IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Heart Stroke Prediction using Machine Learning

B.P. Deepak Kumar[*1], Sagar Yellaram[*2], Sumanth kothamasu[*3], SurendharReddy Puchakayala[*4]

[*1]Assistant Professor,Department of Computer Science and Engineering, CMR Technical Campus, Medchal, Telangana, India

[*2]JNTUH, Computer Science and Engineering, CMR Technical Campus, Medchal, Telangana, India

[*3]JNTUH, Computer Science and Engineering, CMR Technical Campus, Medchal, Telangana, India

[*4]JNTUH, Computer Science and Engineering, CMR Technical Campus, Medchal, Telangana, India

ABSTRACT

In recent times, Heart Stroke prediction is one of the most complicated tasks in medical field. In the modern era, approximately one person dies per minute due to heart Stroke. Data science plays a crucial role in processing huge amount of data in the field of healthcare. As heart stroke prediction is a complex task, there is a need to automate the prediction process to avoid risks associated with it and alert the patient well in advance. This paper makes use of heart stroke dataset. The proposed work predicts the chances of Heart Stroke and classifies patient's risk level by implementing different data mining techniques such as KNN, Decision Tree and Random Forest. Thus, this paper presents a comparative study by analyzing the performance of different machine learning algorithms. The trial results verify that Random Forest algorithm has achieved the highest accuracy of 99.17% compared to other ML algorithms implemented.

Keywords: **KNN, Decision Tree, Random Forest, Heart Stroke Prediction.**

I. INTRODUCTION

The work proposed in this paper focus mainly on various data mining practices that are employed in heart stroke Prediction. Human heart is the principal part of the human body. Any irregularity to heart can cause distress in other parts of body. In today's contemporary world, heart stroke is one of the primary reasons for occurrence of most deaths. Heart stroke may occur due to unhealthy lifestyle, smoking, alcohol and high intake of fat which may cause hyper-tension . According to the World Health Organization more than 10 million die due to Heart stroke every single year around the world. A healthy lifestyle and earliest detection are only ways to prevent the heart stroke. The main challenge in today's healthcare is provision of best quality services and effective accurate diagnosis . The proposed work makes an attempt to detect these heart stroke at early stage to avoid disastrous consequences. Data mining techniques are the means of extracting valuable and hidden information from the large amount of data available. Machine Learning (ML) which is subfield of data mining handles large scale well-formatted dataset efficiently. In the medical field, machine learning can be used for diagnosis, detection and prediction of various diseases. The main goal of this paper is to provide a tool for doctors to detect heart stroke as early stage . This in turn will help to provide effective treatment to patients and avoid severe consequences. ML plays a very important role to detect the hidden discrete patterns and thereby analyze the given data. After analysis of data ML techniques help in heart stroke prediction and early diagnosis. This paper presents performance analysis of various ML techniques such as KNN, Decision Tree, and Random Forest for predicting heart stroke at an early stage

II. LITERATURE SURVEY

Many researchers have already used machine learning based approached to predict heart strokes. Govindarajan et al. [11] conducted a study to categorize heart stroke disorder using a text mining combination and a machine learning classifier and collected data for 507 patients. For their analysis, they used various machine learning approaches for training purposes using ANN, and the SGD algorithm gave them the best value, which was 95%. Amini et al. [4], [12] conducted research to predict stroke incidence, collected 807 healthy and unhealthy subjects in their study categorized 50 risk factors for stroke, diabetes, cardiovascular disease, smoking, hyperlipidemia, and alcohol use. They used two techniques that had the best accuracy from c4.5 decision tree algorithm, and it was 95%, and for K-nearest neighbor, the accuracy was 94%. Cheng et al. [13] published a report on the estimation of the heart stroke prognosis. In their analysis, 82 stroke patient data were used, two ANN models were used to find precision, and 79% and 95% were used. Cheon et al. [14]–[16] performed a study to predict stroke patient mortality. In their study, they used 15099 patients to identify heart stroke occurrence. They used a deep neural network approach to detect heart strokes. The authors used PCA to extract medical record history and predict heart stroke. They have got an area under the curve (AUC) value of 83%. Singh et al.

[17] performed a study on heart stroke prediction applied to artificial intelligence. In their research, they used a different method for predicting stroke on the cardiovascular health study (CHS) dataset. And they took the decision tree algorithm to feature extract to principal component analysis. They used a neural network classification algorithm to construct the model they got 97% accuracy. Chin et al. [18] performed a study to detect an automated early heart stroke. In their study, the main purpose was to develop a system using CNN to automated primary heart stroke. They collected 256 images to train and test the CNN model. In their system image prepossessing remove the impossible area that can't occur of heart stroke, they used the data prolongation method to raise the collected image. Their CNN method has given 90% accuracy.

The main idea behind the proposed system after reviewing the above papers was to create a heart stroke prediction system based on the inputs . We analysed the classification algorithms namely KNN, Decision Tree and Random Forest based on their Accuracy, Precision, Recall and f-measure scores and identified the best classification algorithm which can be used in the heart disease prediction.

III.PROPOSED SYSTEM

The proposed work predicts heart stroke by exploring the above mentioned four classification algorithms and does performance analysis. The objective of this study is to effectively predict if the patient suffers from heart stroke. The health professional enters the input values from the patient's health report. The data is fed into model which predicts the probability of having heart stroke. Fig. 1 shows the entire process involved .

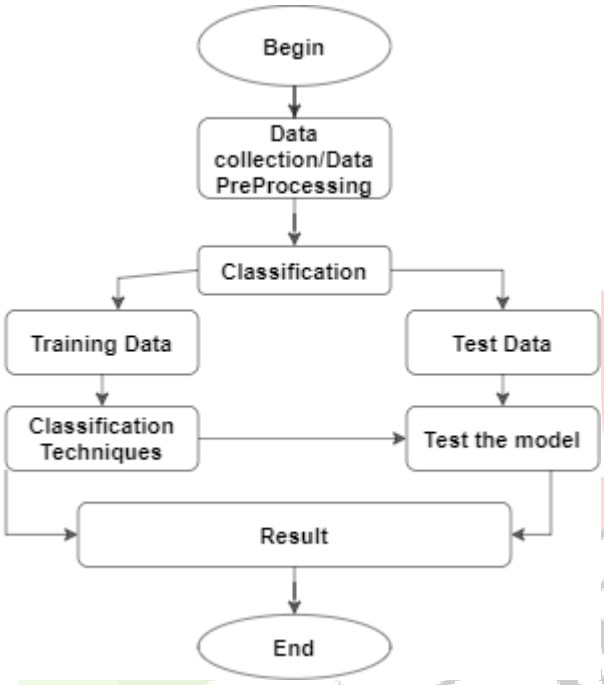

Fig 1: Generic Model Predicting Heart Stroke

METHODOLOGY:

This section is divided into two parts, these are: Data description, machine learning classifiers . These two processes are described below:

A) Data Description:

Here we used the heart stroke dataset that is available in the kaggle website for our analysis. This dataset consists of total 12 attributes. The complete description of the attributes used in the proposed work is given below:

id : This attribute means person's id. It's numerical data.
Age : This attribute means a person's age. It's numerical data.
Gender : This attribute means a person's gender. It's categorical data.
Hypertension : This attribute means that this person is hypertensive or not. It's numerical data.
work type : This attribute represents the person work scenario. It's categorical data.
residence type : This attribute represents the person living scenario. It's categorical data.
heart disease : This attribute means whether this person has a heart disease person or not. It's numerical data.
avg glucose level : This attribute means what was the level of a person's glucose condition. It's numerical data.
Bmi : This attribute means body mass index of a person. It's numerical data.
ever married : This attribute represents a person's married status. It's categorical data.
smoking Status : This attribute means a person's smoking condition. It's categorical data.
Stroke : This attribute means a person previously had a stroke or not. It's numerical data. In this all attribute stroke is the decision class and rest of the attribute is response class.

B)Machine Learning Classifiers:

The attributes mentioned are provided as input to the different ML algorithms such as Random Forest, Decision Tree and KNN. The input dataset is split into 80% of the training dataset and the remaining 20% into the test dataset. Training dataset is the dataset which is used to train a model. Testing dataset is used to check the performance of the trained model. For each of the algorithms the performance is computed and analyzed based on different metrics used such as accuracy, precision, recall and F-measure scores as described further. The different algorithms explored in this paper are listed as below.

i. Random Forest: Random Forest algorithms are used for classification as well as regression. It creates a tree for the data and makes prediction based on that. Random Forest algorithm can be used on large datasets and can produce the same result even when large sets record values are missing. The generated samples from the decision tree can be saved so that it can be used on other data. In random forest there are two stages, firstly create a random forest then make a prediction using a random forest classifier created in the first stage.

ii. Decision Tree: Decision Tree algorithm is in the form of a flowchart where the inner node represents the dataset attributes and the outer branches are the outcome. Decision Tree is chosen because they are fast, reliable, easy to interpret and very little data preparation is required. In Decision Tree, the prediction of class label originates from root of the tree. The value of the root attribute is compared to record's attribute. On the result of comparison, the corresponding branch is followed to that value and jump is made to the next node.

iii.KNN: k-nearest neighbors (KNN) algorithm is a type of supervised ML algorithm which can be used for both classification as well as regression predictive problems. However, it is mainly used for classification predictive problems in industry. The following two properties would define KNN well −

- **Lazy learning algorithm** − KNN is a lazy learning algorithm because it does not have a specialized training phase and uses all the data for training while classification.
- **Non-parametric learning algorithm** − KNN is also a non-parametric learning algorithm because it doesn't assume anything about the underlying data.

IV. RESULTS AND ANALYSIS

The results obtained by applying Random Forest, Decision Tree and KNN are shown in this section. The metrics used to carry out performance analysis of the algorithm are Accuracy score, Precision (P), Recall (R) and F-measure. Precision (mentioned in equation (1)) metric provides the measure of positive analysis that is correct. Recall [mentioned in equation (2)] defines the measure of actual positives that are correct. F-measure [mentioned in equation (3)] tests accuracy.

Precision = (TP) / (TP +FP) (1)

Recall = (TP) / (TP+FN) (2)

F– Measure =(2 * Precision * Recall) / (Precision +Recall) (3)

• TP True positive: the patient has the disease and the test is positive.
• FP False positive: the patient does not have the disease but the test is positive.
• TN True negative: the patient does not have the disease and the test is negative.
• FN False negative: the patient has the disease but the test is negative.

In the experiment the pre-processed dataset is used to carry out the experiments and the above mentioned algorithms are explored and applied. The above mentioned performance metrics are obtained using the confusion matrix. Confusion Matrix describes the performance of the model. The confusion matrix obtained by the proposed model for different algorithms is shown below in Table 1. The accuracy score obtained for Random Forest, Decision Tree and KNN classification techniques is shown below in Table 2.

Table I VALUES OBTAINED FOR CONFUSION MATRIX USING DIFFERENT ALGORITHMS

Algorithm	True Positive	False Positive	False Negative	True Negative
KNN	839	120	4	496
Decision Tree	915	44	0	500
Random Forest	947	12	0	500

Table II ANALYSIS OF DIFFERENT MACHINE LEARNING ALGORITHMS

Algorithm	Precision	Recall	F - measure	Accuracy
KNN	0.92	0.90	0.90	90.15%
Decision Tree	0.97	0.97	0.97	96.25%
Random Forest	0.99	0.99	0.99	99.17%

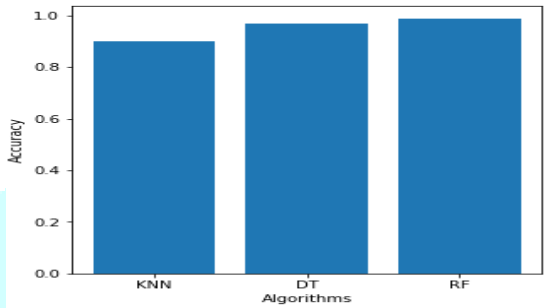

Fig 2:Bar graph representing accuracy of various algorithms used in this proposed work

V. CONCLUSION

With the increasing number of deaths due to heart stroke's, it has become mandatory to develop a system to predict heart stroke effectively and accurately. The motivation for the study was to find the most efficient ML algorithm for detection of heart stroke. This study compares the accuracy score of Random Forest, Decision Tree and KNN algorithms for predicting heart stroke using kaggle dataset. The result of this study indicates that the Random Forest algorithm is the most efficient algorithm with accuracy score of 99.17% for prediction of heart stroke. In future the work can be enhanced by developing a web application based on the Random Forest algorithm as well as using a larger dataset as compared to the one used in this analysis which will help to provide better results and help health professionals in predicting the heart disease effectively and efficiently

ACKNOWLEDGEMENT

Apart from the efforts of us, the success of paper depends largely on encouragement and guidelines of many others. We take this opportunity to express our profound gratitude to CMR Technical Campus College Management for motivating us and for providing all the facilities required for this work. We are deeply indebted to Chairman Shri C. Gopal Reddy, Secretary Smt. C. Vasantha Latha , Director Dr . A. Raji Reddy, HOD CSE Dr. K. Srujan Raju, Project Guide B.P. Deepak Kumar who always has been a constant source of inspiration for us.

REFERENCES

[1] L. Amini, R. Azarpazhouh, M. T. Farzadfar, S. A. Mousavi, F. Jazaieri, F. Khorvash, R. Norouzi, and N. Toghianfar, "Prediction and control of heart stroke by data mining," International Journal of Preventive Medicine, vol. 4, no. Suppl 2, pp. S245–249, May 2013.

[2] S.-F. Sung, C.-Y. Hsieh, Y.-H. Kao Yang, H.-J. Lin, C.-H. Chen, Y.- W. Chen, and Y.-H. Hu, "Developing a heart stroke severity index based on administrative data was feasible using data mining techniques," Journal of Clinical Epidemiology, vol. 68, no. 11, pp. 1292–1300, Nov. 2015.

[3] M. C. Paul, S. Sarkar, M. M. Rahman, S. M. Reza, and M. S. Kaiser, "Low cost and portable patient monitoring system for e-health services in bangladesh," in 2016 International Conference on Computer Communication and Informatics (ICCCI), 2016, pp. 1–4.

[4] S. M. Reza, M. M. Rahman, M. H. Parvez, M. S. Kaiser, and S. Al Mamun, "Innovative approach in web application effort & cost estimation using functional measurement type," in 2015 International Conference on Electrical Engineering and Information Communication Technology (ICEEICT). IEEE, 2015, pp. 1–7.

[5] P. Govindarajan, R. K. Soundarapandian, A. H. Gandomi, R. Patan, P. Jayaraman, and R. Manikandan, "Classification of heart stroke disease using machine learning algorithms," Neural Computing and Applications, vol. 32, no. 3, pp. 817–828, Feb. 2020.

[6] C.V. Krishna Veni, T. R. Shoba, On the classification of imbalanced Datasets, International Journal of Computer Science & Technology 2011; 2:145-148