International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2024.0169
e ISSN: 2584-2854
Volume: 02
Issue: 05 May 2024
Page No: 1250-1253

# Heart Stroke Prediction Using Machine Learning Algorithms

*Thammaligadda Laxmiprasanna[1], Kolluri Hashwitha Teja[2], Gadila Lahari Nandhini[3], Dr. S Satheesh Kumar[4]*
*[1,2,3]Computer Science and Information Technology, Institute of Aeronautical Engineering, Hyderabad, Telangana, India.*
*[4]Associate Professor, ComputerScience and Information Technology, Institute of Aeronautical Engineering, Hyderabad, Telangana, India.*
***Emails:*** *20951a3319@iare.ac.in[1], 20951a3315@iare.ac.in[2], 20951a3318@iare.ac.in[3], s.satheeshkumar@iare.ac.in[4]*

## Abstract
*A Stroke is a disease when there is insufficient blood supply to the brain, which causes cell death. It is currently the world's biggest cause of death. Upon examining the affected individuals, a number of risk variables that are thought to be connected to the cause of stroke have been identified. Numerous studies have been conducted to predict and categorize stroke disorders using the risk variables. The majority of the models are built using machine learning and data mining technologies. In this work, we have employed four machine learning algorithms to identify the type of stroke that may have happened based on medical report data and an individual's physical condition. We have gathered a sizable amount of hospital entries. This study employs many methodologies, including decision trees, Naive Bayes, ANN algorithm, and Random Forest algorithm. Thus, the aim of this study is to evaluate the mentioned algorithms and determine which one does the task more accurately. After completing all of the evaluations, we can conclude that the Random Forest method has the highest accuracy of all the algorithms with 99%.*

***Keywords:*** *KNN, Naive Bayes, Random Forest.*

## 1. Introduction
Poor blood supply to the brain causes cell death, which is the cause of stroke. Hemorrhagic stroke and ischemic stroke are the two primary forms of stroke. Hemorrhagic stroke is caused by bleeding, and ischemic stroke is caused by a reduction in blood flow. Transient ischemic attack is another form of stroke. An embolic stroke happens when a clot forms elsewhere in the body, travels to the brain, and obstructs blood flow there. a thrombotic stroke brought on by a clot that impairs arterial blood flow. Another name for transient is chemic attack is "mini stroke Many individuals lose their lives. polynomial, quadratic, radial basis function and linear functions were applied. The highest accuracy of 91% was found with the linear kernel which gives the balance measure F1-score F-measure 91.7 [1]. Singh and Choudhary developed a model with Artificial Neural Network (ANN) for stroke prediction. They have collected datasets from the Cardiovascular Health Study (CHS) database. During feature selection, the C4.5 decision tree algorithm was used and Principle Component Analysis (PCA) for dimension reduction. In ANN implementation they have used Back Propagation learning method. They have got the accuracy as 95%, 95.2% and 97.7% for the three datasets respectively [2, 3]. Adam et al. used k nearest neighbor (KNN). Their data set was collected from several hospitals and medical centers in Sudan which is the first data set for ischemic disease in Sudan. It contains 15 features and information about 400 patients. The results of the experiment show that the performance of decision tree classification is higher than the performance of KNN algorithm. Their data set contains 1000 records. PCA algorithm was used for dimensional reduction. In ten rounds

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2024.0169
e ISSN: 2584-2854
Volume: 02
Issue: 05 May 2024
Page No: 1250-1253

of each algorithm, they have got the highest accuracy as 92%, 91%, and 94% in Neural Network, Naive Bayes classifier, and Decision tree algorithm respectively. Some of the methods use a very small data set. Govindarajan et al. have predicted only two classes of stroke. Therefore we have proposed a method which uses a large data set with four classes of stroke [4-6].

### 1.1. Scope of the Project

Through the extraction of patient medical history, such as blood pressure, blood sugar levels, and chest pain from a dataset including patient medical history, this initiative predicts individuals who will develop cardiovascular disease.

### 1.2. Objective

The primary goal is to develop a predictive model that can anticipate the likelihood of a heart stroke occurring in individuals based on certain features or risk factors. This aims to assist in early detection or risk assessment.

## 2. Existing System

The current system might involve traditional risk assessment methods or manual evaluation of risk factors related to heart strokes. It may lack the efficiency and accuracy provided by modern machine learning techniques [7, 8].

### 2.1. Existing System Disadvantages

The modeling of input dataset properties, the computing of attribute risk factors, and achieving high prediction accuracy are the primary disadvantages of the existing heart disease prediction systems.

## 3. Proposed System

- Upload Stroke Data set
- Train Naive Bayes Algorithm
- Train J48 Algorithm
- Train KNN Algorithm
- Train Random Forest Algorithm

## 4. Related Work

Naive Bayes classifier are a collection of classification algorithms based on Bayes' Theorem. The bayes theorem finds the conditional probability of an event occurring given the probability of another event that has already occurred. The J48 algorithm is a Java implementation of the C4.5 decision tree algorithm, commonly used in machine learning for classification tasks. J48, being a variant of the C4.5 algorithm, excels in constructing decision trees for classification tasks, providing interpret able models suitable for various domains while requiring less computational complexity compared to certain other Algorithms [9, 10]. k-Nearest Neighbors (k-NN) is a non-parametric method used for classification and regression. Predictions are made for a new instance by searching through the entire training set for the k most similar instances called the neighbors. Majority vote is usually used for choosing the class. Different distance metrics can be used with k-NN like, Euclidean distance, Manhattan/Cityblock distance, Minkowski distance, etc. Random Forest learning method is used for classification and regression. Each classifier in the ensemble is a decision tree classifier (i.e. ID3, C4.5, CART, etc.) so that the collection of classifiers is a forest.

## 5. Methodology

| | id | gender | age | hypertensio | heart_disea | ever_marric | work_type | Residence_t | avg_glucose | bmi | smoking_st | stroke |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 9046 | Male | 67 | 0 | 1 | Yes | Private | Urban | 228.69 | 36.6 | formerly sm | 1 |
| 3 | 51676 | Female | 61 | 0 | 0 | Yes | Self-employ | Rural | 202.21 | N/A | never smok | 1 |
| 4 | 31112 | Male | 80 | 0 | 1 | Yes | Private | Rural | 105.92 | 32.5 | never smok | 1 |
| 5 | 60182 | Female | 49 | 0 | 0 | Yes | Private | Urban | 171.23 | 34.4 | smokes | 1 |
| 6 | 1665 | Female | 79 | 1 | 0 | Yes | Self-employ | Rural | 174.12 | 24 | never smok | 1 |
| 7 | 56669 | Male | 81 | 0 | 0 | Yes | Private | Urban | 186.21 | 29 | formerly sm | 1 |
| 8 | 53882 | Male | 74 | 1 | 1 | Yes | Private | Rural | 70.09 | 27.4 | never smok | 1 |
| 9 | 10434 | Female | 69 | 0 | 0 | No | Private | Urban | 94.39 | 22.8 | never smok | 1 |
| 10 | 27419 | Female | 59 | 0 | 0 | Yes | Private | Rural | 76.15 | N/A | Unknown | 1 |
| 11 | 60491 | Female | 78 | 0 | 0 | Yes | Private | Urban | 58.57 | 24.2 | Unknown | 1 |
| 12 | 12109 | Female | 81 | 1 | 0 | Yes | Private | Rural | 80.43 | 29.7 | never smok | 1 |
| 13 | 12095 | Female | 61 | 0 | 1 | Yes | Govt_job | Rural | 120.46 | 36.8 | smokes | 1 |
| 14 | 12175 | Female | 54 | 0 | 0 | Yes | Private | Urban | 104.51 | 27.3 | smokes | 1 |
| 15 | 8213 | Male | 78 | 0 | 1 | Yes | Private | Urban | 219.84 | N/A | Unknown | 1 |
| 16 | 5317 | Female | 79 | 0 | 1 | Yes | Private | Urban | 214.09 | 28.2 | never smok | 1 |
| 17 | 58202 | Female | 50 | 1 | 0 | Yes | Self-employ | Rural | 167.41 | 30.9 | never smok | 1 |
| 18 | 56112 | Male | 64 | 0 | 1 | Yes | Private | Urban | 191.61 | 37.5 | smokes | 1 |
| 19 | 34120 | Male | 75 | 1 | 0 | Yes | Private | Urban | 221.29 | 25.8 | smokes | 1 |
| 20 | 27458 | Female | 60 | 0 | 0 | No | Private | Urban | 89.22 | 37.8 | never smok | 1 |
| 21 | 25226 | Male | 57 | 0 | 1 | No | Govt_job | Urban | 217.08 | N/A | Unknown | 1 |
| 22 | 70630 | Female | 71 | 0 | 0 | Yes | Govt_job | Rural | 193.94 | 22.4 | smokes | 1 |
| 23 | 13861 | Female | 52 | 1 | 0 | Yes | Self-employ | Urban | 233.29 | 48.9 | never smok | 1 |
| 24 | 68794 | Female | 79 | 0 | 0 | Yes | Self-employ | Urban | 228.7 | 26.6 | never smok | 1 |
| 25 | 64778 | Male | 82 | 0 | 1 | Yes | Private | Rural | 208.3 | 32.5 | Unknown | 1 |
| 26 | 4219 | Male | 71 | 0 | 0 | Yes | Private | Urban | 102.87 | 27.2 | formerly sm | 1 |
| 27 | 70822 | Male | 80 | 0 | 0 | Yes | Self-employ | Rural | 104.12 | 23.5 | never smok | 1 |

**Figure 1 Dataset**

### 5.1. Module Description

This part is divided into two sections: machine learning classifiers and data description. The following outlines these two processes:

#### 5.1.1. Description of the Data

The cardiac stroke dataset from the Kaggle website

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2024.0169
e ISSN: 2584-2854
Volume: 02
Issue: 05 May 2024
Page No: 1250-1253

was used for this study, as shown in Figure 1. This collection has a total of 12 qualities. Below is a comprehensive overview of the features that are employed in the recommended work:

- **ID:** A person's ID is referenced by this property. The data is made up of numbers.
- **Age:** A person's age is indicated by this attribute. The data is made up of numbers.
- **Gender:** An individual's gender is indicated by this attribute. information that is categorized.
- **Hypertension:** Whether or not this person has hypertension is indicated by this attribute. The data is made up of numbers. work type: The setting in which an individual works is reflected [11].

### 5.1.2. Machine Learning Classifiers

- **Random Forest:** Random Forest techniques are used in both classification and regression. Predictions are built upon a treelike organization of the data. When used on large datasets, the Random Forest algorithm can yield identical results even when a significant portion of the record values are missing. The samples produced by the decision tree can be saved and applied to various data sets. There are two steps in random forest: first, create a random forest; next, use the classifier created in the previous stage to create a prediction.
- **Decision Tree:** The Decision Tree algorithm's core node represents the properties of the dataset, while its outer branches produce the outcome. Decision trees are employed because they are incredibly efficient, trustworthy, easy to comprehend, and require very little.
- **KNN:** The supervised machine learning (ML) method known as knearest neighbors, or KNN, can be used to predict regression and handle classification problems. However, it is mostly used in industry to solve classification and forecasting issues.

## 6. Technique Used or Algorithms Used

**Naive Bayes Algorithm:** The bayes theorem finds the conditional probability of an event occurring given the probability of another event that has already occurred.

**Random Forest Algorithm:** Random Forest learning method is used for classification and regression. Each classifier in the ensemble is a decision tree classifier so that the collection of classifiers is a forest. Several works have been carried out to predict the life-threatening diseases using decision tree and proven to be more efficient.

**KNN Algorithm:** k-Nearest Neighbors (k-NN) is a non-parametric method used for classification and regression. Predictions are made for a new instance by searching through the entire training set for the k most similar instances called the neighbors.

**J48 Algorithm:** The J48 algorithm is a Java implementation of the C4.5 decision tree algorithm, commonly used in machine learning for classification tasks. Regularization techniques and validation methods are often used to improve its Generalization capabilities [12]. (as in Figure 2)

## 7. System Architecture

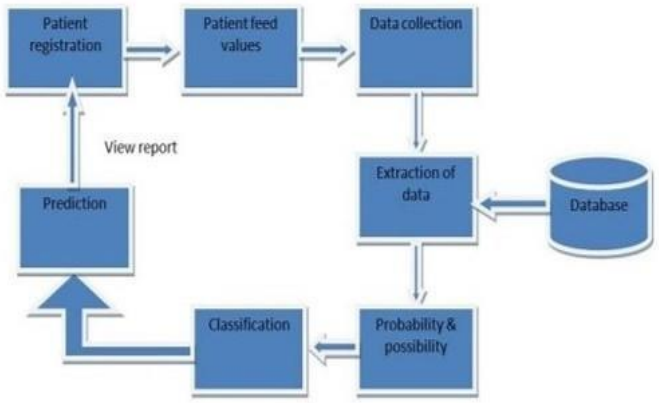


**Figure 2 System Architecture**

## Conclusions

It is essential to create a system that can anticipate heart attacks precisely and effectively given the rise in heart stroke-related fatalities. This study uses the different hospitals dataset to examine the accuracy scores of the Random Forest, Decision Tree, and KNN algorithms for predicting heart attacks. The outcome of this study shows that the Random Forest algorithm, which has an accuracy score of 99.17% for heart attack prediction, is the most effective algorithm. The study can be improved in the future by creating a web application based on the Random Forest method and using a larger dataset than the

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2024.0169
e ISSN: 2584-2854
Volume: 02
Issue: 05 May 2024
Page No: 1250-1253

one used in this analysis, which would help to deliver better results and aid healthcare professionals in accurately and efficiently forecasting cardiac disease.

## References

[1]. S.H. Pahus, A.T.Hansen, and A.M. Hvas, "Thrombophilia testing in young patients with ischemic stroke," Thrombosis research, vol. 137, pp. 108–112, 2016

[2]. P. Govindarajan, R. K. Soundarapandian, A. H. Gandomi, R. Patan, P.Jayaraman, and R.Manikandan, "Classification of stroke disease using machine learning algorithms," Neural Computing and Applications, pp. 1–12.

[3]. L. T. Kohn, J. Corrigan, M. S. Donaldson, et al., To err is human: building a safer health system, vol. 6. National academy press Washington, DC, 2000

[4]. R. Jeena and S. Kumar, "Stroke prediction using svm," in 2016 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT), pp. 600 602, IEEE, 2016.

[5]. P. A. Sandercock, M. Niewada, and A. Członkowska, "The international stroke trial database," Trials, vol. 13, no. 1, pp. 1–1, 2012.

[6]. M. S. Singh and P. Choudhary, "Stroke prediction using artificial intelligence," in 2017 8th Annual Industrial Automation and Electromechanical Engineering Conference (IEMECON), pp. 158–161, IEEE, 2017.

[7]. S. Y. Adam, A. Yousif, and M. B. Bashir, "Classification of ischemic stroke using machine learning algorithms," Int J Comput Appl, vol. 149, no. 10, pp. 26–31, 2016

[8]. A.Sudha, P. Gayathri, and N. Jaisankar, "Effective analysis and predictive model of stroke disease using classification methods," International Journal of Computer Applications, vol. 43, no. 14, pp. 26– 31, 2012.

[9]. G. Kaur and A. Chhabra, "Improved j48 classification algorithm for the prediction of diabetes," International Journal of Computer Applications, vol. 98, no. 22, 2014

[10]. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2016.

[11]. P. Sewaiwar and K. K. Verma, "Comparative study of various decision tree classification algorithm using weka," International Journal of Emerging Research in Management &Technology, vol. 4, pp. 2278– 9359, 2015.

[12]. K. A. Shakil, S. Anis, and M. Alam, "Dengue disease prediction using weka data mining tool," arXiv preprint arXiv:1502.05167, 2015.