International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 06 | June 2021 www.irjet.net p-ISSN: 2395-0072

Stroke Type Prediction using Machine Learning and Artificial Neural Networks

Ms. Gagana M [1], Dr. Padma M C[2]

[1]*Final year PG Student, Department of Computer Science and Engineering, PES College of Engineering, Mandya, Karnataka, India.*

[2]*Professor, Department of Computer Science and Engineering, PES College of Engineering, Mandya, Karnataka, India.*

Abstract - *Today there are plentiful amount of data in case of several diseases in medical science sector. Thereby, it's a great source for using them for physicians to examine and analyze the causes of diseases in early stages. However, among such diseases stroke is one of the diseases which need attention from doctors to treat it at very early stages. Generally, stroke predictive techniques have been widely implemented in clinical decision making. We propose a system that uses machine learning algorithm and artificial neural networks to predict the presence and sub types of strokes. After conducting experiment we try to compare both the techniques and analyze their efficiency. The proposed system can be utilized in clinical predictions to save the time since it has robust features in real time. The results of this research are accurate and efficient than typical systems which are currently in use for treating stroke diseases.*

Key Words: **Stroke prediction, Machine learning, Artificial Neural Networks, Naïve Bayes and Comparative Analysis**

1. INTRODUCTION

Stroke occurs when the blood flow is restricted veins to the brain. When brain cells don't get enough oxygen and nutrients they eventually die within minutes. Stroke ranks as one of the major causes of death everyday globally. The trauma of stroke can be overwhelming to individuals and their families, stealing them of their freedom. It causes disabilities in adults and may even lead to demise of individuals who are not treated at early stages. Stroke disease may lead to other complications like cardiac arrest, loss of vision, inability to walk and loss of speech, etc.,

Each year 1.8million people suffer from stroke in Indian, but our country has only about 2000 neurologists. Following and implementing the typical clinical methods can be challenging in such circumstances. Despite the economic growth in the country, a huge proportion of population in India lies in poverty. Even the treatment option available for stroke disease are a very fewer in developing nations like India. Well organized affordable treatment and emergency transport treatments services are still absent.

Since India is a developing country, with diverse culture, significance economic growth has happened in the recent years including a boom in the Information Technology industry. Many researchers and scientists have conducted several experiments on stroke by implementing sophisticated techniques such as data science, machine learning, deep learning and artificial intelligence. During our research we found that many life style factors such as smoking, alcohol consumption, food/diet in take habits and routine of people affects prominently to the cause of stroke diseases. Early diagnosis of stroke disease based on these parameters can confront death. Data science and artificial intelligence plays the important roles in stroke prediction. The past data is large and complex. They can't be handled by typical analysis methods since it can be in different forms with no uniformity. This massive flow of data can be handled better by sophisticated machine learning and data science models to learn from the past experience and predict the accurate outcomes.

2. RELATED WORKS

Based on the survey carried out, we unveil all the prominent works conducted on stroke disease in this section. Much advancement in rehabilitation and imaging has taken place in the area of stroke prediction. One of the technologies that look promising is electrical cell simulation, which helps to increase the reduced blood flow to the brain. Along with the technologies mentioned earlier, there are much advancement in science that enables the patients to have wearable smart devices to monitor heart rate, blood pressure and other body parameters reading to predict the risk of stroke. Systematic investigations and studies have been carried out utilizing booming technologies of Information Technology which are data science, Machine learning and Artificial Intelligence. Those works and researches have made a tremendous impact on the areas of medical science for disease predictions. We shall explain all of those significant works here in three sections;

2.1 Data mining techniques

In the past decade, many data mining techniques have been implemented in the area of medical science to predict various diseases. Stroke is one of the major area of interest where researchers have utilized many data mining tools and classification techniques.

Leila Aminiet et al, used data mining techniques with K-Nearest Neighbor algorithms and c4.5 Decision trees for classification. To observe the effects they used WEKA tool.

Balar Khalid et al introduced a system for Ischemic stroke prediction utilizing classification techniques, logistic

regression. For pre-processing of data, they made use of WEKA and C4.5 algorithm. Microsoft XLSTAT was used for examine the sample data collected.

Ahmet Kadir et al introduced a stroke detection model using data mining techniques – Stochastic Gradient Boosting (SGB), support vector machine and Penalized Logistic Regression (PLR). The results showed SVM techniques outperformed other techniques.

2.2 Machine Learning Algorithms

Machine learning technology has impacted very positively on stroke prediction. There have been many machine learning approaches used for experiments effectively to predict the diseases at early stages.

Adithya Khosla et al, showed that novel automatic feature selection algorithm approach outperform the typical clinical methods used for stroke detection. They compared Cox proportional hazards system against machine learning methods. They attained greater AUC and concordance index.

Sudha A et al, utilized classification algorithms such as Naïve Bayes, Decision trees and neural networks for predicting the risk and parameters contributing for stroke. They concluded that decision trees showed more promising results in classification with 98.01% accuracy.

Hamed Asadi et al performed a study on Ischemic stroke. They implemented SVM along with classical statistics to classify many predictors into potential end results and poor end results. They used MATLAB, Rapid miner and SPSS tools in the process of constructing the stroke prediction system. Finally, they introduced vigorous machine learning system which possibly optimized the preference process for medical treatment or endovascular treatment of Ischemic stroke.

2.3 Artificial Intelligence and Deep Learning

AI and Deep Learning are emerging technologies in this era. Many researches and clinical studies are being conducted using these technologies.

Chiun-li-chin et al introduced an automated system for the early prediction of Ischemic stroke. They implemented convolution neural networks, deep learning algorithms for analyzing the called CT images to identify the risk of stroke. They used data augmentation method to enhance the missing patches in the images, using CNN technology.

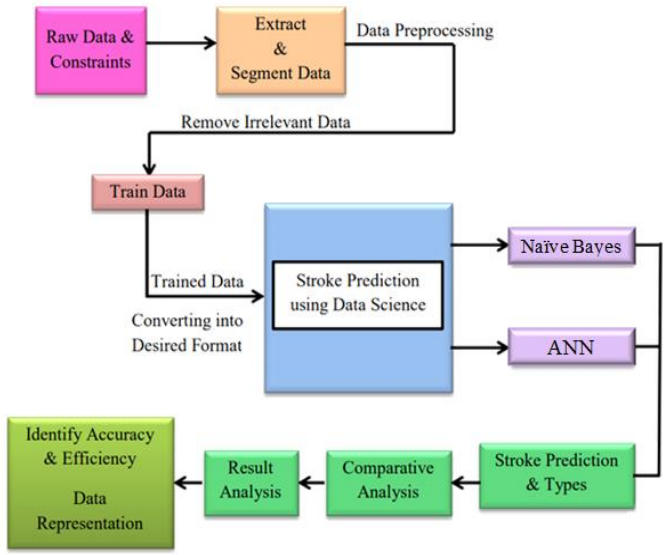
Aishwarya Roy et al, constructed the stroke prediction model using AI decision trees to examine the parameters of stoke disease. They used confusion matrix for producing the results. The system produced 95% accuracy.

Joon Nyung Heo et al built a system that identifies the outcomes of Ischemic stroke. They implemented deep neural networks, logistic regression and Random forest techniques in their system. After the results obtained, they did comparative analysis that showed higher AUC curve for Deep Neural Networks model.

3. METHODOLOGY

For developing the model, we focus on technologies to automate the stroke prediction system. This is achieved by building the application using Machine Learning algorithm and Artificial Neural Networks. We make use of approximately 5000 dataset obtained from UCI Machine Learning website. Around 18 to 19 parameters like, alcohol consumption, family history, lifestyle, red blood cells count, smoking history, work type and other health related data are being used as input to the system to predict the different types of stroke.

Based on these inputs, the model predicts three different types of strokes, which are Ischaemic stroke, Hemorrhagic stroke and Transient Ischaemic stroke. The model is designed for four types of users namely; Administrator, Receptionist, Doctor and Patients. The main goal of this proposed model is to eliminate the complexities for doctors in stroke prediction. This application can be used



in real time to treat patients in early stages without the necessity of any sophisticated medical equipment.

Figure 1: Proposed Model

3.1 Naïve Bayes

It is based on Bayes theorem from the family of probabilistic classifiers. We implemented this in our methodology as it has high accuracy in disease predictions on selected conditional variables. Naïve Bayes algorithm is very quick compared to other machine learning algorithms as it shows the promising results on thousands of dataset which are used for pre-processing. Naïve Bayes assumes the existence of the specific attribute in a class of data which is independent to the existence of any other attribute. It is uncomplicated to build and very useful for large datasets.

Bayes theorem allows a path for calculating posterior probability p(c|x) from p(c), p(x) and p(x|c). The equation is as follows;

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

Where:
P(c|x) = Posterior Probability
P(x|c) = Likelihood
P(c) = Class Prior Probability
P(x) = Predictor Prior Probability

$$P(c|X) = P(x_1|c) \times P(x_2|c) \times \ldots \times P(x_n|c) \times P(c)$$

Naïve Bayes performed well in case of absolute input variables compared to numerical variables in real time prediction.

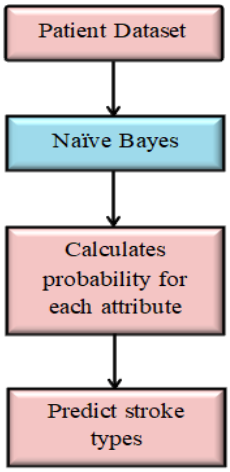


Figure 2: Naïve Bayes Classifier

3.2 Learning Vector Quantization

The Learning Vector Quantization (LVQ) algorithm comes under Artificial Neural Networks (ANN). It lets us to select how many training instances to hold onto and gain an understanding on what exactly those instances should be like. It provides two-class and multi-class classifications to the real world problems. If the problem is a two-class classification, then it is represented by 0 and 1.

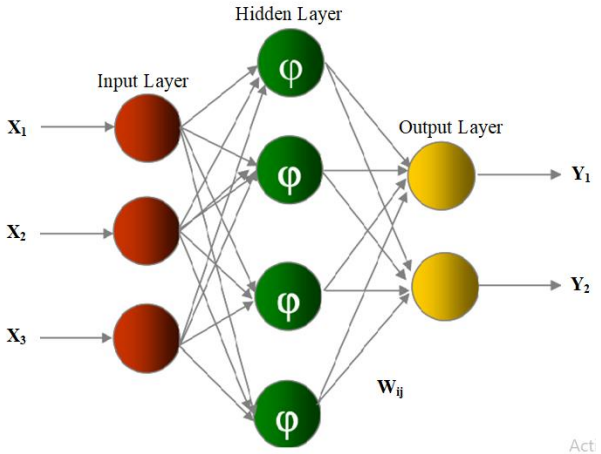


Figure 3: Artificial Neural Networks

It has input layer, hidden layer and output layer. The input layer takes the 18 parameters as input which is the attributes of the disease attached to hidden layers. The hidden layers are inter-connected to the output layer unit. This structure uses the supervised learning method for training and testing the dataset. It works on Euclidean distance and finds the new weights based on the old weights (past learning).

$$D(j) = \sqrt{\sum_{i=1}^{n} (x_i - w_{ij})^2}$$

Where;
D(j) = Distance between two data points
n = number of dimensions
w_{ij} = new weight

3.3 Dataset Collection

The dataset is collected from UCI Machine Learning repository. Around 5000 records are collected which had 18 parameters. The dataset is based on stroke diagnosis of past patients. The parameters includes family history, work type, alcohol habits, smoking, WBC count, blood pressure, age, gender, BMI, heart disease and others. The dataset are error free and their analysis helps in realizing the type of stroke disease.

3.4 Performance Evaluation

Naïve Bayes and Artificial Neural Networks are constructed from the given set of parameters. Naïve Bayes works on probability statistical classification, for each feasible value in the selected/desirable range. The presence of a particular feature in a category is not to the presence of any other feature. Naïve Bayes works on conditional probability since it is built upon Bayes theorem. When the patient details or parameters are given as input, it analyzes all these parameters and finds out the type of stroke disease. Here, we have divided the data into testing dataset and training dataset. Naïve Bayes takes the testing data as input and compares it with the training dataset by analyzing their parameters to give the maximum a posterior, which is used to normalize the result.

On the other hand, Artificial Neural Networks utilizes neural computations. It creates the bridge between input layers and output layers through the hidden layers. In neural networks, codebook vectors are numbers that have some input and output as training dataset. Each codebook vector is known as neuron, each parameter on codebook vector is known as weight and the group of codebook vectors is known as a network. The predictions made by these neural networks are quite similar to the way of K-Nearest Neighbours. The ANN work on basis of Euclidian distance by calculating new weights from the old weights which are 18 parameters are given as input, it calculates Euclidean distance then it analyzes the dataset through hidden layers to select the nearest accurate feature. For each such feature vector, it computes the distance to every prototype using the selected distance measure. Then, it updates the weights of the vector to the closest prototype that has same feature as

the label. Once this is done for each testing dataset, the process repeats several times until in converges.

4. DISCUSSION/RESULT ANALYSIS

We observed that, Naïve Bayes outperformed ANN. Before running these algorithms on the testing dataset, we shuffled the dataset record randomly and kept aside 10% (about 499 dataset out of 4982 dataset) of the set as validation set.

The accuracy is measured on basis of the parameters and their values. Based on patient risk levels with diagnosis, result=0(for no stroke), result=1,2,3 for Ischemic, Haemorraghic and Transient Ischemic respectively.

Since, Naïve Bayes outperformed ANN, we considered Naïve Bayes as efficient algorithm to predict the stroke type in individual patients and added that to the Doctor module, where Doctor can key in the patient details and predict the outcomes of each individual. When we analyzed both the algorithms for their efficiency, we found several factors which influenced their outputs. Naïve Bayes works on probabilistic classification based on conditional probability.

The parameters directly influenced such that more the parameters given for learning better the output or result. And we found that it performed better and require less training dataset. But Naïve Bayes was not able to make predictions in scenarios where the test data set had a categorical variable of a class that was not present in the training dataset. In such situations, Naïve Bayes model assigned it as zero probability. In order to overcome this we have to use smoothing technique. Also, this algorithm assumes that all the attributes or features are independent. In theory it may sound great, but practically we hardly find a set of independent features.

On the other hand, ANN showed less accurate results when compared to Naïve Bayes, ANN works on Euclidean distance and new weight calculations. It learns from the past data and calculates new weights based on the distances among hidden layers. Its intuitive property yields decent results. But, Euclidean distance calculations can cause complications if the data has a lot of dimensions or too noisy for solving complex real time problems. We observed that, the number of parameters influences the outcomes and dimensionality of the hidden layers. Unlike Naïve Bayes, ANN computed distance among hidden layers and selected the nearest or closet prototype. Also, it was observed that ANN showed decent fault tolerance, even though a few data values were missing in the testing dataset is distributed to the entire network.

However, this leads to much computational burden and over fitting. Neural networks in general, show much accuracy in pattern recognition, image processing and clustering. They are less accurate in analyzing textual dataset when compared to images. Also, we found that neural networks processed information faster and had a very less response time when compared to Naïve Bayes model. Neural networks were adaptive, as it updated the weights and changed during training.

Comparative Analysis of Algorithms

Constraint	LVQ	Naive Bayes
Accuracy	91.7995991983968%	99.1983967935872%
Time (milli secs)	95095	130371
Correctly Classified	91.7995991983968%	99.1983967935872%
InCorrectly Classified	8.20040080160321%	0.801603206412821%

Figure 4: Comparative Analysis

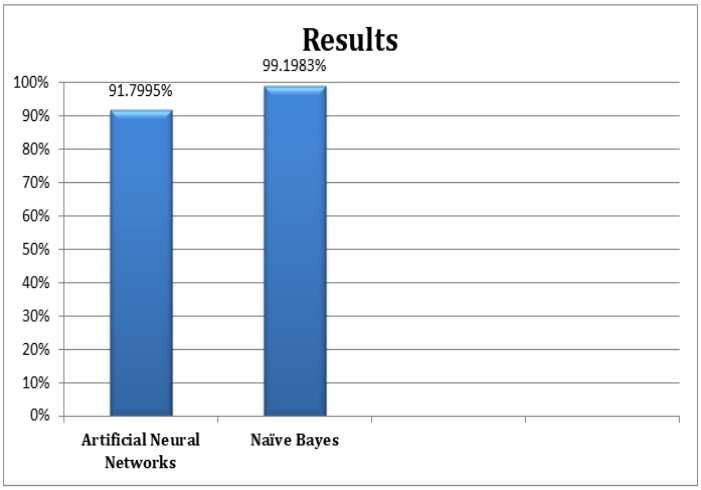


Figure 5: Graphical representation of results

5. CONCLUSION

The Stroke Prediction demands the CT scan and can be diagnosed only by looking at the brain images produced by typical methods. This has posed a challenge to the health care sector since the stroke is hard to diagnose at the early stages by identifying the symptoms. Hence, many scientists and researchers have investigated about the cause of strokes and have conducted many experiments using Machine Learning and Data Science techniques.

In our study we make use of Naive Bayes and Artificial Neural Networks to predict the three different sub-types of strokes which are Ischaemic, Hemorrhagic and Transient ischemic types. After optimization of the test data sets from the training data set it was observed that Naive Bayes Algorithm showed up more accuracy in prediction with 99.1983 % approximately when compared against Artificial Neural Networks which showed only 91.7995% accuracy.

6. DIRECTIONS FOR FUTURE WORK

SMS/Email Module – In the proposed system, the admin creates ID and password credentials for doctors and receptionists. For enhancing the application in the future days we can turn this manual process into an automated process by buying email hosting services or deploying on cloud platforms using its integrated SNS services.

Query Interaction Module - we can add the interactive query module in the days ahead into the application by using Chabot's or form filling where doctors, receptionists, and admin of the application can interact one to one.

REFERENCES

[1] Luis Garcfa-Terriza, Risco Martin, Ayala and Gemma Reig Rosello, "Comparison of different Machine Learning approaches to model stroke subtype classification and risk prediction", Society for Modeling & Simulation International (SCS), 2019 April 29-May2.

[2] JoonNyung Heo, Jihoon G. Yoon, Hyungjong Park, Young Dae Kim, Hyo Suk Nam, Ji Hoe Heo, "Machine Learning-Based Model for Prediction of Outcomes in Acute Stroke", 2019 February 1, doi:10.1161/strokeaha.118.024293

[3] Aishwarya Roy, Anwesh Kumar, Navin Kumar Singh and Shashank D, "Stroke Prediction using Decision Trees in Artificial Intelligence", IJARIIT, Vol. 4, Issue 2, 2018, pp: 1636-1642.

[4] Chiun-Li-Chin, Guei-Ru Wu, Bing-Jhang Lin, Tzu-Chieh Weng, Cheng-Shiun Yang, Rui-Cih Su and Yu-Jen Pan, "An Automated Early Ischemic Stroke Detection System using CNN Deep Learning Algorithm", IEEE 8th International Conference on Awareness Science and Technology, 2017.

[5] Ahmet Kadir Arslan, Cemil Colak, Mehmet Ediz Sarihan, "Different medical data mining approaches based prediction of ischemic stroke", Elsevier, Computer Methods and Programs in Biomedicine 2016 March 18.

[6] Balar Khalid and Naji Abdelwahab, "A model for predicting Ischemic stroke using Data Mining algorithms", IJISET, Vol. 2 Issue 11, Nov 2015, ISSN: 2348-7968.

[7] Hamed Asadi, Richard Dowling and Bernard Yan, "Machine Learning for outcome prediction of acute ischemic stroke", PLOS ONE Vol. 9 Issue 2, Feb 2014.

[8] A. Sudha, P. Gayathri, "Effective analysis & predictive model of stroke disease using classification methods", IJCA(0975-8887), Vol. 43-No. 14, April 2012.

[9] Leila Amini, Reza, Rasul Norouzi & Associates, "Prediction and Control of Stroke by Data Mining", IJPM, 8th Iranian Neurology Congress, Vol. 4, 23 Feb 2013.

[10] Aditya Khosla, Yu cao, Honglak Lee & Associates, "An integrated machine learning approach to stroke prediction", 25-28 July 2010, Washington, DC, USA.

BIOGRAPHIES

Gagana M
Final Year PG Student,
Department of Computer Science and Engineering,
PES College of Engineering,
Mandya, Karnataka, India.

Dr. Padma M C
Ph.D. in Pattern Recognition & Image Processing.
Professor,
Department of Computer Science and Engineering,
PES College of Engineering,
Mandya, Karnataka, India