International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT206429

An Enhanced Novel Dynamic Data Processing (ENDDP) Algorithm for Predicting Heart Disease in Machine Learning

J. Nageswara Rao[1], Dr. R. Satya Prasad[2]

[1]Research Scholar Department of Computer Science and Engineering, Acharya Nagarjuna University, Guntur, Andhra Pradesh, India

[2]Professor, Department of Computer Science and Engineering, Acharya Nagarjuna University, Guntur, Andhra Pradesh, India

ABSTRACT

Article Info
Volume 7, Issue 1
Page Number: 94-104
Publication Issue :
January-February-2021

Article History
Accepted : 10 Jan 2021
Published : 21 Jan 2021

Machine learning (ML) is a rapidly developing field in today's world. Use machine learning to extract data from a wide variety of sources. ML can solve various problems based on complex data sets. The prediction of heart disease is the most complex task in the medical field. It cannot be observed with the naked eye, it can appear immediately anywhere, anytime. Many ML algorithms are more capable of handling various algorithms. Due to complexity, the processing of massive data sets is more complicated. By improving these systems, the quality of medical diagnosis decisions can be improved. They can find patterns hidden in large amounts of data that will avoid the use of traditional statistical methods for analysis. In this article, An Enhanced New Dynamic Data Processing (ENDDP) Algorithm is developed to predict the early stages of heart disease. The results prove the performance of the proposed system.

Keywords : Machine Learning, Massive Datasets, ENDDP.

I. INTRODUCTION

Machine learning (ML) is most widely used in a variety of machine-based applications that can be executed automatically. For everyone, the heart is a very sensitive and important part. Human life is based on heartbeat. Many people suffer from heart disease due to age, habits and other reasons. Analyzing a heart attack in the early stages is very unpredictable. Due to the lack of accurate information and understanding of symptoms, many diseases have been confirmed in the final stage. The diagnosis and treatment of heart disease are more complicated, especially in developing countries. Due to the shortage of doctors and other medical-related equipment, a large number of deaths are occurring [1]. For patients with heart disease, measures must be taken to reduce the risk of serious heart disease and improve heart safety [2].

Machine learning is essential to get the best results in any application. It is also more effective in teaching and testing applications. It is a subset of artificial intelligence (AI), which is one of the broad areas of machine learning that can reproduce human functions; machine learning is a special branch of AI. Based on previously selected data values, it is very useful to predict heart diseases that belong to ML. To overcome various challenges in predicting heart disease, An Enhanced New Dynamic Data Processing (ENDDP) Algorithm has been introduced to achieve more accurate predictions.

II. LITERATURE SURVEY

The extensive literature on the use of machine learning methods to predict heart disease has inspired our work. A detailed overview is presented in this paper.

In everyone's heart, it is the core organ of life. This plays a vital role in absorbing blood and producing oxygen in important parts of the human body, so it is very important to protect blood. In this section, many researchers have proposed various algorithms for predicting heart disease. You can also use existing algorithms to evaluate performance. For each algorithm, performance is calculated based on accuracy [4]. In terms of the performance of NB and KNN, by using these machine learning algorithms, KNN can predict heart disease more accurately.

It is very important that the cores are more difficult to handle, they must be handled very carefully, and otherwise the person may die. Depending on the severity of the heart disease, it is classified based on predetermined methods such as KNN, Decision Tree, General Algorithm, and Naive Bayes [3]. The author wants to combine the two algorithms to get a hybrid method with an improved accuracy of 88.4%, which is a very efficient algorithm.

Data mining (DM) also plays an important role in predicting heart disease. Few researchers use data mining algorithms to predict heart disease. Kaur et al. [6] conducted research on this, explained important patterns, and gained understanding from huge data sets. Use a variety of machine learning algorithms and data mining algorithms to achieve comparison accuracy to observe which algorithm is the best algorithm, and among all these algorithms, SVM has higher performance. Author Kumar et al. [5] Various ML and DM algorithms are used. The data set is selected from the UCI knowledge base learning data set. The data set contains 303 samples and 14 input features, and **SVM** is found to be the best among them. The other different algorithms here are **Naive Bayes, KNN** and **Decision Tree.**

Gavhane et al. [1] developed a multi-layer perceptron model for human heart disease prediction, and showed the accuracy through algorithm CAD technology. Every time more and more heart diseases predict that more and more people will be, as we all know, people's consciousness is also increasing. This can reduce the cause of heart disease and also reduce deaths. Few researchers study one or two disease prediction algorithms.

Krishnan et al. [2] proved that the decision tree is more accurate than the naive Bayes classification algorithm. Machine learning algorithms can be used to predict various diseases. Like Kohali et al. [7] many scientists are also studying this. Introduced a new method of using Logistic regression to predict heart disease, using Support Vector Machine (SVM) based on diabetes prediction, using Adaboot classifier to predict breast cancer through Adaboot classifier, and finally determined that the accuracy of Logistic regression was 87.1%. The accuracy of SVM is 85.71%, while the accuracy of Adaboot classifier is 98.57%, which is very useful for predicate perspective.

III. Data Set Description

3.1. Data collection and pre-processing

The data set used is the "heart disease data set", which is a combination of 4 different databases, but only the UCI Cleveland data set is used. The database has 76 attributes in total, but all published experiments only use a subset of 14 features to represent [9]. Therefore, we used the processed UCI Cleveland dataset provided on the Kaggle website for analysis. Table 1 below lists a complete description of the 14 attributes used in the proposed work

Sl.No.	Attribute Description	Distinct Values of Attribute
1.	Age- represent the age of a person	Multiple values between 29 & 71
2.	Sex- describe the gender of person (0- Female, 1-Male)	0,1
3.	CP- represents the severity of chest pain patient is suffering.	0,1,2,3
4.	RestBP-It represents the patients BP.	Multiple values between 94& 200
5.	Chol-It shows the cholesterol level of the patient.	Multiple values between 126 & 564
6.	FBS-It represents the fasting blood sugar in the patient.	0,1
7.	Resting ECG-It shows the result of ECG	0,1,2
8.	Heartbeat- shows the max heart beat of patient	Multiple values from 71 to 202
9.	Exang- used to identify if there is an exercise induced angina. If yes=1 or else no=0	0,1
10.	Old Peak- describes patients depression level.	Multiple values between 0 to 6.2.
11.	Slope- describes patient condition during peak exercise. It is divided into three segments(Unsloping, Flat, Down sloping)	1,2,3.
12.	CA- Result of fluoroscopy.	0,1,2,3
13.	Thal- test required for patient suffering from pain in chest or difficulty in breathing. There are 4 kinds of values which represent Thallium test.	0,1,2,3
14.	Target-It is the final column of the dataset. It is class or label Colum. It represents	0,1

Table.1 Selected Cleveland Heart Disease Data Set

IV. Machine Learning Algorithms

The purpose of this field is related to whether the patient has heart disease. Expressed in whole numbers. If the value is 0 (absent), the value is less than 4. According to the experiment, the dataset focuses on distinguishing between existence (value 0, 2, 3, 4) and non-existence (value 0).

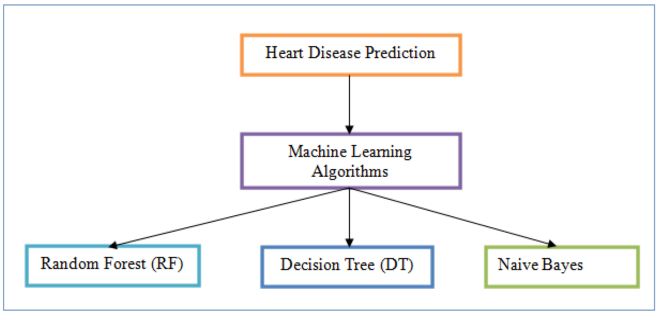


Figure 1: Machine Learning Algorithms for Heart Disease Prediction.

The next step is the prediction of heart disease, which is explained in Figure 1. The comparison results show three machine learning algorithms using the heart disease dataset, such as Naive Bayes (NB), Decision Tree (DT) and Random Forest (RF) and classification models. The following sections will introduce these three algorithms.

4.1. Classification Using Random Forest

Random Forest is most powerful and it is the combination of three predictors by using decision tree and these values are depends on random vector which is separately and with the same distribution for all trees in the forest. This algorithm is combination of classification and the regression based on the problem domain. The following are the steps for random forest algorithm.

- The k features are selected randomly from overall m features, where k << m.
- Adjacent to the k features, calculate the node “d” using the best split point.
- By using the best split, the child nodes are to be split.
- Repeat 1 to 3 steps until l number of nodes has been reached.
- By repeating 1 to 4 steps to build forest for n number times to create n number of trees.

4.2. Classification Using Decision Tree

Decision Tree is very easy and simple classifier to implement. The DT develops the classification or regression model to make the tree structure of a tree making it simple to debug and handle. Datasets are classified as two types such as categorical and numerical data. And DT can handle both types of data. The information gain plays the major role to find the attributes and taking out the attributes for splitting the branches into tree. The equation for information gain is given as:

Eq. (1). E(S) = -P (P) log2P (P)-P (N) log2P (N)

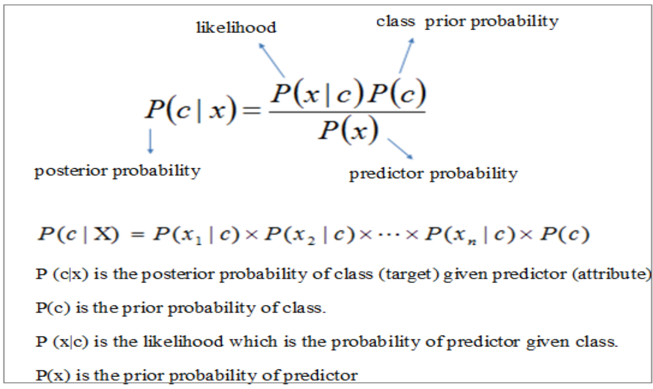
The following are **Decision Tree** algorithm is given as:

- Find the information gain for the attributes in the dataset.
- In the descending order, the sorting is done with the information gain for the heart disease datasets.
- After the processing of step 2, the information gain is assigning the best attribute of the dataset at the root of the tree.
- Using the same formula the information gain is to be calculated.
- Based on the highest information gain the nodes are separated.
- The steps are to be repeated until the each attributes are set as leaf nodes in all the branches of the tree.

4.3.Classification UsingNaïve Bayes Classifier

The naïve Bayes classifier model is simple to create from complex parameters, which makes it especially effective in the prediction of heart illness in the field of medicine. Due to its simplicity, the Naive Bayes classifier has huge impacts andis widely used for its efficiency superior to more complex classification techniques. Bayes' theorem allows you to calculate the posterior probability P (c | x) from P (c), P (x), and P (x | c). A naive Bayesian classifier assumes that the value of a predictor (x) affects a given category (c) and has nothing to do with the values of other predictors. Independence of class conditions

Equation:



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

$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

P (c|x) is the posterior probability of class (target) given predictor (attribute).

P(c) is the prior probability of class.

P (x|c) is the likelihood which is the probability of predictor given class.

P(x) is the prior probability of predictor

4.4. An Enhanced Novel Dynamic Data Processing (ENDDP) Algorithm

The main purpose of analyzing the dynamic data set is to obtain the dynamic information characteristics of the group. To analyze the characteristics of grouping and grouping numerical data according to classification statistics. For everyone with a record, this is Global or local sensitivity is affected by the weight of the entire data set. Payment Sensitivity is also one of the challenges in dynamic data processing. Dynamic data mainly includes incremental data (e.g. timing data, continuous data) and full dynamic data (i.e. all the data inserted, deleted and modified) [9]. An Enhanced Novel Dynamic Data Processing (ENDDP) Algorithm is a statistical classifier that does not imply a prohibition between attributes. This function integrates several Bayesian functions for analyzing properties that are independent of each other. The workflow of the proposed classifier is as follows:

.

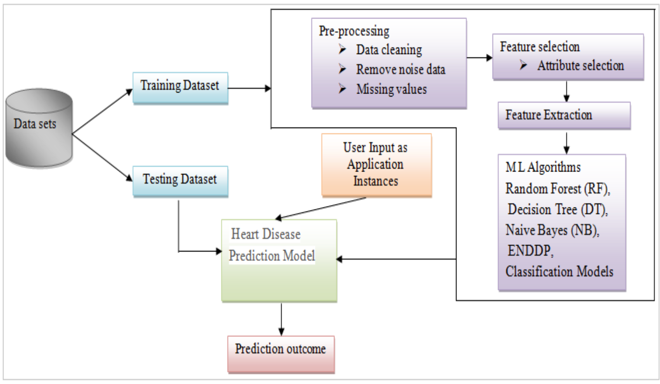


Figure: 2. Working Model –Proposed ENDDP Heart Disease Prediction System

Accuracy: This will calculate the overall accuracy of the abnormal and normal predicted data is calculated by

$$.Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

Performance Evolution

The performance measures namely False Positive Rate (FPR), False Negative Rate (FNR), Sensitivity, Specificity and Accuracy, the performance of the system are estimated. The basic count values such as True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) are used by these measures.

FPR

The percentage of predicted values was classified to normal and abnormal data, but in fact it did not.

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}$$

FNR

The percentage of predicted values was classified to normal and abnormal data, but in fact it did.

$$\text{FNR} = \frac{\text{FN}}{\text{FN} + \text{TN}}$$

Sensitivity

The positives are correctly identified to calculate the sensitivity. This is used to test to identify negative results.

$$\text{Sesitivity} = \frac{\text{No.of TP}}{\text{No. of TP} + \text{No. of TN}}$$

Specificity

The negatives are correctly identified to calculate the specificity. This is used to test to identify negative results.

$$\text{Specificity} = \frac{\text{No. of TN}}{\text{No. of TN} + \text{No. of FP}}$$

V. Results

The parameters based performance is shown in table 1.

Machine Learning Algorithms	**Accuracy**	**Sensitivity**	**Specificity**
Naive Bayes	91.42%	86.43%	79.76%
Random Forest	89.56%	89.76%	84.54%
Proposed System[ENDDP]	97.98%	97.45%	98.54%

Table: 1 Accuracy Calculations

This shows the accuracy of the result based on the data mining techniques.

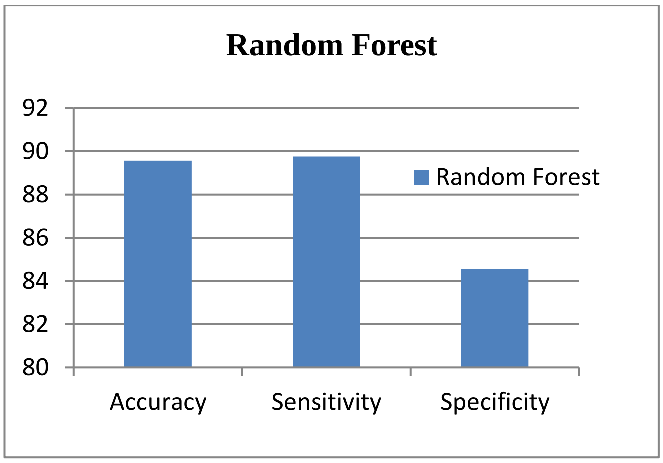


Figure: 3 the performance of the Random Forest

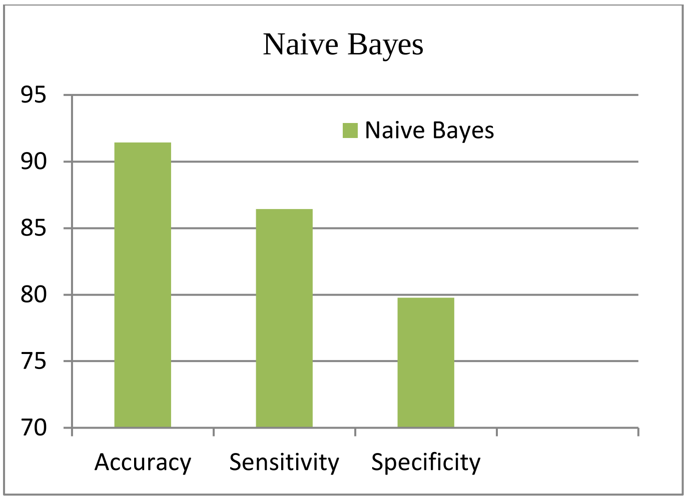


Figure: 4 the performance of the Naïve Bayes

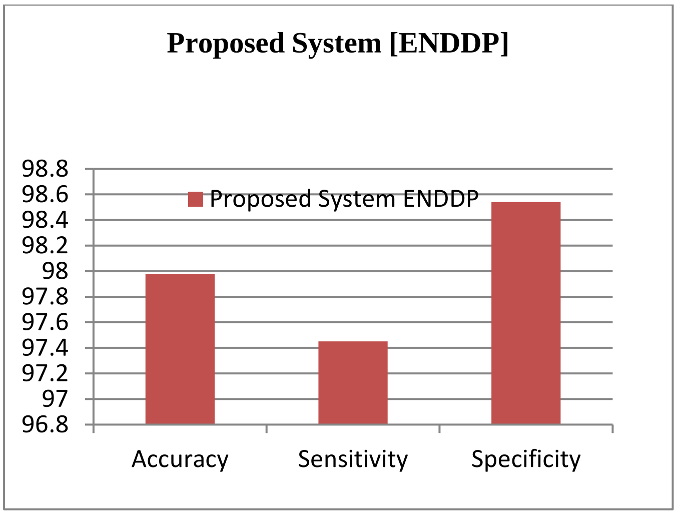


Figure.5.the Performance of the Proposed System [ENDDP]

VI. Conclusion

With the increase in deaths from heart disease, the development of an effective and accurate heart disease prediction system has become a mandatory task. The motivation of this research is to find the most effective ML algorithm to detect heart disease. This study compared the accuracy scores of decision trees, logistic regression, random forest, and naive Bayes algorithm for predicting heart disease using UCI machine learning repository dataset.

In the proposed An Enhanced New Dynamic Data Processing (ENDDP) Algorithm is used to obtain better results. Improve performance by using various techniques. Performance is shown in 3 parameters. Sensitivity, specificity and accuracy. The comparison results are shown in Table 1. The overall performance is calculated based on 3 parameters.

VII. REFERENCES

[1]. "S. Ghwanmeh, A.Mohammad, and A.Al-Ibrahim, “Innovative artificial neural networks-based decision support system for heart diseases diagnosis,” Journal of Intelligent Learning Systems and Applications, vol. 5, no. 3, pp. 176–183, 2013.

[2]. Q.K.Al-Shayea, “Artificial neural networks in medical diagnosis,” International Journal of Computer Science Issues, vol. 8, no. 2, pp. 150–154, 2011.

[3]. Senthil kumar mohan, chandrasegar thirumalai and Gautam Srivastva, “Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques” IEEE Access 2019.

[4]. Himanshu Sharma and M A Rizvi, “Prediction of Heart Disease using Machine Learning Algorithms: A Survey” International Journal on Recent and Innovation Trends in Computing and Communication Volume: 5 Issue: 8 , IJRITCC August 2017.

[5]. M. Nikhil Kumar, K. V. S. Koushik, K. Deepak, “Prediction of Heart Diseases Using Data Mining and Machine Learning Algorithms and Tools” International Journal of Scientific Research in Computer Science, Engineering and Information Technology ,IJSRCSEIT 2019.

[6]. Amandeep Kaur and Jyoti Arora,“Heart Diseases Prediction using Data Mining Techniques: A survey” International Journal of Advanced Research in Computer Science , IJARCS 2015-2019.

[7]. Pahulpreet Singh Kohli and Shriya Arora, “Application of Machine Learning in Diseases Prediction”, 4th International Conference on Computing Communication And Automation(ICCCA), 2018.

[8]. M. Akhil, B. L. Deekshatulu, and P. Chandra, “Classification of Heart Disease Using K-Nearest Neighbor and Genetic Algorithm,” Procedia Technol., vol. 10, pp. 85–94, 2013.

[9]. H. Y. Kanga, *, Y. L. Mab , X. M. Sia” An Enhanced Algorithm for Dynamic Data Release Based on Differential Privacy” Procedia Computer Science 174 (2020) 15–21.

Cite this article as :

J. Nageswara Rao, Dr. R. Satya Prasad, "An Enhanced Novel Dynamic Data Processing (ENDDP) Algorithm for Predicting Heart Disease in Machine Learning", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 7 Issue 1, pp. 94-104, January-February 2021. Available at
doi : https://doi.org/10.32628/CSEIT206429