International journal of basic and applied research

www.pragatipublication.com

ISSN 2249-3352 (P) 2278-0505 (E)

Cosmos Impact Factor-5.86

DETECTION OF STROKE DIESEASE USING MACHINE LEARNING

Deepa D , Professor, Department Of AIML, SICET, Hyderabad

G Saketh Reddy, G.Pavan Kumar Reddy,Tiparthi Sakshith Reddy,G Karthik, Kasani Saidulu

UG Student, Department Of AIML, SICET, Hyderabad

Abstract

A stroke is a medical disorder in which blood vessels in the brain burst and damage the brain. When the supply of blood and other nutrients to the brain is interrupted, symptoms may develop. According to the World Health Organization (WHO), stroke is the leading cause of death and disability worldwide. Early recognition of the various warning signs of a stroke can help reduce the severity of a stroke. Various machine learning (ML) models have been developed to predict the probability of stroke in the brain. This research uses a variety of physiological parameters and machine learning algorithms such as logistic regression (LR), decision tree (DT) classification, random forest (RF) classification, and voting classifier to train four different models for reliable prediction. Random Forest was the best performing algorithm for this task with an accuracy of around 96 percent. The dataset used in the development of the method was the open access Stroke Prediction dataset. The accuracy percentage of the models used in this investigation is significantly higher than in previous studies, indicating that the models used in this investigation are more reliable. Numerous model comparisons have proven their robustness, and a scheme can be derived from the analysis of the study.

1. Introduction

A stroke occurs when blood flow to different areas of the brain is disrupted or reduced, resulting in cells in those areas of the brain not getting the nutrients and oxygen they need and dying. A stroke is a medical emergency that requires immediate medical attention. Early detection and appropriate treatment are necessary to prevent further damage to the affected area of the brain and further complications in other

International journal of basic and applied research

www.pragatipublication.com

ISSN 2249-3352 (P) 2278-0505 (E)

Cosmos Impact Factor-5.86

parts of the body. The World Health Organization (WHO) estimates that 15 million people suffer from stroke worldwide each year, with one person dying every four to five minutes in the affected population. According to the Centers for Disease Control and Prevention (CDC), stroke is the sixth leading cause of death in the United States [1]. A stroke is a non-communicable disease that kills approximately 11% of patients population. In the United States, approximately 795,000 people regularly suffer from the disabling effects of stroke [2]. It is the fourth leading cause of death in India.

Many academics have previously used machine learning to predict stroke. Govindarajan et al. [3] used text mining and a machine learning classifier to classify stroke disorders in 507 individuals. They tested various machine learning methods for training purposes, including an artificial neural network (ANN), and found that the SGD algorithm provided the highest value at 95 percent. Amini et al. [4, 5] conducted research to predict the occurrence of stroke. They classified 50 risk variables for stroke, diabetes, cardiovascular disease, smoking, hyperlipidemia, and alcohol consumption in 807 healthy and unhealthy individuals. They used the two most accurate methods: the c4.5 decision tree algorithm (95 percent accuracy) and the K-nearest neighbor algorithm (94 percent accuracy). Cheng et al. [6] presented a study on the estimation of ischemic stroke prognosis. In their study, they used 82 datasets of ischemic stroke patients, two ANN models, and accuracy values of 79 and 95 percent. Cheon et al. [7–9] conducted research to determine the predictability of stroke patient death. They identified stroke incidence using 15,099 individuals in their research. They detected strokes using a deep neural network method. The authors used PCA to extract information from medical records and predict stroke. They have an 83 percent area under the curve (AUC). Singh et al. [10] conducted research using artificial intelligence to predict stroke. In their research, they used a new technique to predict stroke using a cardiovascular health (CHS) dataset. In addition, they used a decision tree method to perform feature extraction followed by principal component analysis. In this case, the model was built using a neural network classification method and achieved 97 percent accuracy.

Chin et al. [11] conducted research to determineaccuracy of automatic early detection of ischemic stroke. The main goal of their research was to create a method for automating primary ischemic stroke using a

International journal of basic and applied research

www.pragatipublication.com

ISSN 2249-3352 (P) 2278-0505 (E)

Cosmos Impact Factor-5.86

convolutional neural network (CNN). They collected 256 images for training and testing the CNN model. They used the data stretching technique to increase the collected image while preparing the image of their system. Their CNN technique achieved 90 percent accuracy. Sung et al. [12] conducted research to establish a stroke severity index. They collected data on 3577 patients who had an acute ischemic stroke. They used a variety of data mining methods, including linear regression, to create their predictive models. Their predictive ability outperformed the k-nearest neighbor method (95% confidence interval). Monteiro et al. [13] used machine learning to predict

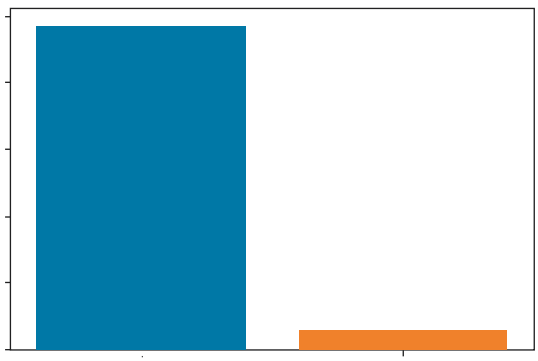
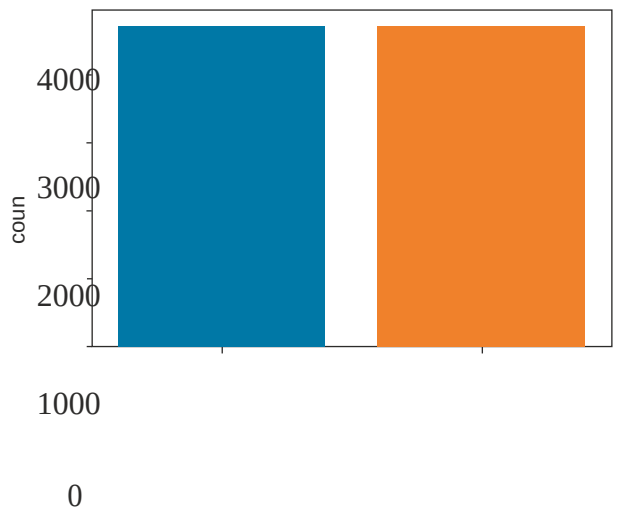


FIGURE 2: Total number of stroke and normal data.



International journal of basic and applied research

www.pragatipublication.com

ISSN 2249-3352 (P) 2278-0505 (E)

Cosmos Impact Factor-5.86

Figure 3: Pre-

processed output line. After segmentation, the model is trained using various classifiers. The classificatio n methods used in this study include random forest, decision tree, voting and logistic regression. Practical algorithm. The most common disease in the medical field is stroke, and its incidence is increasing every y ear. This study used public stroke data to evaluate four machine learning methods for predicting stroke re currence:

(i) Random Forest

(ii) Decision Making

(iii) Voting Classifier

(iv) This Logistic regression performed by the algorithm

. Each CE in this method must vote for one of two exit groups (in this case, strike or no strike). The final prediction is determined by the RF method, which selects the category with the most votes. The block dia gram of the random forest distribution is shown in Figure 4. It can be used for repetitive detection and gro uping, and the overall weight given to information is clear. It's also a useful feature because the preset hyp erparameters it uses generally provide a clear expectation. It is important to understand hyperparameters b ecause there are some of them in the first place. Overfitting is good

Figure 2: Total number of blood vessels and normal data.

A known problem in machine learning, although it rarely occurs in random forest classifiers. If there are e nough trees in the forest, the classifier will not overwrite the model. Decision tree. Both regression and cl assification problems can be solved using DT classification [18]. Moreover, these models are conservativ e models since different strategies already have different results. It resembles a tree. In this way, data is co ntinuously segmented according to certain parameters. Order nodes and leaf nodes are two parts of the or der tree. While data is distributed in the first node, the second is the node that produces the results. The ba sic structure of the DT classifier0 1 is shown in Figure 5. After segmentation, the model is trained using v arious classifiers. The classification methods used in this study include random forest, decision tree, votin g and logistic regression. Practical algorithm. The most common disease in the medical field is stroke, an

International journal of basic and applied research

www.pragatipublication.com

ISSN 2249-3352 (P) 2278-0505 (E)

Cosmos Impact Factor-5.86

d its incidence is increasing every year. This study used public stroke data to evaluate four machine learni ng methods for predicting stroke recurrence:

2.4.1. Random forest. The chosen classification algorithm is RF classification [17]. RF consists of several independent decision trees trained separately on random data. These trees were created during training an d the outputs of the decision trees were collected. A process called voting is used to determine the final es timateDT is easy to understand because it replicates the stages a person goes through when responding to real-

world decisions. This can be very useful in decision making. Consider all possible solutions to the proble m. Unlike other methods that require data cleaning. Election vote. Candidate voter is a classification mod el that learns from an ensemble of multiple models and predicts the outcome (category) based on the cate gory most likely to be selected as the output [19]. It is used to predict voting results. An example voting p rocess is shown in Figure 6. There are two voting methods as follows:

(i) Soft voting: In this stage, estimate the probability of adding and averaging each sample. The category with the highest value is considered the winner and its content is accepted as output. While this may seem like a fair and reasonable idea, it is only advisable if a group of people are properly evaluated. This is sim ilar to calculating the weighted average of a series of numbers, except that each sample contributes the sa me amount to the final output vector.

International journal of basic and applied research

www.pragatipublication.com

ISSN 2249-3352 (P) 2278-0505 (E)

Cosmos Impact Factor-5.86

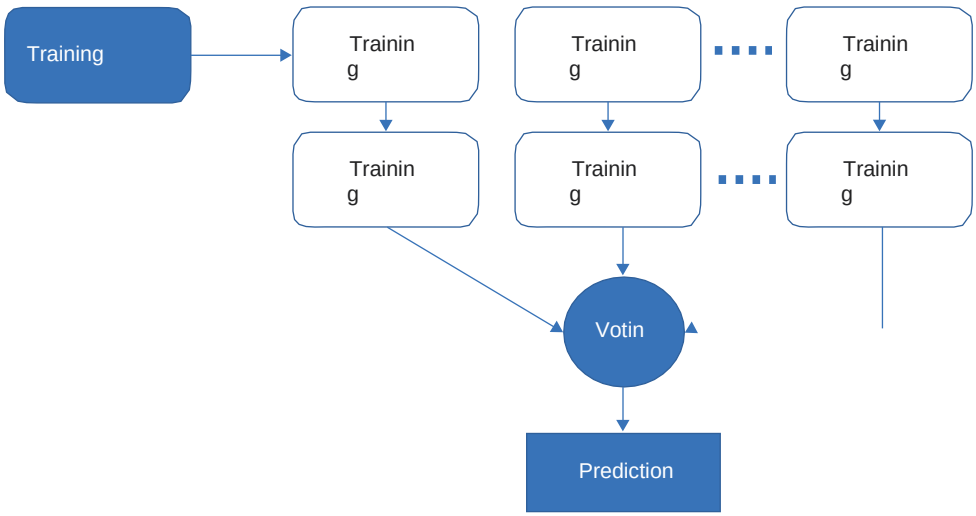


FIGURE 4: Block diagram of Random Forest Classifier.

The final output value is the output type. This formula is similar to the arithmetic calculation of numbers i n that certain values associated with each formula are ignored. Consider the output of each model separate ly. Logistic regression. The flow chart of the logistic regression model is shown in Figure 7. It is a forecas ting method that uses a collection of independent variables to predict the categorical dependent variable. Therefore the output must be discrete or categorical in nature. True or false, 0 or 1, true or false, etc. may be possible, but the value to be given is between 0 and 1. Logistic regression and linear regression are sim ilarly used in a positive sense. The classification problem was solved using LR, and the regression proble m was solved using linear regression. We use the sigmoid logistic function instead of the regression line t o estimate the two maximum values (0 or 1). Evaluation matrix. Figure 8 shows the confusion matrix or e valuation matrix. Confusion matrix is a tool to evaluate the performance of machine learning classificatio n algorithms. Confusion matrices were used to evaluate the performance of each design. The confusion m atrix shows how much our model predicted correctly and how much it predicted incorrectly. Negatives an d negatives are given incorrectly predicted values, while true positives and negatives are given expected v alues. After placing all predictions in the matrix, evaluate the model's performance using its accuracyThis section examines the model's capabilities, model predictions, findings, and final results. data display. The

International journal of basic and applied research

www.pragatipublication.com

ISSN 2249-3352 (P) 2278-0505 (E)

Cosmos Impact Factor-5.86

histogram shows the recursive distribution with infinite sets. It is a square expression field with its base at the class boundary and an area proportional to the comparison class frequency. Square because it fills the center of the center class boundary

1. Result Analysis

The models' capacities, model forecasts, investigation, and eventual outcomes are examined in this part.

1.1. Data Visualization. A histogram depicts a recurrence dispersion with infinite classes. It is a region outline made of square shapes with bases at class boundary spans and regions proportionate to the comparing classes' frequencies. As the base fills in the spaces between the class borders, the square

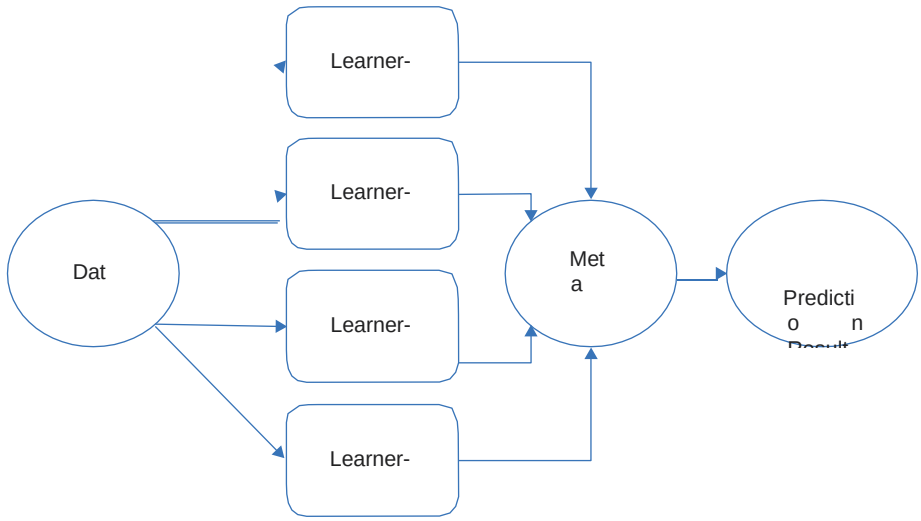


Figure 6: Flowchart of a voting classifier.

International journal of basic and applied research
www.pragatipublication.com
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

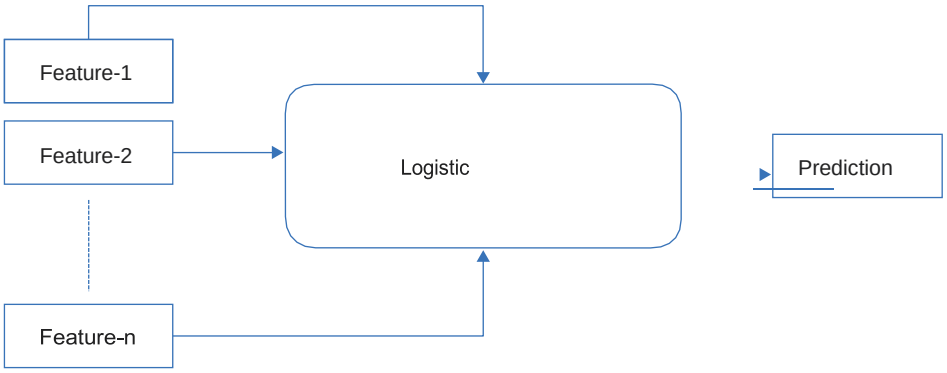


FIGURE 7: Structure of a logistic regression classifier.

The pictures are all connected to each other. The height squares are proportional to the compa rison category frequencies and the occurrence densities of different categories. Figure 9 show s some important features of histograms. Histograms show proportions of datasets. In the gen der feature, 0 represents male and 1 represents female. There are more female models than m ale models in this set. However, when looking at the age distribution, it is clear that the avera ge age of the model is around 40 years old, while the upper limit is around 60 years old. All h ealthy individuals without a history of heart disease were included in this study. Regarding B MI and average blood glucose, Figure 10 shows the relationship between one and the target f eature. Married and stroke, average blood sugar and stroke, BMI and stroke. Visualization of feature selection. The feature selection process is shown in Figure 11. However, gender was a ssociated with stroke. Evaluation of the model

3.3.1. Random Forest (RF). Figure 12 shows the distribution map of the RF model. The F1 sc ore of healthy individuals was 96%, and the F1 score of paralyzed patients was 96%. The mo del has been optimized to achieve maximum accuracy. Before correction, the model was 92% accurate. The prediction results and performance of the computational models are shown in t he confusion matrix. There were 2,707 correct guesses and 113 incorrect guesses.

International journal of basic and applied research

www.pragatipublication.com

ISSN 2249-3352 (P) 2278-0505 (E)

Cosmos Impact Factor-5.86

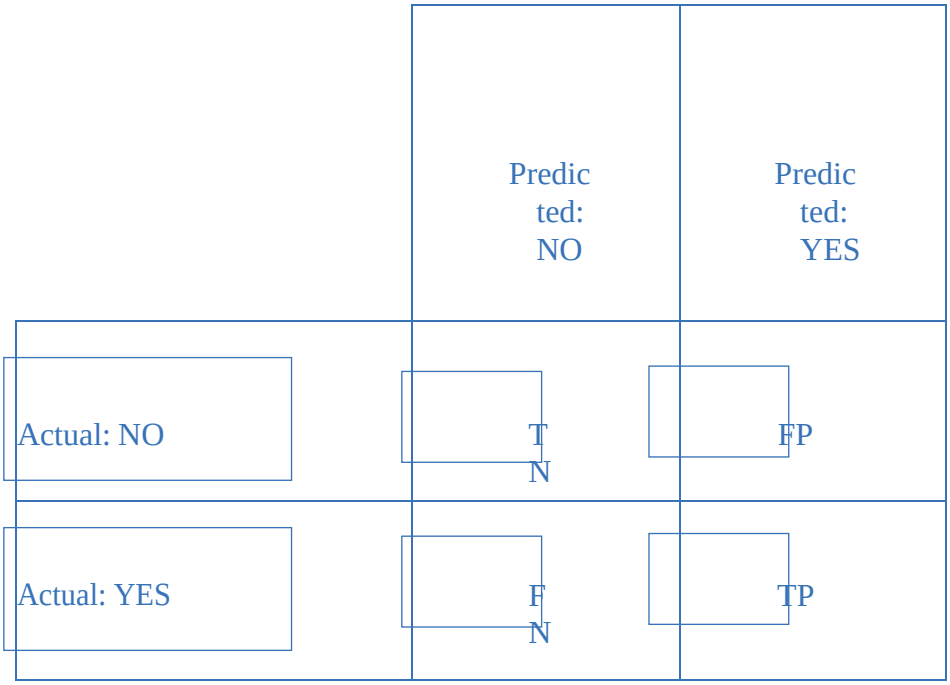


FIGURE 8: Block diagram of confusion matrix.

REFERENCES

[1] Sachin Kumar, Durga Toshniwal, "A data mining approach to characterize road accident locations", J. Mod. Transport. 24(1):62–72..

[2] Tessa K. Anderson, "Kernel density estimation and Kmeans clustering to profile road accident hotspots", Accident Analysis and Prevention 41,359–364.

[3] Shristi Sonal and Saumya Suman "A Framework for Analysis of Road Accidents" Proceedings of International Conference on Emerging Trends and Innovations in Engineering and Technological Research (ICETIETR)

[4] Analysis of road accidents in India using data mining classification algorithms- E. Suganya,S. Vijayrani.

[5] Hao, W., Kamga, C., Yang, X., Ma, J., Thorson, E., Zhong, M., & Wu, C., (2016), Driver injury severity study for truck involved accidents at highway-rail grade crossings in the United States, Transportation research part F: traffic psychology and behavior, 43, 379-386.

International journal of basic and applied research

www.pragatipublication.com

ISSN 2249-3352 (P) 2278-0505 (E)

Cosmos Impact Factor-5.86

[6] Li, L., Shrestha, S., & Hu, G., (2017), Analysis of road traffic fatal accidents using data mining techniques, In Software Engineering Research, Management and Applications (SERA), IEEE 15th International Conference on (pp. 363-370). IEEE.

[7] El Tayeb, A. A., Pareek, V., & Araar, A. (2015). Applying association rules mining algorithms for traffic accidents in Dubai. International Journal of Soft Computing and Engineering.

[8] Bahram Sadeghi Bigham ,(2014),ROAD ACCIDENT DATA ANALYSIS: A DATA MINING APPROACH, Indian Journal Of Scientific Research 3(3):437-443.

[9] Divya Bansal, Lekha Bhambhu, "Execution of Apriori algorithm of data mining directed towards tumultuouscrimes concerningwomen", International Journal of AdvancedResearch in Computer Science and Software Engineering, vol. 3, no. 9, September 2013.

[10]S. Krishnaveni, M. Hemalatha, "A perspective analysis of traffic accident using data mining techniques", International Journal of Computer Applications, vol. 23, no. 7, pp. 40-48, June 2011.