Analysis of Diabetes mellitus using Machine Learning Techniques

Salliah Shafi Bhat[1]
B. S. AbdurRahman Crescent Institute of Science and Technology,Chennai-48, India.
Salliahshafi678@gmail.com

Venkatesan Selvam [1]
B. S. AbdurRahman Crescent Institute of Science and Technology,Chennai-48, India.
dean_scis@crescent.education

Gufran Ahmad Ansari [2]
Faculty of Science, MIT World Peace University (MITWPU) Pune- 411 038, India.
gufran.ansari@mitwpu.edu.in

Mohd Dilshad Ansari[3]
Dept. of Computer Science & Engineering, Guru Nanak University, Hyderabad, India
m.dilshadcse@gmail.com

2022 5th International Conference on Multimedia, Signal Processing and Communication Technologies (IMPACT) | 978-1-6654-7647-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/IMPACT55510.2022.10029058

Abstract— **Diabetes Mellitus (DM) often known as hyperglycemia is caused by high blood sugar levels. Although DM is a metabolic chronic disease. Treatment and early detection are essential to reducing the risk of serious outcomes. The World Health Organization (WHO) reports that diabetes has a significant mortality rate causing 1.5 million deaths worldwide. The disease can be identified early because to technology tremendous improvements. In order to build a model with a few variables based on the PIMA dataset this research focuses on evaluating diabetes patients as well as diabetes diagnosis using various Machine Learning Techniques (MLT). Exploratory data analysis is the first step in our process after which the information is transferred for data pre-processing and feature selection. The relevant features are chosen and the data is then training and testing using three different MLT such as Support Vector Machine (SVC), Random Forest (RF) and K-Nearest Neighbors (KNN). Amongst all of the classifiers Random Forest has the highest accuracy of 97.75% followed by Support Vector Machine (82.25%) and K-Nearest Neighbors (86.25%).**

Keywords: **Diabetes Mellitus, Feature Selection, Machine Learning Techniques, Support Vector Machine, Random Forest, K-Nearest Neighbors**

1. Introduction

Diabetes often known as diabetes mellitus is a metabolic disorder (DM) is defined by a high blood sugar level that is either caused by insufficient insulin production [1]. Diabetes mellitus is one of the most frequent chronic diseases in the world. A metabolic disorder called diabetes is defined by hyperglycemia or the abnormal rise in glucose levels. Hyperglycemia or the abnormal growth in glucose levels is the feature of the metabolic condition known as diabetes [2]. Hyperglycemia elevated blood sugar levels can seriously harm the heart rate and have a negative impact on vital organs such as the eyesight, lungs and brain [3]. DM is now considered a non-communicable disease and has become a highly serious condition in the majority of developing world. Statistics show that 425 million people had diabetes diagnoses in 2017 and that two to five million people die from the disorder each year [4]. According to the National Diabetes Statistics Report 2020 type-1 and type-2 diabetes now affect one in ten Americans and the prevalence of both diseases has rapidly grown among young people. Using the abilities of techniques and technologies like artificial intelligence, machine learning etc. is essential because the medical sector health care system is a crucial pillar of society [5]. According to the International Diabetes Federation 6.6 percent of the adult population or 382 million individuals had diabetes mellitus in 2013. The population is projected to increase to 490 million by 2030. The medical area was significantly impacted by advances in technology. Health implications for those who can't go to a clinic or get emergency care could be identified in a few seconds. All those who gain from technology can overcome the limitations of money and time. Regarding resources, the healthcare systems acquire a lot of data in the form of databases with both structured and unorganized data further the diabetes is defined into different types.

Prediabetes: It refers to a condition in which the blood sugar level rises from normal to a critical level but is insufficient for a diagnosis.it can cause heart attack. Lifestyle and maintaining a healthy weight can aid in reducing future health risks.

Type 1 diabetes (T1D): It is a diabetic that needs treatment. The type of diabetes develops when an organism immune system kills the cells that produce insulin. The illness can occur at any age although children and adolescents are the ones most often infected.

Type 2 diabetes (T2D): Although it is frequently seen in adults this kind of diabetes has become more prevalent in the younger generation over the past 20 years due to being overweight or obese. This type of diabetes develops when the immune system fights the production of the hormone insulin causing a rapid rise in blood sugar levels. Reduce the risk factors by maintaining a healthy weight, eating habits, exercising each day and using the right medicine.

Gestational Diabetic: When certain hormones in the body do not even create enough insulin during pregnancy the mother's blood sugar level rises. This type of diabetes is frequently discovered in the middle or final phase of pregnancy [6-7]. The child is at greater risk from this type of diabetes than the mother. Appropriate medication and sufficient nutrition can aid in lowering the health risks. Furthermore diabetes is a possibly separate risk factor for micro-vascular entanglements. Due to the increased risk of micro vascular damage that diabetic patients face, long-term complications from cardio-vascular disease are the main cause of death in this population. As a result the micro vascular damage and rapid cardiovascular disease, retinopathy, kidneys and neuropathy eventually develop. The diseases can be controlled and human lives can be saved with early disease diagnosis. This research mainly analyses the early diagnosis of diabetes mellitus by taking into consideration a variety of risk factors connected to this illness. For this purpose we collected data from UCI data repository having 9 attributes and 2000 instances. These attributes are BMI, Diabetes Pedigree Function, Age, Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin and outcome. Several attributes and their corresponding values are discussed in the next section. Based on these attributes author build a prediction model using several machine learning techniques depending on these features to predict diabetes mellitus. Using PIDD dataset MLT rapidly extract knowledge by building predictive models. To forecast diabetic individuals information obtained from data will be useful. It is possible to forecast diabetes mellitus using a variety of MLT. By selecting the best method to predict DM based on these variables is really difficult. Furthermore, we use three well-

978-1-6654-7647-8/22/$31.00 ©2022 IEEE

known MLT in our research such as SVC, KNN and RF on data set to predict diabetes mellitus.

The main objectives of this paper are

- Three Machine Learning Algorithms were used viz: RF, SVC, and KNN.
- Author Proposed Methodology for Diabetes Mellitus using Machine Learning Techniques.
- Feature selection has been done on two methods i) information method ii) Traditional linear correlation.
- We analyses each algorithm's performance in terms of accuracy, recall, precision, and Fi score.

The remaining sections of this research study are organized as: Section 2 includes the Literature Review of previous work. Section 3 introduces the Proposed Methodology for Diabetes Mellitus using Machine Learning Techniques. Moreover the experimental result is used in section 4. Section 5 concludes by outlining the conclusion and Future work.

2. Literature Review

Every patient has a unique set of issues and risk factors for diabetes. The use of MLT to predict various diseases has grown in popularity in recent years. Numerous software tools and algorithms have been created by researchers. These factors have shown tremendous potential in the medical care sector. PIMA Indians Diabetes Dataset (PIDD) a well-known dataset from the "Kaggle" repository is used by a number of researchers [8]. The Waikato (WEKA) tool was programmed with 78.42% accuracy using the PIDD Dataset and two MLT such as the modified K-means method and the logistic regression algorithm. They concentrated on the primary difficulties of the classification such as accuracy of the prediction model and generating the adaptability for two or more sets of data at the same time [9]. When k was set to 0 a model using the Pima Indians Dataset was able to obtain a maximum receiver operating characteristic accuracy of 74%. This was achieved using the k-nearest neighbor (KNN) technique with a range of k-values between 1 and 100[10]. A method for diagnosing diabetes using machine learning algorithms including Decision Tree, KNN, Naive Bayes and SVM was proposed by MiteshWarke et al. [11]. Additionally, a database from the Pima India Diabetes dataset was used to evaluate the effectiveness of the developed model. In comparison to the Decision Tree algorithm's accuracy of 68%, the SVM algorithm accuracy of 62%, and the KNN algorithm accuracy of 66%, the suggested method for records categorization using NB attained an accuracy of 72%. The system that used SVM, Naive Bayes, KNN and C4.5 algorithms for Performance Analysis of ML methods for predicting DM. The recommended method for classifying records with NB had an accuracy of 68%, whereas the C4.5 algorithm had an accuracy of 73%, the SVM algorithm had an accuracy of 70%, and the KNN methodology had an accuracy of 71%. Using machine learning techniques, the performance analysis of RF, KNN, C5.0, and SVM to predict diabetes the performance of the developed model was evaluated using a database from the Pima India Diabetes dataset. The accuracy of the suggested method for classifying records using KNN was 73.57%, whereas that of the Random Forest algorithm was 74.67%, the C5.0 algorithm was 74.63%, and the SVM was 72.17% [12]. Proposed system that predicts diabetes mellitus using the C4.5 algorithm. The effectiveness of the established model is accessible through a database from the Pima India Diabetes dataset. The accuracy of this algorithm is 72.08 percent [13].Author suggested a method to categories whether a diabetes diagnostic test result was positive or negative utilizing a variety of DM techniques including NB, Sequential Minimal Optimization (SMO) and Simple Logistic Regression. The efficacy of the developed model is available through a database from the Pima India Diabetes dataset. These methods provide accuracy for Naive Bayes of 73.60%, whereas they provide accuracy for Simple Logistic Regression of 75.70% an accuracy for Rep Tree of 75.10%, and an efficiency for Sequential Minimal Optimization of 74%. (SMO). A system that predicts diabetes using the Multilayer Perception, Neural Network model has been proposed. Utilizing a database from the Pima India Diabetes dataset the performance of the created model was examined. 82% accuracy is provided by this model [14]. According to the literature review each of these researchers looked into certain methods and improved them to their best ability. The key objectives of this develop model that can accurately classify and predict data diabetes mellitus using Machine Learning Techniques and various feature selection strategies [15-20].

3. Proposed Methodology for Diabetes Mellitus using Machine Learning Techniques

The Proposed methodology in this paper has been summarized in the diagram shown in Figure 1.

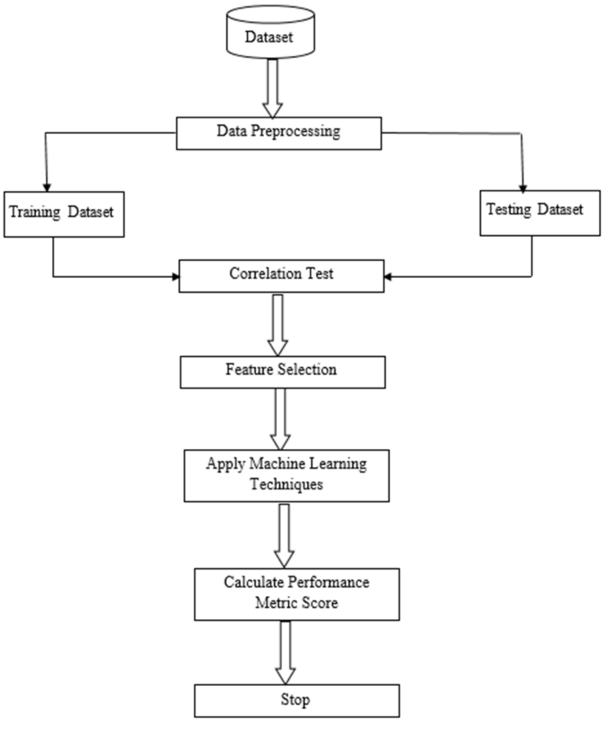


Figure 1 Diagnostic Analysis of Diabetes Mellitus using Machine Learning Techniques

Data Set: To identify diabetes mellitus the PIDD data is gathered from the Kaggle website. The data set consist of 9 attributes and 2000 instances. With the help of the nine features this classification algorithm determines whether a patient has diabetes or not. Figure 2 displays cases of diabetes and healthy patients where the diabetic Patients are 960 and the healthy patients are 1040.

978-1-6654-7647-8/22/$31.00 ©2022 IEEE

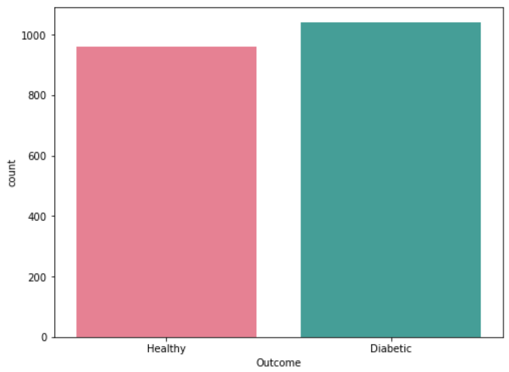


Figure 2: Total no of Diabetic and Non Diabetic Patient Distribution

Data Pre-processing: Data pre-processing is a crucial step that makes the data better and improves the extraction of useful information. Data pre-processing is a critical initial step that is given to raw data to get them ready for the analysis. If there are impurities in the data such as missing data, analytical tools could produce inaccurate results. Therefore in pre-processing the data is essential before beginning the analysis of the data.

Split the data set: The performance of the techniques is evaluated using a test dataset in order to choose the best classifier for predicting diabetes mellitus. The dataset is separated into two parts after pre-processing: 1) Training dataset 2) Testing dataset. In the entire dataset 80% of the data are utilized for training and 20% are used for testing.

Applying Correlation: One of most popular and essential method used by researchers is correlation. Finding a dynamic link between two variables in a dataset is useful. This relationship shows whether the factors are positively or negatively related to one another. This method also works when there is no relationship at all. Figure 3 displays the correlation test for this dataset using Pearson correlation [15]. If one variable is impacted by another we can primarily determine the link between the two variables.

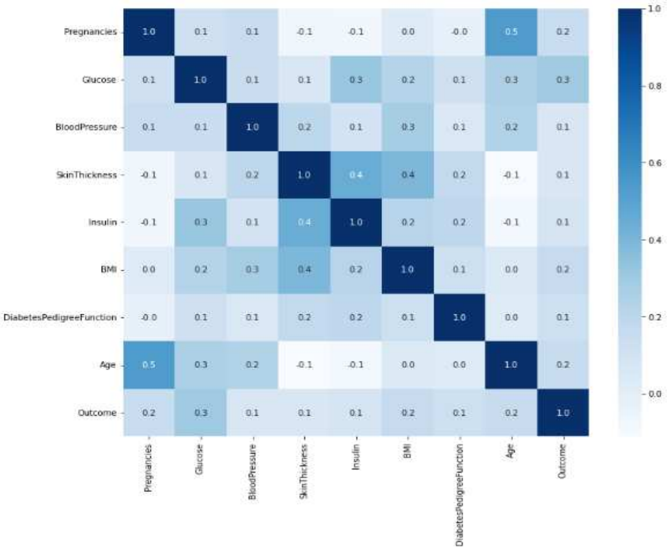


Figure 3 displays the correlation test for this dataset using Pearson's correlation.

The dataset has six significant predictive variables, as seen in the correlation matrix plot i.e. Insulin, DiabetesPedigreeFunction, Pregnancies, Blood Pressure, Skin Thickness, and Age. Figure 4 shows the relationship between all the variables.

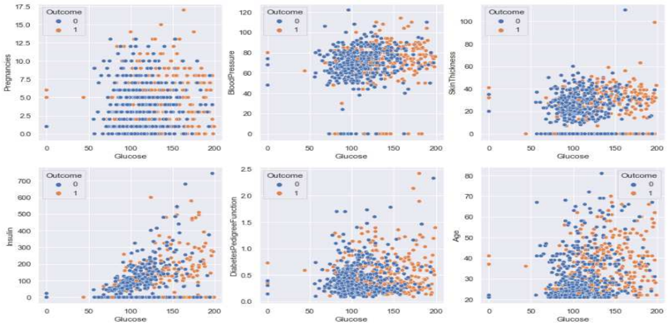


Figure 4 shows the relationship between all the important variables.

Feature Selection: By using correlation feature selection assesses performance. The correlation between the same collections of features can be measured using two different sorts of methods. Information theory and traditional linear correlation are the two methods. Since it defines and measures the random components of features. The linear correlation technique is primarily taken into account in our feature selection. Simulated correlation coefficient (a, b), as determined by Eq. (1)

$$Sin(a,b) = \frac{Cov(a,b)}{\sqrt{Var(a)Var(b)}} \qquad (1)$$

Where cov (a, b) is the covariance between and var () is the feature variance where a is the input the result. The values of sim (a, b) range from -1 to 1. When they are entirely connected, the sim (a, b) value takes values from 1 or -1 and if independent, the sim (a, b) value takes the value 0. Diabetes Pedigree Function and Pregnancies Glucose, BMI, Age are the chosen features in accordance with the correlation mapped with the dataset purpose.

4. Apply Machine Learning Techniques: We have utilized three MLT for the building model namely SVM, RF and KNN once the data is prepared for modelling. As a result we provide an overview of various techniques.

Support Vector Machine (SVC): SVM is one of the classification techniques used in regression and analysis to identify patterns in data. It divides data by finding the most appropriate hyperplane that separates all data points from one class to other. The data is converted using a method called Kernel Trick using SVM. Data conversion is used to determine the best splitting line among the expected outcomes. The margin can range from a simple narrow margin for binary classes to a more difficult separation involving multiple classes [16]. First take a look at a two-dimensional space initially. Two-dimensional data can be separated using a linear equation of a line in which the data points along either side of the line indicate the corresponding classes. Then the Eqn of the line is

$$y = az + by \qquad (2)$$

Considering y and z as features z_1, z_2..., z_n it can be written as

$$az1 - z2 + b = 0 \qquad (3)$$

If we define z (z1 and z2) and w (a,-1) we get

$$w.z + b = 0 \qquad (4)$$

Random Forest (RF): it is a variant of the Decision Tree (DT) algorithm. The variance component of the model is reduced by averaging decision trees which have high variance and low bias. It is possible to generate the unknown samples by averaging the prediction.

$$k = \frac{1}{m}\sum_{m=1}^{m} \text{f(x)} \qquad (5)$$

Where uncertainty is

$$\alpha = \frac{\sqrt{\sum_{m=1}^{m}(f(x) - f)2}}{\text{m} - 1} \qquad (6)$$

978-1-6654-7647-8/22/$31.00 ©2022 IEEE

The Random Forest (RF) algorithm applies different decision trees to data collecting predictions from each of them and determining the optimum way to proceed. Additionally it is built on an ensemble learning method that uses a bagging algorithm and can handle data with missing values [17].

K-Nearest Neighbors (KNN): The Learning division uses the K Nearest Neighbor algorithm for sorting and regression. It is a flexible approach that is used to assign missing values and resample datasets. The analysis uses K Nearest Neighbors as the title suggests to predict the class or related value for the original Data point. Equation 7 shows

$$\sqrt{(y_1 - y_2)^2 + (z_1 - z_2)^2} \qquad (7)$$

Calculate Performance metric Score: In this section we define all the measures used in our experiments. Such factors are to be analyzed by classifying the algorithms based on their performance. Equation 8-11 displays the performance measurements which are accuracy, precision, recall, and fi score.

Accuracy (Acc): Accuracy provides performance and observing the facts of the appropriately predicted Classifier and expresses as

$$Acc = \frac{TP + TN}{TP + FP + TN + FN} \qquad (8)$$

Precision (Pr): Precision is stated as the ratio of expected positive results to the total target positive outcomes as follows:

$$Pr = \frac{TP}{TP + FP} \qquad (9)$$

Recall (Re): The definition of recall is as follows where FN stands for false negative rate**.**

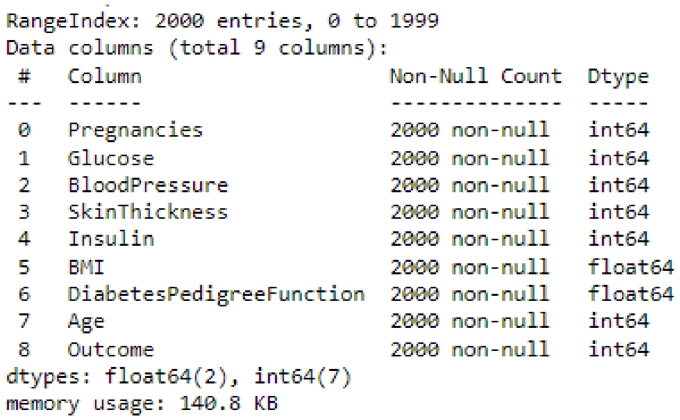
$$Re = \frac{TP}{TP + FN} \qquad (10)$$

Fi score (Fi): The precision and recall harmonic mean is known as the F1-Score, and the range is [0, 1].The mathematical formulation is as follows, and the F1-Score suggests the classifier reliability.

$$Fi = 2^* = \frac{1}{\frac{1}{Precision} + \frac{1}{Recall}} \qquad (11)$$

5. Experimental Results

This section describes and shows the expected outcome from the experimental study that used MLT to predict diabetes mellitus.PIDD data set is used to from the UC Irvine ML collection to determine if the candidate patient has diabetes or not. Data set has been splitted into 80% of the data are utilized for training and 20% are used for testing. MLbased models were used to predict the occurrence of diabetes mellitus [18]. The diagnostic dataset contained 2000 records and nine parameters eight of which were predicate qualities and one of which outcome was a target variable. Figure 5 displays the dataset information which comprises the name of the attribute, non-null, count, and data type [19]. The performance of the techniques is evaluated after applying various ML classification algorithms to the PIDD dataset. Table 1 shows the Performance of various MLT.



```
RangeIndex: 2000 entries, 0 to 1999
Data columns (total 9 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Pregnancies               2000 non-null   int64
 1   Glucose                   2000 non-null   int64
 2   BloodPressure             2000 non-null   int64
 3   SkinThickness             2000 non-null   int64
 4   Insulin                   2000 non-null   int64
 5   BMI                       2000 non-null   float64
 6   DiabetesPedigreeFunction  2000 non-null   float64
 7   Age                       2000 non-null   int64
 8   Outcome                   2000 non-null   int64
dtypes: float64(2), int64(7)
memory usage: 140.8 KB
```

Figure 5 Dataset parameters information

According to the experimental results Random Forest achieved the best accuracy (97.75%), K Nearest Neighbor (86.25%) and Support Vector Machine achieved accuracy (82.25%). The ROC curves for ML classifiers are receiver operating characteristic curves are shown in Figures 6-8 and they are created by situating the true positive rate (TPR) in relation to the false positive rate (FPR) at various thresholds. The covered surface area under the curve (AUC) values is calculated for curves.

Table 1Performance Indicators of Machine Learning Techniques

Algorithms	Accuracy (%)	Precision	Recall	Fi Score
SVC	82.25	0.84	0.91	0.87
RF	97.75	0.97	0.99	0.98
KNN	86.25	0.89	0.91	0.90

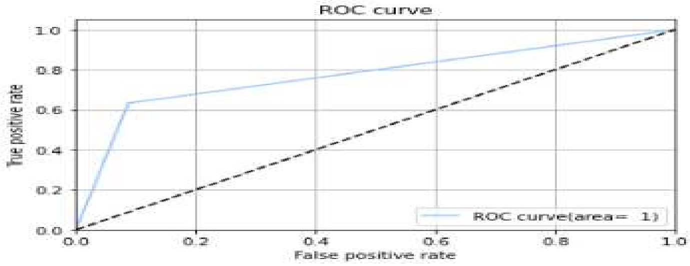


Figure 6 ROC Curve for Support Vector Classification

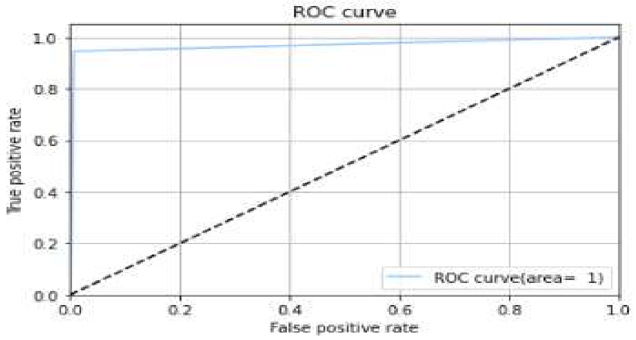


Figure 7 ROC Curve for Random Forest.

978-1-6654-7647-8/22/$31.00 ©2022 IEEE

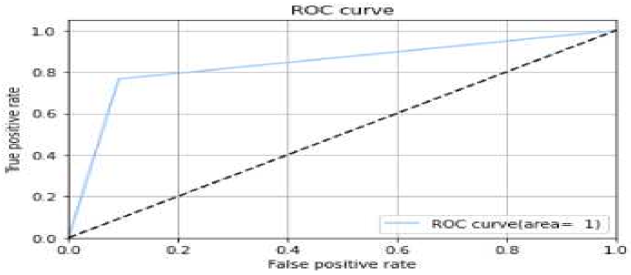


Figure 8 ROC Curve for K Nearest Neighbor.

6. Conclusion and Future work

In the traditional healthcare system doctors used diagnostic tests to assess diabetes patients. The diagnostic test technique in the healthcare system involves large financial expenditures and delays the patient's ability to identify diabetes. As a result our work suggests the ability to diagnose and forecast diabetes. The PIDD dataset is trained and tested using a variety of MLT in this research. The obtained experimental results indicate that RF achieved the best accuracy (97.75%), KNN (86.25%) and SVM achieved accuracy of (82.25%). using the diabetes dataset for PIDD is presented and implemented. using the diabetes dataset for PIDD is presented and implemented in Proposed Methodology for Diabetes Mellitus using MLT. Various Performance metric Score have been calculated to verify the diabetes prediction such as Accuracy, Precision, Recall and Fi score. A framework employed in this research will be used to further the current research. MLT must be employed for ensemble and hybridization for better diabetes mellitus forecast. In the future work is to use a large dataset and also apply the machine learning for better performance.

References

[1]. Rahman, S. F. A., Rafiee, N. S. M., Yaakob, A. M., Shafie, S., & Ibrahim, I. (2022, August). Multiclass classification scheme for diagnosis of diabetes mellitus based on type-1 fuzzy systems. In *AIP Conference Proceedings* (Vol. 2472, No. 1, p. 030001). AIP Publishing LLC.

[2]. Rahman, M. F., Sharma, G. K., & Kishore, K. Diabetes Mellitus: A Review of Current sTrends. *Journal homepage: www. Ijrpr. Com ISSN*, *2582*, 7421.

[3]. Unal, D., Onbaşı, K., & Kilit, T. P. (2022). Osteoporosis and Related Factors in Patient with Type 2 Diabetes and Prediabetes. *Turk J Osteoporos*, *28*, 97-103.

[4]. Miriyala, N. P., Kottapalli, R. L., Miriyala, G. P., Lorenzini, G., Ganteda, C., & Bhogapurapu, V. A. (2022). Diagnostic Analysis of Diabetes Mellitus Using Machine Learning Approach. *Revue intelligence Artificielle*, *36*(3), 347-352.

[5]. Manne, R., & Kantheti, S. C. (2021). Application of artificial intelligence in healthcare: chances and challenges. *Current Journal of Applied Science and Technology*, *40*(6), 78-89.

[6]. Kampmann, U., Knorr, S., Fuglsang, J., & Ovesen, P. (2019). Determinants of maternal insulin resistance during pregnancy: an updated overview. *Journal of diabetes research*, *2019*.

[7]. Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science*, *2*(3), 1-21.

[8]. Suyanto, S., Meliana, S., Wahyuningrum, T., & Khomsah, S. (2022). A new nearest neighbor-based framework for diabetes detection. *Expert Systems with Applications*, *199*, 116857.

[9]. Bundasak, S. (2020). Student Behaviours Analysis Affecting Learning Achievement of Information Technology and Computer Science Students. *International Journal of Machine Learning and Computing*, *10*(2).

[10]. Abdulhadi, N., & Al-Mousa, A. (2021, July). Diabetes detection using machine learning classification methods. In *2021 International Conference on Information Technology (ICIT)* (pp. 350-354). IEEE.

[11]. Warke, M., Kumar, V., Tarale, S., Galgat, P., & Chaudhari, D. J. (2019). Diabetes diagnosis using machine learning algorithms. *Diabetes*, *6*(03), 1470-1476.

[12]. Gunman, V. K., Kumar, S., Ansari, M. D., & Vijayalata, Y. (2022). Prediction of Agriculture Yields Using Machine Learning Algorithms. In Proceedings of the 2nd International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications (pp. 17-26). Springer, Singapore.

[13]. Agarwal, M., Bohat, V. K., Ansari, M. D., Sinha, A., Gupta, S. K., & Garg, D. (2019, December). A convolution neural network based approach to detect the disease in corn crop. In 2019 IEEE 9th international conference on advanced computing (IACC) (pp. 176-181). IEEE.

[14]. Gupta, H., Varshney, H., Sharma, T. K., Pachauri, N., & Verma, O. P. (2021). Comparative performance analysis of quantum machine learning with deep learning for diabetes prediction. *Complex & Intelligent Systems*, 1-15.

[15]. Wan, Z., Wang, Q. D., Wang, B. Y., & Liang, J. (2022). Development of machine learning models for the prediction of laminar flame speeds of hydrocarbon and oxygenated fuels. *Fuel Communications*, *12*, 100071.

[16]. Nilashi, M., bin Ibrahim, O., Ahmadi, H., & Shahmoradi, L. (2017). An analytical method for diseases prediction using machine learning techniques. *Computers & Chemical Engineering*, *106*, 212-223.

[17]. Shaikhina, T., Lowe, D., Daga, S., Briggs, D., Higgins, R., & Khovanova, N. (2019). Decision tree and random forest models for outcome prediction in antibody incompatible kidney transplantation. *Biomedical Signal Processing and Control*, *52*, 456-462.

[18]. Gaddam, D. K. R., Ansari, M. D., Vuppala, S., Gunjan, V. K., & Sati, M. M. (2022). A Performance Comparison of Optimization Algorithms on a Generated Dataset. In ICDSMLA 2020 (pp. 1407-1415). Springer, Singapore.

[19]. Bhat, S. S., Selvam, V., Ansari, G. A., Ansari, M. D., & Rahman, M. H. (2022). Prevalence and Early Prediction of Diabetes Using Machine Learning in North Kashmir: A Case Study of District Bandipora. *Computational Intelligence and Neuroscience*, *2022*.

[20]. Gaddam, D. K. R., Ansari, M. D., Vuppala, S., Gunjan, V. K., & Sati, M. M. (2022). Human facial emotion detection using deep learning. In ICDSMLA 2020 (pp. 1417-1427). Springer, Singapore.