Early Stroke Prediction Using Machine Learning

[1]Chetan Sharma, [2]Shamneesh Sharma, [3]Mukesh Kumar, [4]Ankur Sodhi
[1]Chitkara University, Himachal Pradesh (INDIA)
[2]UpGrad Campus, upGrad Education Private Limited (INDIA)
[3]School of Computer Application, Lovely Professional University, Phagwara, Punjab (INDIA)
[4]upGrad Education Private Limited (INDIA)

chetanshekhu@gmail.com, shamneesh.sharma@gmail.com, mukesh.27406@lpu.co.in, ankursodhi@gmail.com

2022 International Conference on Decision Aid Sciences and Applications (DASA) | 978-1-6654-9501-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/DASA54658.2022.9765307

Abstract: Stroke is one of the most severe diseases globally, and it is directly or indirectly responsible for a considerable number of deaths. A variety of data mining techniques are employed in the health care industry to aid in diagnosing and early detection of illnesses. Several elements that lead to stroke are considered in the current investigation. First, we're looking into the characteristics of those who are more likely to suffer from a stroke than others. The dataset is obtained from a freely available source, and multiple classification algorithms are used to predict the occurrence of a stroke shortly. By employing the random forest algorithm, it has been possible to obtain an accuracy of 98.94 percent. Finally, various preventative steps such as quitting smoking, avoiding alcohol, and other factors are recommended to reduce the risk of having a stroke.

Keywords: Machine learning, Stroke, Classification, Supervised Learning, Data Mining

1. INTRODUCTION

Human life is based on various body parts and their functionality. The heart is considered the essential part of the body that pumps blood to other organs. Stroke is a dangerous disease that causes to end up human life. This disease is commonly found after the age of 65 years. Stroke affects the brain as heart attack, which affects the heart's working. Strokes are caused by the rupture and bleeding of brain blood vessels or the restriction of blood supply to the brain, resulting in one of these two conditions. Blood and oxygen cannot reach the brain's tissues if there is a rupture or a blockage. It is currently the third leading cause of death in developed and developing countries alike[1]. Strokes are caused by either the rupture and bleeding of a blood artery in the brain or the blood flow restriction to the brain. As blood and oxygen are blocked from reaching the brain's tissues, the brain is damaged. Today, it is the third leading cause of death in developed and developing countries.

A blood clot in a transient ischemic attack (TIA) is a sign of the condition.

Ischemic stroke occurs when a blood clot or plaque accumulation blocks an artery, resulting in a loss of blood flow to the brain. As opposed to a transient ischemic attack (TIA), symptoms and effects of an ischemic stroke can remain much longer and can be permanent.

A hemorrhagic stroke ensues whenever a blood vessel ruptures or leaks into the brain.

In addition to the cost of medical care and chronic disability, a stroke can also lead to death. The brain's tissues are damaged when blood flow to the brain is disrupted. When a segment of the nervous system is injured, symptoms might be seen in the bodily parts controlled by that region. The sooner a stroke victim receives medical assistance, the better their chances of making a full recovery. Consequently, you must know the warning signs of a stroke so that you can act quickly. Stroke symptoms include the following:

- Numbness or weakness in the arm, face, and leg
- Problem in communicating
- Problem with response
- Change in Behaviour
- Problem with vision
- Trouble in walk
- Dizziness
- Headache
- Nausea or vomiting

A stroke requires immediate medical attention. Otherwise, it will cause brain damage, long-term disability, and death. Human condition after stroke depends upon the type of stroke, and it is categorized in three categories:-

1. Transient Ischemic Attack (TIA)
2. Ischemic Stroke
3. Hemorrhagic Stroke

When blood flow to the brain is temporarily disrupted, it's called a transient ischemic attack (TIA). It is common for patients to recover from this type of stroke within a few minutes. The most common cause of TIA is blood clots, and it serves as a warning sign for the person experiencing it. Centers for Disease Control and Prevention (CDC) data shows [2], Within a year, a stroke occurs in one-third of those who have suffered a transient ischemic attack (TIA).

During an ischemic stroke, the blood vessels carrying oxygen and nutrients to the brain constrict or close completely. Because of blood clots and particles of plaque that were broken, blood vessels were obstructed. The Centers for Disease Control and Prevention (CDC) say that [2], The ischemic form of stroke affects 87% of stroke victims. A hemorrhagic stroke occurs when a brain artery ruptures and blood pours out. Damage to brain cells and tissues occurs when the artery's blood pressure is greater than the skull's. Based on the American heart association [3], 13% of strokes are hemorrhagic type.

There are various reasons due to which patients got stoke. According to National Heart, Lung, and blood institute [4], diet, inactivity, tobacco, alcohol, personal history, health history, and complications are the major factors that lead the patient to stroke. Today people's diet is unbalanced in which people are taking highly salted, saturated fats, trans fats, and cholesterol food items. Physical activity is deficient; as suggested by the center for disease control and prevention [2], people have to do 2.5 hours of exercise, but stroke risk is very high due to lack of exercise. Apart from diet and exercise, people are more fond of alcohol and tobacco products. Heavy use of alcohol and tobacco is one of the significant reasons for stroke in people. People's personal history also plays a vital role in predicting the cause of stroke. In personal history, sex, family history in stroke, age, and geographical area are the significant areas which play a vital role in identifying the stroke patient. The health history of the patient is also playing an important role. There are certain medical conditions of the patients that may lead to stroke. Medical conditions like TIA history, high blood pressure, high cholesterol, increased body weight, heart valve defect, diabetes, and other diseases may lead patients to stroke.

Machine learning is the backbone of the modern era, used to predict various problems in earlier stages. For example, multiple diseases can be prevented if predicted early, as stroke is one of the major diseases that can be cured if expected in the early stage. As a whole, machine learning is essential in the health care industry when it comes to disease prediction and diagnosis. A large amount of medical data need robust data analysis techniques. Hospitals store the amount of data in patients' medical records is constantly increasing.

On the other hand, medical researchers will find a wealth of information in these datasets. Medical decision-making, which can be highly challenging for various reasons, particularly in diseases with similar symptoms or rare diseases, is at the center of our research project. Artificial Intelligence (AI) in medicine is a significant focus of research in this area. The patient's data would be analyzed by an artificial intelligence system, providing a set of relevant forecasts. The technology can learn from a patient's medical history and predict which patients are at risk of developing the condition. The technology can estimate the likelihood of developing an illness by looking at their medical profiles, including age, blood pressure, sugar levels, and more. Classification algorithms are used to predict disease when there are many variables.

For complex inquiries, each type of machine learning could offer advantages in terms of model interpretation, access to a wide range of information, as well as accuracy. After a thorough review of relevant and unrelated studies, machine learning outperforms other categorization algorithms in terms of accuracy. Therefore, it was necessary to compare the various algorithms for classification used in this study to discover the optimum method for stroke prediction.

2. PROPOSED METHODOLOGY

Using the procedures mentioned in this section, the proposed work will be carried out as stated. It is shown in Figure 1 that data collection, pre-processing, feature selection, classification, and analysis are used in this work. Detailed explanations of each step can be found in the following sections.

2.1 Dataset

This study's dataset for stroke prediction was obtained from a publicly accessible site [5]. The data in this set pertains to strokes. When determining whether a patient is at risk for a stroke, this dataset considers factors such as gender, age, numerous diseases, and smoking status. Each entry in the data table provides essential information about the patient's condition. Table 1 shows the dataset's features, as well as their description.

Feature	Description
ID	It belongs to the unique number provided to the patient
Gender	Male = ‘0’ Female = ‘1’ and Other =’3’
Age	Age is mentioned in years, which belongs to the patient's age.
Hypertension	Hypertension = ‘1’ and No Hypertension = ‘0’
Heart Disease	No Heart Disease = ‘0’ and Heart Disease = ‘1.’
Marital Status	It is represented by Yes or No
Work Status	Children = ‘0’ Government Job=’1’ Never Worked=’2’ Private=’3’ Self Employed=’4’
Residential Area	Rural =’0’ Urban = ‘1’
Average Glucose Level	It is represented in Numeric
Body Mass Index	It is represented in Numeric
Smoking Status	Formerly Smoked=’0’ Never Smoked=’1’ Smokes=’2’ Unknown=’3’ (No Information)
Stroke	It is considered the target variable in which 0 is regarded as no stroke, and 1 represents stroke.

The dataset contains 5110 patients record, and there are 12 attributes for each patient. In 12 features, 10 are the habit or activities related to the patient. This dataset is best for prediction as in the things

dataset, out of 5110 patients, 59% of the patient record is female, and 41% are male.

2.2 Data Pre-Processing

This process is carried out to clean the data. The real-world data contains many redundant values and significant noise [6]. Therefore, most data we see in real life have duplicate, redundant, or missing values. When this type of data is used to construct a model, it produces incorrect results. Consequently, data must be cleaned before being fed into the model to give more accurate results. Duplicate and redundant data values should be removed from the data set throughout the cleaning process. In addition, it is necessary to correct any missing values. The data is changed after preprocessing and after the noise and missing values have been eliminated. The modified data can now be utilized to generate the classification model, which is the final step

2.3 Classification Algorithm

After going through the available techniques used for classification for early stroke detection, The author has considered supervised learning algorithms like decision trees, random forest, and naïve Bayes algorithms to check for the features that contribute to detecting stroke.

- **Naive Bayes Classification**

In machine learning, naive Bayes is a type of supervised learning based on the naive theorem and is a work-based learning method. The Naive Bayes algorithm is based on the assumption that the existence of a feature/parameter does not affect the presence of another feature, i.e., that the fact of one feature is independent of the existence of the others. When it comes to mathematics, the Bayes theorem is a conditional probability theorem that determines the likelihood that a particular event will occur under the premise that a specific condition has previously been met. Conditional probability is contrasted with the help of the Bayes theorem [7][8]; it is the likelihood that a specific event has occurred, given the assumption that some event has already happened. For example, P(y|z) can be derived from P(y), P(z), and P(z|y), among other things. The following is the formula for calculating the posterior probability:

$$P(y|z)=P(z|y)*P(y)/P(z|y) \quad (1)$$

Where:

P(y|z) is defined as the conditional probability which occurs when x has already happened.

P(z) is defined as the known probability of the class.

P(z|y) is defined as the conditional probability of x condition in which c has occurred.

P(y) is defined as the known probability of the class.

This method gains importance in the current work since we are examining if the existence of the traits is independent of each other and the presence/absence of one or more is contributing to the development of diabetes in a given individual.

- **Decision Tree**

It is again a supervised learning algorithm that creates a model to predict the class of a target variable based on the features of the target set. The algorithm assumes that the presence or absence of a particular feature in the dataset depends on each other and contributes to classifying the target towards a specific class [9][8].

- **Random Forest**

Ensemble learning techniques like Random Forests (also known as random decision forests) can solve classification and regression issues. They function by training numerous decision trees in a distributed manner. A random forest's output is the class most trees choose while dealing with classification issues. When it comes to the specifics of ensemble learning, which is the practice of using multiple classifiers to solve complicated problems and increase the performance of a model, it is built. According to the Random Forest classifier's name, "combining a large number of decision trees on different subsets of a given dataset and taking an average to enhance the projected accuracy" The random forest considers the forecasts from all of the trees as opposed to just one. The most popular forecasts are used to determine the outcome [10].

- **Multi-layer Perceptron Algorithm**

The Multi-layer Perceptron (MLP) differs from a linear perceptron in that it has multiple layers and does not activate linearly. Therefore, it can separate data that is not linearly separable, among other things. Multilayer Perceptrons (MLPs) are feedforward artificial neural networks that generate outputs from a collection of inputs. A directed graph combines the input and output layers of an MLP's input nodes to form a single unit known as a directed graph unit. MLP uses backpropagation to train the network and improve its performance. MLP is a deep learning method that uses machine learning [11]. MLP is classified as a deep learning technique since it uses multiple layers of neurons. MLP is frequently utilized in computational neuroscience and parallel distributed processing research. It is also commonly used for supervised learning tasks, so it is popular. Speech and image recognition and machine translation are just a few of the many applications available today.

- **JRip Algorithm**

This class represents this as the implementation of a propositional rule learner. This method, which is an optimized variant of the IREP, was devised by William W. Cohen. It is named after him. It performs Repeated Incremental Pruning to reduce the number of errors (RIPPER). With its bottom-up approach to rule identification and learning, JRip classifies instances in the training data into groups and then discovers the set of rules that apply to all members of each group. Techniques such as cross-

validation and having a specific quantity of words are employed to avoid overfitting the model[10].

3. IMPLEMENTATION

The dataset is downloaded from the open-source [5] in this study, freely available on public platforms. In this experiment, python language is used to implement various classification algorithms shown in Table 1, Table 2. Table 1 represents the results when 70% of data is used for training, and 30% of information is used for validation, while in Table 2, 10 fold cross validation is used.

Table 1: Accuracy achieved for Stroke Prediction Dataset using 70-30 Ration

Classification Algorithms	Accuracy	Precision	Recall	F-Measure
Decision Tree Algorithm	95.28%	0.957	0.953	0.953
Random Forest Algorithm	98.63%	0.987	0.986	0.986
Naïve Bayes Algorithm	75.57%	0.761	0.756	0.756
Multi-layer Preceptron Algorithm	79.10%	0.791	0.791	0.791
JRip Algorithm	93.68%	0.945	0.937	0.937

Table 2: Accuracy achieved for Stroke Prediction Dataset using 10 Fold Cross-Validation

Classification Algorithms	Accuracy	Precision	Recall	F-Measure
Decision Tree Algorithm	96.15%	0.965	0.962	0.962
Random Forest Algorithm	98.94%	0.990	0.989	0.990
Naïve Bayes Algorithm	75.74%	0.764	0.757	0.758
Multi-layer Preceptron Algorithm	80.80%	0.814	0.808	0.809
JRip Algorithm	94.94%	0.955	0.949	0.950

From the above accuracy achieved, it is concluded that 98.94% is the highest accuracy achieved in the experiment for random forest algorithm using ten-fold cross validation. To enhance or to see the critical feature ranker method is applied to the dataset. After applying the ranker algorithm for feature selection, two features named gender and residence type do not affect the accuracy; these two features are removed from the dataset. Table 3 represents the accuracy achieved after removing the features that have no high impact on stroke prediction.

Table 3: Feature Selection with Ranker Method using Information Gain Attribute Evaluator

Classification Algorithms	Accuracy	Precision	Recall	F-Measure
Decision Tree Algorithm	96.45%	0.963	0.960	0.960
Random Forest Algorithm	98.94%	0.990	0.989	0.990
Naïve Bayes Algorithm	77.80%	0.770	0.767	0.768
Multi-layer Preceptron Algorithm	83.70%	0.844	0.821	0.819
JRip Algorithm	93.94%	0.945	0.939	0.940

4. RESULTS AND DISCUSSION

In this experiment, the author applied various classification algorithms for the early prediction of stroke. The random forest algorithm achieved the highest accuracy rate among all the classification algorithms. Using a 70-30 percent ratio, 98.63% accuracy is performed while applying ten-fold cross validation 98.94% accuracy is achieved. As some features do not play a vital role among all features so to see the low ranking feature ranker algorithm is applied and conclude that two features have low impact. Finally, 98.94% of accuracy is achieved using the random forest algorithm.

5. CONCLUSION

According to this study's findings, it is possible to forecast stroke prediction using historical data mining approaches. Three well-known classification algorithms, such as Decision Tree, Random Forest, Nave Bayes, Multilayer Perceptron, and JRip, were tested for their accuracy rates. The Information Gain Attribute Evaluator strategy was used to improve classification performance but concluded that the same accuracy rate had been achieved. Lifestyle changes can't avert all strokes. When it comes to reducing your risk of stroke, though, many of these modifications can make a significant difference. Quitting smoking today will minimize your stroke risk. To develop a strategy for quitting smoking, you can speak to your doctor. Alcohol use should be limited. Alcohol can elevate your blood pressure, putting you at greater risk of stroke. Reach out to your doctor if you find it tough to reduce your intake.

Maintain a healthy weight. They are at an increased risk of stroke because they are overweight or obese. To keep your weight in check, consume a well-balanced diet and engage in regular physical activity. Both methods can also help lower cholesterol and blood pressure. In addition, ensure that you have regular checks. The frequency of blood pressure and cholesterol checks will depend on your doctor's recommendations, so be sure to discuss this information with them. They can help you make these alterations to your way of life by offering encouragement and support. To prevent a stroke, you should do all of these steps.

REFERENCES

[1] NLM, "National Library of Medicine," 2022. https://pubmed.ncbi.nlm.nih.gov/19075105/ (accessed Jan. 29, 2022).

[2] CDC, "Center For Disease Control And Prevention," 2022. https://www.cdc.gov/ (accessed Jan. 29, 2022).

[3] A. S. Association, "American Stroke Association," 2022. https://www.stroke.org/en/ (accessed Jan. 30, 2022).

[4] L. and B. I. National Heart, "National Heart, Lung and Blood Institute," 2022. https://www.nhlbi.nih.gov/health-topics/stroke (accessed Feb. 01, 2022).

[5] Fedesoriano, "Stroke Prediction Dataset," 2021. https://www.kaggle.com/fedesoriano/stroke-prediction-dataset (accessed Jan. 28, 2022).

[6] C. Sharma, S. Shambhu, P. Das, and S. Jain, "Features Contributing Towards Heart Disease Prediction Using Machine Learning," 2021.

[7] P. Das, S. Jain, C. Sharma, and S. Shambhu, "Prediction of Heart Disease Mortality Rate Using Data Mining," 2021.

[8] P. Das, S. Jain, S. Shambhu, C. Sharma, and S. Ahuja, "Prediction of Diabetes Rate Using Data Mining," in *2021 International Conference on Decision Aid Sciences and Application (DASA)*, 2021, pp. 463–465.

[9] S. R. Safavian and D. Landgrebe, "A Survey of Decision Tree Classifier Methodology," *IEEE Trans. Syst. Man Cybern.*, vol. 21, no. 3, pp. 660–674, 1991, doi: 10.1109/21.97458.

[10] S. Gautam, C. Sharma, V. Kukreja, and others, "Handwritten Mathematical Symbols Classification Using WEKA," in *Applications of Artificial Intelligence and Machine Learning*, Springer, 2021, pp. 33–41.

[11] C.-F. Tsai, C.-T. Tsai, C.-S. Hung, and P.-S. Hwang, "Data mining techniques for identifying students at risk of failing a computer proficiency test required for graduation," *Australas. J. Educ. Technol.*, vol. 27, no. 3, 2011.