Comparison of Classification Methods for the Diabetic Prediction

Parshad H. Suthar, Research Scholar

Abstract- The diabetic prediction is the major challenge of data sciences due to complex dataset. The dataset of diabetic is collected from the UCI repository and it is further pre-processed to remove missing values. In the second phase, the feature extraction is applied in which relationship between attribute and target set is established. The method of classification is applied in this last phase for the prediction analysis. To apply classification whole dataset get divide into training and testing. The methods of SVM, KNN and decision tree are applied for the diabetic prediction. The three classifiers are implemented in python and it is analyzed that decision tree is best performing classifier among SVM and KNN for the diabetic prediction

Keywords- Diabetic prediction, SVM, KNN, Decision tree

I. INTRODUCTION

An illness that impacts the ability of a body to produce hormone insulin due to which abnormal metabolism of carbohydrate is caused and also, the levels of glucose in the blood are increased is known as diabetes. Blood sugar level generally remains high in a diabetic person. Every year large number of people all over the world lost their lives because of this deadly disease. The severity of this disease can be gauged from the fact that this disease generates various other diseases as well. Some steps such as visiting a diagnostic center, consulting with the doctor, and wait for a day or more to obtain the reports are included in general detection process [1]. Also, the patient has to waste a lot of money all the time for getting his/her diagnosis report. A group of metabolic diseases is defined as Diabetes Mellitus (DM). Abnormal insulin secretion and/or action are the main cause of these disorders. High blood sugar levels (hyperglycemia) and damaged metabolism of carbohydrates, fat and proteins are caused due to insufficiency of insulin. Diabetes Mellitus is a very common endocrine disease. Around 200 million people all over the world are suffering from this disease. A dramatic growth in diabetic patients is expected in the nearby future [2]. There are different types of Diabetes Mellitus. But, two types are important from medical point of view. On the basis of etiopathology of this disease, these types are called type 1 diabetes (T1D) and type 2 diabetes (T2D).

Machine learning is a technical field. This field deals with the methods in which machines learn from experience. Some researchers consider both “machine learning” and “artificial intelligence” same. They also provide a proof in the favor of their claim that the possibility of learning is the key trait of an object known as intelligence in the broader sense of the word. Constructing computer systems with the ability of adapting and learning from experience is the main objective of machine learning. It is possible to find a solution of issues related to diabetes mellitus using Machine Learning algorithms [3]. In recent times, several systems have been designed with the help of data mining for making predictions about Diabetes Mallitus. These systems can make predictions about a patient being diabetic or not. In addition, the treatment of patient can be started before becoming this disease more dangerous by predicting this disease in early stage. Hidden knowledge from large number of diabetes data can be extracted using data mining. Therefore, it can be said that data mining plays an important role in diabetes detection. There are mainly two types of machine learning algorithms. These are called supervised and unsupervised machine learning algorithms. In supervised learning, it is required for a system to “learn” a specific function in inductive manner [4]. This function is known as target function. This function is an expression of a model that describes the data. The value of a variable is predicted using objective function. This variable is known as dependent variable or output variable. The value is predicted from a set of variables known as independent variables or input variables. Instances are defined as a set of possible input values of the function, i.e. its domain. A set of features define every case. On the other hand, training data or examples are defined as a subset of all cases, for which the output variable value is identified. Classification and regression are the two sorts of learning tasks in supervised learning. Classification models make an attempt for predicting different classes, such as blood groups while regression models make predictions about mathematical values. The system makes an attempt to determine the hidden structure of data or relations among variables in unsupervised learning [5]. In such case, training data is made up of instances without any matching labels. In contrast to machine learning, Association Rule Mining developed after a long time. Unsupervised learning algorithms learn few features from the data. Unsupervised learning uses earlier learned features for recognizing the class of the data after the introduction of novel data. This approach is generally utilized for clustering and reducing features. It is a key task of exploratory data mining, and a frequent method for statistical data analysis. This approach is utilized in various areas. These areas include machine learning, pattern recognition, image analysis, information extraction, bioinformatics, data compression, and computer graphics and so on. One more learning type is known as Reinforcement Learning. The term Reinforcement Learning is a common term. This term represents a family of techniques [6]. Here, the system tries to learn by establishing direct communication with environ for maximizing the concept of growing return. In this type of learning, the system does not have any previous information regarding the behavior of the environment. The only way of discovery is via trial and failure. This type of learning is generally implemented in independent systems. Some popular machine learning algorithms include Support Vector Machine, Naive Bayes, and Decision Tree and so on. Support Vector Machine is one of the most commonly utilized state-of-the-art machine learning algorithms. Support vector machine is a standard set of supervised machine learning model [7]. This approach is generally utilized for classification purpose. The working of support vector machine is based on the principle of margin computation. On the whole, this algorithm draws margins among the classes. The margins are drawn for maximizing the distance between the margin and the classes. This phenomenon minimizes the classification error. Naive Bayes is a classification algorithm. This algorithm is based on the concept that all features are independent and do not relate to each other. This classifier describes that status of a particular feature within a class and causes no influence on the status of another feature. This algorithm depends on conditional probability [8]. Naive Bayes is considered as a robust algorithm for performing classification task. Decision Tree is a supervised machine learning algorithm. This algorithm is generally utilized for providing solution of complex classification issues. Predicting target class with the help of decision rule taken from prior data is the main of this classifier. This classifier makes use of nodes and internodes for performing prediction and classification tasks.

II. LITERATURE REVIEW

Hasan Abbas, et.al (2019) used machine learning for making predictions about the future researches based on type-2 diabetes [9]. Ten features and support vector machines were used in this work for constructing the prediction model. The employed ten features were considered as powerful predictors of future diabetes mellitus. This work employed 10-fold cross-validation for the training of classifier because of the unbalanced nature of the dataset in terms of the class labels. A hold-out set was used for testing purpose. The analytic outcomes revealed that accuracy and recall rate of 84.1% and 81.1% were achieved respectively. The results of this analysis could be used to identify the people having more risk of occurring type-2 diabetes.

Debadri Dutta, et.al (2018) stated that fatness and high sugar level were the root causes of Diabetes mellitus [10]. In this work, an effort had been made to find out the crucial factors behind the occurrence of diabetes mellitus. Variable and feature choice had become the center of attraction in various diabetes related researches. In these researches, datasets with tens or large volume of factors were available. In the same way, the most essential features were considered in this work for making predictions about a person being diabetic in nearby future.

Priyanka Sonar, et.al (2019) presented a comparative analysis of different machine learning algorithms for diabetes prediction [11]. These algorithms included support vector machine, artificial neural network, naïve bayes, decision tree etc. Developing a system for making predictions about the diabetes risk level in a patient in more accurate way was the major purpose of this analysis. The prediction model was developed on the basis of different classification algorithms. The tested results depicted that Decision Tree, Naive Bayes and Support Vector Machine showed accuracy rate of 85%, 77% and 77.3% respectively in diabetes prediction.

Ayman Mir, et.al (2018) recognized Machine Learning as an extremely efficient approach for diabetes prediction. It was possible to use this approach for diagnosis of diabetes disorder in early stage [12]. This could provide support to the doctors in efficient decision making for diagnosis. The main objective of this work was to construct a classifier with the help of WEKA software for predicting diabetes mellitus using different learning algorithms. The research works for suggesting the optimum algorithm were based on good performance result for predicting diabetes disorder. The tested results depicted that Support Vector Machine outperformed the other classification algorithms in terms of accuracy in diabetes prediction.

Aparimita Swain, et.al (2016) presented a comparative analysis of Artificial Neural Network (ANN) and hybrid Adaptive Neuro-Fuzzy Inference System (ANFIS) for making predictions about the risk level of diabetes mellitus [13]. Diabetes Mellitus was also categorized in different categories in this work using these approaches. The data of 100 people having average 42 year age with equal ratio of men and women was used for the training of network. Moreover, a discussion had been made regarding the performance of both algorithms in terms of accuracy. The tested outcomes depicted that the ANFIS (Adaptive Neuro-Fuzzy Inference System) algorithm outperformed the Artificial Neural Network in terms of accuracy.

Muhammad Azeem Sarwar, et.al (2018) used six different machine learning algorithms for discussing the predictive analytics in medical domain [14]. A dataset of patient's medicinal record was used for carrying out tests. On this dataset, six different machine learning algorithms were implemented. A discussion was made regarding the performance and accuracy of the implemented algorithms.

Afterward, these algorithms were compared for evaluating them on the basis of their performance. The comparative results discovered the optimum algorithm for diabetes prediction. The main objective of this work was to provide support to doctors and practitioners for predicting diabetes disease in early stage with the help of machine learning algorithms.

Md. Faisal Faruque, et.al (2019) used machine learning algorithms for discovering several risk factors related to diabetes mellitus [15]. Machine learning algorithms constructed prediction models using diagnostic clinical datasets for providing good results in information extraction. These datasets were gathered from the patients suffering from diabetes. Diabetic patients could be predicted easily by extracting information from the gathered data. In this work, four prominent machine learning algorithms were used on adult population data for making prediction about diabetic mellitus. The tested results revealed that C4.5 decision tree outperformed the other machine learning algorithms in terms of accuracy.

V. Swathi Lakshmi, et.al (2019) recommended a novel technique for detecting risk level in diabetic patients [16]. Ontology Based Machine Learning System had been used in this work for detecting the diabetic patient’s risk level. Diabetes symptoms, causes and treatments were included in Ontology. Nave base algorithm was utilized for decision making decision on the basis of patient’s medical data. This algorithm described the probabilities of risk level as well. The tested results depicted that the proposed technique outperformed the other existing techniques in terms of accuracy.

Table 1: Table of Comparison

Author	Year	Description	Outcome
Hasan Abbas, Lejla Alic, Marelyn Rios, Muhammad Abdul-Ghani, Khalid Qaraqe	2019	Used machine learning for making predictions about the future researches based on type-2 diabetes. Ten features and support vector machines were used in this work for constructing the prediction model.	The analytic outcomes revealed that accuracy and recall rate of 84.1% and 81.1% were achieved respectively. The results of this analysis could be used to identify the people having more risk of occurring type-2 diabetes
Debadri Dutta, Debpriyo Paul, Parthajeet Ghosh	2018	Stated that fatness and high sugar level were the root causes of Diabetes mellitus. In this work, an effort had been made to find out the crucial factors behind the occurrence of diabetes mellitus.	The most essential features were considered in this work for making predictions about a person being diabetic in nearby future.
Priyanka Sonar, K. JayaMalini	2019	Presented a comparative analysis of different machine learning algorithms for diabetes prediction.	The tested results depicted that Decision Tree, Naive Bayes and Support Vector Machine showed accuracy rate of 85%, 77% and 77.3% respectively in diabetes prediction.
Ayman Mir, Sudhir N. Dhage	2018	Recognized Machine Learning as an extremely efficient approach for diabetes prediction. It was possible to use this approach for diagnosis of diabetes disorder in early stage.	The tested results depicted that Support Vector Machine outperformed the other classification algorithms in terms of accuracy in diabetes prediction.
Aparimita Swain, Sachi Nandan Mohanty, Ananta Chandra Das	2016	Presented a comparative analysis of Artificial Neural Network (ANN) and hybrid Adaptive Neuro-Fuzzy Inference System (ANFIS) for making predictions about the risk level of diabetes mellitus.	The tested outcomes depicted that the ANFIS (Adaptive Neuro-Fuzzy Inference System) algorithm outperformed the Artificial Neural Network in terms of accuracy.
Muhammad Azeem Sarwar, Nasir Kamal, Wajeeha Hamid, Munam Ali Shah	2018	Used six different machine learning algorithms for discussing the predictive analytics in medical domain. A dataset of patient's medicinal record was used for carrying out tests.	The main objective of this work was to provide support to doctors and practitioners for predicting diabetes disease in early stage with the help of machine learning algorithms.

Md. Faisal Faruque, Asaduzzaman, Iqbal H. Sarker	2019	Used several machine learning algorithms for discovering several risk factors related to diabetes mellitus.	The tested results revealed that C4.5 decision tree outperformed the other machine learning algorithms in terms of accuracy.
V. Swathi Lakshmi, V. Nithya, K. Sripriya, C. Preethi, K. Logeshwari	2019	Recommended a novel technique for detecting risk level in diabetic patients. Ontology Based Machine Learning System had been used in this work for detecting the diabetic patient’s risk level.	The tested results depicted that the proposed technique outperformed the other existing techniques in terms of accuracy.

III. Research Methodology

This research work is related to diabetic prediction. The diabetic prediction has various steps which are based on the certain steps which are described below:-

1. Data Collection and pre-processing:- The dataset of diabetes will be collected from the UCI repository. The dataset get pre-processed and missing, redundant data will be removed from the dataset

2. Feature Extraction- In the second phase, technique of feature extraction will be applied in the network. The feature extraction technique will establish relationship between the attribute and target set. The PCA algorithm is applied on the input dataset which will simply the dataset. The principle component analysis (PCA) algorithm will select the most frequent attribute from the dataset for the prediction

3. Classification: - The classification approach is the last phase of the prediction analysis. In the classification approach, input dataset will be divided into training and test sets. The training set will be 60 percent of the whole dataset and test set will be 40 percent. The approach of voting based classification is applied for the diabetes prediction. The three classifiers are implemented for the prediction analysis which are described below

3.1. Support vector machine: - Support Vector Machine classifier is a supervised statistical learning algorithm. This approach is utilized for linear and non-linear deterioration scrutiny and prototype categorization. SVM approach segregates the two classes with an utmost fringe amid them with the help of a hyper-linear plane for linear separable categorization. The characteristic vectors are planned to a novel characteristic space in a non-linear separable manner for non-linear separable categorization. After this, image classification is performed on the basis of linear SVM segregation.

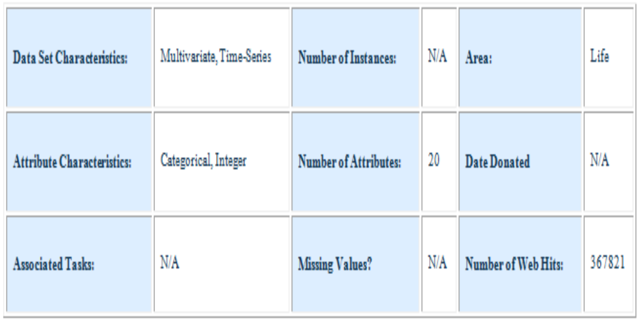
3.2. K-Nearest Neighbor:- This classification model is used for the classification of x pattern. Every pattern is assigned a class label. In a picture, the class label is signified commonly. In this algorithm, the patterns are assigned within the k nearest patterns. The class with minimum standard distance is given the test pattern when a tie occurs amid two patterns. Therefore, this method is based on distance function as well. The minimal average distance is measured using Euclidean distance. This algorithm performs normalization of whole features in the same range. A conventional non- parametric classifier called k-nearest neighbor computes the good performance of best values of k.

3.3. Decision Tree: - The decision tree classifier uses a layered or hierarchical method for performing classification. At each level of the accessible tree, the properties of a dimension are matched to the amount of jointly exclusive nodes. Merely leaf nodes can assign classes to the measurement purpose. The measurements are classified by including the test sequences. This phenomenon reduces the interpretation of every succeeding test. The tests’ sequences are driven for the classifier in a training period.

Result and Discussion

The dataset is collected from the UCI repository. The three scenarios are implemented for the diabetic prediction. The first scenario is based on the SVM classifier, the second scenario is based on KNN classifier, the third scenario is based on the random forest. The dataset is collected from the UCI repository which is described in table 1. The performance of these three classifiers are compared in terms of recall, precision, accuracy and execution time

Table 1: Dataset Description



Data Set Characteristics:	Multivariate, Time-Series	**Number of Instances:**	N/A	**Area:**	Life
Attribute Characteristics:	Categorical, Integer	**Number of Attributes:**	20	**Date Donated**	N/A
Associated Tasks:	N/A	**Missing Values?**	N/A	**Number of Web Hits:**	367821

The parameters for the performance analysis are described below:-

1. Precision: The ratio of number of relevant samples among the total retrieved samples is known as precision.

Precision= (True Positive) / (True Positive + False Positive)

2. Recall: The ratio of relevant samples retrieved to the total number of relevant samples is called recall.

Recall= (True Positive) / (True Positive + False Negative)

3. Accuracy: The ratio of total number of points that are classified correctly to the total number of points multiplied by 100 is called accuracy.

$$\text{Accuracy} = \frac{\text{Number of points correctly classified}}{\text{Total Number of points}} * 100$$

4. *Execution Time:* Execution time is defined as difference of end time when algorithm stops performing and start time when algorithm starts performing

Execution time = End time of algorithm- start of the algorithm

Table 2: Performance Analysis

Parameters	SVM	KNN	Decision Tree
Precision	84.56 percent	78.67 percent	89.89 percent
Recall	86.78 percent	79.67 percent	91.90 percent
Accuracy	88.67 percent	76.90 percent	92.34 percent
Execution Time	2.45 second	3.4 second	1.78 second

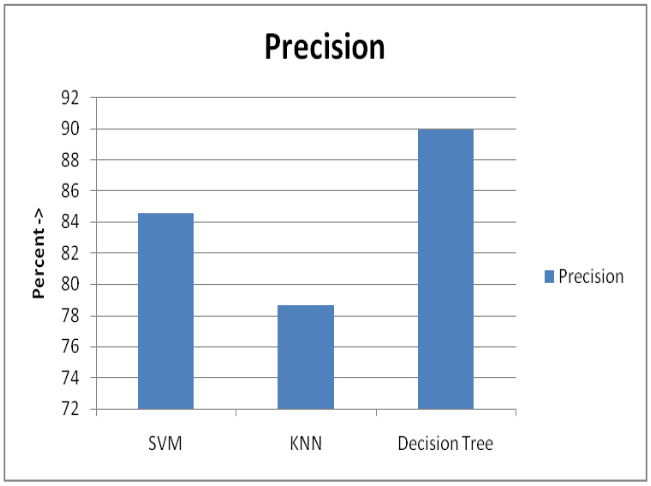


Fig.1: Precision Analysis

As shown in figure 1, the precision value of three classifiers which are SVM, KNN and decision tree is analyzed for the diabetic prediction. It is analyzed that decision tree classifier has the maximum value as compared to SVM and KNN.

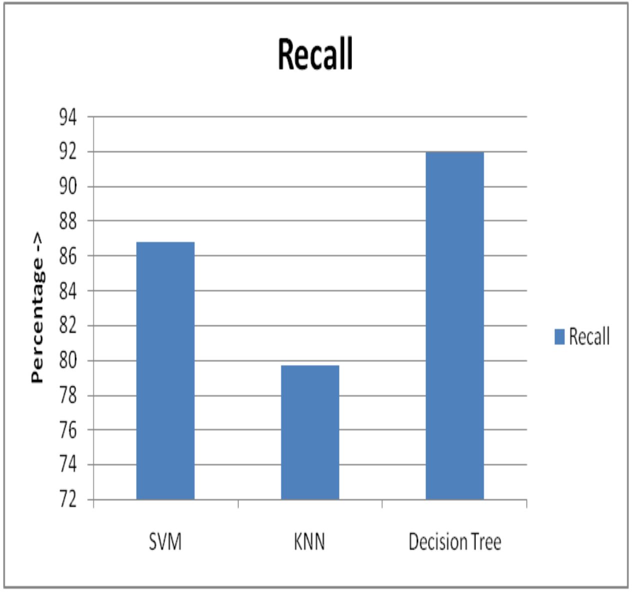


Fig.2: Recall Analysis

As shown in figure 1, the recall value of three classifiers which are SVM, KNN and decision tree is analyzed for the diabetic prediction. It is analyzed that decision tree classifier has the maximum value as compared to SVM and KNN.

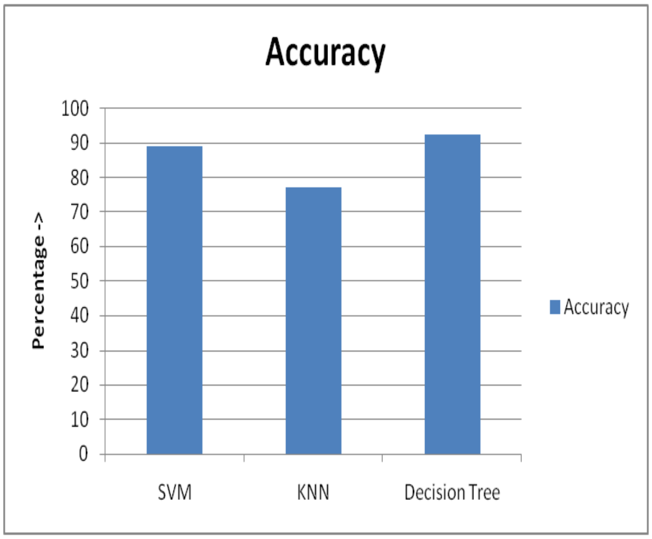


Fig.3: Accuracy Analysis

As shown in figure 3, the accuracy value of three classifiers which are SVM, KNN and decision tree is analyzed for the diabetic prediction. It is analyzed that decision tree classifier has the maximum value as compared to SVM and KNN.

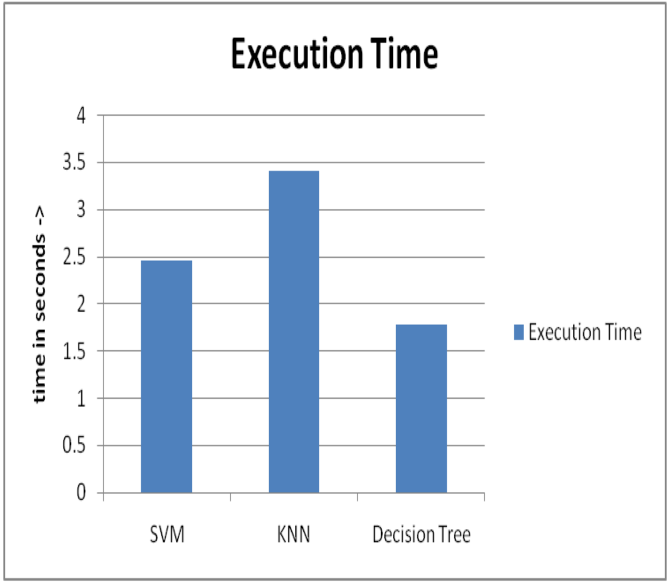


Fig.4: Execution Time

As shown in figure 4, the execution time of the SVM, KNN and decision tree is compared for the performance analysis. It is analyzed that decision tree has least execution time as compared to SVM and KNN for the diabetic prediction.

IV. CONCLUSION

In this paper, it is concluded that diabetic prediction is the major challenge of prediction analysis. The diabetic prediction has the three steps which are pre-processing, feature extraction and classification. The data set is collected from the UCI repository and in the feature extraction phase relationship gets established between attribute and target set. In the last phase, to predict target set of test set SVM, KNN and decision tree classifiers are applied for the prediction. In future hybrid classification model will be designed for the diabetic prediction.

V. REFERENCES

[1]. V. Veena Vijayan, C. Anjali, “Prediction and diagnosis of diabetes mellitus- A machine learning approach”, IEEE Recent Advances in Intelligent Computational Systems (RAICS), 2015

[2]. K Sowjanya, Ayush Singhal, Chaitali Choudhary, “MobDBTest: A machine learning based system for predicting diabetes risk using mobile devices”, IEEE International Advance Computing Conference (IACC), 2015

[3]. Quan Zou, Kaiyang Qu, Yamei Luo3, Dehui Yin, Ying Ju and Hua Tang, “Predicting Diabetes Mellitus With Machine Learning Techniques”, Frontiers in Genetics, November 2018, Volume 9

[4]. Tejas N. Joshi, Prof. Pramila M. Chawan, “Diabetes Prediction Using Machine Learning Techniques”, Int. Journal of Engineering Research and Application, Vol. 8, Issue 1, (Part -II) January 2018, pp.-09-13

[5]. Priyanka Indoria, Yogesh Kumar Rathore, “A Survey: Detection and Prediction of Diabetes Using Machine Learning Techniques”, International Journal of Engineering Research & Technology (IJERT), Vol. 7 Issue 03, March-2018

[6]. Rahul Joshi, Minyechil Alehegn, “Analysis and prediction of diabetes diseases using machine learning algorithm: Ensemble approach”, International Research Journal of Engineering and Technology (IRJET), Volume: 04 Issue: 10 | Oct -2017

[7]. Uswa Ali Zia, Dr. Naeem Khan, “Predicting Diabetes in Medical Datasets Using Machine Learning Techniques”, International Journal of Scientific & Engineering Research Volume 8, Issue 5, May-2017

[8]. Minyechil Alehegn, Rahul Joshi& Dr. Preeti Mulay, “Analysis and Prediction of Diabetes Mellitus using Machine Learning Algorithm”, International Journal of Pure and Applied Mathematics, Volume 118 No. 9 2018, 871-878

[9]. Hasan Abbas, Lejla Alic, Marelyn Rios, Muhammad Abdul-Ghani, Khalid Qaraqe, “Predicting Diabetes in Healthy Population through Machine Learning”, IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS), Year: 2019 | Conference Paper | Publisher: IEEE

[10]. Debadri Dutta, Debpriyo Paul, Parthajeet Ghosh, “Analysing Feature Importances for Diabetes Prediction using Machine Learning”, IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), Year: 2018 | Conference Paper | Publisher: IEEE

[11]. Priyanka Sonar, K. JayaMalini, “Diabetes Prediction Using Different Machine Learning Approaches”, 3rd International Conference on Computing Methodologies and Communication (ICCMC), Year: 2019 | Conference Paper | Publisher: IEEE

[12]. Ayman Mir, Sudhir N. Dhage, “Diabetes Disease Prediction Using Machine Learning on Big Data of Healthcare”, Fourth International Conference on Computing Communication

Control and Automation (ICCUBEA), Year: 2018 | Conference Paper | Publisher: IEEE

[13]. Aparimita Swain, Sachi Nandan Mohanty, Ananta Chandra Das, "Comparative risk analysis on prediction of Diabetes Mellitus using machine learning approach", International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT), Year: 2016 | Conference Paper | Publisher: IEEE

[14]. Muhammad Azeem Sarwar, Nasir Kamal, Wajeeha Hamid, Munam Ali Shah, "Prediction of Diabetes Using Machine Learning Algorithms in Healthcare", 24th International Conference on Automation and Computing (ICAC), Year: 2018 | Conference Paper | Publisher: IEEE

[15]. Md. Faisal Faruque, Asaduzzaman, Iqbal H. Sarker, "Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus", International Conference on Electrical, Computer and Communication Engineering (ECCE), Year: 2019 | Conference Paper | Publisher: IEEE

[16]. V. Swathi Lakshmi, V. Nithya, K. Sripriya, C. Preethi, K. Logeshwari, "Prediction of Diabetes Patient Stage Using Ontology Based Machine Learning System", IEEE International Conference on System, Computation, Automation and Networking (ICSCAN), Year: 2019 | Conference Paper | Publisher: IEEE