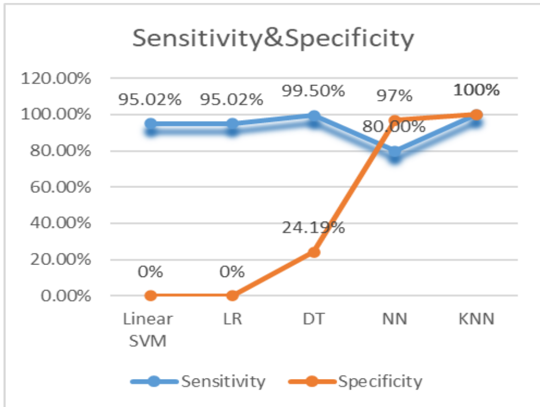

Figure 11. Comparing the five classifiers' sensitivity and specificity

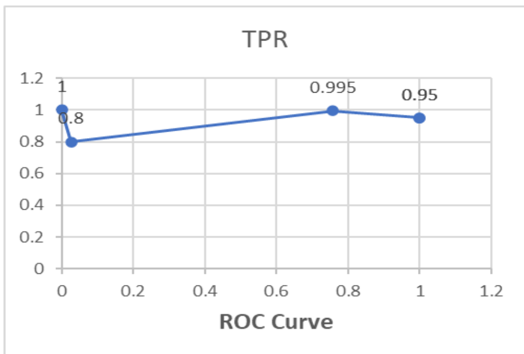

Figure 12. Assessing training measures

The decision tree algorithm is utilized from the construction of the fifth model, is provided result with an accuracy of 95.8%. For this particular model, confusion matrix represented in Figure 9.

Table 4 produces a summary of the entire potential performance evaluation indicators that are predicted by the classifiers. A number of different metrics, including mean absolute error, accuracy, sensitivity, specificity, and false positive rate (FPR), are included in your evaluation.

Figure 10 shows an easily understood representation of the comparison accuracy among classifiers. It is clear that for our proposed models, KNN achieved the highest accuracy and linear SVM with logistic regression achieved the lowest accuracy. On the other hand, Figure 11 shows a comparison of the sensitivity and specificity values between the decision tree, KNN, NN, linear SVM, and logistic regression classifiers. Here, the KNN classifier gets the highest sensitivity and specificity of 100%. The Figure 12 shows the ROC curve for the algorithms used in this study.

Evaluation of this work in relation to its predecessors is explained in Table 5, which shows the comparison between the proposed works and previous works. As shown in their proposed work, Akter et al. [8], Bandi et al. [12] and Saleh et al. [18] used several types of machine learning (ML) techniques for prediction and early handing and detection of stroke, but their results are not accurate enough, and they used a restricted data set. While the proposed model obtained more accurate results, and used a wide data set with a larger number of people. Table 6 lists the benefits and drawbacks of the methods employed in this study.

Table 4. Metrics for evaluating five classifiers performance

Classifier	Accuracy	Precision	Sensitivity	Specificity	F1 Score	FPR	FNR	Mean Absolute Error
SVM	95.2%	95.02%	95.02%	0%	95.01	100%	0%	4.8%
LR	95.02%	95.02%	95.02%	0%	95.01	100%	0%	4.8%
DT	95.8%	96.16%	99.5%	24.19%	97.80	75.80%	0.422%	4.2%
NN	96.8%	47.5%	80%	97%	59.60	2.68%	19.7%	3.2%
KNN	100%	100%	100%	100%	100	0%	0%	0%

Table 5. Evaluation of this work in relation to its predecessors

Reference	Object	Dataset Source	Dataset Size	Applied Classifier	Best Classifier	Obtained Accuracy
This work	Brain stroke prediction	Kaggle	4982	SVM, LR, DT, NN, KNN	KNN	100%
Akter et al. [8]	Classification of stroke disease	Hospital using tools equipped with tagging and maximum entropy algorithms.	507	LR, DT, NN, SVM, Bagging and boosting	NN	95.3%
Bandi et al. [12]	Prediction of brain stroke	The National Institutes of Health Stroke Scale (NIHSS)	4,799	LR, KNN, DT, CV, NB, RF, AdaBoost	RF	96.97%
Saleh et al. [18]	Predicted stroke	A dataset collected from TurgutOzal Medical Centre, Inonu University, Malatya, Turkey	112 healthy people 80 people with ischemic stroke	Combination of SVM, SGB, PLR	SVM	98%

Table 6. Pros and cons of different machine learning techniques and neural network

Technique	Pros	Cons
LR	Easy understanding and execution	Uses the premise that attributes and targets are linearly related
SVM	Works well in areas with many dimensions	Needs careful choice of kernel and tuning of factors
DT	Simple to understand and figure out	Easy to overfit, particularly with large and complicated datasets
NN	Works well with consecutive data sets	Might experience disappearing or bursting gradients
KNN	There is no required training period and the implementation process is straightforward.	Data should be appropriately scaled (normalised and standardised) across all dimensions for feature scaling to work

5. CONCLUSION

The inability to use one’s brain due to a stroke is a devastating medical emergency. By using machine learning models, we can reduce the likelihood of stroke by predicting them early on. In order to provide a high-accuracy prediction of Stroke based on many features, this study shows that we took a similar route, but with a better and more creative approach, as well as a bigger dataset for the model's training. The dataset that was used includes 4982 patients' observation problems with 11 brain stroke-related attributes. processing method has been used to increase the dataset's flexibility for training and testing the five classifiers. It is evident that the KNN algorithm outperforms other classifiers are it has a high features impact, and a high substantially accuracy of 100%. Diagnosis with the technology mentioned in this research using machine learning is considered an additional means to facilitate the doctor’s work and an indicator of the disease condition in addition to the patient’s vital and clinical indicators and supports them in accurately diagnosing the condition. The results of this method of diagnosis, medical data, medical history of the patient, and medical information of the specialist doctor in this field make the accurate diagnosis of the condition with high efficiency of up to 100%. Despite the accuracy of the results provided by this technology, the doctor remains the final decision maker, and what we provide is facilities for performing his work and diagnosing the disease condition better.

6. FUTURE SCOPE

This system can be improved and expended at future by enhancing the outcomes and the user experience. This will save time and will be better preparing to act based on the results. For future work objectives can be included a larger number of dataset and to use this model on disease cases with diverse data. The artificial intelligence structure can support the general public estimation of a stroke occurring to patient of young age groups and with adults, the level of risk that associated with it, and the likelihood of happening again, in other hand it is providing some basic information. In an ideal, it help patients to receiving a stroke treatment at timely manner and recover it. The technology mentioned in this research can be used to detect cardiac arrest, nerve weakness, nerve response speed, cancerous tumors, hepatitis, and others early.

REFERENCES

[1] Kumar, G.R., Vyshnavi, P., Prasanna, S., Reddy, T.H., Charanya, C., Chandrababu, P. (2022). Brain stroke detection using machine learning. International Journal of Research in Engineering, Science and Management, 5(3): 34-36. https://journal.ijresm.com/index.php/ijresm/article/view/1823.

[2] Akter, B., Rajbongshi, A., Sazzad, S., Shakil, R., Biswas, J., Sara, U. (2022). A machine learning approach to detect the brain stroke disease. In 2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, pp. 897-901. https://doi.org/10.1109/ICSSIT53264.2022.9716345

[3] Sirsat, M.S., Fermé, E., Camara, J. (2020). Machine learning for brain stroke: A review. Journal of Stroke and Cerebrovascular Diseases, 29(10): 105162. https://doi.org/10.1016/j.jstrokecerebrovasdis.2020.105162

[4] Rahman, S., Hasan, M., Sarkar, A.K. (2023). Prediction of brain stroke using machine learning algorithms and deep neural network techniques. European Journal of Electrical Engineering and Computer Science, 7(1): 23-30. https://doi.org/10.24018/ejece.2023.7.1.483

[5] Singh, M.S., Choudhary, P. (2017). Stroke prediction using artificial intelligence. In 2017 8th Annual Industrial Automation and Electromechanical Engineering Conference (IEMECON), Bangkok, Thailand, pp. 158-161. https://doi.org/10.1109/IEMECON.2017.8079581

[6] Kanna, R.K., Reddy, C.V.R., Panigrahi, B.S., Behera, N., Mohanty, S. (2024). Machine learning based stroke predictor application. EAI Endorsed Transactions on Internet of Things, 10. https://doi.org/10.4108/eetiot.5384

[7] Kaggle. (2022). Brain stroke dataset. https://www.kaggle.com/datasets/jillanisofttech/brain-stroke-dataset.

[8] Akter, B., Rajbongshi, A., Sazzad, S., Shakil, R., Biswas, J., Sara, U. (2022). A machine learning approach to detect the brain stroke disease. In 2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, pp. 897-901. https://doi.org/10.1109/icssit53264.2022.9716345

[9] Al-Zubaidi, H., Dweik, M., Al-Mousa, A. (2022). Stroke prediction using machine learning classification methods. In 2022 International Arab Conference on Information Technology (ACIT), Abu Dhabi, United Arab Emirates, pp. 1-8. https://doi.org/10.1109/ACIT57182.2022.10022050

[10] Govindarajan, P., Soundarapandian, R.K., Gandomi, A.H., Patan, R., Jayaraman, P., Manikandan, R. (2020) Classification of stroke disease using machine learning algorithms. Neural Computing and Applications, 32(3): 817-828. https://doi.org/10.1007/s00521-019-04041-y

[11] Tazin, T., Alam, M.N., Dola, N.N., Bari, M.S., Bourouis, S., Monirujjaman Khan, M. (2021). Stroke disease detection and prediction using robust learning approaches. Journal of Healthcare Engineering, 2021(1): 7633381. https://doi.org/10.1155/2021/7633381

[12] Bandi, V., Bhattacharyya, D., Midhunchakkravarthy, D. (2020). Prediction of brain stroke severity using machine learning. Revue d'Intelligence Artificielle, 34(6): 753-761. https://doi.org/10.18280/ria.340609

[13] Sakri, S., Basheer, S., Zain, Z.M., Ismail, N.H.A., Aldehim, G., Alharaki, M. (2024). An improved concatenation of deep learning models for predicting and interpreting ischemic stroke. IEEE Access, 12: 53189-53204. https://doi.org/10.1109/ACCESS.2024.3386220

[14] Merdas, H.M. (2024). Elastic Net-MLP-SMOTE (EMS)-Based model for enhancing stroke prediction. Medinformatics, 1(2): 73-78. https://doi.org/10.47852/bonviewmedin42022470

[15] Patereha, Y., Melnyk, M. (2024). Prediction of the occurrence of stroke based on machine learning models. Computer Design Systems Theory and Practice, 6(1): 17-27.

[16] Mridha, K., Ghimire, S., Shin, J., Aran, A., Uddin, M.M., Mridha, M.F. (2023). Automated stroke prediction using machine learning: An explainable and exploratory study with a web application for early intervention. IEEE Access, 11: 52288-52308. https://doi.org/10.1109/ACCESS.2023.3278273

[17] Purwono, P., Ma'arif, A., Negara, I.S.M., Rahmaniar, W., Rahmawan, J. (2021). Linkage detection of features that cause stroke using feyn qlattice machine learning model. Jurnal Ilmiah Teknik Elektro Komputer dan Informatika, 3(7): 423-432. https://doi.org/10.26555/jiteki.v7i3.22237

[18] Saleh, H., Ghany, S.F.A., Younis, E.M.G., Omran, N. (2019). Stroke prediction using distributed machine learning based on Apache spark. International Journal of Advanced Science and Technology, 28(15): 89-97. https://doi.org/10.13140/RG.2.2.13478.68162

[19] Chavhan, Y., Dhore, M.L., Yesaware, P. (2010). Speech emotion recognition using support vector machine. International Journal of Computer Applications, 1(20): 6-9. https://doi.org/10.1007/978-3-642-21402-8_35

[20] Sharma, H., Kumar, S. (2016). A survey on decision tree algorithms of classification in data mining. International Journal of Science and Research, 5(4): 2094-2097. https://doi.org/10.21275/v5i4.NOV162954

[21] Mienye, I.D., Sun, Y., Wang, Z. (2019). Prediction performance of improved decision tree-based algorithms: A review. Procedia Manufacturing, 35: 698-703. https://doi.org/10.1016/j.promfg.2019.06.011

[22] Patel, H.H., Prajapati, P. (2018). Study and analysis of decision tree based classification algorithms. International Journal of Computer Sciences and

Engineering, 6(10): 74-78. https://doi.org/10.26438/ijcse/v6i10.7478

[23] Kilincer, I.F., Ertam, F., Sengur, A. (2021). Machine learning methods for cyber security intrusion detection: Datasets and comparative study. Computer Networks, 188: 107840. https://doi.org/10.1016/j.comnet.2021.107840

[24] Chomboon, K., Chujai, P., Teerarassamee, P., Kerdprasop, K., Kerdprasop, N. (2015). An empirical study of distance metrics for k-nearest neighbor algorithm. In Proceedings of the 3rd International Conference on Industrial Application Engineering, pp. 280-285. https://doi.org/10.12792/iciae2015.051

[25] Gupta, N. (2013). Artificial neural network. network and complex systems. In International Conference on Recent Trends in Applied Sciences with Engineering Applications, pp. 24-28. https://iiste.org/Journals/index.php/NCS/article/view/6063/6019.

[26] Dongare, A.D., Kharde, R.R., Kachare, A.D. (2008). Introduction to artificial neural network. European Journal of Gastroenterology & Hepatology, 19(12): 1046-1054. https://doi.org/10.1097/MEG.0b013e3282f198a0

[27] Svozil, D., Kvasnieka, V., Pospichal, J. (1997). Introduction to multi-layer feed-forward neural networks. Chemometrics and Intelligent Laboratory Systems, 39(1): 43-62. https://doi.org/10.1016/S0169-7439(97)00061-0

[28] Shiruru, K. (2016). An introduction to artificial neural network. International Journal of Advance Research and Innovative Ideas in Education, 1(5): 27-30.