Journal of Theoretical and Applied Information Technology
15th August 2021. Vol.99. No 15

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

DETECTION OF HEART DISEASE BY USING RELIABLE BOOLEAN MACHINE LEARNING ALGORITHM

Dr. M. BHEEMALINGAIAH[1], Dr. G. RAMA SWAMY[2], Dr. P. VISHVAPATHI[3], P. VENU BABU[4], E. NAGESWARA RAO[5], Dr. P. NAGESWARA RAO[6]

[1]Professor, Department of CSE, Malineni Lakshmaiah Women's Engineering College, A.P, India

[2]Professor, Department of CSE, Malineni Lakshmaiah Women's Engineering College, A.P, India.

[3]Professor, Department of CSE, Deccan College of Engineering & Technology, Telangana, India.

[4]Associate Professor, Department of CSE, Malineni Lakshmaiah Women's Engineering College, A.P, India.

[5]Associate Professor, Department of CSE, Malineni Lakshmaiah Women's Engineering College, A.P, India

[6]Director, Malineni Lakshmaiah Women's Engineering College, A.P, India

E-mail: [1]bheemasiva2020@gmail.com, [2]gramaswamy@gmail.com, [3]vishvapathi@deccancollege.ac.in, [4]venupanchumarthi@gmail.com, [5]eluri76@gmail.com, [6]pnrao33@gmail.com

ABSTRACT

Artificial Intelligence (A.I) is one of most exciting fields of computer engineering today. It is the science and technique used to make machine intelligent and it is vast and truly universal field. However, tremendous growth has been observed in this filed in past two decade owing to valuable contributions from variety of domains. It has numerous potential applications such as computer vision, medicine, philosophy, psychology, linguistics, automatic programming, natural language processing, speech processing and robotics, etc. Machine Learning takes training from natural events and helps in predicting any type of event and is a branch of Artificial Intelligence (AI). Over the past two decades, Machine Learning became a major source for information technology in developing applications, such as manufacturing industry for automation in assembly line, biometric recognition, handwriting recognition, medical diagnosis, speech recognition, text retrieval, natural language processing and Machine Learning is widely using in Data Science (DS), it is predominant and hotcake field of 21st century. Today all of use machine learning several times a day, without knowing it. Examples of such "ubiquitous" or "invisible" usage include search engines, customer-adaptive web services, email managers (spam filters), computer network security, and so on. Since last few decades Cardiovascular(Heart) Diseases (CVDs) has emerged as the most life-threatening diseases and proved to be fatal not only in India but throughout the whole world. In time detection, diagnosis and treatment of the disease needs a reliable, accurate and feasible system. In this paper we proposed Reliable Boolean Machine Learning Algorithm (RBMLA) by using novel approach to predict heat disease. Finally performance of RBMLA is measured by using various performance metrics like accuracy, precision, recall, sensitivity, specificity, reliability, F-score and ROC curve. It is shown that it gives better performance for given any new test data and new real time data. It has given better accuracy of 86%.

Keywords: *Support Vector Machine ,Naive Bayes algorithm, k-Nearest Neighbor Algorithm, Decision Tree Algorithm, Random Forest Algorithm, , Reliable Boolean Machine Learning Algorithm.*

1. INTRODUCTION

Over the past two decades, Machine Learning became a major source for information technology. Many machine Learning applications have been developed, such as: machine vision (image processing) in the manufacturing industry for automation in assembly line, biometric recognition, handwriting recognition, medical diagnosis, speech recognition, text retrieval, natural language processing, and so on. Today all of use machine learning several times a day, without knowing it Examples of such "ubiquitous" or "invisible" usage include search engines, customer-adaptive web services, email managers (spam filters), computer

network security, and so on. We are rethinking on everything we have been doing, with the aim of doing it differently using tools of machine learning for better success [1]. Huge volumes of historical data describing their Operations, products, and customers is being collected routinely by different organizations and at the same time complex datasets are captured by scientists and engineers. For example, banks are collecting huge volumes of customer data to analyze how people spend their money; hospitals are recording what treatments patients are on, for which periods (and how they respond to them); engine monitoring systems in cars are recording information about the engine in order to detect when it might fail; high-resolution images of night sky are being stored by world's observatories; medical science is storing the outcomes of medical tests from measurements as diverse as Magnetic Resonance Imaging (MRI) scans and simple blood tests; bioinformatics is storing massive amounts of data with the ability to measure gene expression in DNA microarrays, and so on. To improve the process of decision making machine learning will use this historical data to discover recurrent patterns.. Terminology in the field of Learning is exceptionally diverse, and very often similar concepts are variously named. The term machine Learning has been mostly used to describe various concepts, though the terms: artificial intelligence, machine intelligence, pattern recognition, statistical Learning, data mining, soft computing, data analytics (when applied in business contexts), also appear at various places. The advances in theory and algorithms has laid the foundations of machine learning field [1].

2. RELATED WORK

Experiments on medical data sets have used multiple classifiers and features selection techniques but there is little research done on the classification of the heart disease dataset.

A Novel Approach based on Significant Feature and Ensemble Learning Model has been proposed by Muhammad Affan Alim et al. (2020)[12] for early prediction of heart disease .Essentially, the aims of the paper are to find those features by correlation which can help robust prediction results.

Joshua Emakhuet al.(2020)[13] proposed Prediction System for Heart Disease Based on Ensemble Classifiers. It is designed based on Random Forest, AdaBoost, Bagging, Voting Ensemble, SVM, Logistic Regression, Decision and performance measures, such as accuracy, sensitivity, and specificity, were used to evaluate the proposed methods' performance. The proposed method achieved an accuracy of 87.04%.

Rutuja Gujare et al.(2020)[14] proposed Enhanced Heart Disease Prediction Using Ensemble Learning Methods ,which aims at enhancing the accuracy of the deficient algorithms. The predictive analytics model is used to diagnose the various stages of heart patients by ensemble Learning methods like Bagging, Boosting and Voting.

An homogeneous ensemble is created from different CART models using an accuracy based weighted aging classifier ensemble, a modification of the weighted aging classifier ensemble (WAE) by Ibomoiye Domor Mienye et al.(2020)[15].

A paper on Heart Disease Prediction Using Machine Learning Algorithms by Archana Singh and Rakesh Kumar (2020) [16] calculates accuracy by different machine Learning algorithms like k-nearest neighbor, decision tree, linear regression and support vector machine(SVM) with dataset from UCI repository.

The proposed work by Apurb Rajdhan et al.(2020)[17] predicts the chances of Heart Disease and classifies patient's risk using Machine Learning algorithms Naive Bayes, Decision Tree, Logistic Regression and Random Forest on dataset from UCI repository..

Logistic regression model of machine Learning is used to improve the Heart Disease prediction by Montu Saw et al(2020)[18] in their proposed work.

Jian Ping LI et al (2020)[19] proposed heart disease identification method using machine Learning classification in E-healthcare. It is based on classification algorithms includes Support vector machine, Logistic regression, Artificial Neural Network, K-nearest neighbor, Naïve bays, and Decision tree while standard features selection algorithms have been used such as Relief, Minimal redundancy maximal relevance, Least absolute shrinkage selection operator and Local Learning for removing irrelevant and redundant features decision tree, KNN, K-mean grouping and AdaBoost

Decision tree, KNN, K-mean grouping and AdaBoost have been used by B.KeerthiSamhitha et al (2020)[20] to improve accuracy in prediction heart disease.

A novel feature reduction (NFR) model that is aligned with the ML and DM algorithms is proposed to reduce the error rate and further improve the performance by

Syed Javeed Pasha and E. Syed Mohamed (2020)[21].

A hybrid random forest with a linear model (HRFLM) by Senthil kumar Mohan et al. (2019)

[22] with different combinations of features and several known classification techniques achieved an accuracy level of 88:7%.

A study by Noor Basha et al(2019)[23] for Early Detection of Heart Syndrome by using five machine Learning algorithms KNN, Decision Tree, Random, Forest, SVM and Naive Bayes has been proposed.

An Improved Linear Model (RERFILM) for Heart Disease Detection on the Internet of Medical Things Platform has been proposed by ChunyanGuo et al(2019)[24].

X. Liu et al. [25] presented a study to assist in the diagnosis of heart disease using a hybrid classification system based on the Relief F and Rough Set (RFRS) method. The proposed system consists of two subsystems: the RFRS feature selection system and a classification system with an overall classifier.

Combining the K-means clustering algorithm and the artificial neural network a hybrid algorithmic approach for predicting heart disease is proposed by A. Malav et al. [26] and the proposed model achieves an accuracy of 97%.

In [27] authors proposed ensemble models such as Random Forest for classification and Genetic algorithm as feature subset selection. They compared these with other classification algorithms like decision trees and proved that Random Forest improved the accuracy by removing less ranked features which in turn helped the patients indirectly by reducing the number of diagnostic tests.

In [28] authors proposed an intelligence computational method for predicting the heart disease by enhancing Naïve Bayes method and considering some feature subset selection measure like Gini-gain, Chi-square, Relief –F and Genetic Search. A One-R feature selection measure applied on Navie-Bayes classifier and obtained an accuracy of 86.29%.

Resht Agrawal[29], proposed an approach for prediction of heart disease by using clustering algorithm (DTRS) and fuzzy set. Authors observed that DTRFCS approach is more efficient than DTRS algorithm.

A.F Otoom et al. [30] presented a system for analysis and follow-up. Coronary artery disease is detected and monitored by the proposed system. Cleveland Heart data are taken from the UCI. This dataset consists of 303 cases and 76 attributes/features. 13 features are used out of 76 features.

Vembandasamy et al. [31] diagnosed heart disease using the Naive Bayes algorithm. Bayes' theorem is used in Naive Bayes. WEKA is used as a tool and performs classification using 70% of the Percentage Split. Naive Bayes offers 86.419% accuracy.

V Krishnaiah et.al [32] proposed a hybrid approach for prediction of heart disease using the Fuzzy K-NN method. This method removed the uncertainty in classical K-NN classifier and measured appropriate values to build a new model.

In [33] authors analyzed heart disease data set using AdaBoost and feature subset selection method PCA. This combination improved prediction rates by 2.11% over classification accuracy of J4.8 and 7.33% over 10 cross validations.

Hlaudi Daniel Masethe et.al [34] compared various classification algorithms such as J4.8, REPTREE, Bayescart, Naïve Bayes, and Simple CART for prediction of heart disease. Results showed that NAÏVE BAYES achieved an accuracy of 99%.

3. PROBLEM STATEMENT

Lack of (suitable) data, lack of access to the data, data bias, privacy problems, badly chosen tasks and algorithms, wrong tools and people, lack of resources, and evaluation problems are some of the reasons for which machine-Learning programs often fail to deliver expected results. Many researchers have proposed different techniques to predict heat disease. By using single machine learning algorithm to predict any type of diseases is not efficient solution. The efficient solution is by using combination of multiple machine learning algorithms. Example there is three types of COVID-19 tests, and which one is the most accurate? Molecular test (aka RNA or PCR test),

Antigen test (aka rapid test), Antibody test (aka serology test or blood test) tests may not give same result for given symptoms of any person. However it may not be found that there is a need of reliable, accurate and feasible system to predict heart diseases and it was found that most of proposed systems only consider the accuracy as primary performance metric during model evaluation. But there is limitation of accuracy in some cases. In those cases precision, recall, F-score are considered for model evolution.

To address this problem the Reliable Boolean Machine Learning Algorithm (RBLA) has been proposed by considering high reliability.

4. PROPOSED WORK

In Machine Learning Ensemble learning most popular technique, it helps improve machine learning results by combining several models.

Journal of Theoretical and Applied Information Technology
15th August 2021. Vol.99. No 15

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

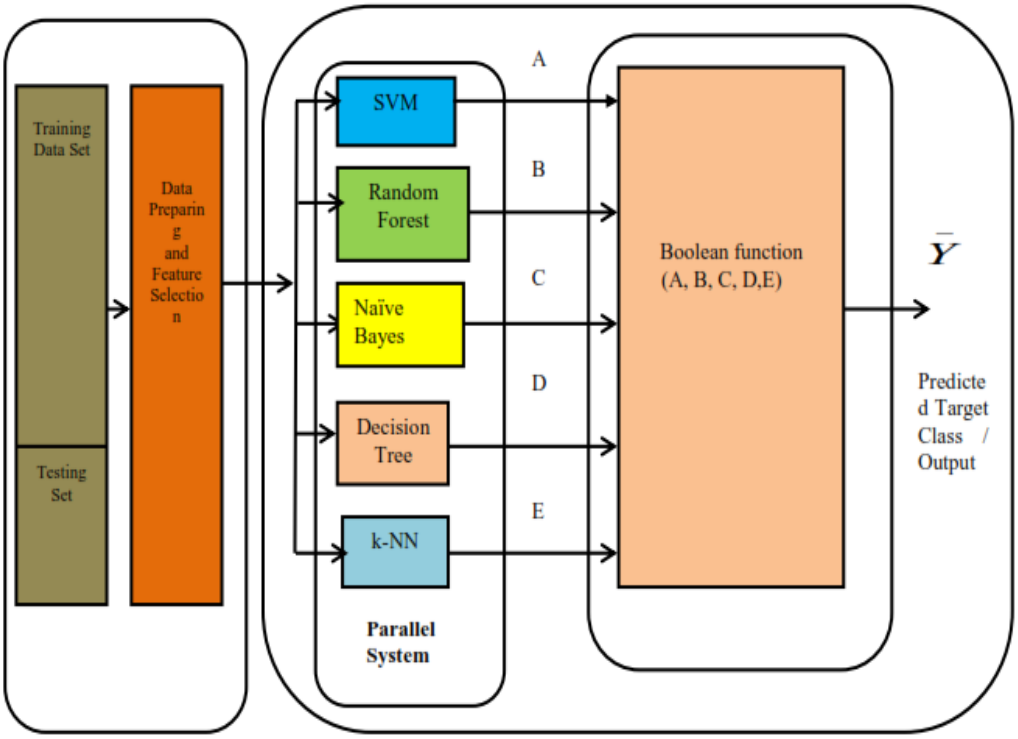

Figure 1: Block Diagram of RBMLA

The most popular ensemble methods are boosting, bagging, and stacking. Maximum Voting, Averaging, Weighted Averaging. The maximum voting method is generally used for classification problems. In this technique, multiple models are used to make predictions for each data point. The predictions by each model are considered as a 'vote'. The predictions which we get from the majority of the models are used as the final prediction. Similar to the maximum voting technique, multiple predictions are made for each data point in averaging. In this method, we take an average of predictions from all the models and use it to make the final prediction. Averaging can be used for making predictions in regression problems or while calculating probabilities for classification problems. This is an extension of the averaging method. All models are assigned different weights defining the importance of each model for prediction. Reliable Boolean Machine Learning Algorithm (RBMLA) is different type of ensemble learning technique; it is developed in novel approach by using well-known five standard supervised machine Learning algorithms and Boolean functions and maximum (majority) voting technique.

- Support Vector Machine (SVM)
- Naive Bayes algorithm
- k-Nearest Neighbor (kNN) Algorithm
- Decision Tree Algorithm
- Random Forest Algorithm

The block diagram of Reliable Boolean Machine Learning Algorithm (RBMLA) is shown fig.1 Reliable Boolean Machine Learning Algorithm (RBMLA) accepts five different outputs from five selected above machine learning algorithms as its inputs. It generates or predicts output $\bar{Y}$ based on majority of inputs. It is represented by Boolean function. If all its inputs are zeros then it predicts $\bar{Y}$ output/target class as –ve class, if all its inputs are ones then it predicts output/target class as +ve class. If number of 1's is greater than 0's then it predicts output/ target class as +ve class otherwise –ve class by using corresponding Boolean function as shown table 1. In the truth table has five inputs variables and one output variable. Maximum number of entries in is 2^5=32 different possible combinations from 0 to 31. First all five machine learning algorithms were trained with heart disease data set, it is clearly explained in section 5. Performance analysis of RBMLA.

Journal of Theoretical and Applied Information Technology
15th August 2021. Vol.99. No 15

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 1: Truth Table

S No	E	D	C	B	A	$\bar{Y}$
0	0	0	0	0	0	0
1	0	0	0	0	1	0
2	0	0	0	1	0	0
3	0	0	0	1	1	0
4	0	0	1	0	0	0
5	0	0	1	0	1	0
6	0	0	1	1	0	0
7	0	0	1	1	1	1
8	0	1	0	0	0	0
9	0	1	0	0	1	0
10	0	1	0	1	0	0
11	0	1	0	1	1	1
12	0	1	1	0	0	0
13	0	1	1	0	1	1
14	0	1	1	1	0	1
15	0	1	1	1	1	1
16	1	0	0	0	0	0
17	1	0	0	0	1	0
18	1	0	0	1	0	0
19	1	0	0	1	1	1
20	1	0	1	0	0	0
21	1	0	1	0	1	1
22	1	0	1	1	0	1
23	1	0	1	1	1	1
24	1	1	0	0	0	0
25	1	1	0	0	1	1
26	1	1	0	1	0	1
27	1	1	0	1	1	1
28	1	1	1	0	0	1
29	1	1	1	0	1	1
30	1	1	1	1	0	1
31	1	1	1	1	1	1

All 1's of output function are entered into five variables K-map as shown in fig.2

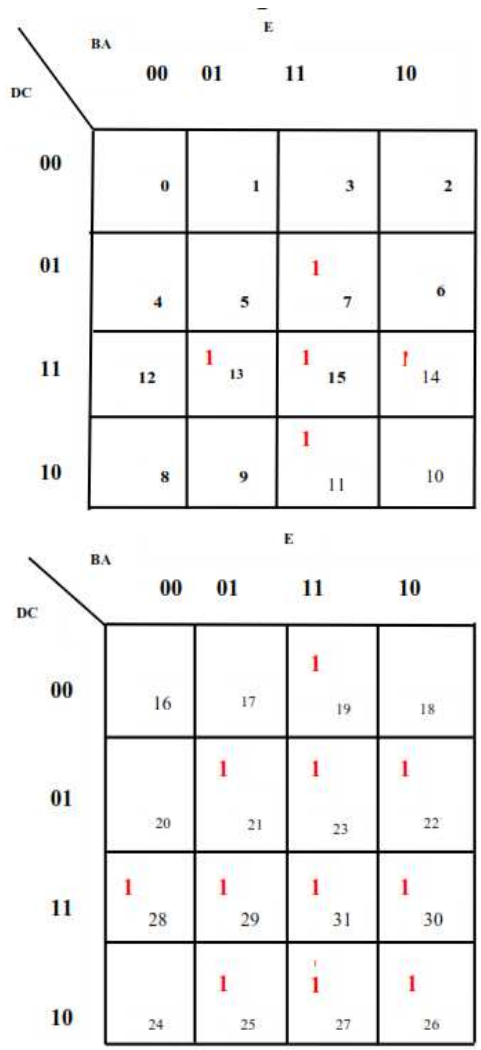

Figure 2: K-map

The following ten are products terms also called Min terms

Quad1 : $\prod (m_{19}, m_{23}, m_{27}, m_{37}) = EBA$
Quad2 : $\prod (m_{28}, m_{29}, m_{30}, m_{31}) = EDC$
Quad3 : $\prod (m_{13}, m_{15}, m_{29}, m_{31}) = DCA$
Quad4 : $\prod (m_{11}, m_{15}, m_{27}, m_{31}) = DBA$
Quad5 : $\prod (m_{14}, m_{15}, m_{30}, m_{31}) = DCB$
Quad6 : $\prod (m_{7}, m_{15}, m_{23}, m_{31}) = CBA$
Quad7 : $\prod (m_{21}, m_{23}, m_{29}, m_{31}) = ECA$
Quad8 : $\prod (m_{22}, m_{23}, m_{30}, m_{31}) = ECB$
Quad9 : $\prod (m_{25}, m_{27}, m_{29}, m_{31}) = EDA$
Quad10 : $\prod (m_{26}, m_{27}, m_{30}, m_{31}) = EDB$

After mapping resulting K-map is shown following fig.3 in next page.

Journal of Theoretical and Applied Information Technology
15th August 2021. Vol.99. No 15

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

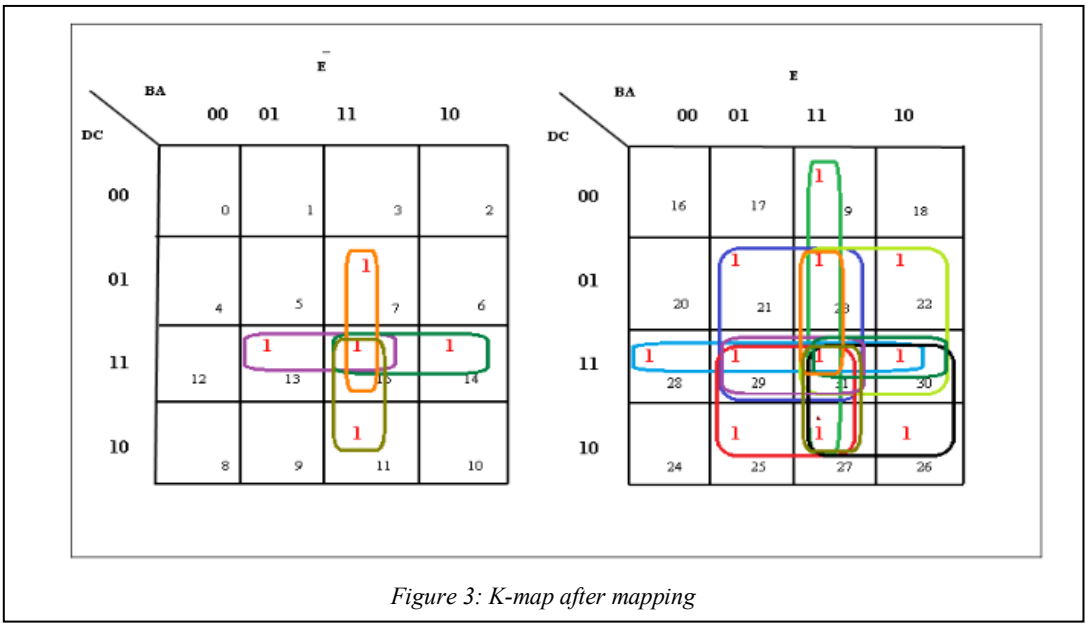

Figure 3: K-map after mapping

The output Boolean function $\bar{Y}$ is sum of products

$$\bar{Y} = EBA + EDC + DCA + DBA + DCB + CBA + ECA + ECB + EDA + EDB$$

The output Boolean function can be implemented by using ten AND gates and one OR gate as shown as fig.4

Manual Testing of Output Boolean function

Case 1: Randomly select 4th entry of truth table and substitute input values in Boolean output function

S No	E	D	C	B	A	$\bar{Y}$
4	0	0	1	0	0	0

$\bar{Y}$ = EBA + EDC+ DCA+ DBA + DCB + CBA + ECA + ECB + EDA + EDB

0.0.0 + 0.0.1 + 0.1.0 + 0.0.0 + 0.1.0 + 1.0.0 + 0.1.1 + 0.1.0 + 0.0.0 + 0.0.0 = 0+0+0+0+0+0+0+0+0+0 = 0

$\bar{Y}$ = 0, predicted target class is negative

Hence design is correct

Case 2: Randomly select 28th entry of truth table and Substitute input values in the out function

S No	E	D	C	B	A	$\bar{Y}$
28	1	1	1	0	0	1

=1.0.0 + 1.1.1 + 1.1.0 + 1.0.0 + 1.1.0 + 1.0.0 + 1.1.0 + 1.1.0 + 1.1.0 + 1.1.0 = 0+1+0+0+0+0+0+0+0+0 = 1

$\bar{Y}$ = 1, predicted class in positive, hence design is correct.

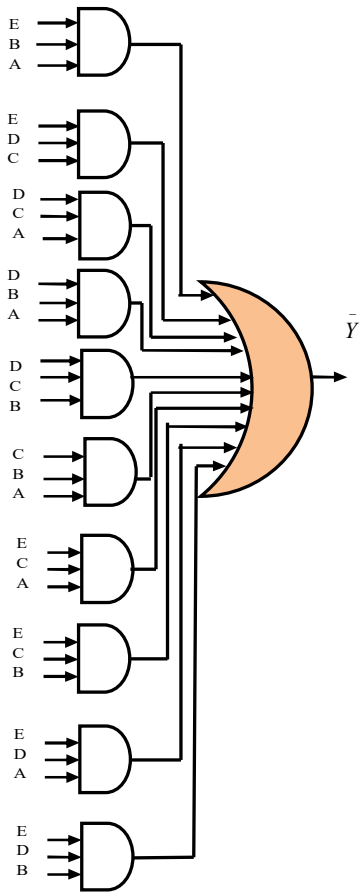

Fig.4. Logic Circuit

4.2 Reliability Estimation of RBMLA

Let X be the life time or the time of failure of a component. The probability that the component survives until sometime t is called as the reliability R(t) of component [10] [11] [12]. This R(t) = P(X > t) = 1 – f(t) Where F is the distribution function (df) life time X. Note that the components is assumed to be working properly at time t =0 then R(0)=1. Also no component can work forever without failure i.e $li_{mn\to\infty} R(t) = 0$.

Note: 1. R(t) a monotonically increasing function of t.

2. For t<0 reliability has no meaning. But we let R(t) = 1 for t<0. F(t) will then be called unreliability. Consider fixed number of identical components under the test. After time t_0.$N_f(t)$. components failed and $N_s(t)$ components have survived with

$$N_f(t)+N_s(t)=N_0$$

$$R(t)=P(survival)=\frac{N_f(t)}{N_0}$$

$$\frac{N_0-N_s(t)}{N_0}=1-\frac{N_f(t)}{N_0}$$

The total number of components N_0 is a constant while $N_s(t)$ the number of failed components increases with time. Taking derivatives on both sides we get

$$R(t)=\frac{-N_f(t)}{N_0}$$

Here $N_f(t)$ is the rate at which the components fail. A $N_0\to\infty$ the RHS is interpreted as the negative of the failure density function $f_x(t) i.e. R(t)=-f_x(t)$.

The conditional probability that the component does not survive for an additional interval of duration x given that it has survived until time t can be written as $\left(P(X)>t+\frac{x}{x}>t\right)$

$$G_f(x)=\frac{P(<x<t+x)}{R(t)}$$

RBMLA accepts the five inputs from Parallel System (S). It conations five subsystems (Machine Learning Algorithms).as shown in fig .5

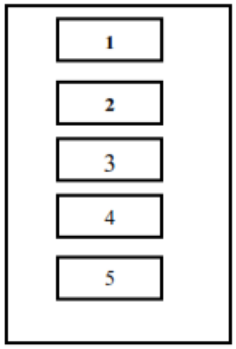

Fig. 5. Parallel System

Here reliability of machine learning algorithm is considered as successful operation or correctly predicted output without generating wrong output at time t. Now we can estimate the reliability of Parallel System by considering individual reliability of each subsystem. Let X_1, X_2, X_3, X_4, X_5 represents the successful operation of five subsystems respectively at time t with their corresponding probabilities $P(X_1)$, $P(X_2)$, $P(X_3)$, $P(X_4)$, $P(X_5)$. For i=1, 2,3,4,5. Let $P(\overline{X_i})=1-P(X_i)$ denotes the probability of failure that indicates that system i wrongly predicts the output at time t. Hence, for the complete failure of system all the five subsystems have to fail simultaneously at time t. If $P(\bar{s})$ is the probability of failure of the system at time t. Then

$$P(\overline{S})=P(\overline{X}_1)x\,P(\overline{X}_2/\overline{X}_1)x$$
$$P(\overline{X}_3/\overline{X}_2\ and\ \overline{X}_1)\,x$$
$$P(\overline{X}_4/\overline{X}_3\ and\ \overline{X}_2/\overline{X}_1)\,x$$
$$P(\overline{X}_5/\overline{X}_4\ and\ \overline{X}_3\,\overline{X}_2\,\overline{X}_1)$$

Hence the subsystem failures are independent then

$$P(\bar{S}) = P(\bar{X}_1)\ P(\bar{X}_2)\ P(\bar{X}_3)P(\bar{X}_4)P(\bar{X}_5)$$

$$P(\bar{S}) = 1 - P(S)$$

$$P(S) = 1 - P(\bar{S})$$

$$P(S) = 1 - \left(P(\bar{X}_1)\ P(\bar{X}_2)\ P(\bar{X}_3)P(\bar{X}_4)P(\bar{X}_5)\right)$$

$$P(S)=1-\left((1-P(X_1))\ (1-P(X_2))(1-P(X_3))(1-P(X_4))(1-P(X_5))\right)$$

In general i=1 to n then

$$P(S)=1-\prod_{i=1}^{n}(1-P(X_1))$$

This is called product law of unreliability. Also if n subsystems are identical and if the subsystem failures are independent of one another th**en**

$$P(\bar{S}) = 1 - P(S)$$

$$P(\bar{S}) = 1 - (1 - P(X))^n$$

We can also denotes the reliability R_i by $P(X_i)$ and let Rs denotes the reliability of the system . The probability law of unreliabilities

$$P(S) = 1 - \prod_{i=1}^{n} (1 - Rs)$$

The following subsections from 4.3 to 4.7 describe existing standard supervised machine learning algorithms with mathematical equations, their merits and demerits and their applications

4.3 Support Vector Machine (SVM)

To predict and classify data various machine learning algorithms are used according to the dataset. A linear model for classification and regression problems is Support Vector Machine (SVM)[1]. It can solve linear and non-linear problems and work well for many practical problems. SVM creates a line or a hyperplane which separates the data into classes.Fig.6 shows four hyperplane.

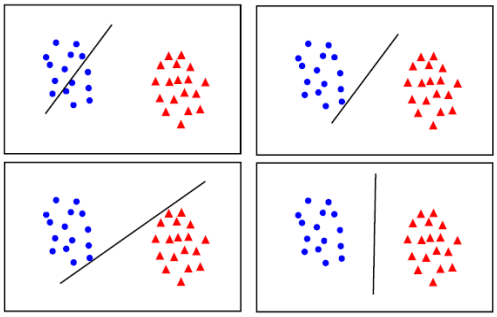

Fig.6. Four Hyperplanes

An optimal hyperplane (i.e., decision boundary, see Fig. 7) separates the tuples of one class from another in classification. SVM takes care of dimensionality problem as it works well with higher dimensional data. Even though the training time is high and is extremely slow, the result, is however highly accurate. SVM is less prone to over fitting than other methods. The main key concept in SVM is maximum margin hyperplane. Linear SVM is a classification technique when training data are linearly separable. Non-linear SVM is a classification technique when training data are linearly non-separable.

4.3.1 Linear discriminant function:

Linear discriminant function for n-dimensional feature space in [1] , $R^n : g(X) = w^T x + w_0 = 0$

$x = [x_1\, x_2\, x_3 \ldots . x_n]^T$ is the feature vector and $W = [w_1\, w_2\, w_3 \ldots . w_n]^T$ the *weight vector*. w_0 is the *Bias* parameter. Discriminant function hyperplane H.[1]. For discriminant function $g(x)$, two-category classifier has the decision rule: Decide Class 1 if $g(x) > 0$ and Class 2 if $g(x) < 0$ [1]. In fig 7, the location of any point **x** maybe considered relative to H. Defining x_p as the normal projection of **x** onto H as shown in fig.7

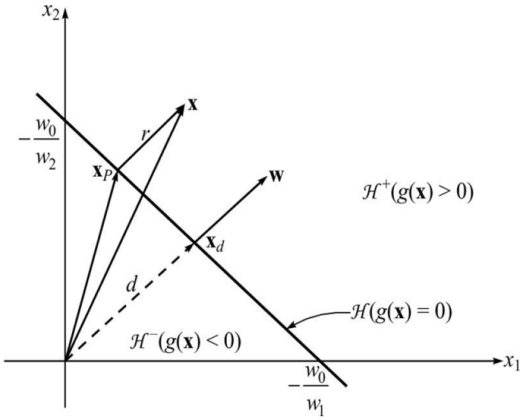

Fig.7. Linear decision boundary between two classes

$$x = x_p + r\frac{w}{\|w\|}$$

Where $\|w\|$ is the Euclidean norm of w and $\frac{w}{\|w\|}$ is a unit vector. It can be shown that

$$r = \frac{g(x)}{\|w\|}$$

$|g(x)|$ is a measure of the Euclidean distance of the point **x** from the decision hyperplane $\mathcal{H}$

$$g(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0 \begin{cases} > 0 \text{ if } \mathbf{x} \in \mathcal{H}^+ \\ = 0 \text{ if } \mathbf{x} \in \mathcal{H} \\ < 0 \text{ if } \mathbf{x} \in \mathcal{H}^- \end{cases}$$

Perpendicular distance d from coordinate origin to $\mathcal{H} = w_0 / \|\mathbf{w}\|$

Fig. 8 shows Geometry for 3-dimensions (n=3) [1]

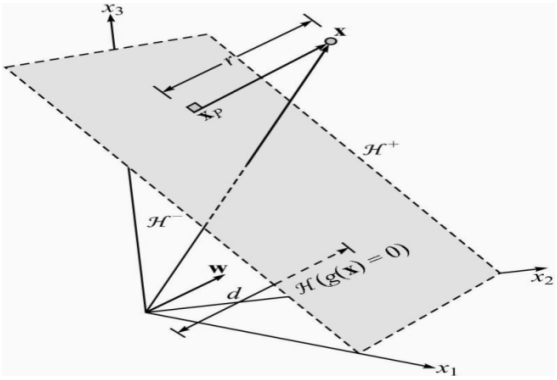

Fig.8. Hyperplane $\mathcal{H}$ separates the feature space into two half space $\mathcal{H}^+$ and $\mathcal{H}^-$

4.3.2 Linear Maximal Margin Classifier for Linearly Separable Data:

Linearly separable, many hyper planes exist to perform separation[1]. SVM framework tells which hyperplane is best[1] . Hyperplane with the largest margin which minimizes training error. Select the decision boundary that is far away from both the classes [1]. Large margin separation is expected to yield good generalization as shown fig.9.

Two parallel hyperplanes $\mathcal{H}_1$ and $\mathcal{H}_2$ that pass through $x^{(i)}$ and $x^{(k)}$ respectively as shown fig.10. $\mathcal{H}_1$ and $\mathcal{H}_2$ are parallel to the hyperplane[1]

$W^T x + w_0 = 0$

$\mathcal{H}_1$: $W^T x + W_0 = +1$

$\mathcal{H}_2$: $W^T x + W_0 = -1$

Such that

$W^T X^{(i)} + w_0 \geq 1$ if $y^{(i)} = +1$

$W^T X^{(i)} + w_0 \leq -1$ if $y^{(i)} = -1$

or equivalently,

$y^{(i)} (W^T X^{(i)} + W_0) \geq 1$

Distance d between the two Hyperplanes=margin M where $M = \frac{2}{\|W\|}$ [1]

4.3.3 Learning problem in SVM:

Linearly separable training examples

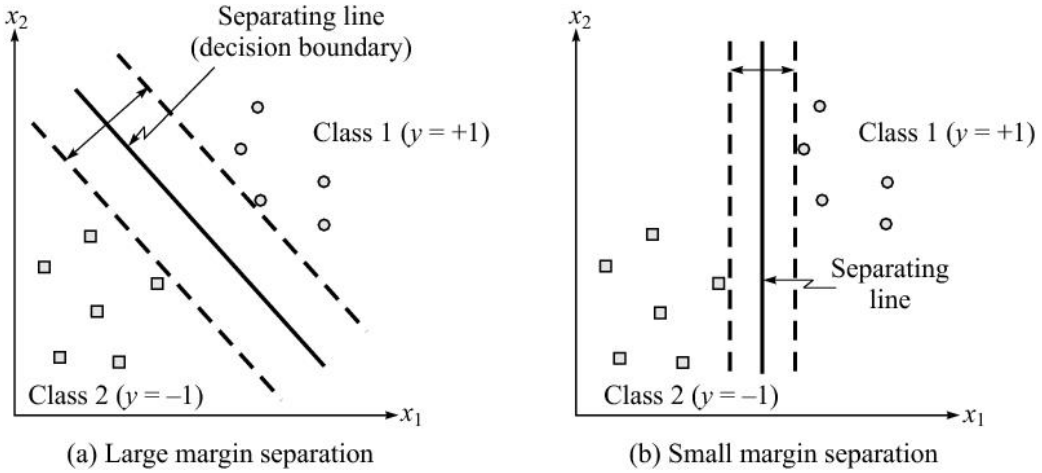

Fig.9.Large margin and small margin separation

$$D = \{(x^{(1)}, h^{(1)}), (x^{(2)}, y^{(2)}) \ldots\ldots, (x^{(N)}, y^{(N)})$$

Problem: Solve the following constrained minimization problem [1]

$$\min imize\ f(w) = \frac{1}{2} w^T w$$

$$subject\ to\ y^{(1)} (W^T x^{(i)} + W_0) \geq 1\ ;\ i = 1, \ldots N$$

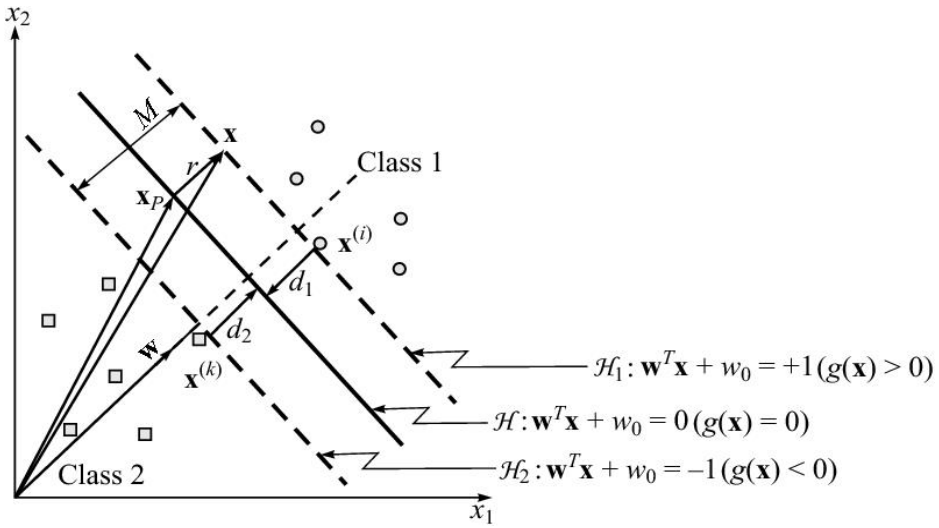

Fig.10. Two parallel hyperplanes $\mathcal{H}_1$ and $\mathcal{H}_2$

This is the formulation of *hard-margin* SVM.
Dual formulation of constrained optimization problem [1]:

Lagrangian is constructed:

$$L(w, w_0, \lambda) = \frac{1}{2} w^T\ w - \sum_{i=1}^{N} \lambda_i \left[y^{(1)} \left(w^T\ x^{(i)} + w_0 \right) - 1 \right]$$

The KKT conditions are as follows [1]:

$$1.\ \frac{\partial L}{\partial w} = 0 \Rightarrow w = \sum_{i=1}^{N} \lambda_i\ y^{(i)}\ x^{(i)}$$

$$\frac{\partial L}{\partial w_0} = 0 \Rightarrow \sum_{i=1}^{N} \lambda_i\ y^{(i)} = 0$$

$$2.\ y^{(i)} \left(w^T\ x^{(i)} + w_0 \right) - 1 \geq 0;\ i = 1, \ldots, N$$

$$3.\ \lambda_i \geq 0;\ i = 1, \ldots, N$$

$$4.\ \lambda_i \left(y^{(i)} \left(w^T\ x^{(i)} + w_0 \right) - 1 \right) = 0;\ i = 1, \ldots N$$

After solving the dual problem numerically, the resulting optimum λ_i values are used to compute $w\ and\ w_0$ using the KKT conditions [1].

Hence it all boils down to optimization problem

$$\min imize\ \frac{1}{2} w^T\ w + C \sum_{i=1}^{N} \varsigma_i$$

Subject to

$$y^{(i)} \left(w^T\ x^{(i)} + w_0 \right) \geq 1 - \varsigma_i;\ i = 1 \ldots\ldots, \varsigma_i \geq 0, i = 1, \ldots, N$$

This formulation is the soft margin SVM.
Lagrangian [1]

$$L(w, w_0, \varsigma, \lambda, \mu) = \frac{1}{2} w^T\ w +$$

$$C \sum_{i=1}^{N} \varsigma_i - \sum_{i=1}^{N} \lambda_i [\left(w^T\ x^{(i)} + w_0 \right) - 1 + \varsigma_i] - \sum_{i=1}^{N} \mu_i \varsigma_i$$

Where μ_i, $\lambda_i \geq 0$ are the dual variables?

4.3.4 Non-linear classifiers SVM

For training examples which cannot be linearly separated [1]. In the feature space, they can be separated linearly with some transformations as shown in fig.11

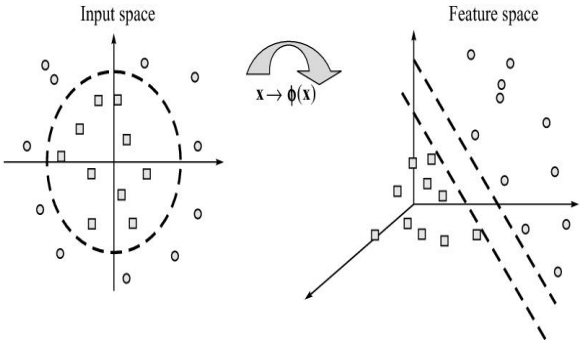

Fig.11.Transformation from input space to feature space

The new optimization problem becomes [1]

$$\min imize \frac{1}{2} w^T \ w + C \sum_{i=1}^{N} \varsigma_i$$

Subject to

$$y^{(i)} \left(w^T \phi \left(x^{(i)} \right) + w_0 \right) \geq 1 - \varsigma_i ; i = 1, \ldots, N \varsigma_i \geq 0; i = 1, \ldots, N$$

The corresponding dual is [1]

$$\min imize \ L_* (\lambda) =$$

$$\sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{k=1}^{N} \lambda_i \lambda_k y^{(i)} \left[\phi \left(x^{(i)} \right) \right]^T \phi \left(x^{(k)} \right)$$

$$Subject \ to \sum_{i=1}^{N} \lambda_i \ y^{(i)} = 0$$

$$0 \leq \lambda_i \ 0 \leq \lambda_i \leq C; i = 1, \ldots, N$$

The decision boundary becomes [1]:

$$\sum_{i=1}^{N} \lambda_i \ y^{(1)} \left[\phi \left(x^{(1)} \right) \right]^T \ \phi (x) + w_0 = 0$$

Is there a need to know the mapping of ϕ? No.

In SVM, this is done through the use of *kernel function*, denoted by *K*.

$$K \left(x^{(1)}, x \right) = \left[\phi \left(x^{(1)} \right) \right]^T \ \phi (x)$$

There is no explicit need to know what ϕ is.

Constructing Kernels [1]:

Does any kernel work? No, only valid kernel functions work. Identification of ϕ is not needed if it can be shown whether the function is a kernel or not without the need of mapping. [1] .Function satisfying Mercer's theorem can work as kernel function.Mercer's theorem, which provides a test whether a function $K \left(x^{(1)}, x^{(k)} \right)$ constitutes a valid kernel without having to construct the function $\phi (x)$ [1].

$$K \left(x^{(1)}, x^{(k)} \right) = \left[\phi \left(x^{(1)} \right) \right]^T \ \phi \left(x^{(k)} \right)$$

Common kernel functions used *as Polynomial kernel of degree d*

$$K \left(x^{(1)}, x^{(k)} \right) = \left(x^{(i)^T} \ x^{(k)} + c \right)^a ; c > 0, \ d \geq 2$$

Gaussian radial basis function kernel (RBF) [1]

$$K \left(x^{(1)}, x^{(k)} \right) = \exp \left(\frac{\left\| x^{(i)} - x^{(k)} \right\|^2}{2 \sigma^2} \right), \sigma > 0$$

4.3.5 Validation of SVM

- SVM gives us resultant classes not the probability.
- Sensitivity, Specificity, cross validation, ROC and AUC are the validation methods
- An optimal hyperplane with a maximized margin need to be defined
- Data is mapped to a high dimensional space to classify with linear decision surfaces
- Problem is reformulated so that data is mapped implicitly into this space

4.3.6 Real Life Applications of SVM

- Face detection classify between face and non-face areas on images
- Text and hypertext categorization
- Images Classification
- Bioinformatics: protein, genes, biological or cancer classification.
- Handwriting recognition
- Drug Discovery for Therapy
- Protein Structure Prediction
- Intrusion Detection
- Handwriting Recognition
- Detecting Steganography in digital images
- Breast Cancer Diagnosis
- Almost all the applications where ANN is used
- SVM has played a very important role in cancer detection and its therapy with its application in classification.

4.3.7 Advantages and Disadvantages of SVM

Advantages:

- SVM works relatively well when there is a clear margin of separation between classes.
- Effective in high dimensional spaces.
- Effective in cases where the number of dimensions is greater than the number of samples.
- Relatively memory efficient
- Very good when we have no idea on the data.
- Works well with even unstructured and semi structured data like text, Images and trees.
- The kernel trick is real strength of SVM which can solve any complex problem.
- Unlike in neural networks, SVM is not solved for local optima.
- It scales relatively well to high dimensional data.
- SVM models have generalization in practice; the risk of over-fitting is less in SVM.
- SVM is always compared with ANN. When compared to ANN models, SVMs give better results.

Disadvantages of SVM

- Not suitable for large data sets.
- Does not perform very well when the data set has more noise i.e. target classes are overlapping.
- In cases where the number of features for each data point exceeds the number of training data samples, the SVM will underperform.

- As the support vector classifier works by putting data points, above and below the classifying hyperplane there is no probabilistic explanation for the classification.
- Choosing a "good" kernel function is not easy.
- Long training time for large datasets.
- Difficult to understand and interpret the final model, variable weights and individual impact.
- Since the final model is not so easy to see, we cannot do small calibrations to the model hence its tough to incorporate our business logic.
- The SVM hyper parameters are Cost -C and gamma. It is not that easy to fine-tune these hyper-parameters. It is hard to visualize their impact

4.4 Naive Bayes algorithm

A supervised Learning algorithm based on Bayes theorem is Naïve Bayes algorithm mainly used in text classification that includes a high-dimensional training dataset. A simple and most effective Classification algorithm helps in building the fast machine learning models that can make quick predictions. It is a probabilistic classifier, which means it predicts on the basis of the probability of an object and used in spam filtration, Sentimental analysis, and classifying articles.

4.4.1 Why is it called Naïve Bayes?

The meaning of Naïve in this algorithm is that it assumes that the occurrence of a certain feature is independent of the occurrence of other features. Such as if the fruit is identified on the bases of color, shape, and taste, then red, spherical, and sweet fruit is recognized as an apple. Hence each feature individually contributes to identify that it is an apple without depending on each other. And Bayes' Theorem is included in this algorithm

4.4.2 General Bayes Theorem

Let $\{Y_1, \ldots Y_M\}$ be the finite set of M classes in data D [1] , $Y_q; q = 1, \ldots, M$ is a variable that must be described probabilistically. Feature vector $x \; in \; D$ be an $n-$component random variable. $p(x)$ be the probability density function that x will be observed. $p(x|y_q)$ be the conditional probability density function for **x** conditioned on y_q being its class[1]. $p(y_q)$ be the conditional probability that **x** belongs to class y_q [1].

Posterior probability $p(y_k \mid x)$ is calculated from $p(x|y_q)$ and $p(y_q)$ using Bayes formula [1][2]:

$$p(y_k \mid x) = \frac{p(x \mid y_k)\, p(y_k)}{p(x)}$$

Where

$$p(x) = \sum_{q=1}^{M} p(x|y_q)\, p(y_q)$$

Bayes theorem provides a way to calculate the probability of a class k based on its prior $p(y_k)$, the probability density function $p(x \mid y_k)$

In terms of probability mass functions, the posteriors can be calculated as [1]

$$P(y_q \mid x) = \frac{P(y_k) P(x \mid y_a)}{\sum_{q=1}^{M} P(x \mid y_q) P(y_q)}$$

Learner wants to find the most probable class k given the observed **x**.

Any such maximally probable class is called a *Maximum A Posteriori* (MAP) class determined by [1]:

$$Class \; k \; if \; P(Y_k \mid x) = \max_q P(y_q \mid X)$$

y_{MAP} Corresponds to MAP class provided

$$y_{MAP} \equiv \arg\max_q P(y_q|x)$$

$$\equiv \arg\max_q \frac{P(y_q) P(x \mid y_q)}{P(x)}$$

$$\equiv \arg\max_q \; P(y_q) P(x|y_q)$$

Sometimes every class is equally probable *a priori* $(P(y_q) = P(y_k); \forall_{q,k})$[1]. In this case, only consider the term $P(x \mid y_q)$ to find the most probable class [1]. Since $P(x \mid y_q)$ represents the likelihood of the data x given class y_q, any class that maximizes $P(x \mid y_q)$ is called *Maximum Likelihood* (ML) class [1]. Thus y_{ML} corresponds to ML class provided [1]

$$y_{ML} \equiv \arg\max_q P(x|y_q)$$

4.4.3 Naive Bayes Classifier

The practicability of Bayes theorem lies in the fact that conditional probability function $P(y_q \mid x)$ can be calculated from $P(x \mid y_q)$ and $P(y_q)$, which can be estimated from data[1].

The *naïve Bayes classifier* uses data-based probability estimation. In typical supervised

pattern classification problems, each of $P(y_q)$ may be estimated simply by counting the frequency y_q occurs in the training data [1]

$$P(y_q) = \frac{Number\ of\ data\ with\ class\ y_q}{Total\ number\ (N)\ of\ dat}$$

Class-conditional probabilities $(x \mid y_q)$:

$$P(x \mid y_q) = \frac{Number\ of\ times\ pattern\ x\ appears\ in\ y_q\ class}{Number\ of\ times\ y_q\ appears\ in\ the\ data}$$

The naïve Bayes classifier is based on assumption that given class, probability of observing conjunction $x_1, x_2, \ldots\ldots, xn$[1];

$$P(x_1, x_2, \ldots x_n \mid y_q) = \prod_j P(x_j \mid y_q)$$

Substituting this, we have the naïve Bayes algorithm:

$$y_{NB} = \arg\max_q P(y_q) \prod_j P(x_j \mid y_q)$$

Where y_{NB} denotes the class output. With conditional independence, the MAP classifier becomes the NB classifier [1]. Let the value of x_j be v_{ixj}. Then

$$P(x_j \mid y_q) = N_{qv_{ixj}} / N_q$$

Where N_{qvixj} is the number of training samples of class y_q having the value v_{ixj} for attribute x_j, and N_q is the total number of training samples with class y_q [1].Class prior probabilities may be calculated as $P(y_q) = \frac{N_q}{N}$

Where N is the total number of training samples, and N_q is the number of samples of class y_q [1].

4.4.4 Advantages of Naïve Bayes Classifier:

- Fast and easy ML algorithm to predict a class of datasets.
- Applied to Binary as well as Multi-class Classifications.
- Compared to the other Algorithms it performs well in Multi-class predictions.
- Mostly used for text classification problems.
- When assumption of independence holds, a Naive Bayes classifier performs better compare to other models like logistic regression and you need less training data.
- It performs well in case of categorical input variables compared to numerical variable(s). For numerical variable, normal distribution is assumed (bell curve, which is a strong assumption.

4.4.5 Disadvantages of Naïve Bayes Classifier:

- Naive Bayes assumes that all features are independent or unrelated, so it cannot learn the relationship between features.[2][3][4]
- If categorical variable has a category (in test data set), which was not observed in training data set, then model will assign a 0 (zero) probability and will be unable to make a prediction. This is often known as "Zero Frequency". To solve this, we can use the smoothing technique. One of the simplest smoothing techniques is called Laplace estimation.
- On the other side naive Bayes is also known as a bad estimator, so the probability outputs from predict probabilities are not to be taken too seriously.
- Another limitation of Naive Bayes is the assumption of independent predictors. In real life, it is almost impossible that we get a set of predictors which are completely independent.

4.4.6 Applications of Naïve Bayes Classifier

- Credit Scoring.
- Medical data classification.
- Text classification such as Spam filtering and Sentiment analysis.
- Real time Prediction: .
- Multi class Prediction:
- Text classification/ Spam Filtering/ Sentiment Analysis:
- Recommendation System:

4.4.7 Types of Naïve Bayes Model:

There are three types of Naive Bayes Model, which are given below:

- Gaussian: The Gaussian model assumes that features follow a normal distribution.
- Multinomial: The Multinomial Naïve Bayes classifier is used when the data is multinomial distributed.
- Bernoulli: The Bernoulli classifier works similar to the Multinomial classifier,

4.5 k -Nearest Neighbor (k -NN) Algorithm

One of the simplest Machine Learning algorithms based on Supervised Learning technique is K-Nearest Neighbor. K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. K-NN algorithm stores all the available data and classifies

a new data point based on the similarity [1] [2] [3] [4] [5]. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm. K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.

k-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data. It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset. KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

4.5.1. How does k-NN work?

The k-NN algorithm:
Step-1: Select the number k of the neighbors
Step-2: Calculate the Euclidean distance of k number of neighbors
Step-3: Take the k nearest neighbors as per the calculated Euclidean distance.
Step-4: Among these k neighbors, count the number of the data points in each category.
Step-5: Assign the new data points to that category for which the number of the neighbor is maximum.
Step-6: Our model is ready.

Fig.12 shows k-NNN classification

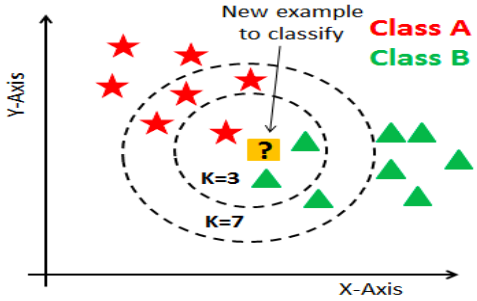

Fig.12. k-KN with k=7

4.5.2 How to select the value of k in the k-NN Algorithm?

Below are some points to remember while selecting the value of l in the k-NN algorithm:

- There is no particular way to determine the best value for "k", so we need to try some values to find the best out of them. The most preferred value for k is 5.
- A very low value for k such as k=1 or k=2, can be noisy and lead to the effects of outliers in the model.
- Large values for k are good, but it may find some difficulties

4.5.3 Mathematical model of k-NN

Consider points that are nearest to using certain distance matrix. Classification of class label found in majority among neighbors. Neighbors have equal vote and class having the maximum number of votes among the neighbors is chosen. Ties are broken arbitrarily or weighted vote is taken. – odd number to minimize ties .Key component - distance/similarity function chosen based on application and nature of data. Number of neighbors?.Imperially determined from validation set. No right number. Number depends on distribution of data and dependent on the problem.

Classification and Discriminant function

Classification decisions based on x may be stated using a set of explicitly defined discriminant function $g_q(x);\ q=1,\ 2,\ \ldots\ldots,\ M$

Where each discriminant is associated with a particular recognized class $y_q;\ q=1,2,\ldots,\ M$

The classifier designed assigns a pattern with feature vector x to class y_k such that corresponding value g_k is the largest:

$$gl(x)>g_q(x)\forall_q=1,2,\ldots.,\ M;_q\neq k$$

In the case of binary classification, instead of two discriminant functions applied, it is more common to define a single function and to use the following decision rule:

$$Decide\ y_1,\ if\ g(x)>0\ ;\ otherwise\ decide\ y_2$$

4.5.4Advantages of k-NN Algorithm

- Simple to implement.
- Robust to the noisy training data
- More effective if the training data is large.

4.5.5 Disadvantages of k-NN Algorithm

- k value determining takes time.
- The cost of computing the distance between the data points for all the training samples is high.

4.6 Decision Tree Algorithm

Decision-tree Learning is generally best suited to problems with the following characteristics [1]

Patterns are described by a fixed set of attributes x_j ; j =1, 2 , . . , n and each attribute x_j takes on a small number of disjoint possible values (categorical or numeric) $v_{ix_j};\ l=1,\ 2,\ldots,d_j$ .The output variable y is Boolean-valued function (binary classification problems) defined over the set S or patterns $\{S^{(i)}\}=\{x^{(i)}\};i=1,\ 2,\ldots,\ .$ That is, y take on values $y_q;\ q=1,\ 2$. For instance, if $y_1=0,$ and $y_2=1,$ then $y;S\rightarrow[0,\ 1]$. The training

data is described by the dataset D of N patterns with corresponding observed outputs[1]:

$$D = \{s^{(i)}, y^{(i)}\} = \{x^{(i)}, y^{(i)}\}; i = 1, 2, , \ldots , N$$

Extension of the basic decision-tree Learning algorithm allow handling of continuous- valued attributes, and Learning functions with more than two possible output values (multiclass classification problems).

Basic characteristics of decision-tree building:

- Hierarchical model for supervised Learning.
- Training set is portioned into smaller subsets in a sequence of recursive splits as the tree is being built. The building follows a top-down hierarchical approach.
- Tree-Learning algorithm is greedy; best split (non backtracking) is desired.
- Tree with no error and smallest number of nodes is desired.
- Divide and conquer a frequently used heuristic, is the tree building strategy.

The tree always starts from the root node and grows by splitting the data at each level into new nodes (daughter nodes) [1]. The root node (parent node) contains the entire data and daughter nodes (internal nodes) hold respective subsets of the data. All the nodes are connected by branches shown by the line segments. The nodes that are at the end of the branches are called terminal nodes or leaf nodes, shown by boxes[1]. The leaf nodes in this figure are class labels as shown in fig.13

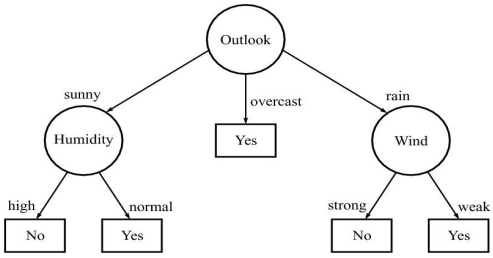

Fig.13 Decision Tree

4.6.1 Measures of Impurity for Evaluating Splits in Decision Trees

An impurity is a heuristic for selecting the splitting criterion that "best" separates a given dataset D of class labeled training tuples into individual classes. If D were split into smaller partitions according to the outcome of the splitting criterion, ideally each partition would be pure[1].

Information gain/entropy reduction

- The expected information needed to classify a pattern in D is given by[1]

$$\text{info}(\text{D}) = -\sum_{q=1}^{2} \text{P}_q \log_2 (\text{P}_q)$$

$$Entropy(D) = -\sum_{q=1}^{2} P_q \log_2 (P_q)$$

- $\text{info}(\text{D})$ is between 0 and 1.
- It is a measure of impurity of the collection of examples.
- More the impurity (more the heterogeneity in the dataset), more the entropy, more the expected amount of information that would be needed to classify a new amount of information that would be needed to classify a new pattern [1].
- Attribute x_j has distinct values $v_{ix_j}; = 1, \ldots , d_j$, as observed from the training data D [1].
- Attribute x_j can be used to split data into $l; l = 1, \ldots . , d_j$, partitions or subsets $\{D_1, D_2, \ldots , D_{d_j}\}$ where D_i contains those patterns in D that have values v_{ix_j} *of* x_j.
- These partitions would correspond to branches grown from the node.
- It is quite likely that partitions will be impure.
- This amount of more information required, is measured by[1]

$$\text{info}\ (\text{D},\ \text{x}_\text{j}) = \sum_{i=1}^{d_j} \frac{|D_i|}{D} \times \text{info}\ (\text{D}_\text{i})$$

The term $\frac{|D_i|}{D}$ acts as the weight of l^{th} partition.

$\text{info}\ (\text{D}_\text{i})$ is given by:

$$\text{info}\ (\text{D}_\text{i}) = -\sum_{q=1}^{2} p_{ql} \log_2 (P_{qi})$$

Where P_{qi} is the probability that the arbitrary sample in subset D_i belongs to class y_q and is estimated as[1]

$$P_{qi} = \frac{freq(y_q, D_i)}{|D_i|}$$

$\inf o\ (D, x_j)$ is expected information required to classify a pattern from D based on the partitioning by x_j. The smaller the expected information (still) required, the greater the purity of the pattern [1].

4.6.2 Gain Ratio: For ID3, decision-tree tool developed by Quinlan, the selection of partitioning was made on the basis of the information

gain/entropy reduction [1]. C4.5, a successor of ID3, uses an extension of information gain known as gain ratio [1].

$$GainRatio\left(D, x_j\right) = \frac{Gain\left(D, x_j\right)}{\text{Splitinfo}\left(\text{D, x}_\text{J}\right)}$$

$$\text{Splitinfo}\left(\text{D, x}_\text{j}\right) = -\sum_{i=1}^{d_j} \frac{|D_i|}{|D|} \times \log_2 \frac{|D_i|}{|D|}$$

4.6.3 Gini Index: Gini Index is used in CART [1].

$$Gini\left(D\right) = 1 - \sum_{q=1}^{M} P_q^2$$

Where P_q is the probability that a tuple in D belongs to class y_q and is estimated by[1]

$$P_q = \frac{freq\left(y_q, D\right)}{|D|}$$

Gini index considers a binary split for each attribute. First consider the case where x_j is continuous-valued attribute having d_j distinct values $l = 1, 2, \ldots, d_j$. It is common to take mid-point between each pair of (sorted) adjacent values as a possible split-point [1]. The point giving the minimum Gini index for the attribute x_j is taken as its split[1].

4.6.4 To avoid overfitting in decision tree Learning.

Two main groups:

- Prepruning: Growing of the tree is stopped before it reaches the point where it perfectly classifies the training data.
- Postpruning: The tree is allowed to grow to perfectly classify the training examples, and then post-prune is done.
- The second approach of Post pruning over fit trees has been found to be more successful in practice.

4.7 Random forest algorithm

A supervised learning algorithm which is used for both classification as well as regression. But however, it is mainly used for classification problems. As forest is made up of trees and more trees means more robust forest. Similarly, random forest algorithm creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution by means of voting[1][2][3][4][5]. It is an ensemble method which is better than a single decision tree because it reduces the over-fitting by averaging the result.

We can understand the working of Random Forest algorithm with the help of following steps

Step 1: First select random samples from a given dataset.

Step 2: Next a decision tree is constructed for every sample. Then it will get the prediction result from every decision tree.

Step 3: In this step, voting will be performed for every predicted result.

Step 4: At last, select the most voted prediction result as the final prediction result

Fig.14 shows Random Forrest with three decision trees

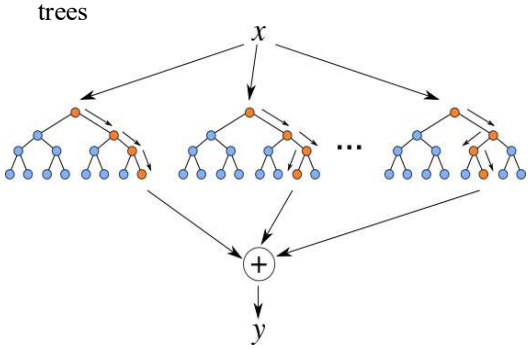

Fig.14. Random Forrest

4.7.1 Regression Problems

When using the Random Forest Algorithm to solve regression problems, you are using the mean squared error (MSE) to how your data branches from each node.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{y_1} - y_i\right)^2$$

Classification Problems

When performing Random Forests based on classification data, you should know that you are often using the Gini index, or the formula used to decide how nodes on a decision tree branch. [1]

$$Gini\left(D\right) = 1 - \sum_{q=1}^{M} P_q^2$$

This formula uses the class and probability to determine the Gini of each branch on a node, determining which of the branches is more likely to occur. Here, pi represents the relative frequency of the class you are observing in the dataset and c represents the number of classes.You can also use entropy to determine how nodes branch in a decision tree.

Entropy: it uses the probability of a certain outcome in order to make a decision on how the node should branch[1]. Unlike the Gini index, it is

more mathematical intensive due to the logarithmic function used in calculating it[1].

$$Entropy(D) = -\sum_{q=1}^{2} P_q \log_2 (P_q)$$

4.7.2 Advantages of Random Forest

- It reduces over fitting in decision trees and helps to improve the accuracy
- It is flexible to both classification and regression problems
- It works well with both categorical and continuous values
- It automates missing values present in the data
- Normalizing of data is not required as it uses a rule-based approach.

4.7.3 Disadvantages of Random Forest

- It requires much computational power as well as resources as it builds numerous trees to combine their outputs.
- It also requires much time for training as it combines a lot of decision trees to determine the class.
- Due to the ensemble of decision trees, it also suffers interpretability and fails to determine the significance of each variable.

5. Performance Analysis of RBMLA

Reliable Boolean Machine Learning Algorithm (RBMLA) has been implemented by using Python Anaconda Navigator Jupyter Note book a well known and popular tool for Data Science and Machine Learning. It is provides a web-interactive framework to create and share documents that contain live code, equations, visualizations and narrative text [7][8][9] . Uses include data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more.

Data set from UCI repository is used. It contains total 300 samples with total fourteen features age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal, target. The description of attributes is shown in following table. 200 samples are taken for and for testing 100 samples were taken. The detail of all features in data set is described in following Table.2.

Table.2 Features of Heart disease

S No	Attribute Description	Distinct Values of Attribute
1	Age- age of a person	Multiple values between 29 & 71
2	Sex-gender of person (0-Feamle, 1-Male)	0,1
3	CP-severity of chest pain patient is suffering.	0,1,2,3
4	Rest BP- patient's BP	Multiple values between 94& 200
5	Chol-cholesterol level of the patient.	Multiple values between 126 & 564
6	FBS-fasting blood sugar in the patient.	0,1
7	Resting ECG-the result of ECG	0,1,2
8	Heartbeat- max heartbeat 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	OldPeak-patient's depression level.	Multiple values between 0 to 6.2
11	Slope- patient condition during peak exercise. It is divided into three egments(Unsloping, Flat, Down sloping)	. 1,2,3.
12	CA- fluoroscopy.	0,1,2,3
13	Thal- Thallium test shows patient suffering from pain in chest or difficulty in breathing.	0,1,2,3
14	Target-It is class or label Colum. It represents the number of classes in dataset. This dataset has binary classification i.e. two classes (0,1).In class "0" represent there is less possibility of heart disease whereas "1" represent high chances of heart disease. The value "0" Or "1" depends on other 13 attribute.	0,1

Heart disease set from UCI website [35] is in MS–Excel format contains 304 records and following 14 columns (Age, sex,cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal, target) among them 13 are that are related to heart disease symptoms last column is target class it contains 0 or 1. 0 indicates –ve class and 1 indicates +Ve class. The columns are called features and rows are called instances .Each feature has different attributes values that are explained in performance analysis section. The five selected above machine learning algorithms were trained by using instances 204 instances from the data set and remaining 100 instances were used for testing five selected above machine learning algorithms

5.1Cross Validation:

- In Machine learning, we usually divide the dataset into Training dataset, Validation dataset, and Test dataset.

Journal of Theoretical and Applied Information Technology
15th August 2021. Vol.99. No 15

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

- Training data set is used to train the model, it can vary but typically we use 60% of the available data for training.
- Validation data set: Once we select the model that performs well on training data, we run the model on validation data set. This is a subset of the data usually ranges from 10% to 20%. Validation data set helps provide an unbiased evaluation of the model's fitness. if the error on the validation dataset increases then we have an over fitting model.
- Test dataset: Also called as holdout data set. This dataset contains data that has never been used in the training. Test data set helps with final model evaluation. Typically would be 5% to 20% of the dataset.

We applied well the K-fold validation technique for cross volition model by taking K value =10

5.2 Analyzing selected features in the data set

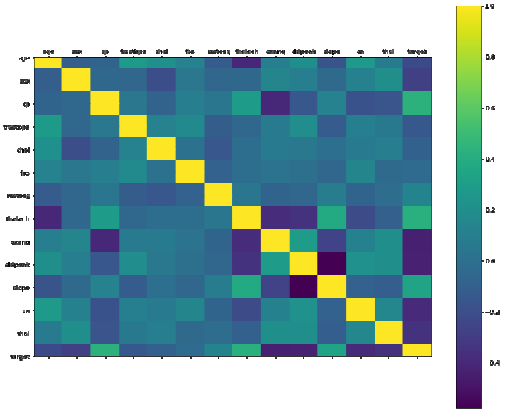

Fig.15. Correlation matrix

Fig.15 shows that a few features have negative correlation with the target value while some have positive

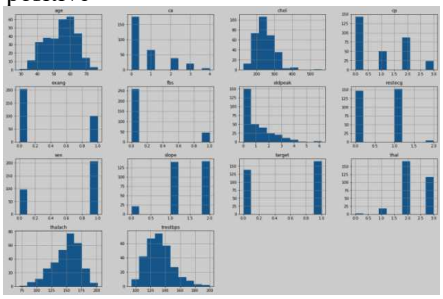

Fig.16. Features in Data Set

Fig.16 shows 13 features in Data Set

Fig.17. 0 is female and 1 is male patients

Fig.17 shows the two classes are not exactly 50% each but the ratio is good enough to continue without dropping/increasing our data.

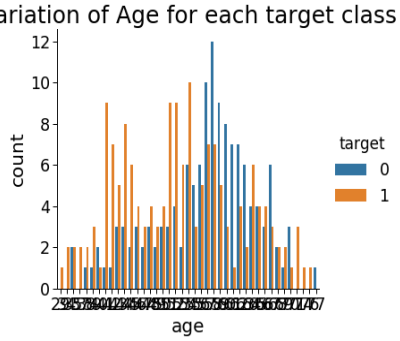

Fig.18. Variation of age for each target class

Fig.18. shows variation of age for each target class in data set.

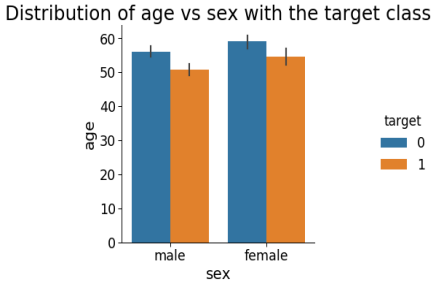

Fig.19. Distribution of age Vs sex in target class

Fig.19 shows distribution of age Vs sex in target class

Journal of Theoretical and Applied Information Technology
15th August 2021. Vol.99. No 15

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

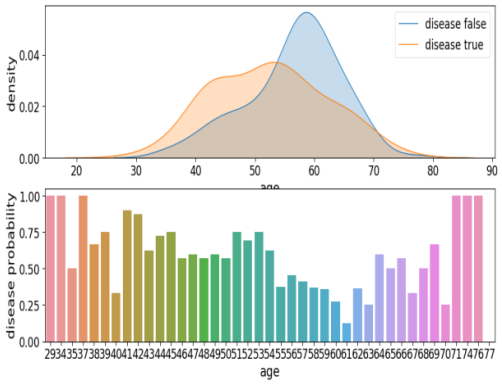

Fig.20 Density and probability of disease

Fig.20 shows density and probability of disease

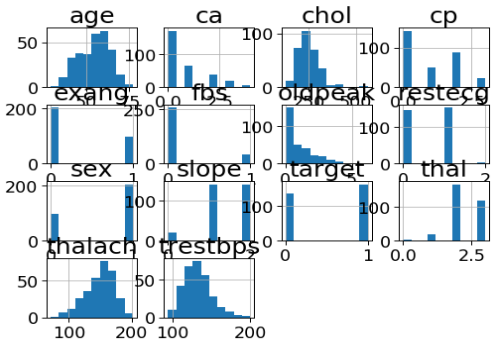

Fig.21.Feature has a different range of distribution.

Fig.21 shows feature has a different range of distribution.

5.3 Confusion Matrix

To evaluate the performance of machine learning model an N×N Confusion matrix is used and N is the number of target class labels. The predicted values are compared with the actual target values. The format of 2 x 2 Confusion matrix is shown in the following fig.22.

In this confusion matrix four values are used for binary classification problem, predicted values of machine learning models are represented by rows and actual values of target variable are represented in the columns. Another format of confusion matrix as follows. .The binary classification target variable has two values either positive or negative. A single prediction on the test set has four possible outcomes.

- The True Positive (TP) means that the actual value and predicted values are both positive.
- True Negative (TN) means that the actual value and predicted values are both negative.
- False Positive (FP) means that the actual value is negative and predicted value is positive, it is called Type-1 error. A False Positive (FP) occurs when the outcome is incorrectly predicted as positive when it is actually negative.

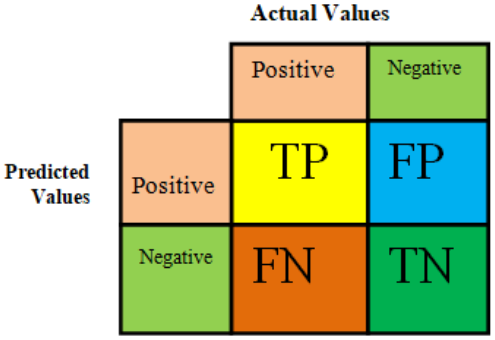

Fig.22.The format of Confusion matrix of size 2x2

False Negative (FN) means that actual value is positive and predicted value is negative, it is called Type -2 errors, A False Negative (FN) occurs when the outcome is incorrectly predicted as negative when it is actually positive. The true positive (TP) and true negative (TN) are correct classifications. The False Positive (FP) and False Negative (FN) are incorrect classifications

5.4 The Performance Metrics

To evaluate the machine learning model are the following seven performance metrics are used

- Accuracy ,
- Precision
- Recall
- F_1 score
- F_β score
- ROC AUC
- Log Loss
- Reliability

5.4.1Accuracy: It describes how many observations; both positive and negative have been correctly classified. It is not used for imbalanced problems (imbalanced data set) because all observations are classed majority class and it leads to high accuracy. It is used for balanced problems (balanced data set) Defined mathematically as

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

5.4.2 Precision: Precision is a measure that tells us what proportion of patients was diagnosed as having heart disease to that of actual heart disease

patients .The predicted positives (People predicted as heart disease is FP) and the people actually having heart disease is TP. Mathematically it is defined as follows:

$$\Pr ecision = \frac{TP}{TP + FP}$$

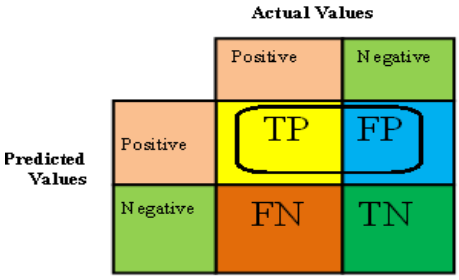

Fig. 23 confusion matrix for Precision

Fig. 23 shows confusion matrix for Precision

5.4.3. Recall or Sensitivity

It is the proportion of actual positives as was identified correctly. It is also called sensitivity. Recall is a measure that tells us what proportion of patients that actually had heart disease was diagnosed by the algorithm as having heart disease. The actual positives (People having heart disease are TP and FN) and the people diagnosed by the model having a heart disease are TP. (Note: FN is included because the Person actually had a heart disease even though the model predicted otherwise).Recall is the number of true positive results is divided by the number of all samples that should have been identified as positive. Mathematically is defined as follows:

$$\Re call = \frac{TP}{TP + FN}$$

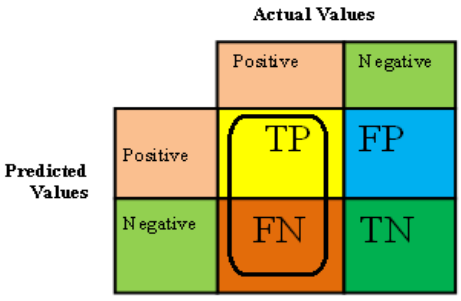

Fig.24. Confisution Matrix for Recall

5.4.4 F1 score

F1 score is a single score that represents both Precision (P) and Recall(R). F1 score is the harmonic mean of the precision and recall. It is defined as mathematically as follows uscripts must be

$$F_1 = 2x \frac{(\Pr ecision\ x\ \Re call)}{(\Pr ecision + \Re call)}$$

5.4.5 F_β Score

The F_β Score is more general and applies additional weights by giving more value to either precision or recall. F_1 score cares equally recall and precision. Higher beta you should choose if you care about recall over precision. When recall is considered β times as important as precision then it is defined mathematically

$$F_\beta = (1 + \beta)^2\ x\ \frac{(\Pr ecision\ x\ \Re call)}{(\beta^2\ x\ (\Pr ecision + \Re call))}$$

F_β is used in every binary classification problem where you care more about the positive class.

5.4.6 ROC AUC

ROC curve (Receiver Operating Characteristic curve) is most popular measure; three are quite a number of other measures to evaluate the performance of a supervised learning model. However , visualization is an easier and more effective way to understand the model performance. It also helps in comparing the efficiency of models.

ROC curve helps in visualizing the performance of a classification model. It shows the efficiency of a model in the detection of true positive while avoiding the occurrence of false positive

True Positive Rate (TPR) and False Positive Rate (FPR) are defined as follows

True Positive Rate (TPR) is a same as recall and is therefore defined as follows:

$$TPR = \frac{TP}{TP + FN}$$

False Positive Rate (FPR) is defined as follows:

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

In the ROC curve, False Positive Rate (FPR) is plotted True Positive Rate (TPR) (in horizontal axis) against (in vertical axis) at different classification thresholds . If we assume a lower value of classification threshold, the model classifies more items as positive. Hence value of both False Positive Rate and True Positive Rate increase.

The area under curve (AOU) value as shown in fig.25

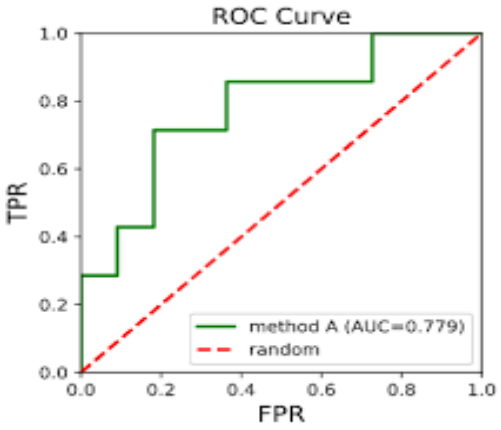

Fig.25 ROC curve

When the classification threshold is decreased the classifier classifies more items as positive, thus increasing both False Positives and True Positives. The Following figure shows a typical ROC curve as shown fig.26

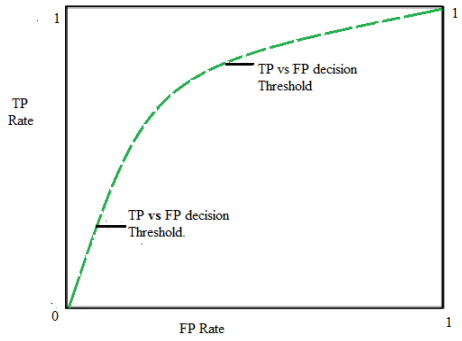

Fig.26 shows the ROC curve

Fig.26. TP vs. FP rate at different classification thresholds values . The area of two-dimensional space under the curve extending from (0, 0) to (1, 1). Where each point on the curve gives a set of true and false positive values at specific classification threshold. This curve gives an indication of positive of the predictive quality of model. AUC value ranges from 0 to 1 with an AUC of less than 0.5 indicating that the classifier has no predictive ability.

The area of two-dimensional space under the curve extending from (0, 0) to (1, 1). Where each point on the curve gives a set of true and false positive values at specific classification threshold. This curve gives an indication of positive of the predictive quality of model. AUC value ranges from 0 to 1 with an AUC of less than 0.5 indicating that the classifier has no predictive ability.

A quick indicative interpretation of the predictive values from 0.5 to 1.0 is given below

Case 1: 0.5 to 0.6: Almost no predictive ability

Case 2: 0.6 to 0.7: Weak predictive ability

Case 3: 0.7 to 0.8: Fair predictive ability

Case 4: 0.8 to 0.9: Good predictive ability

Case 5: 0.9 to 1.0: Excellent predictive ability

5.4.7. Log Loss classification metric is based on probability for comparing models when it is hard to understand and interpret raw values. A less value means better prediction. It is the negative average of the log of corrected predicted probabilities for each instance. To maintain a common convention that lower loss scores are better, the negative average of these negative log values are taken. Log-loss is indicative of how close the prediction probability is to the corresponding actual/true value (0 or 1 in case of binary classification). The higher is the log-loss value The more the predicted probability diverges from the actual value.

$$Log\ loss = -\frac{1}{N}\sum_{i=1}^{N} \log loss_i$$

$$Log\ loss = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i\ \ln p_i + (1-y_1)\ln(1-p_i)\right]$$

where N is the number of observations , i is the given observation/record, y is the actual/true value, p is the prediction probability, and ln refers to the natural logarithm (logarithmic value using base of e) of a number

5.2.8. Reliability of RBMLA

Let reliability be considered the accuracy and if the reliability of SVM, Random Forest, Naive Bayes, Decision Tree, KNN are 0.82, 0.84, 0.80, 0.71, 0.83 respectively then reliability of RBMLA is

$$R(RBMLA)=1-((1-0.82)\ x\ (1-0.84)x(1-0.80)x(1-0.7)x(1-0.83))$$

$$R(RBMLA)=1-(0.19\,x\,0.16\,x\ 0.2\,x\,0.29\,x\,0.17)$$

$$R(RBMLA)=1-(0.0002)$$

$$R(RBMLA)=0.998$$

5.3 Compression of Performance using graphs

Performance Comparison of Six Machine Learning models is shown in the table 3

5.3.1 Support Vector Machine (SVM) Vs RBMLA

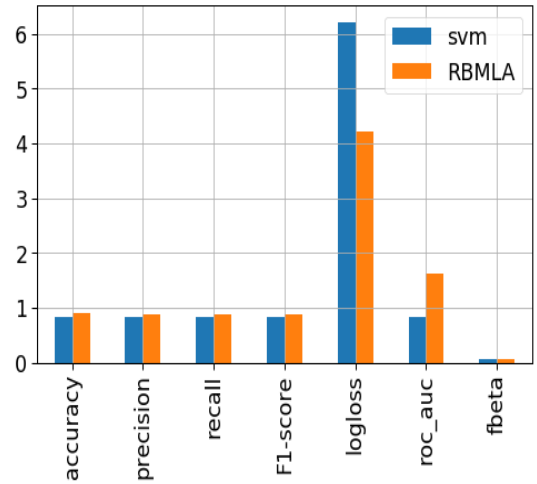

Fig.27 SVM vs RBMLA

Fig.27 shows SVM vs RBMLA. RBMLA yields better performance as compared to SVM

5.3.2 Random Forest Algorithm Vs RBMLA

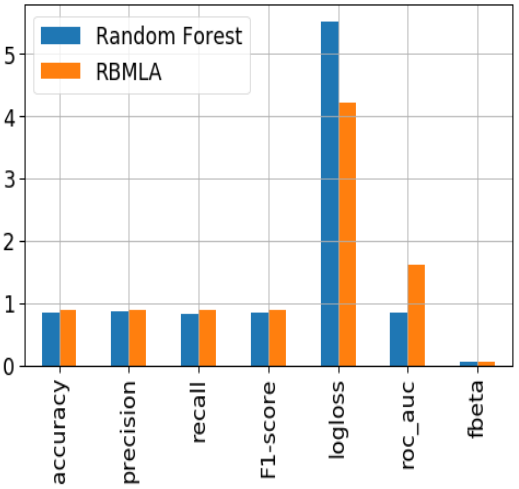

Fig.28 Random Forest Vs RBMLA

Fig.28 shows the Random Forest Vs RBMLA.

RBMLA yields better performance as compared to Random Forest

Fig.30 shows the Decision Tree Algorithm Vs RBMLA. RBMLA yields better performance as compared to Decision Tree Algorithm

Fig.31 shows the KNN Vs RBMLA. RBMLA yields better performance as compared to KNN

Fig.29. shows Naive Bayes algorithm Vs RBMLA. RBMLA yields better performance as compared to Naive Bayes algorithm

5.3.3 Naive Bayes algorithm Vs RBMLA

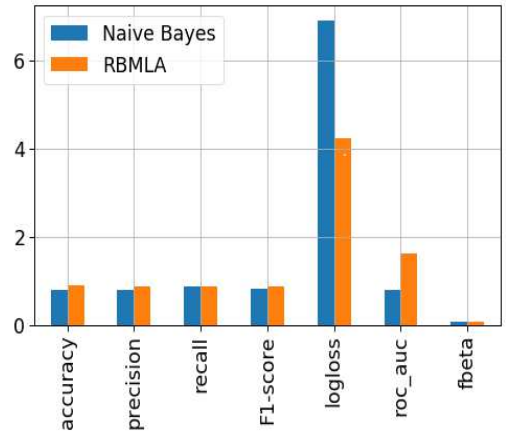

Fig.29. Naive Bayes algorithm Vs RBMLA

5.3.4 Decision Tree Algorithms RBMLA

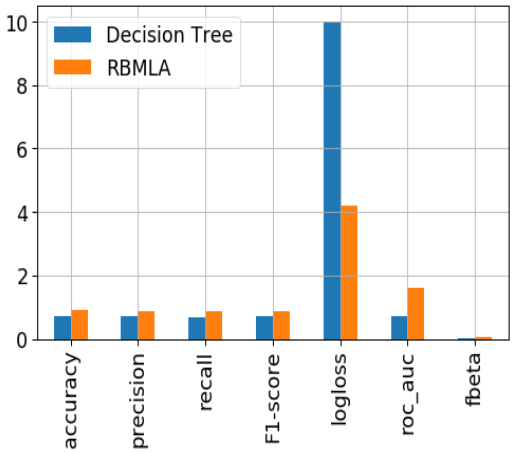

Fig.30 Decision Tree Algorithm Vs RBMLA

5.3.5 K-Nearest Neighbor (-NN) Vs RBMLA

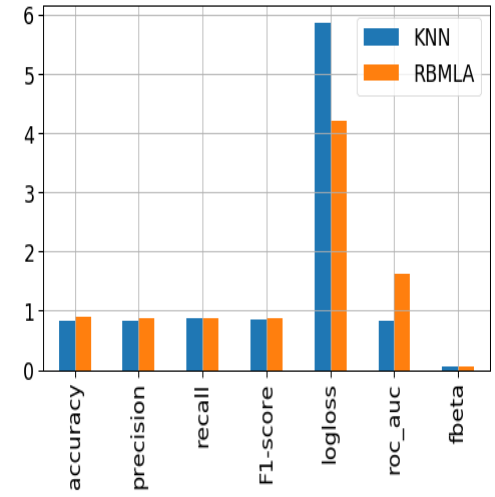

Fig.31 K-Nearest Neighbor (-NN) Vs RBMLA

The following Table 3 shows results of existing and proposed machine learning algorithms

Table 3: Performance Comparison of Six Machine Learning Models

S NO	Performances Metric	SVM	Random Forest	Naive Bayes	Decision Tree	KNN	RBMLA
1	Accuracy (%)	82.00	84.00	80.00	71.00	83.00	86.00
2	Precision	0.8	0.86	0.77	0.73	0.81	0.89
3	Recall	0.82	0.82	0.86	0.69	0.86	0.87
4	F1-Score	0.82	0.84	0.81	0.71	0.84	0.87
5	Roc AUC Score	0.81	0.84	0.79	0.71	0.82	1.61
6	F_β score	5.65	5.99	5.50	3.62	5.96	5.96
7	Log loss	6.21	5.52	6.90	10.01	5.87	4.21
8	Reliability	0.82	0.84	0.80	0.71	0.83	0.99

5.3.6 RBMLA Vs (SVM, Random Forest, Naïve Bayes, Decision Tree and KNN)

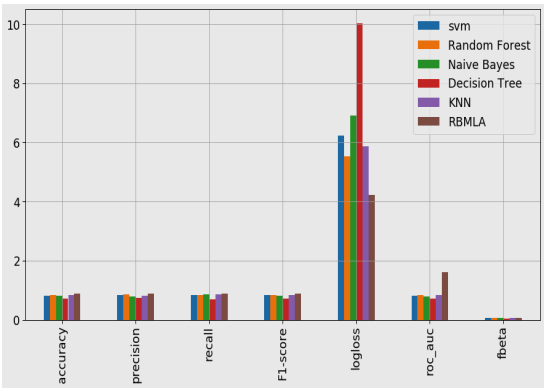

Fig.32. RBMLA Vs (SVM, Random Forest, Naïve Bayes, Decision Tree and KNN)

Fig.32 shows RBMLA Vs (SVM, Random Forest, Naïve Bayes, Decision Tree and KNN). RBMLA yields better performance as compared to five algorithms (SVM, Random Forest, Naïve Bayes, Decision Tree and KNN)

5. Conclusions , Limitations and Future Scope

In Machine Learning Ensemble learning most popular technique, it helps improve machine learning results by combining several models. The most popular ensemble methods are boosting, bagging, and stacking. There are three technique Maximum Voting Averaging, Weighted Averaging. for final result for prediction . The maximum voting method is generally used for classification problem. In this paper, the Reliable Boolean Machine Learning Algorithm has been proposed by using novel approach by using five standard existing supervised machine learning models Support Vector, Random Forest, Naïve Bayes, Decision tree and k-NN. majority voting by using Boolean algebra. It has been applied for prediction of Heart disease . It uses five standard existing machine learning algorithms It is a binary classification algorithm. It gives better performance metrics like Accuracy, Precision, Recall, F1 score, score, ROC AUC, Log Loss and Reliability as compared to standard existing five machine learning algorithms. Its accuracy is 86%; Precision is 0.8969, Recall is 0.8769, F1 score is 0.8782,ROC AUC is 1.6197 , is, 5.960, Log Loss is 4.217, and Reliability is 0.998

Advantages and Disadvantages of Machine Learning

Advantages of Machine learning; Easily identifies trends and patterns, No human intervention needed (automation), Continuous Improvement, Handling multi-dimensional and multi-variety data, Wide Applications.

Disadvantages of Machine Learning: Data Acquisition, Time and Resources, Interpretation of Results, High error-susceptibility, enable to solve very complex problems (Deep learning is used for solving complex problems, it subset of machine leaning).

The limitations of proposed work: The proposed Reliable Boolean Machine Learning Algorithm (RBMLA) was designed based on five well-known supervised machine learning algorithms such as Support Vector Machine (SVM), Random Forest, Naïve Bayes, Decision Tree, k-Nearest Neighbor (k-NN) using majority voting method. None of machine learning algorithm may not give 100% accuracy and not perfect model due above said disadvantages. The results of the proposed Reliable Boolean Machine Learning Algorithm (RBMLA) are purely depending and reflecting on results of existing algorithms. Hence the proposed Reliable Boolean Machine Learning Algorithm (RBMLA) does not give 100% accuracy and performance.

Future Scope of proposed work: The proposed Reliable Boolean Machine Learning Algorithm (RBMLA) is applicable to predict various diseases like kidney cancer, diabetes, liver, breast cancer, etc. in medical heath care applications.

REFRENCES:

[1] Text book title Applied Machine Learning by Gopal, Mc Graw Hill Education, 1st edition

[2] Text book title Machine Learning By Tom M. Mitchell, Mc Graw Hill Education

[3] Text book title Machine Learning the art and science of algorithms that make sense of data by Peter Flach, Cambridge University Press.

[4] Text book title Machine Learning by Anuradha Srinivasaraghavan and Vincy Joseph, WILEY

[5] Text book title Machine Learning by Dr. Ravi Chopra, 2nd edition, Khanna Publishing.

[6] Text book title Machine Learning with Python by Abhishek Vijayvargia, BPB Publications

[7] Text book title Data Science with Jupyter by Prateek Gupta, BPB Publications

[8] Text book on Data Science from Scratch by Joel Grus, O' Reilly

[9] Text book title Statistics Random Processes and Queuing Theory by Prabha , R .Sujatha, SCITECH Publication PVT LTD.

[10] Text Reliability Engineering by E.Balagurusamy, Tata Mc-Graw Hill

[11] Text Reliability Engineering and Life Testing by V.N.A Naikan, PHI

[12] Muhammad Affan Alim, Shamsheela Habib, Yumna Farooq and Abdul Rafay, "Robust Heart Disease Prediction: A Novel Approach based on Significant Feature and Ensemble Learning Model", 2020 3rd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET), 2020, 978-1-7281-4970-7/20/$31.00 ©2020 IEEE

[13] JoshuaEmakhu , Sujeet Shrestha and Suzan Arslanturk, "Prediction System for Heart Disease Based on Ensemble Classifiers", Proceedings of the 5th NA International Conference on Industrial Engineering and Operations Management Detroit, Michigan, USA, pp: 2337 -2347, August 10 - 14, 2020.

[14] RutujaGujare, D.Viji, Simran Bhatt "Enhanced Heart Disease Prediction Using Ensemble Learning Methods", International Journal of Advanced Science and Technology, Vol. 29, No. 6,), pp. 76-85, 2020

[15] Ibomoiye DomorMienye , Yanxia Sun, Professor , Zenghui Wang , "An improved ensemble Learning approach for the prediction of heart disease risk", Elsevier Journal of Informatics in Medicine Unlocked", vol 20, pp: 1-5, 2020.

[16] Archana Singh and Rakesh Kumar, "Heart Disease Prediction Using Machine Learning Algorithms", 2020 International Conference on Electrical and Electronics Engineering (ICE3-2020), pp: 452- 457, 2020. (978-1-7281-5846-4/20/$31.00 ©2020 IEEE)

[17] Apurb Rajdhan, Milan Sai, Avi Agarwal and Dundigalla Ravi," International Journal of Engineering Research & Technology (IJERT)", Vol. 9 , Issue 04, pp:659- 662, April-2020.

[18] Montu Saw, TarunSaxena, SanjanaKaithwas, Rahul Yadav and Nidhi Lal," Estimation of Prediction for Getting Heart Disease Using Logistic Regression Model of Machine Learning", 2020 International Conference on Computer Communication and Informatics (ICCCI -2020), Jan. 22-24,Coimbatore, India,2020.

[19] Jian Ping Li , Amin UlHaq , Salah Ud Din , Jalaluddin Khan , Asif Khan and AndAbdusSaboor "Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare:,IEEEACCESS,June 19, Volume 8, 2020.

[20] B.Keerthi Samhitha, SarikaPriya. M. R, Sanjana. C, Suja Cherukullapurath Mana and JithinaJose, "Improving the Accuracy in Prediction of Heart Disease using Machine Learning Algorithms", International Conference on Communication and Signal Processing, July 28 - 30, pp:1326- 1330 2020, India

[21] SyedJaveed Pasha and E. Syed Mohamed," Novel Feature Reduction (NFR) Model WithMachineLearning and Data Mining Algorithms for Effective Disease Risk Prediction", IEEE ACCESS, Volume 8, 2020 October 5, 2020,

[22] Senthilkumar mohan, , Ghandrasegar thirumala and Gautam Srivastava," Effective [Heart Disease Prediction Using Hybrid Machine Learning Techniques.", IEEE ACCESS special section on smart caching, communications, computing and cybersecurity for information-centric internet of things', Volume 7, 2019

[23] NoorBasha, Ashok Kumar P S Gopal Krishna C and Venkatesh P " Early Detection of Heart Syndrome Using Machine Learning Technique", 2019 4th International Conference on Electrical, Electronics, Communication, [Computer Technologies and Optimization Techniques (ICEECCOT), pp: 387-391,2019

[24] Chunyan Guo, Jiabing Zhang, Yang Liu, YayingXie,Zhiqiang Han and Jianshe Yu1,"Recursion Enhanced Random Forest With An Improved Linear Model (RERFILM) for Heart Disease Detection on the Internet of Medical Things Platform", IEEE ACCESS,2019

[25] X. Liu, X. Wang, Q. Su, M. Zhang, Y.Zhu, Q. Wang, and Q. Wang, "A hybrid classification system for heart disease diagnosis based on the rfrs method", Computational and Mathematical Methods in Medicine, Vol.2017, Article ID 8272091, 11 pages, 2017.

[26] A. Malav, K. Kadam, and P. Kamat, "Prediction of heart disease using k-means and artificial neural network as a hybrid approach to improve accuracy", International Journal of Engineering and Technology, Vol.9, No.4, 2017

[27] M.A.Jabbar, B.L.Deekshatulu, Priti Chandra," Intelligent heart disease predictionsystem using random forest and evolutionary approach", Journal of Network and Innovative Computing ISSN 2160-2174Volume4(2016) pp. 175-184.

[28] M.A.Jabba and LDeekshatulu, "Computational Intelligence Technique for early Diagnosis of Heart Disease", IEEE International Conference on Engineering and Technology (ICETECH) 2015.

[29] Sresht Agrawal and B.K.Tripathy," A Decision Theoretic Rough Fuzzy c-means Algorithm", IEEE (ICRCICN) 2015,pp.192-196.

[30] A. F. Otoom, E. E. Abdallah, Y. Kilani, A. Kefaye, and M. Ashour, "Effective diagnosis and monitoring of heart disease", International Journal of Software Engineering and Its Applications, Vol.9, No.1, pp. 143-156, 2015.

[31] K. Vembandasamy, R. Sasipriya, and E. Deepa, "Heart Diseases Detection Using Naive Bayes Algorithm", IJISET-International Journal of Innovative Science, Engineering & Technology, Vol.2, pp.441-444, 2015. 2014

[32] V Krishnaiah, M Srinivas, "Diagnosis of Heart Disease Patients Using Fuzzy Classification Technique", IEEE International Conference on Computer and Communications Technologies (ICCCT), 2014

[33] M.A.Jabbar,B.L andDeekshatulu,"Alternating decision trees for early diagnosis of heartdisease", Proceedings of International Conference on Circuits, Communication, Control andComputing (I4C 2014),pp-322-328

[34] Hlaudi Daniel Masethe, Mosima Anna Masethe," Prediction of Heart Disease using Classification Algorithms", Proceedings of the World Congress on Engineering and Computer Science 2014 Vol II WCECS 2014, PP.22-24.

[35] UCI Machine Learning Repository, https://archive.ics.uci.edu