Detection and Prediction of Diabetes Using Machine Learning Techniques

[1]Priyanka Indoria, M.Tech., Dept. of CSE, Raipur Institute of Technology, Raipur, Chhattisgarh, India
[2]Yogesh Kumar Rathore, Assistant Professor, Dept. of CSE, Raipur Institute of Technology, Raipur, Chhattisgarh, India

Abstract— Diabetes is a one of the main source of visual deficiency, kidney disappointment, removals, heart disappointment and stroke. When we eat, our body transforms sustenance into sugars, or glucose. By then, our pancreas should discharge insulin. Insulin fills in as a "key" to open our cells, to enable the glucose to enter - and enable us to utilize the glucose for vitality. However, with diabetes, this framework does not work. A few noteworthy things can turn out badly – causing the beginning of diabetes. Sort 1 and sort 2 diabetes are the most well-known types of the malady, however there are additionally different sorts, for example, gestational diabetes, which happens amid pregnancy, and also different structures. This paper centreson late advancements in machine learning which have had critical effects in the discovery and determination of diabetes. In this paper, we explored Support Vector Machine, Decision Tree and Logistic Regression classifiers to detect the Diabetes. We obtained confusion matrix and ROC from different classifiers and generated True Positive (TP), False Negative (FN), False Positive (FP) and True Negative (TN). The accuracy of different classifiers when Training and testing ratio is kept 90 and 10. After applying the classifiers Accuracy of Decision Tree is 99.74, accuracy of Regression Analysis is 96.37 and Accuracy of Linear SVM is 97.62. Specificity of decision Tree is 99.687, Specificity of Regression Analysis is 98.283 and Specificity of Linear SVM is 99.688 And Sensitivity of Decision Tree is 100, Sensitivity of Regression Analysis is 88.60 and Sensitivity of Linear SVM is 89.24%. It is observed that 100% sensitivity is achieved by Decision Tree. In the case of appropriately chosen parameters, we explored that Decision Tree is 100% with that heavy betterments in case of sensitivity, and has the enough betterment in accuracy, Specificity. This way is computationally tractable and scales perfectly to huge high-dimensional data sets also.

***Keywords*-Diabetes, Type 1 Diabetes, Type 2 Diabetes, Insulin, Machine Learning, Support Vector Machine, Regression, decision Tree.**

I. INTRODUCTION

Diabetes is a chronic disease, it is associated with risk for heart disease, kidney disease, blindness, and serious health challenge in India. It needs continued medical aid and patient self-management education to forestall severe issues and to scale back the danger of long-run issues. Diabetes Mellitus implies elevated amounts of sugar (glucose) in the circulatory system and in the pee. Signs or Symptoms of Diabetes: Frequent Urination,Increased thirst, Increased yearning, Tired/Sleepiness, Weight misfortune, Blurred vision. Emotional episodes, Confusion and trouble concentrating, frequent infections/poor recuperating. Type 1 diabetes : In Type 1 diabetes the beta cells of the pancreas have been harmed or assaulted by the body's own safe framework (auto - invulnerability). Because of this assault, the beta cells bite the dust and are along these lines unfit to make the required measure of insulin to move glucose into the cells, causing high glucose (hyperglycaemia). Type 1 diabetes happens in around 5 - 10% of those with diabetes and as a rule in individuals under 30 years old, yet can happen at any age. The signs and manifestations have a fast beginning and are generally serious in nature. As Type 1 diabetes is caused by an absence of insulin, individuals need to supplant what the body can't create itself. Type 2 diabetes commonly occurs in adults who are obese. There are number of factors that resulting to the high blood glucose levels. One of the important factor is the body's resistance to insulin in the body, essentially ignoring its insulin secretions. Another factor is the falling production of insulin by the beta cells of the pancreas. Diagnosis disease is one of the applications where machine learning techniques and algorithms areproving successful results. In recent years, many researches have been done to develop decision support system to help in diagnosing the diabetes. Machine learning techniques is used to solve the real world problems by building a model that is good and useful approximation to the data. Machine learning is the logical field managing the routes in which machines gain as a matter of fact. For some researchers, the term "machine learning" is indistinguishable to the expression "artificial intelligence", given that the likelihood of learning is the primary normal for an element called smart in the broadest feeling of the word. The reason for machine learning is the development of PC frameworks that can adjust and gain from their experience. Learning process in machine learning is divided into two steps:

1. Training, and
2. Testing

In training process, samples in training data are taken as input in which attributes/features are learned by different learning algorithm and build the learning model. In the testing process, machine learning model uses the execution to make the prediction for the test or production data. Resulting data is the output of learning model which gives the prediction or classified data.Machine learning undertakings are normally ordered into three general classifications:a) Supervised Learning, in which the framework derives a capacity from marked preparing information, b) Unsupervised Learning, in

which the learning framework tries to surmise the structure of unlabeled information, andc) Reinforcement learning, in which the framework communicates with a dynamic domain.Support Vector Machine (SVM) is a supervised machine learning technique used in medical diagnosis for characterization or regression problem where the dataset shows SVM about the classes with the goal that SVM can group any new information. It works by ordering the information into various classes by finding a line which isolates the training dataset into classes. As there are numerous such straight hyperplanes, SVM calculation tries to expand the separation between the different classes that are included and this is referred as margin maximization. On the off chance that the line that maximizes the separation between the classes is distinguished, the likelihood to sum up well to unseen information is expanded.SVM's are characterized into two classifications:Linear Support Vector Machine, in linear Support Vector Machine the training data are divided by a hyperplane, andNon-Linear Support Vector Machine, in non-linear linear Support Vector Machine, it is not possible to separate the training data using a hyperplane.A decision tree is a graphical portrayal that makes utilization of branching philosophy to embody every conceivable result of a choice, in view of specific conditions. In a decision tree, the inside node speaks to a test on the quality, each branch of the tree speaks to the result of the test and the leaf node speaks to a specific class label i.e. the decision made subsequent to registering the majority of the properties. The classification rules are spoken to through the path from root to the leaf node.Decision tree machine learning calculations enable an information researcher to catch if an alternate choice was taken, at that point how the operational idea of a circumstance or model would have changed strongly. Decision tree help settle on ideal decision by enabling an information researcher to navigate through forward and in reverse computation ways.

Regression techniques predict continuous responses—for example, changes in temperature or fluctuations in power demand, electricity load forecasting and algorithmic trading.Regression is concerned about displaying the relationship between factors that is iteratively refined utilizing a measure of mistake in the prediction made by the model. Regression is the undertaking of predicting the value of a persistently varying variable, for example, a value, a temperature if given some info factors like highlights and regressor. Regression is in almost every field such as engineering, physics, economics, management, environmental sciences, biology and social sciences is needed for estimation and forecasting. It can be said regression analysis, the most commonly used method among statistical techniques.

This paper presentsan importance of machine learning techniques and classifiers Support Vector Machine, Regression analysis and Decision Tree in the disease diagnosis through the collected data for diabetes.

II. LITERATURE SURVEY

Machine learning techniques for classification of diabetes and cardiovascular diseases. Berina et al. [3], in this paper creators intended to play out an audit of Artificial Neural Network and Bayesian Network and their application in order of diabetes and CVD sicknesses. The reason for existing is to demonstrate the correlation of machine learning procedures and to find the best choice for accomplishing the most astounding yield exactness of the arrangement.This paper speaks to the examination of utilization of two machine learning methods, Artificial Neural Network and Bayesian Network in order of diabetes and cardiovascular infections. In the examination of use of Artificial Neural Network and Bayesian Network for characterization of diabetes and CVD, diverse esteems for the system precision have been accomplished. The consequences of prepared ANN and BN for order of diabetes from chose papers [5-9, 15-19]. Creators watched that the precision of diabetes order utilizing ANN fluctuates in the vicinity of 72.2% and 99%. What's more, the exactness of diabetes arrangement utilizing BN shifts in the vicinity of 71% and 99.51%. As indicated by thought about outcomes, the most noteworthy precision was accomplished in Bayesian Network yet in addition the littlest exactness was appeared in Bayesian Network.One of the greatest reasons for death worldwide are diabetes and cardiovascular malady. The early grouping of these sicknesses can be accomplished creating machine learning models, for example, Artificial Neural Network and Bayesian Network. In correlation of mean precision of 10 logical papers about diabetes grouping and 10 papers about CVD order it was reasoned that the higher exactness was accomplished with ANN in the two cases (87.29 for diabetes and 89.38 for CVD). The utilized Naive Bayesian system, because of the suspicion of autonomy among watched hubs, may be less precise than ANN approach. Thus, in agreement to get result it can be reasoned that the higher probability to get better exactness in grouping diabetes as well as CVD is the point at which it is connected to Artificial Neural Network.

III.SYSTEM STUDY

In the detection of diabetes dieses using machine learning techniques literature, it is quite common to assume that observations come from different classes and data sets. One represented by a trained data set and the other by testing data set. We observe independent and identically measurements from the mixture of these dataset. We have to compute all the information needed to compute density levels, to perform hypothesis testing, or to make other statistical arguments. Consequently, some form of density estimation (e.g., kernel density estimation) seems to be a natural prerequisite for our task. However, kernel density estimation is itself an unnecessary intermediate step which estimates the continuous density for the whole data domain from discrete points, after which one level set parameter is calculated for each data point. The quantities we are actually interested in are the properties of the discrete observations, not the continuous space around them. As a result, specifying the full distribution throughout the whole space is inessential, introducing computational burden while accumulating estimation errors.

In this paper we propose detection of diabetes by applying different classification algorithm. Our algorithm takes as input a set of 70 set of data of diabetes patients. Database contains the distribution for seventy sets of information recorded on

diabetes patients. The dataset is obtained from UCI Machine Learning Repository within which diabetes patient records were obtained from 2 sources: associate automatic electronic recording device and paper records. The automated device had an indoor clock to timestamp events, whereas the paper records solely provided "logical time" slots (breakfast, lunch, dinner, bedtime). The info contains the info of quite 5000 persons. We have a tendency to square measure taking knowledge samples of one thousand and 800 folks for our testing purpose wherever each field contains twenty five feature of a patient

In the detection of diabetes using machine learning technique literature, the observations come from trained data set. We have explored some problems analysis and results in the existing system and the results comes from different machine learning classification. The details of methods and techniques employed in the investigation have been stated. In this paper component details and theoretical consideration to design/proof/theory has been be described.

Diabetes risk Prediction Model will support medical professionals and practitioners in predicting risk standing based on the clinical information records. In medical specialty field data processing and its techniques plays an important role for prediction and analysing totally different kind of health problems. The care trade offers vast amounts of care information and that got to be deep-mined to establish hidden info for valuable decision selection. Determinant hidden patterns and relationships could typically terribly powerful and unreliable. The health record is classed and foreseen if they need the symptoms of malady risk and victimization risk factors of disease. It's indispensable to seek out the most effective work algorithmic program that has greater accuracy, speedy and memory utilization on prediction within the case of Diabetes.

IV. SYSTEM DEVELOPMENT

Algorithm and Properties

System Modules:

Module 1: Access the database

The datasets are stored in the access file. The files cannot be used for direct comparison. The file is pre-processed to identify the features from each datasets.

Module 2: Applying Classifiers

The module identifies the features for each data applying Support Vector Machine Classifier, applying Logistic Regression Classifier, and applying decision Tree Classifier

Module 3: Obtaining Confusion Matrix, ROC

After Applying different classifiers confusion matrix and ROC is obtained. Confusion matrix is used to generate True Positive (TP), False Negative (FN), False Positive (FP) and True Negative (TN).

Module 4: Calculation of Accuracy

Calculation of Accuracy is obtained for SVM, Logistic Regression and Decision Tree.

Accuracy is calculated as:

Accuracy = Sum of All Diagonal Elements in Confusion Matrix

Accuracy = ((TP+TN) / (TP+TN+FP+FN)) * 100

Module 5: Calculation of Specificity

Specificity is calculated as:

Specificity = (TN/(TN+FP))*100

Module 6: Sensitivity Calculation

Specificity is calculated as:

Sensitivity = (TP/(TP+FN)) * 100

Database contains the distribution for 70 sets of data recorded on diabetes patients. The dataset is obtained from UCI Machine Learning Repository in which diabetes patient records were obtained from two sources: an automatic electronic recording device and paper records. The automatic device had an internal clock to timestamp events, whereas the paper records only provided "logical time" slots. The database contains the data of more than 5000 persons. We are taking data samples of 1000 and 800 people for our testing purpose where every field contains 25 feature of a patient.

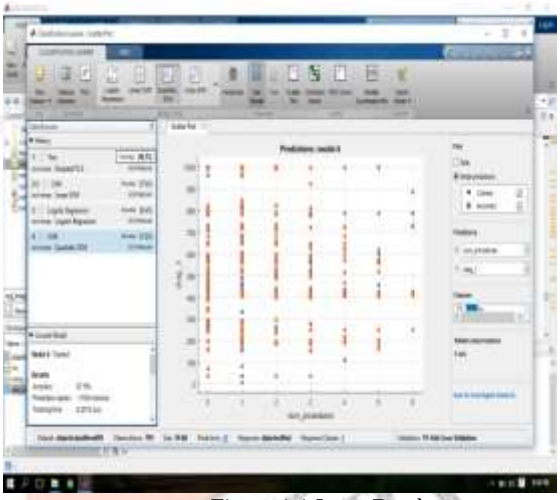

Figure 1.1 Input Database.

On applying different classification algorithms following confusion matrix and ROC has been obtained which is shown in figure 1.4, 1.5 and figure 1.6, here confusion matrix is used to generate True Positive (TP), False Negative (FN), False Positive (FP) and True Negative (TN) classification results and Receiver Operating Characteristic curve (or ROC curve is a plot of the true positive rate against the false positive rate for the different possible cut points of a diagnostic test Figure 1.2 shows the different feature selected for classifiers and figure 1.6 shows the accuracy of all applied classifiers on using cross 10 validation.

Figure 1.2 Select different features as Input.

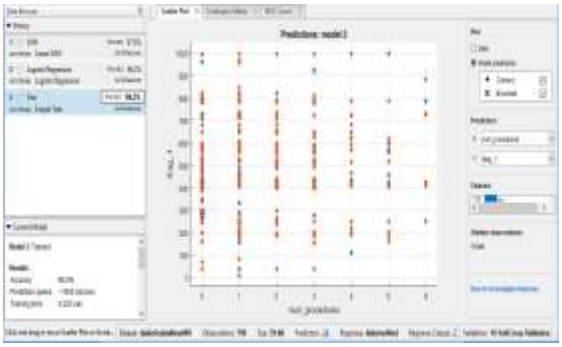

Figure 1.3 Output Showing Accuracy of DifferentClassifiers.

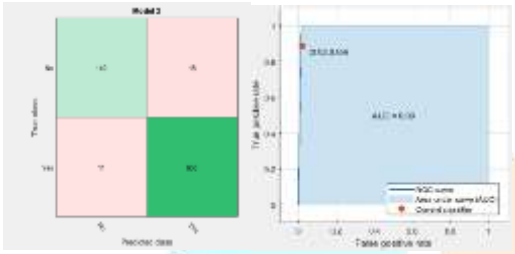

Figure 1.4: Confusion matrix and ROC curve of Linear Regression Analysis.

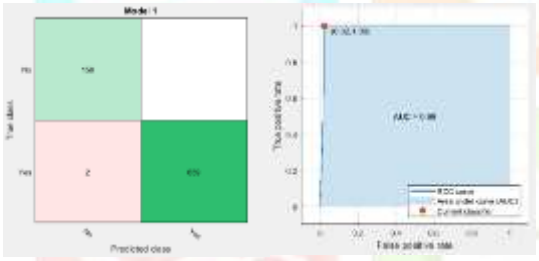

Figure 1.5: Confusion matrix and ROC curve of Decision Tree Algorithm.

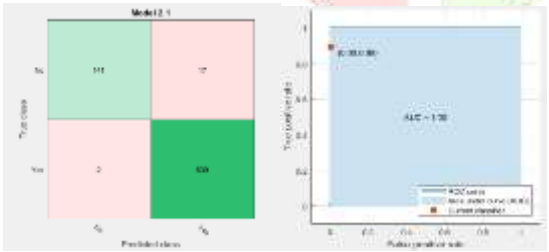

Figure1.6: Confusion matrix and ROC curve of Linear Support Vector Machine.

Above Figures the accuracy of different classifiers when Training and testing ratio is kept 90 and 10.

Output Comparison:

From above confusion matrix we can find out True Positive (TP), False Negative (FN), False Positive (FP) and True Negative (TN) as:

A. TP would be the value in the main diagonal.
B. FN for each class would be the sum of all values in the corresponding row excluding (TP).
C. FP for each class would be the sum of all values in the corresponding column excluding the main diagonal element (TP).
D. TN for each class would be the sum of all the values of the confusion matrix excluding that class's row and column.
E. First confusion matrix is created then following things are calculated:

1. Accuracy:
   Accuracy = Sum of All Diagonal Elements in Confusion Matrix

Accuracy = ((TP+TN) / (TP+TN+FP+FN)) * 100

For Decision Tree:
Accuracy = (158+639)/ (158+639+2+0) * 100
Accuracy= 99.74 %

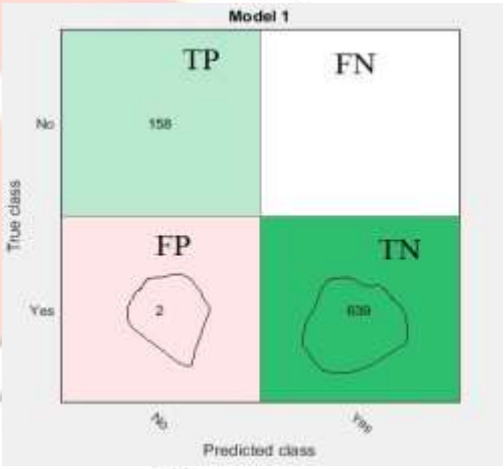

Figure 1.7 Accuracy for decision Tree.

Similarly other accuracy can be calculate and show in table 1.1 below as:

S. No.	Algorithm	Accuracy
1	Decision Tree	99.74%
2	Regression Analysis	96.37%
3	Linear SVM	97.62%

Table 1.1 Accuracy of Different Classifiers.

2. Specificity:

Specificity of a class can be find out by using following formula

Specificity = TNvalues of a particular class / TNvalues of a particular class + FP values of class

Specificity = (TN/(TN+FP))*100 (2)

For Decision Tree Algorithm:

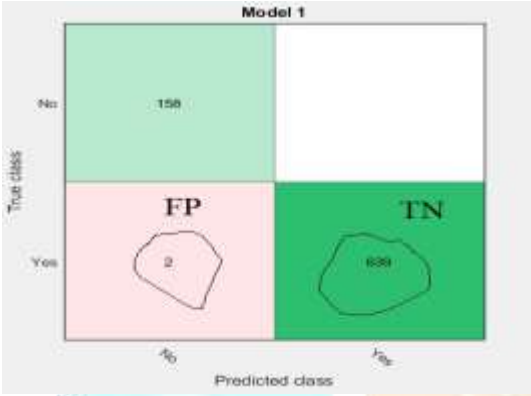

Figure 1.8 Specificity for Decision Tree.

Specificity = (TN/(TN+FP))*100

Specificity = 639/(639+2) * 100 = 99.687

Similarly other Specificity can be calculate and show in table 1.2 below as:

Table 1.2 Specificity of Different Classifiers.

S. No.	Algorithm	Specificity
2	Decision Tree	99.687
3	Regression Analysis	98.283
4	Linear SVM	99.688

3. Sensitivity:

Sensitivity = (TP/(TP+FN)) * 100

For Decision Tree:

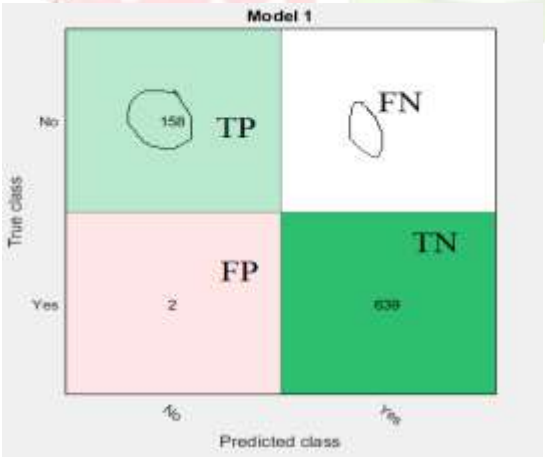

Figure 1.9 Sensitivity for decision Tree

Sensitivity = (TP/(TP+FN)) * 100

Sensitivity = [158/(158+0)] * 100 = 100

Similarly other Specificity can be calculate and show in table 1.3 below as:

Table 1.3 Sensitivity of Different Classifiers.

S. No.	Algorithm	Sensitivity
1	Decision Tree	100
2	Regression Analysis	88.60
3	Linear SVM	89.24

V. CONCLUSION

In this work, we explored Support Vector Machine, Decision Tree and Logistic Regression classifiers to detect the Diabetes. We obtained confusion matrix and ROC from different classifiers and generated True Positive (TP), False Negative (FN), False Positive (FP) and True Negative (TN). The accuracy of different classifiers when Training and testing ratio is kept 90 and 10. After applying the classifiers Accuracy of Decision Tree is 99.74, accuracy of Regression Analysis is 96.37 and Accuracy of Linear SVM is 97.62. Specificity of decision Tree is 99.687, Specificity of Regression Analysis is 98.283 and Specificity of Linear SVM is 99.688 And Sensitivity of Decision Tree is 100, Sensitivity of Regression Analysis is 88.60 and Sensitivity of Linear SVM is 89.24%. It is observed that 100% sensitivity is achieved by Decision Tree. In the case of appropriately chosen parameters, we explored that Decision Tree is 100% with that heavy betterments in case of sensitivity, and has the enough betterment in accuracy, Specificity. This way is computationally tractable and scales perfectly to huge high-dimensional data sets also.

VI. REFERENCES

[1] www.diabetesresearch.org/document.doc?id=284

[2] D. Yu, and L. Deng, 2011, “Deep learning and its applications to signal and information processing,” IEEE Signal Process. Mag., vol. 28, no. 1, pp. 145-154.

[3] Habibi, N., Hashim, S. Z. M., Norouzi, A., & Samian, M. R. (2014). A review of machine learning methods to predict the solubility of overexpressed recombinant proteins in Escherichia coli. BMC bioinformatics, 15(1), 134.

[4]Langarizadeh, M., & Moghbeli, F. (2016). Applying Naive Bayesian Networks to Disease Prediction: a Systematic Review. Acta Informatica Medica, 24(5), 364.

[5] Olaniyi, E. O., & Adnan, K. (2014). Onset diabetes diagnosis usingartificial neural network. International Journal of Scientific andEngineering Research, 5(10).

[6] Jayalakshmi, T., & Santhakumaran, A (2010, February). A novelclassification method for diagnosis of diabetes mellitus using artificialneural networks. OSDE, 159-163. (2010)

[7] Pradhan, M., & Sahu, R. K. (2011). Predict the onset of diabetes diseaseusing Artificial Neural Network (ANN). International Journal ofComputer Science & Emerging Technologies (E-ISSN: 2044-6004).
[8] Sejdinovic, Dijana, et al. "Classification of Prediabetes and Type 2Diabetes using Artificial Neural Network." Springer. CMBEBIH 2017.
[9] Soltani, Z., & Jafarian, A (2016). A New Artificial Neural NetworksApproach for diagnosing Diabetes Disease Type II. InternationalJournal of Advanced Computer Science & Applications, 1(7),89-94.
[10] Atkov, O. Y., Gorokhova, S. G., Sboev, A G., Generozov, E. Y.,Muraseyeva, E. v., Moroshkina, S. Y, & Cherniy, N. N. (2012).Coronary heart disease diagnosis by artificial neural networks includinggenetic polymorphisms and clinical parameters. Journal of cardiology,59(2), 190-194.
[11] Olaniyi, E. O., Oyedotun, O. K., & Adnan, K. (2015). Heart diseasesdiagnosis using neural networks arbitration. International Journal ofIntelligent Systems and Applications, 7(12), 72.
[12] Colak, M. C. et. al., Predicting coronary artery disease using differentartificial neural network modelslkoroner arter hastaliginin degisik yapay
sinir agi modelleri lie tahmini. The Anatolian Journal of Cardiology(Anadolu Kardiyoloji Dergisi), 8(4), 249-255, (2008).
[13] Can, M. (2013). Diagnosis of cardiovascular diseases by boosted neuralnetworks.
[14] Sayad, A T., & Halkarnikar, P. P. Diagnosis of heart disease usingneural network approach. In Proceedings of IRF InternationalConference, 13th April-2014, Pune, India, ISBN (pp. 978-93).
[15] Guo, Y, Bai, G., & Hu, Y (2012, December). Using bayes network forprediction of type-2 diabetes. In Internet Technology and Secured
Transactions, 2012 International Conference for (pp. 471-472). IEEE.
[16] Kumari, M., Vohra, R., & Arora, A (2014). Prediction of DiabetesUsing Bayesian Network.
[17] N. Sarma, S. Kumar, AK. Saini, A Comparative Study on Decision Treeand Bayes Net Classifier for Predicting Diabetes Type 2, 2014, ISSN:2278-0882, ICRTIET-2014.
[18] Dewangan L. A., & Agrawal, P. Classification of Diabetes MellitusUsing Machine Learning Techniques.
[19] Nai-arun, N., & Moungmai, R. (2015). Comparison of Classifiers for theRisk of Diabetes Prediction. Procedia Computer Science, 69,132-142.
[20] Elsayad, A, & Fakr, M. (2015). Diagnosis of cardiovascular diseaseswith Bayesian classifiers. 1. Comput. Sci., II (2), 274-282.
[21] K. P. Exarchos, et al. Prediction of coronary atherosclerosis progressionusing dynamic Bayesian networks. IEEE EMBC, 2013.
[22] D.S. Medhekar, M.P. Bote & Deshmukh, S. D., Heart disease predictionsystem using naive bayes. Int. J. Enhanced Res. Sci. Technol. (2013).
[23] Patil, R. R., Heart disease prediction system using naive bayes andjelinek-mercer smoothing. International Journal of Advanced Researchin Computer Science and Communication Engineering, (2014).
[24] E. Miranda et. al., Detection of CYD Risk's Level for Adults UsingNaive Bayes Classifier. Healthcare Informatics Research, (2016).
[25] Dinu A.J., Ganesan R, Felix Joseph and Balaji V, "A study on Deep Machine Learning Algorithms for diagnosis of diseases." International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 17 (2017) pp. 6338-6346.
[26] R. Catherine Silvia, R. Vijayalakshmi, 2013, "Detection of Non-Proliferative Diabetic Retinopathy in fundus images of the human retina", International Conference on Information Communication and Embedded Systems (ICICES).
[27] Iyer A., Jeyalatha S. and Sumbaly R, 2015, "Diagnosis of Diabetes Using Classification Mining Techniques", International Journal of Data Mining & Knowledge Management Process (IJDKP), 5, 1-14.
[28] Sen S.K. and Dash S, 2014, "Application of Meta Learning Algorithms for the Prediction of Diabetes Disease", International Journal of Advance Research in Computer Science and Management Studies, 2, 396-401.
[29] Zahed Soltani and Ahmad Jafarian, "A New Artificial Neural Networks Approach for Diagnosing Diabetes Disease Type II." International Journal of Advanced Computer Science and Applications, Vol. 7, No. 6, 2016
[30] Pima Indians Diabetes Data Set, https://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes [Last Available: February 2016].
[31] E. Miranda et. al., Detection of CYD Risk's Level for Adults UsingNaive Bayes Classifier. Healthcare Informatics Research, (2016).

Author's Profile

1. Mrs. Priyanka Indoria is pursuingMaster in Technology in Computer Science & Engineering from Raipur Institute of Technology, Raipur, Chhattisgarh, Affiliated from Chhattisgarh Swami Vivekanand Technical University, Bhilai, Chhattisgarh, India. Her area of interest includes Digital image processing and Computer Grphics,

2. Mr.Yogesh Rathore is Assistant Professor in Department of Computer Science & Engineering, Raipur Institute of Technology, Raipur, Chhattisgarh, India. He is M. Tech. in Computer Science& Engineering. His area of interest include Digital image processing and Computer Graphics.