Survey on Predictive Analysis of Diabetes Disease Using Machine Learning Algorithms

A.Jamuna[1]; Dr. R. Jemima Priyadarsini[2]; Dr. S.Titus[3]

[1]Department of Computer Science, Bishop Heber College (Autonomous), Affiliated to Bharathidasan University, Tiruchirappalli, India
[2]Department of Computer Science, Bishop Heber College (Autonomous), Affiliated to Bharathidasan University, Tiruchirappalli, India
[3]Department of Electrical and Electronics Engineering, K.Ramakrishnan College of Engineering, Tiruchirappalli, India
[1] jamu1793@gmail.com; [2] jemititus@gmail.com; [3] drs.titus@gmail.com
DOI: 10.47760/IJCSMC.2020.v09i10.003

Abstract— Big data analysis is predicated on large amount of data. Diabetes is caused due to the excessive amount of sugar condensed into the blood. One of the most critical chronic healthcare problems is diabetes.Undiagonosed diabetes problem may leads to damage eyes, heart, kidneys and nerves of diabetes patients. If improper medication taken is done which also lead to death. Early detection of diabetes is very important to maintain healthy life. Machine learning algorithm to identify a best predicting algorithm based various matrices such us accuracy, precision, recall, F-measure, sensitivity and specificity. This paper discusses about various ML techniques to predict the Diabetes disease by using dataset .Machine learning algorithm namely Decision Tree, SVM, Naive Bayes, Random forest, k-NN, K-mean clustering and LR algorithms are used in their experiment to detect diabetes at an early stage. Experiments are performed on Pima Indian diabetes dataset (PIDD) which is sourced from UCI Machine Learning repository. Result obtained show SVM outperform with high accuracy of 82% comparatively other algorithms.

Keywords— Diabetes, Decision Tree, SVM, Naive Bayes, Random forest, k-NN and LR

I. INTRODUCTION

Big Data Analytics is the emerging approach which is dealing with many sectors, thus applying it to the healthcare domain will result in improved healthcare services. While big data approach in healthcare is still in developing phase, it is clear that the designing of a healthcare platform for a better tomorrow will be of great help in enhancing the quality delivery of healthcare. According to the WHO report India is ranked Number one

with 31.7 million number of diabetic patient in 2000 and is likely to increase up to 79.4 millionth[1]. Since there is a huge risk involved in the likelihood of increasing number of diabetic patient where accurate diagnosis will be the need of the hour [2]. Hence this statistic provides the motivation to carry out the research to find the best performing algorithm

Machine learning is another emerging and trending approach which closely works to solve the real time problems[4]. Currently in the health care domain various data mining methods are used to find interesting pattern of disease using statistical medical data with the help of machine learning algorithms. Machine learning approach can be applied for prediction of diseases and provide automated diagnosis under the validation of professional doctor. Machine gaining knowledge of generally categorized into two categories, these categories are based on how gaining knowledge of is obtained to the system developed. Two of the most widely used device getting to know techniques are Supervised studying, Unsupervised learning [5].

In supervised gaining knowledge of personal computer is supplying inputs that are labeled with preferred outputs. The cause of the algorithm is in a position to 'learn' via comparing real output with "taught" output to discover errors. In unsupervised learning, records are unlabeled; Algorithm is left to locate its input statistics. Unlabeled data are more accrue than labeled ones. It is commonly used for transactional data like large datasets of customers to purchase it. Machine learning algorithms are good at health care to prevent it.

Machine learning algorithms are good at health care to prevent it[5]. In this paper, discussing about insulin-dependent and non insulin dependent diabetes as follows. Insulin-dependent diabetes is a common disease that can be characterized by excessive increase in blood glucose level. It mainly occurs in adults. In insulin dependent diabetes, pancreas doesn't releases insulin due to the damage of pancreatic beta germ. High blood glucose ends in each acute and persistent complication bring about failure of various organs.

Non insulin-dependent diabetes (diabetes mellitus) is a metabolism sickness that causes glucose to build up in the body. The severity level of diabetes vary too high, some patients must make small adjustments to their life-style after they are diagnosed. [6] Reducing weight and getting workout may be enough for them to control their diabetes. Other people who have non insulin dependent diabetes want durable remedy that entails taking drugs or injections. This is especially significant for them to have an excellent understanding of their illness and follow what they are able to do to be more active.

People who have non insulin-dependent diabetes, the pancreas simply give sufficient sebum; however now it is not an effect at the frame material and germs. Doctors refer to as "insulin resistance". Pancreas can give amends by producing extra insulin. At few point we cannot maintain, and then blood glucose levels start to rising up.

Table 1: The main terminologies mentioned in this paper (include abbreviations, indicators, parameters, variables)

KNN	K-Nearest neighbor
SVM	Support Vector Machine
RF	Random Forest
DT	Decision Tree
ANN	Artificial Neural Network
LR	Logistic Regression
PIDD	Pima Indians Diabetes Dataset

Rest of the paper is organized as follows: Section II describes about the diabetes disease. In Section III Related works is described including dataset description. Section IV includes different algorithms of machine learning. After that Result in Section V which includes performance comparison of the algorithms and at last paper is concluded in Section VI.

II. DIABETES DISEASE

Diabetes is an illness which affects the ability of the body in producing the hormone insulin, which in turn makes the metabolism of carbohydrate abnormal and raise the levels of glucose in the blood. In Diabetes a person generally suffers from high blood .Intensify thirst, Intensify hunger and frequent urination are some of the symptoms caused due to high blood sugar. Many complications occur if diabetes remain untreated. This difficulty leads to death. Diabetic patient is at a greater risk of developing cardiovascular disease, visual impairment and undergo limb amputations, as compared to a non-diabetic person. Severe difficulties lead to cardiovascular disease foot sores, and eye blurriness. When there is a rise within the sugar level within the

blood, it is referred to as prior diabetes. The prior diabetes isn't therefore great than the traditional worth.

Diabetes can be classified into three categories such as

- ➢ Type 1 diabetes
- ➢ Type 2 diabetes
- ➢ Type 3 diabetes

Type 1 diabetes is an autoimmune disease. In this case, the body destroys the cells that are essential to produce insulin to absorb the sugar to produce energy. This kind of diabetes can cause obesity. The obesity is the increase in body mass index (BMI) than the normal level of BMI of an individual. Type 1 diabetes can occur in childhood or during adolescence age. This type was spoken as "insulin-dependent polygenic disease mellitus" (IDDM) or "juvenile diabetes". The reason is unidentified. The type one polygenic disease found in children beneath twenty years old. The diabetic patients must often follow workouts and fit regime which are recommended by doctors.

Type 2 diabetes usually affects the adults who are obese. In this type, the body resists observing insulin or fails to produce insulin. Type 2 generally occurs in the middle or aged groups. This type was spoken as "non-insulin-dependent polygenic disease mellitus". The usual cause is extreme weight. Moreover, there are other causes for diabetes such as bacterial or viral infection, toxic or chemical contents in food, autoimmune reaction, obesity, bad diet, change of lifestyles, eating habit, environment pollution, etc. The obesity is one of the main reasons for type-2 diabetes. The type-2 polygenic diseases are under control by proper workout and taking appropriate regime. Type 3 Gestational diabetes occurs when a woman is pregnant and develops the high blood sugar levels without a previous history of diabetes. Therefore, it is found that in total 18% of women in pregnancy have diabetes. So in the older age there is a risk of emerging the gestational diabetes in pregnancy.

Machine learning algorithms are good at health care to prevent it. Many researchers are conducting experiments for diagnosing the diseases using various classification algorithms of machine learning approaches like J48, SVM, Naive Bayes, Decision Tree, Decision Table etc. as researches have proved that machine-learning algorithms works better in diagnosing different diseases.

This paper discusses about Decision Tree, SVM, Naive Bayes, Random forest, k-NN, K-mean clustering and LR algorithms machine learning classification algorithms are used and evaluated on the PIDD dataset to find the prediction of diabetes in a patient. Experimental performances of all the algorithms are compared on various measures and achieved good accuracy.

III. RELATED WORKS

Veena Vijayan V. Anjali C. [1] proposed that uses AdaBoost algorithm with Decision Stump as base classifier for classification. Additionally Support Vector Machine, Naive Bayes and Decision Tree are also implemented as base classifiers for AdaBoost algorithm for accuracy verification. The accuracy obtained for AdaBoost algorithm with decision stump as base classifier is 80.72% which is greater compared to that of Support Vector Machine79.8%, Naive Bayes 79.8% and Decision Tree 77.6%.

N. Yuvaraj · K. R. SriPreethaa [2] proposed which automates the working model of health care systems and enhances the accuracy of disease prediction. Hadoop cluster based distributed work proposes the novel implementation of machine learning algorithms in hadoop based clusters for diabetes prediction. The results show that the machine learning algorithms Random forest algorithm produces highest accuracy [80%] than decision tree [78%] and naïve bayes [75%] algorithm in 4 node hadoop cluster environment

Thanga Prasad. S [3] This system values the predictive investigation of algorithm in Hadoop/Map Reduce environment to predict and classify the type of Diabetic Mellitus, Type-1 diabetes does not produce the insulin in our body. hadoop implementation present methodical approach designed for reach in good health result similar to availability and affordability of healthcare once-over on the way to every residents.

P. Suresh Kumar , S. Pranavi [4] Made a comprehensive study on diabetes dataset with Random Forest (RF), SVM, k-NN, CART and LDA algorithms. The achieved results shows that RF is giving more accurate

predictions with compared to other algorithms.

Harleen Kaur [5] Classified the patients into diabetic and non-diabetic we have developed and analyzed five different predictive models using R data manipulation tool. For this purpose we used supervised machine learning algorithms namely linear kernel support vector machine (SVM-linear), radial basis function (RBF) kernel support vector machine, k-nearest neighbour (k-NN), artificial neural network (ANN) and multifactor dimensionality reduction (MDR).The experimental results suggested that all the models achieved good results; SVM-linear model provides best accuracy of 0.89 and precision of 0.88 for prediction of diabetes as compared to other models used. On the other hand k-NN model provided best recall and F1 score of 0.90 and 0.88. As our dataset is an example of imbalanced class, F1 score may provides better insight into performance of our models. F1 score provides balance between precision and recall. Further it can be seen that AUC value of SVM- linear and k-NN model are 0.90 and 0.92 respectively

Avinash Golande, Pavan Kumar T [6] made use of experimental results carried out using dissimilar classification methods for heart disease dataset. The different classification systems which were used and tested by him are Decision Tree, KNN, K-mean clustering, Adaboost. This investigation tells us about dissimilar technologies that are used in dissimilar papers with dissimilar count of attributes with different accuracies depending on the tools designed for execution.

Shaik MD Ansari, Kiruthika Devi[7] Made a detailed study of segmenting and detection of tumor based on computed tomography images. The main part of work is to determine the kidney tumor by applying the algorithm on the test result acquired from the patient medical report and segmentation of the tumor location. proposed method, the obtained images move on to a pre processing stage which is of four folds namely: Image enhancement; Filtering; Classification and Segmentation.

S. Pitchumani Angayarkanni [8] The Proposed methods that validates the dataset against the thumb rule when working with less number of data points used for classification and the classifier is validated against under fit, over fit conditions. The performance of the classifier is evaluated using accuracy and F-Score. The main focus of this research work is to analyze the dataset of chronic kidney failure and perform the classification of CKD and Non CKD cases. The proposed architecture indicates that LR, RF and ABR provides a very high accuracy and F-Score.RNN algorithm helps in automatic detection of CKD with high accuracy of 98.99%.

Ayman Mir,Sudhir N. Dhage[9] Proposed a classifier model using WEKA tool to predict diabetes disease by employing Naive Bayes, Support Vector Machine, Random Forest and Simple CART algorithm. It is observed that Support Vector Machine performed best in prediction of the disease having maximum accuracy 0.7913.The Accuracy of Naive Bayes is 0.77 better than Random Forest and Simple CART. The accuracy of Random Forest and Simple CART is almost equal with value 0.765. The training time of Naive Bayes is less than SVM. The training time of Simple CART is the highest. Overall according to classification Accuracy SVM outperformed all other classifier.

Aditi Gavhane,Gouthami Kokkula[10] proposed system we used the neural network algorithm multi-layer perception (MLP) to train and test the dataset.MLP provides its users with a prediction result that gives the state of a user leading to CAD.the machine learning algorithms are evolved a lot and hence we use Multi Layered Perceptron (MLP) in the proposed system because of its efficiency and accuracy. Also, the algorithm gives the nearby reliable output based on the input provided by the users.

Deepti Sisodiaa, Dilip Singh Sisodia [11] The tedious identified process results in visiting of a patient to a diagnostic center and consulting doctor. But the rise in machine learning approaches solves this critical problem. proposed system used three machine learning classification algorithms namely Decision Tree, SVM and Naive Bayes are used in this experiment to detect diabetes at an early stage. Results obtained show Naive Bayes outperforms with the highest accuracy of 76.30% comparatively SVM(65.10%),Naive Bayes(73.82%) algorithms.

Muhammad Azeem Sarwar, Nasir Kamal[12] These six algorithms are K Nearest Neighbours (KNN), Naive Bayes (NB), Support Vector Machine (SVM), Decision Tree (DT), Logistic Regression (LR) and Random Forest (RF).LR gives 74% accuracy, SVM gives 77% accuracy, 74% accuracy was achieved by using NB, DT

and RF achieved 71% accuracy and KNN achieved 77%. So SVM and KNN achieved highest accuracy which is 77%.

IV. MACHINE LEARNING TECHNIQUES USED FOR PREDICTION

Machine Learning Techniques

A. *k-Nearest* Neighbor *(k-NN)*

k- Nearest neighbor [5] is a simple algorithm but yields very good results. It is a lazy, nonparametric and instance based learning algorithm. This algorithm can be used in both classification and regression problems. In classification, k-NN is applied to find out the class, to which new unlabeled object belongs. For this, a 'k' is decided (where k is number of neighbors to be considered) which is generally odd and the distance between the data points that are nearest to the objects is calculated by the ways like Euclidean's distance, Hamming distance, Manhattan distance or Minkowski distance. After calculating the distance, 'k' nearest neighbor are selected the resultant class of the new object is calculated on the basis of the votes of the neighbors. The k-NN predicts the outcome with high accuracy

B. *Random forest:*

Random forests [2] or random decision forests are ensemble learning method for classification, regression that operates by constructing a multi node of decision trees at training time and the mode of the classes or mean prediction of the individual trees. Habit of over fitting the training set is corrected by Random format. Random forest algorithm is a supervised classification algorithm. The pseudo code for random forest algorithm can split into two stages. One is Random forest. creation pseudo code and the other is Pseudo code to perform prediction from the created random forest classifier. Random Forest pseudo code is as follows:

1. Randomly select "k" features from total "m" features. Where k << m
2. Among the "k" features, calculate the node "d" using the best split point.
3. Split the node into daughter nodes using the best split.
4. Repeat 1–3 steps until "l" number of nodes has been reached.
5. Build forest by repeating steps 1–4 for "n" number times to create "n" number of trees.

C. *Support Vector Machine:*

SVM [5]is a supervised machine learning algorithm. It is mostly used in classification problems. This has a training phase in the beginning stage. SVM used a maximum margin strategy that transformed into solving a complex quadratic programming problem. Due to the high performance of SVM in classification, various applications widely applied it [8]. In a binary classification problem, the instances are separated with a hyper plane

$$w^T x + b = 0$$

where *w* and T are dimensional coefficient vectors, which are normal to the hyper plane of the surface, *b* is offset value from the origin, and *x* is data set values.

$$w = \sum_{i=1}^{n} \alpha_i y_i x_i$$

where *n* is the number of support vectors and y_i are target labels to *x*.

D. *K-mean clustering*

K-mean clustering [6] is an unsupervised learning which is used when class label is not known or you have unlabeled data. The main focus of this algorithm is finding the groups in the data with that number of groups that represent the variable K. This algorithm iteratively allocating the k groups to the point. Data points here are clustered based on feature of similarity. The consequences of the K-means clustering algorithm are:

1) We can use centroid of the K clusters, to tag new data
2) The training data are tagged (A single data point is allocated to a single cluster)

Clustering defines groups beforehand observing at the obtainable data, and also allows us to diagnose and examine the groups that have been designed naturally. Each centroid of the accessible clusters is a group of feature ideals that defines the subsequent groups. By studying the centroid eye, weights can easily be used to qualitatively understand that the cluster fits to which group.

E. *Adaboost*

Ad boost [6] is a fine technique that is used to increase the performance of decision tree on binary classification problems. Adaboost was previously known as AdaBoost.M1. Currently it is also discussed to as discrete. Adaboost as it is used mainly for classification relatively than regression. We can increase the presentation of every machine learning algorithm using Ad boost. It is Finest to use when the beginners are weak. These models gain the accuracy level just above the random chance on a given classification problem. The common algorithm that is used with Adaboost is decision tree but with one level. As these trees are tiny and can contain exactly one decision for classification, they are mostly called as decision stumps. Each occurrence that is available in the training dataset should be weighed .The original weights are set to:

weight xi= i/n

Where xi is the ith training occurrence and n is the count of training occurrences.

F. *Naive Bayes Classifier*

Naive Bayes[2] is a classification technique with a notion which defines all features are independent and unrelated to each other. It defines that status of a specific feature in a class does not affect the status of another feature. Since it is based on conditional probability it is considered as a powerful algorithm employed for classification purpose .It works well for the data with imbalancing problems and missing values. Naive Bayes [6] is a machine learning classifier which employs the Bayes Theorem. Using Bayes theorem posterior probability $P(C|X)$ can be calculated from $P(C)$,$P(X)$ and $P(X|C)$.

Therefore, $P(C|X) = (P(X|C) P(C))/P(X)$

Where,

$P(C|X)$ = target class's posterior probability .

$P(X|C)$ = predictor class's probability.

$P(C)$ = class C's probability being true.

$P(X)$ = predictor's prior probability.

The evaluated performance of Naive Bayes algorithm using Confusion Matrix is as follows:

G. *Decision Tree Classifier*

Decision Tree[11] is a supervised machine learning algorithm used to solve classification problems. The main objective of using Decision Tree in this research work is the prediction of target class using decision rule taken from prior data. It uses nodes and internodes for the prediction and classification. Root nodes classify the instances with different features. Root nodes can have two or more branches while the leaf nodes represent classification. In every stage, Decision tree chooses each node by evaluating the highest information gain among all the attributes [11].

The evaluated performance of Decision Tree technique using Confusion Matrix is as follows:

	A	B
A-Tested Negative	407	93
B-Tested Positive	108	160

Table 2 : Confusion Matrix of Decision Tree

H. *Logistic Regression*

Logistic regression [11] is a machine getting to know classifier. This set of rules is used to split the observations for discrete classes. The outputs given by using the logistic regression is based totally on the opportunity feature. It uses the fee function that's known as sigma characteristic. Sigma function is more complex than the normal linear function. Logistic regression limit the cost function value between 0 to 1.

A-Tested Positive	72

Table 3: Confusion matrix for Logistic Regression

Dataset Used:

In this work WEKA tool is used for performing the experiment. WEKA is a software which is designed in the country New Zealand by University of Waikato, which includes a collection of various machine learning methods for data classification, clustering, regression, visualization etc. One of the biggest advantages of using WEKA is that it can be personalized according to the requirements. The main aim of this study is the prediction of the patient affected by diabetes using the WEKA tool by using the medical database PIDD.

PIDD-Pima Indians Diabetes Dataset :

The proposed methodology is evaluated on Diabetes dataset namely (PIDD) , which is taken from UCI Repository. This dataset comprises of medical detail of 768 instances which are female patients. The dataset also comprises numeric-valued 8 attributes where value of one class '0' treated as tested negative for diabetes and value of another class '1' is treated as tested positive for diabetes.

Attribute Description

NO	Attribute	Abbreviation of Attributes
1	Number of times pregnant	pr
2	Plasma glucose concentration	pl
3	Diastolic blood pressure (mm Hg)	pr
4	Skin fold thickness (mm)	sk
5	2-Hour serum insulin (mu U/ml)	in
6	BMI (weight in kg/(height in m)2)	ma
7	Diabetes pedigree function	pe
8	Age in years	ag
9	Class '0' or '1'	cl

Table 4: Attribute Description

Accuracy Measures Definitions and Formula:

Accuracy (A):

Accuracy determines the accuracy of the algorithm in predicting instances.

$$A = (TP+TN) / (Total\ no\ of\ samples)$$

Precision (P):

Classifiers correctness/accuracy is measured by Precision.

$$P = TP / (TP+ FP)$$

Recall (R):

To measure the classifiers completeness or sensitivity, Recall is used.
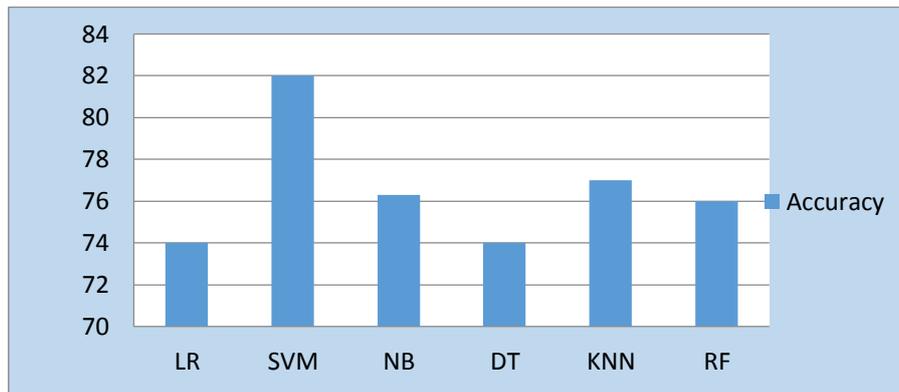
$$R = TP / (TP+FN)$$

F-Measure:

F-Measure is the weighted average of precision and recall.
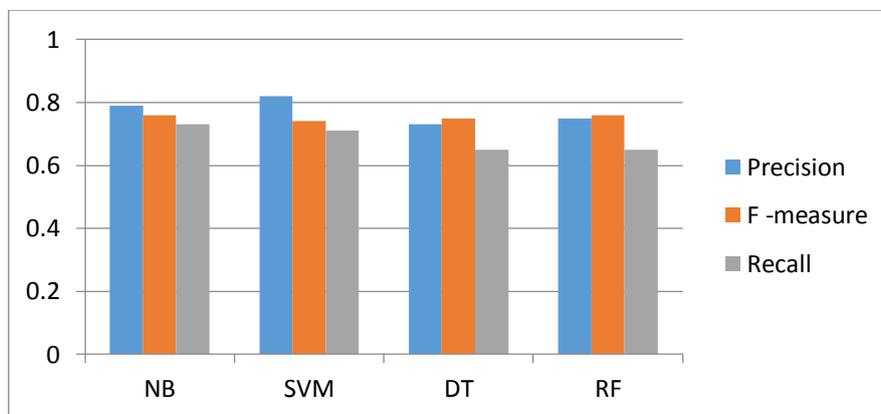
$$F = 2*(P*R) / (P+R)$$

V. RESULTS AND DISCUSSION

In this section will furnish the comparison results of various Machine Learning Algorithm for each techniques and ensemble methods of algorithms.

Graph 1. Represents the accuracy level of each algorithm

Comparison Major Accuracy Measure Values:

Graph 2. Represents the comparison major accuracy measure values

VI. CONCLUSION

A healthcare system plays a main role nowadays in monitoring the health related aspects of the humans around the globe. The researchers used various techniques to predict the data saved in healthcare database.In that way, machine learning techniques is one of the techniques used to predict the database stored. Here discussed about different algorithms and the accuracy recorded by the techniques. SVM algorithm produces highest accuracy than other algorithms.

REFERENCES

[1]. N. Yuvaraj ,K.R. SriPreethaa1 "Diabetes prediction in healthcare systems using machine learning algorithms on Hadoop cluster", Springer – 7 December2017

[2]. P.Suresh Kumar1, S. Pranavi2 "Performance Analysis of Machine Learning Algorithms on Diabetes Dataset using Big Data Analytics", IEEE, 978-1-5386-0514-1/17/$31.00 ©2017

[3]. Thanga Prasad, S Sangavi. S" Diabetic Data Analysis In Big Data With Predictive Method", ICAMMAET icammaet.2017.8186738

[4]. Ayman Mir ,Sudhir N. Dhage" Diabetes Disease Prediction using Machine Learning on Big Data of Healthcare" (ICCUBEA)-2018

[5]. Deepti Sisodiaa, Dilip Singh Sisodiab "Prediction of Diabetes using Classification Algorithms" Elsevier Procedia Computer Science, 132 (2018) 1578–1585

[6]. Muhammad Azeem Sarwar, 2Nasir Kamal, "Prediction of Diabetes Using Machine Learning Algorithms in Healthcare", International Conference on Automation & Computing, Newcastle University, Newcastle upon Tyne, UK, 6-7 September 2018

[7]. Priyanka Sonar "Diabetes Prediction Using Different Machine Learning Approaches", IEEE Xplore Part Number: CFP19K25-ART; ISBN: 978-1-5386-7808-4 (ICCMC 2019)

[8]. Lejla Alic, Hasan T. Abbas,"Predicting Diabetes in Healthy Population through Machine Learning", IEEE (CBMS), 2019.

[9]. Samrat Kumar Dey , Ashraf Hossain,"Implementation of a Web Application to Predict Diabetes Disease: An Approach Using Machine Learning Algorithm", 2018 21st International Conference of Computer and Information Technology (ICCIT), 21-23 December, 2018

[10]. Avinash Golande, Pavan Kumar T ,"Heart Disease Prediction Using Effective Machine Learning Techniques ", International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8, Issue-1S4, June 2019.

[11]. S. Pitchumani Angayarkanni ,"Predictive Analytics of Chronic Kidney Disease using Machine Learning Algorithm", International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8 Issue-2, July 2019.

[12]. Aditi Gavhane, Gouthami Kokkula ," Prediction of Heart Disease Using Machine Learning" 2nd International conference on Electronics, Communication and Aerospace Technology, IEEE (ICECA 2018)

[13]. D.Ramana kumar1, Dr. S.Krishnamohan Rao," Health care System: Stream Machine Learning Classifier for features Prediction in Diabetes Therapy" International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 1 (2018)

[14]. Mrs. Ashwini Abhale ,Shruti Gulhane, "Predictive analysis of Diabetic Patient Data Using Machine Learning and Big Data", International Journal of Research in Advent Technology (IJRAT) Special Issue E-ISSN: 2321-9637, 27th February 2019.

[15]. B. Suvarnamukhi, M. Seshashayee ," Big Data Processing System for Diabetes Prediction using Machine Learning Technique", International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-12, October 2019.