ISSN: 0975-3583, 0976-2833 VOL 12, ISSUE 03, 2021

An Ensemble Feature Extraction and Deep Learning Approach for Prediction of Heart Disease

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

[1]Research Scholar Department of Computer Science and Engineering,

[2]Professor, Department of Computer Science and Engineering,

[1,2]Acharya Nagarjuna University, Guntur, Andhra Pradesh, India.

Abstract: Cardiovascular disease is commonly known as heart disease. Heart disease prediction in early stages is more complex to get the accurate results. Data mining (DM), Machine Learning (ML) and Deep Learning (DL) many domains doing huge research on medical data especially in heart disease prediction. Heart is most important part in human body. Heart attack is one of the heart diseases. Deep Learning (DL) plays the major role in prediction of heart disease. In this paper, An Ensemble Feature Extraction Learning Approach (EFEDLA) is introduced for the early prediction of heart diseases with the patient data. The dataset is collected from Kaggle website. To overcome the difficulties in dataset the proposed system utilized the enhanced pre-processing technique. The proposed system follows the few steps to predict the heart disease in the early stages. The performance of proposed system is compared with several existing systems.

Keywords: Feature Extraction, Machine Learning, Deep Learning, EFEDLA.

1. Introduction

Coronary heart disease is the leading cause of disappearances worldwide and the main driving force for hospitalization in the United States and Europe. Year [2, 3]. Cardiovascular rupture is a true reformist clinical disease in which the capacity of the focal ventricles is weakened, leading to basic hypo perfusion. Usually, the analysis of cardiovascular degeneration is based on medical history, actual evaluation, basic laboratory tests and imaging considerations [4]. However, when the cause of cardiovascular abnormalities is unclear, endocardial biopsy (EMB) represents the highest level of quality in evaluating and organizing coronary artery disease [5]. The main problems with manual understanding of OS are moderate differences between raters [6] and limited clinical signs [5, 7]. Planned examinations and cardiovascular histopathological examinations can be used as objective information for auxiliary readings to reduce instability.

The heart is one of the largest and most important organs in the human body, so it is very important to care for the heart. Most diseases are related to the heart, so heart disease must be predicted. Therefore, comparative research in this field is needed. Today, most patients die due to the inaccuracy of the instrument, so their disease can be confirmed in the final stage. Therefore, algorithms that better predict disease are needed.

AI is one of the feasible testing techniques that rely on learning and testing. It is part of artificial intelligence (AI), which is one of the broad areas where machines can simulate human capabilities, and AI is a unique part of AI. The artificial intelligence framework is once again ready to figure out how to measure and use information, which is why the hybrid of the two advances is also called artificial consciousness.

By the definition of machine learning, it benefits from unique miracles and common things. Therefore, in this article, we use natural boundaries such as cholesterol, blood circulation, gender, age, and other test information, and focus on calculations such as deep learning under this premise.

2. Literature Survey

Ahmed M. Alaa proposed an AI strategy to address the risk of cardiovascular infection. In any case, they have reached the highest accuracy of 77%. Due to the unbalanced data set, testing methods need to be applied. Nonetheless, they applied the machine learning model directly to the data set. Stephen F. Weng studied the use of AI calculations to improve the risk expectations of cardiovascular diseases. They proved that machine learning calculations have achieved fruitful results in improving the accuracy of cardiovascular disease risk expectations, but that more patient records are needed to obtain better results. Rine Nakanishi [7] and others evaluated machine learning strategies to increase the expected incidence of coronary heart disease (CHD). They applied the AI method to the records of 6,814 patients and achieved a high accuracy rate.

Poornima Singh et al. [8] proposed a framework for predicting coronary artery disease that depends on nerve tissue. The proposed technique takes 15 credits into consideration for prediction. Preparing the model utilizes the multi-layered insight of neural tissue with reverse proliferation capabilities. The data set considered is organized information, and the model can be selected with 100% accuracy.

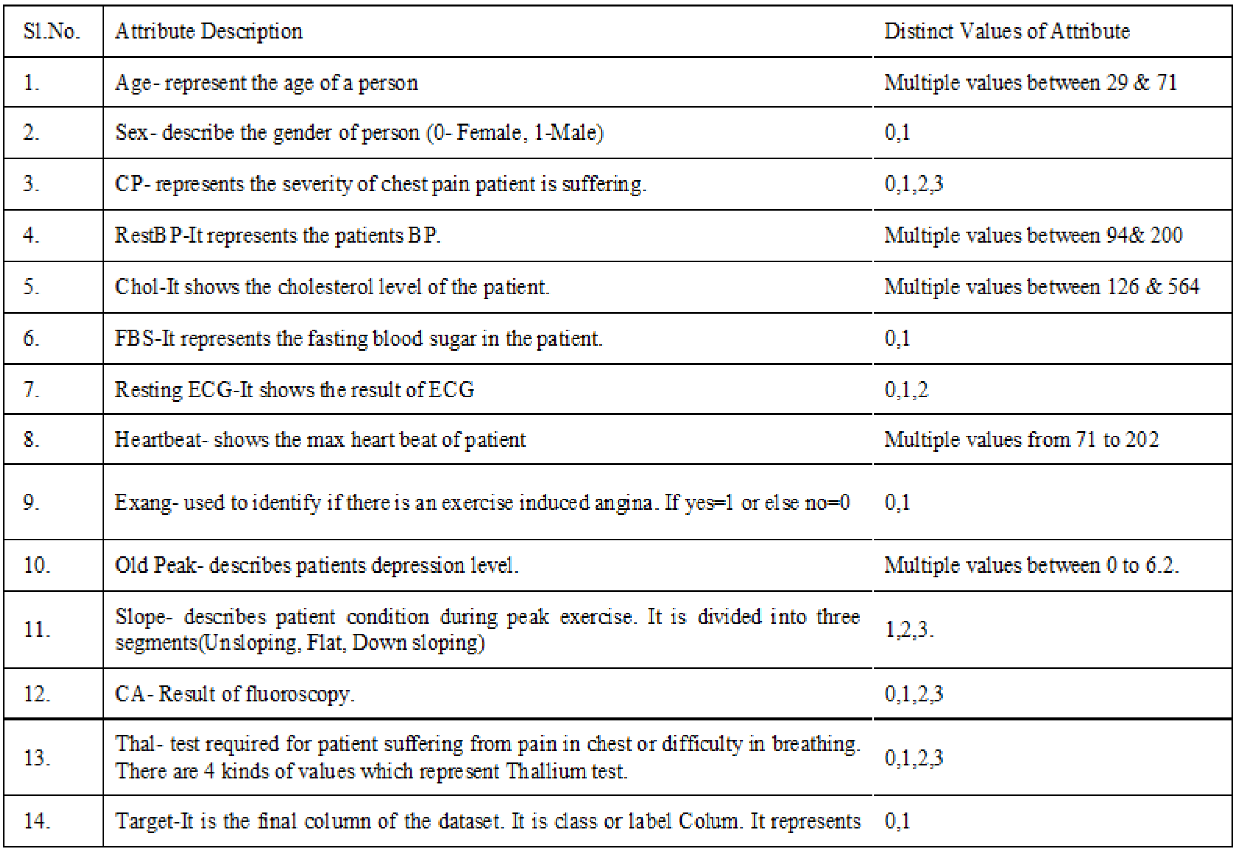
Gomathi et al. [9] used Naive Bayes and selection tree information mining methods to predict various infections. They mainly focus on the prediction of heart disease, diabetes and malignant tumors. The results come from chaotic measurements.

Miranda et al. [10] proposed a naive Bayesian classifier method to predict cardiovascular disease. The creators did not consider choosing many of the main risk factors for coronary heart disease. The proposed ideas can increase accuracy, impact and clarity by 85%.

3. Dataset Description

The dataset is collected from "heart disease dataset", and this dataset is the combination of 4 different databases, among this UCI Cleveland data set is utilized. This dataset consists of 76 attributes and 14 features [9]. Table 1 explains about the every attribute of the dataset.

ISSN: 0975-3583, 0976-2833 VOL 12, ISSUE 03, 2021



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

4. Methods

Data Cleaning and Data Pre-Processing:

Data cleaning is the Process of deleting or altering data that is missing, insufficient, redundant, decrypted, or improperly structured in order to prepare it for analysis.

Data pre - processing is the way of transforming a training dataset into a visual format. Data processing is an important phase in machine learning that improve quality performance of the data. Every statistical method's performance is hugely affected by data training strategies.

Feature selection: After the data is clean and correct, we can select relevant features from the entire data set to improve the accuracy of the classifier. Irrelevant functions are those that do not play a role in the classification. We have used the MRMR algorithm to select relevant functions from the database.

Minimum redundancy and maximum relevance (MRMR) is an element determination calculation that is used to identify the attributes of clinical data sets and limit its importance in a manner similar to that described in the combination of related element selection calculations. It is calculated by selecting bright spots that are usually far away and have a "high" connection with the characteristic variable. In AI inclusion, extraction is an important subfield. This subfield selects a subset of information related to a specific problem area, called the maximum correlation. This subset of information contains repeated paragraphs here and here, and MRMR intends to eliminate these redundant parts.MRMR has many uses, such as voice confirmation and cancer diagnosis. Using this calculation, we can remove highlights from multiple angles. One strategy for highlight extraction is to select the most basic feature variables, which is the most important choice. Another plan is to select bright spots, which are usually far away from each other and have a high relationship with ranking factors. This strategy is called "Minimum Redundancy and Maximum Correlation" and finds that the highlighted subset is more powerful than the most important process. In many cases, links can be replaced by measurable dependencies between factors. In this case, MRMR will be filled in to expand the dependence between the selected and highlighted shared transportation and order variables. The main goal of this calculation is to select bright spots by using public data, connection or distance scores. Come on, the bright spots are usually connected together and cover a very small space. The relevance of the function "S" list of class "c" is characterized by the normal inferred from all public data between the singular component "fi" and class "c", which is characterized by a given condition .

$$D(S,c) = \frac{1}{|S|}\sum_{fi\in S} I(fi,c)$$

Redundancy of all features in 'S' is the average of all mutual information between feature 'f_i' and 'f_j' is shown below and defined by the given equation[21]:

$$R(S) = \frac{1}{|S|^2}\sum_{fi,fj\in S} I(fi,fj)$$

MRMR is the combination of above mentioned equations (4.8) and (4.9) and is defined as [21]:

$$MRMR = \max_S \left[\frac{1}{|S|}\sum_{fi\in S} I(fi,c) - \frac{1}{|S|^2}\sum_{fi,fj\in S} I(fi,fj)\right]$$

Suppose there is a set of "n" functions. Assuming that "xi" is the membership index function set of the function "f_i", $x_i = 1$ (indicating existence) and $x_i = 0$ (indicating not existing) in the global optimal function set. Let us consider $c_i = I\,(f_i, c)$ and aij = I (f_i, f_j). Then, the above content can be written as an optimization problem.

Logistic Regression

This is mainly focused on predicting (1/0, yes/no), these are also considered as independent variables. Replace the variables with dummy variables to reflect binary/classification results. When dependent this variable is binary data, and we use the likelihood probability distribution as the dependent variable. In short, by adapting data to logistics the regression equation is used to predict the probability of an event.

To illustrate the doubling/direct result, we used the error factor. When the outcome variable is an absolute variable, you can also treat strategic relapse as a rare case of direct relapse. In this case, we use the logarithm of chance as the required variable. Fundamentally, it predicts the likelihood of events occurring under specific circumstances by adapting information to logical work.

It is very important for a larger range of calculations called the general linear model (GLM). The authors in 1972 proposed this model and proposed a method of using direct recursion to solve problems that are not suitable for direct recursion.

The essential condition of the summed up direct model is:

$$g(E(y)) \text{ and } a + bx_1 + cx_2$$

The link function is represented as g(), the expected variable E(y) and $a + bx_1 + cx_2$ is consider as linear predictor (a, b, c are to predict). The 'link' function is represented as y to linear predictor.

GLM does not accept the direct connection between demand factors and autonomous factors. However, it expects a direct connection between the connection work in the logit model and the autonomous factors.

The dependent variable does not have to have a normal distribution.

Let's take an example with 1,000 customers. We need to foresee the possibility of customers buying (some kind of) magazine. Obviously, we have a complete outcome variable and we will use strategic relapse.

First, starting from the calculated loop, I will use the dependent variables included in the connection work to form a direct loop condition:

$$g(y) = bo + b(Age)$$

Here the "age" is a free factor.

In strategic recurrence, we only worry about the possibility of low-level variables of outcome (achievement or disappointment). As mentioned above, g() is the connection work. Setting this function uses two things: p is represented as probability of success and (1-p) represented as probability of failure. The following are represented as:

$(p >= 0)$- represents positive

$(p <= 1)$- Never equal to 1

It must always be less than or equal to 1 (because $p <= 1$)

At present, we will meet these two conditions and enter the strategic recurrence center. To assemble the connection work, we first mark g() as p, and finally send it out. Since the probability should always be positive, the direct condition is set to an excellent structure. For any tilt and ward variable values, the condition example will never be negative.

$$p = \exp(\beta o + \beta(Age)) = e\hat{}(\beta o + \beta(Age))$$

Make the probability less than 1, p should be divide by total greater than p. This is easily done by:

$$p = \exp(\beta o + \beta(Age)) / \exp(\beta o + \beta(Age)) + 1$$
$$= e\hat{}(\beta o + \beta(Age)) / e\hat{}(\beta o + \beta(Age)) + 1$$

Using above equations, we can reclassify the possibilities as:

$$p = e\hat{}y / 1 + e\hat{}y$$

'p' represents the probability progress. (D) Logit work

'p' is the probability with success, then 1-p represents the failure probability of the point, which can be expressed as:

$q = 1 - p = 1 - (e\hat{}y / 1 + e\hat{}y)$

Where q is the probability of disappointment

Divided by (d) /(e), we get

$$\log\left[\frac{p}{1-p}\right] = y$$

The log (p / 1-p) is the connection work. The non-linear relationship is represented with straight line based on the change of the extracted variables.

The estimated value of y after subtraction operation applied, we will get:

$$\log \frac{p}{1 - p} = b_0 + b(Age)$$

Performance Evolution

The performance is estimated by using False Positive Rate (FPR), False Negative Rate (FNR), Sensitivity, Specificity and Accuracy, the performance of the system are estimated.

True Positive (TP)	True Negative (TN)
False Positive (FP)	False Negative (FN)

FPR

The overall positive predicted values were classified to normal data which is represented as:

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

FNR

The overall negative predicted values were classified to normal data which is represented as:

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

Sensitivity

The overall positives are accurately identified to calculate the sensitivity.

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

Specificity

The overall negatives are accurately identified to measure the specificity.

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

Accuracy: The overall accuracy of the result is measure with the below equation.

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

Results and Discussion

The experiments are conducted by using the java programming language with Net Beans 8.0.2. The hardware is totally based on 4 GB ram and 1 TB hard disk with I3 or I5 processor can be used. In table-1 the performance is shown.

Algorithms	Accuracy	Sensitivity	Specificity
Naive Bayes	91.42%	86.43%	79.76%
Random Forest	89.56%	89.76%	84.54%
Decision Tree	92.32	87.12	86.67
ENDDP	97.98%	97.45%	98.54%
EDDA	98.12%	98.45%	99.12%
HFEDA	98.98%	99.10%	99.60%

Table: 1 comparative result

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

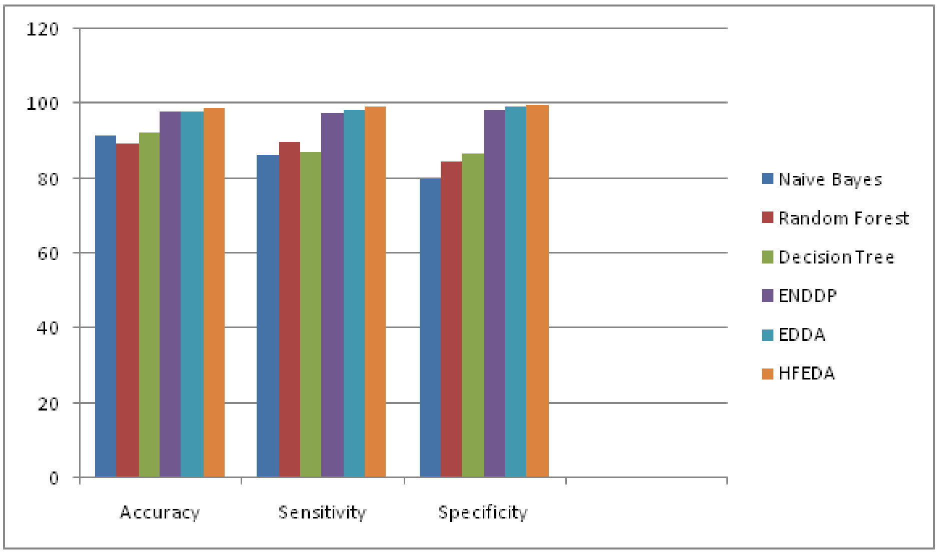


Figure: 4 the performance of the various machine learning and proposed methodolo

Conclusion

In this paper, we apply various machine learning and deep learning algorithms and to predict the heart disease very efficiently. The proposed methodology can show the improved accuracy upto 99.60%. HFEDA is a means for early heart disease risk determination using structured data.

References

[1]. J. Nageswara Rao, Dr. R. Satya Prasad, "An Enhanced Novel Dynamic Data Processing (ENDDP) Algorithm for Predicting Heart Disease in Machine Learning", IJSRCSEIT, ISSN : 2456-3307, Volume 7 Issue 1, pp. 94-104, January-February 2021.

[2]. J. Nageswara Rao, Dr. R. Satya Prasad, "An Ensemble Deep Dynamic Algorithm (EDDA) to Predict the Heart Disease", IJSRSET, Online ISSN: 2394-4099, Print ISSN : 2395-1990, Volume 8 Issue 1, pp. 105-111, January-February 2021.

[3].J. N. Rao and M. Ramesh, "A Review on Data Mining & Big Data Machine Learning Techniques", JRTE, vol. 7, no. 6S2, pp. 914-916, April 2019.

[4] Kochanek KD, Xu J, Murphy SL, Minino AM, Kung HC. Deaths: final data for 2009. Natl Vital Stat Rep. 2011;60(3):1–116. Epub 2011/12/29.

[5] Ambrosy AP, Fonarow GC, Butler J, Chioncel O, Greene SJ, Vaduganathan M, et al. The global health and economic burden of hospitalizations for heart failure: lessons learned from hospitalized heart failure registries. J Am Coll Cardiol. 2014;63(12):1123–33.

[6] Mozaffarian D, Benjamin EJ, Go AS, Arnett DK, Blaha MJ, Cushman M, et al. Heart Disease and Stroke Statistics-2016 Update: A Report From the American Heart Association. Circulation. 2016;133(4):e38–360. Epub 2015/12/18.

[7] Rine Nakanishi, Damini Dey, Frederic Commandeur, PiotrSlomka, —Machine Learning in Predicting Coronary Heart Disease and Cardiovascular Disease Events: Results from The Multi-Ethnic Study of Atherosclerosis (Mesa), JACC Mar- 20, 2018, Volume 71, Issue 11

[8] Singh P, Singh S, Pandi-Jain GS. Effective heart disease prediction system using data mining techniques. Int J Nanomed. 2018;13 (T-NANO 2014 Abstracts):121–4.

[9] Gomathi K, ShanmugaPriyaa D. Multi disease prediction using data mining techniques. Int J Syst Softw Eng. 2016;4(2):12–4.