AN EFFICIENT DIABETES MELLITUS PREDICTION USING ENTROPY-BASED DLMNN CLASSIFIER

Asma Ahmad Abokhzam , N. K. Gupta, Sam Higginbottom University of Agriculture Technology and Science - Pryagraj , Shuats 211007 Uttar Pradesh India.

Abstract—

The pancreas is one of the vital organs of a normal human as it manufactures or generates insulin that has a direct effect on the amount of glucose in the blood which in turn causes a chronic disease called Diabetes Mellitus. It is the most occurring harmful and dangerous disorder affecting an enormous number of people regularly. It is very difficult to prevent and control, because of its asymptomatic nature. The cause of this harmful disease is still unknown. It might be obesity or lack of exercise or the glucose level in human blood. The diagnosis of this disease during the initial phase is significant so that it could be controlled effectively. This chronic disease could also result in a poor quality of life as well as the high cost of medicines for treatment and an increase in mortality rate. People with this disease are usually young. As the number of patients increases, it results in overloading patients in hospitals. A person could be affected by this chronic disease for a long period without knowing the fact that they are suffering from diabetes which would cause difficulties in the future. Thus, in this paper, an effective method for predicting diabetes mellitus is proposed by using the DLMNN Algorithm. In this paper, a dataset named “Pima Indians Diabetes Database” is used. Further steps including preprocessing, feature selection, instance evaluation, and disease prediction was also done. The DLMNN Algorithm, Artificial Neural Network Algorithm, and Adam Optimizer are used in the disease prediction phase. In the preprocessing phase, the problem of missing data imputation present in previous studies was solved, and further are as follows: sort instances by age, removal of duplication, missing data imputation, and normalization. The maximum accuracy achieved by using the entropy-based DLMNN classifier is almost 79%, which is quite higher when compared to previous studies.

Index Terms—Diabetes Mellitus, DLMNN, Feature Selection, Classification, Disease Prediction.

INTRODUCTION

DIABETES Mellitus can be called as a most dangerous chronic disorder, the causes and cure of which is still unknown except in certain situations when the glucose’s amount in human blood remains under observation regularly and the amount of glucose in human blood are retained near the normal level of glucose in the blood without resulting in hypoglycemia. But on the other hand, it can be controlled by a certain course of actions like, by taking proper diet, by doing proper exercise regularly or by using certain specified medications [1]. Diabetes is common among the 382 million human beings in the entire world and it will be twice at the end of the year 2035 as said or illustrated by the International Diabetes Federation [2].

Diabetes mellitus is also dependent upon some other health factors like obesity so it is very difficult for a medical person to predict diabetes mellitus at a very early stage. It is also harmfulto other organs like kidney, foot, or eyes [3]. It is a disease which could usually occur in well-developed as well as underdeveloped countries and states. More than almost 20M population which also consists of adults as well as children had been diagnosed with diabetes in the year 2007 in the United States [9]. The expected number of patients having diabetes in India will most probably rise from 30 to 80 million in the year 2000 to the year 2030 and 80% will be the people from well-developed countries in the year 2030 [10]. The number of people who are infected or diagnosed with such diseases that are chronic in the whole world is increasing day by day dangerously [4]. Almost 70% of total deaths well as account 78% of total expenses for healthcare in the U.S are caused by such chronic diseases [5]. This chronic disease such as diabetes mellitus is highly linked with other health issues like cardiovascular, any issues related to kidney, non-traumatic amputations as well as some vision-related issues [6]. Diabetes mellitus is causing morbidity as well as mortality at a greater rate. Almost 9% of humans in the United States are detected with diabetes (any kind) which is also increasing [7], [8]. This chronic disease can also cause a poor quality of life as well as a high cost for medicines for treatment and an increase in the rate of mortality. People having this disease are usually young. As the number of patients increases, it results in the overloading of patients in hospitals also. A person can have this chronic disease from a long period without knowing the fact that he or she is suffering from diabetes which can cause difficulties in the future. Almost 29.1 million people in America in 2014, there were 29.1 million Americans are recorded which also consists of almost 8.1 million people who did not know about their disease [11].

Continuous as well as through checkup of few important symptoms is a difficult and important step as this can prevent the happening of such chronic diseases like diabetes mellitus. But it is very crucial for a person having diabetes to stick to the process of management of their self. As we know that a lot of people do not even check their level of glucose in blood regularly so it will be very difficult for them to diagnose diabetes mellitus [12]. Only 50% to 70% of the people of America received the eye examination recommendation with the fact that they also have diabetes so that vision loss can be prevented as discussed in the survey [13]. Patients having diabetes mellitus and people who are not suffering from such a disease also required to self-monitor and self-control themselves as suggested by chronic disease management [7].

The pancreas is one of the vital organs of a normal human as it manufactures or generates insulin that has a direct effect on the amount of glucose in the blood which in turn causes this chronic disease. It is becoming the most occurring harmful as well as a dangerous disorder that is affecting a greater number of people daily [14]. It is very difficult to prevent as well as to control this dangerous disease as it is a silent disease. As the cause of this harmful disease is still unknown, it can be the obesity of lack of exercise or it can be associated with the glucose's amount in human blood. That is why the diagnosis of any disease during the initial phase is significant so that it can be controlled effectively. It is one of the most harmful diseases which is affecting human life badly and damage the human organs like kidney etc. [15]. Causes

of this chronic disease i.e. diabetes mellitus are still unknown, it can be said that it is caused due to certain other health conditions like kidney problems, lack of exercise, etc. The data management about such disease and the record of a patient's history can act as an important knowledge in controlling and preventing such diseases. With the smart analysis method's rise [16], using intelligence to diagnose some medical problem is an unmatched important concern [17].

The next parts of the research are ordered in the following way. The literature review is elaborated in Section II. Section III describes the adopted methodology, basic steps of the proposed method for prediction, and information about the dataset. Section IV describes the results which were obtained using the proposed methodology for the prediction of diabetes mellitus. Research is discussed in Section V whereas research is concluded in Section VI.

RELATED WORK

The diabetes mellitus prediction methods can be divided into two according to the recent studies i.e, by using a machine learning algorithm and by using data mining. The data mining technique can be elaborated in a way that a procedure of introducing correlations, trends as well as patterns for searching through a huge number of data saved in repositories, data warehouses as well as databases [18]. To get information about the diabetes data which is present in a large amount, algorithms of machine learning as well as the data mining technique is playing a significant role in research about diabetes mellitus. Diabetes mellitus is becoming an important issue worldwide due to its significant effect on human life as well as a growing number of patients which in turn producing a large amount of patient data. Due to which, the algorithms of machine learning as well as the data mining technique is playing a major role in managing the data of patients [19]. Specifically, data mining [20] and algorithms of machine learning reached the high strength as well as the high capability to handle a large number of data and to make better and best predictions [21]. It is proved by researchers the algorithms of machine learning [22] like algorithms for classification like Bayes net, decision tree, naive Bayes, support vector, etc. works best for predicting such cases. The method of combing such algorithms for increasing the accuracy of the prediction of diabetes is also adopted [23]. Instead of the fact of advanced research, diabetes is too strenuous to detect and treat it in very early stages [24].

In literature, different studies for predicting diabetes mellitus have been proposed. A data mining-based technique or scheme was introduced by Han Wu et al. [25] for predicting type 2 diabetes. In the first step of this methodology, preprocessing was done on the dataset (Pima Indians Diabetes Dataset) which consists of the following steps: 1) Applying K- Means (improved) Algorithm 2) Applying logistic regression. The results were compared with previous works by using an analysis toolkit named Waikato Environment for Knowledge Analysis toolkit. After which it was shown that the proposed model or scheme is meant for managing diabetes.

N. Sneha and Tarun Gangil [26] presented an early diabetes mellitus system using optimal feature selection. This approach was introduced for making use of important features, finding the best classifier for giving true results as well as for making a best ML algorithm for

prediction. The best specificity for analyzing the data of diabetes was the result of using random forest (98.00) and decision tree (98.20). Whereas, the best accuracy was achieved by using Naive Bayes which is 82.30%. P. C. Sherimon and Reshmy Krishnan [27] focused on the use of Web Ontology Language 2 for modeling and implementing the clinical basic guidelines. This system evaluates score and based on that score, it predicts how much risky smoking, cardiovascular, alcohol, and other diseases are like all of these causes this chronic disease significantly. J. Jayashree and S. Ananda Kumar [28] introduced the new technique named as "swarm smart redundancy relevance (RR)" for detecting or predicting diabetes as well as NN (neural network) having a compositional design which was trained with convolution. This technique consists of different stages, one of them constitutes the following: acquiring of data of diabetes, removing noise as well as any inconsistency, and selection of features for getting the best results. In the next stage, the appertaining of approach named "CTCPNN". After this step, the performance was calculated or evaluated by observing the results or outcomes of the experimental procedure as well as discussions. But it needs more improvement as this scheme only includes a few basic steps like selection features, preprocessing, or classification.

N. Yuvaraj and K. R. SriPreethaa [29] proposed the novel implementation algorithms of ML in clusters depending upon Hadoop for classification. Originally the features were selected from the available dataset the dataset contains a large number of attributes about the patients. Next, the Information Gain (IG) was used for measures information in bits about class prediction. After that, predicting the diabetes patient using machine learning algorithms. The best outcomes will be generated using algorithms of machine learning (ML) as represented in outcomes. Changsheng Zhu et al. [30] represents a classification model that is made using concepts of data mining which to preprocess as well as classification as major steps by making use of the same dataset i.e. PIMA. By amalgamating the concepts of K-mean, PCA as well as logistic regression. Then for transforming the early features, PCA was used due to which the issue of correlation was solved that was making classification difficult. In literature, there are few problems which have been solved in this study is given as follows:

1) In the Existing preprocessing step, the missing value was replaced by the mean value. So, all the missing values of a single attribute have the same value and the classification accuracy becomes low, 2) In Existing work, the classification was done with the duplicate attribute. So, the classification accuracy becomes low, 3) In Existing work, the classification was done with a miscellaneous way. Because k- means algorithm may wrongly classify the data due to distance metrics. So, the correctly classified data was removed from the dataset. So, the training accuracy becomes low, 4) In Existing Neural Network algorithm takes more training time due to random weight generation and back propagation technique. So, either the classification accuracy or the training accuracy is low. Similarly, in some studies, the training time is very high. In this research, a new efficient method for predicting diabetes mellitus was proposed by using an entropy-based DLMNN classifier. In this research, an entropy for improving the novelty of the prediction of disease is introduced. In this new scheme, we have used the Diabetes Disease Dataset. And also, the above-mentioned problem of data manipulation was overcome in the proposed scheme of

diabetes prediction.

PROPOSED METHODOLOGY

Fig 1 presents the working model or the block diagram of our proposed methodology for the prediction of diabetes. The methodology consists up of data collection, preprocessing, feature selection, instances evaluation, and disease prediction. Each step of the proposed methodology is described below:

Data Collection

A dataset named "Pima Indians Diabetes Database" is used in the proposed scheme to predict diabetes mellitus. The disease data of Diabetes was collected from the National Institute of Diabetes and Digestive and Kidney Diseases. This dataset consists of more than 700 instances and 9 attributes with class. The attributes are, number of times the person is pregnant, Body-Mass-Index, Skinfold thickness, Insulin, Diastolic blood pressure, Class, Glucose, Age as well as Diabetes pedigree function. In this 80% of the data was given to the first phase which is training whereas in the second phase which is testing only 20% was given.

Preprocessing

Preprocessing is the main step before starting any further procedure. The main purpose of this phase is increasing classification accuracy and decreasing training time. The preprocessing phase consists of further four steps which are as follows: Sort instances by age, Removal of duplication, Missing data imputation, and Normalization.

Sort instances by age: The instance is sorted according to the age in this step. The main purpose of this step in the preprocessing phase is the gradual training of the dataset.

Removal of duplication: In the second step of the preprocessing phase, the duplicate data was removed. The performance was faster since the previous step was done.

Missing data imputation: This is the third step of the preprocessing phase, the missing data imputation is performed by replacing the missing values with values of comparison of relevant features such as age, glucose, blood pressure, etc. of the instances. In such a way, the problem of missing data imputation present in the literature discussed was overcome in the previous section.

Normalization: Normalization is the last step of the pre-processing phase. In this step, the values calculated or measured on a distinct scale was adjusted to a hypothetically common scale, commonly before averaging.

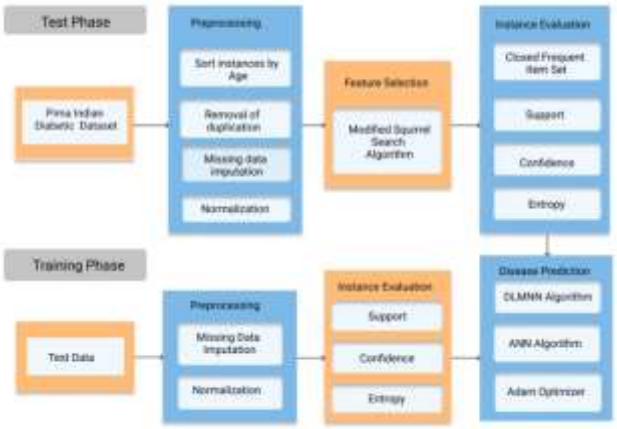

Fig. 1. The proposed methodology for predicting diabetes mellitus using entropy-based DLMNN Classifier

Feature Selection

After the preprocessing phase, the next phase is the feature selection. In this phase, the features are selected by using the Modified Squirrel Algorithm whereas the attribute of the used dataset was contemplated as features for selecting the features that are least associated or connected among each other whereas highly associated with the class labels. By using the Genetic Algorithm, the modification was done in the Modified Squirrel Algorithm. A genetic algorithm can be called as one of the most-newest algorithms used for selecting the desired features. This algorithm does optimization based on natural genetics as well as evolution. It consists of the following steps: initialization, fitness assignment, selection, crossover, and mutation. In the table below, the attributes of the dataset or features were discussed.

S.No	FEATURES
1	age
2	Diastolic_{bp}
3	BMI
4	Insulin
5	Target
6	glucose_{con}
7	Diab_{pred}
8	Thickness
9	num_{preg}

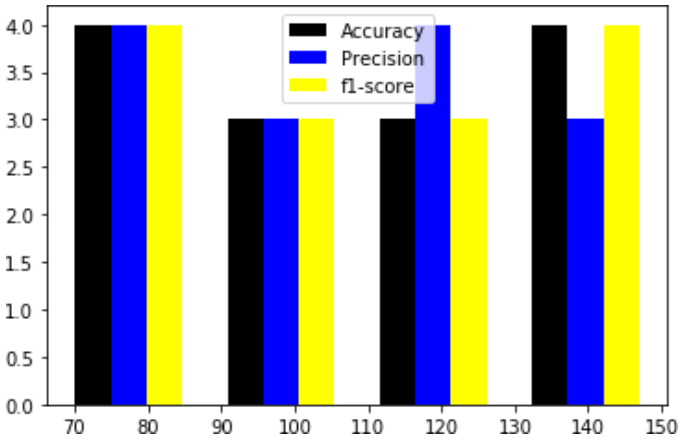

Fig. 2. Performance of the proposed methodology or scheme in terms of accuracy, kappa statics, and F-measure

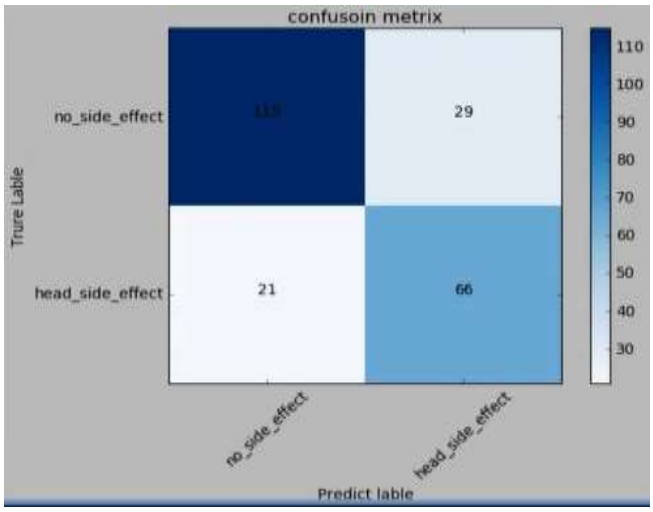

Fig. 3. Confusion matrix obtained as a result of using entropy-based DLMNN

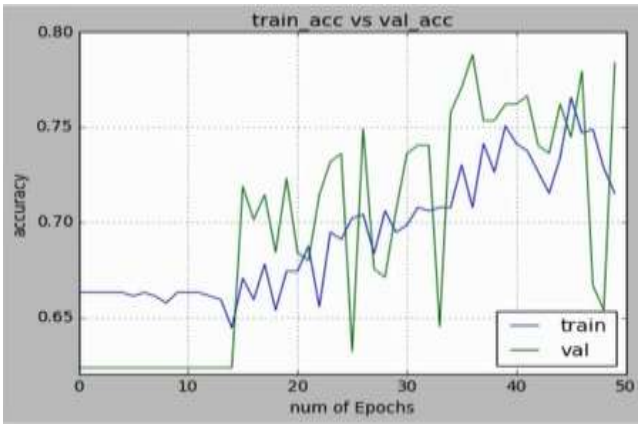

Fig. 4. Entropy-based DLMNN model classification accuracy visualization

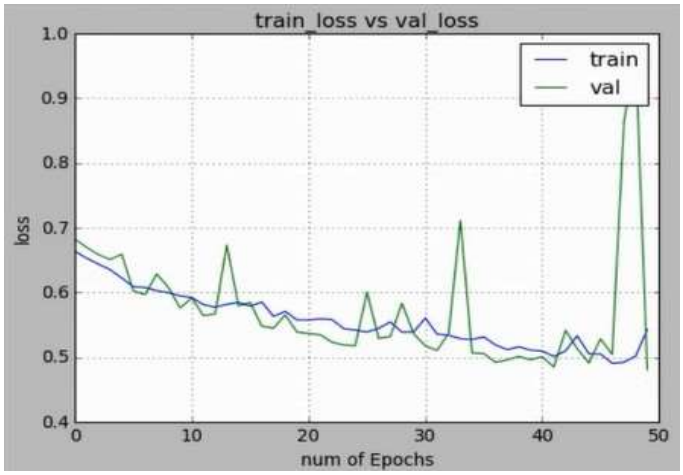

Fig. 5. DEntropy based LMNN model loss visualization

DISCUSSION

The proposed methodology obtained the classification accuracy is quite high when compared to all the existing researches. Figure 6, represents the classification accuracy achieved using the ANN model. In this case, the maximum accuracy achieved is almost 75% which was also high. In this case, as the number of Epochs increases, the classification accuracy achieved was also increased gradually. Figure 6 also shows the comparison of train-acc (training accuracy) and val-acc.

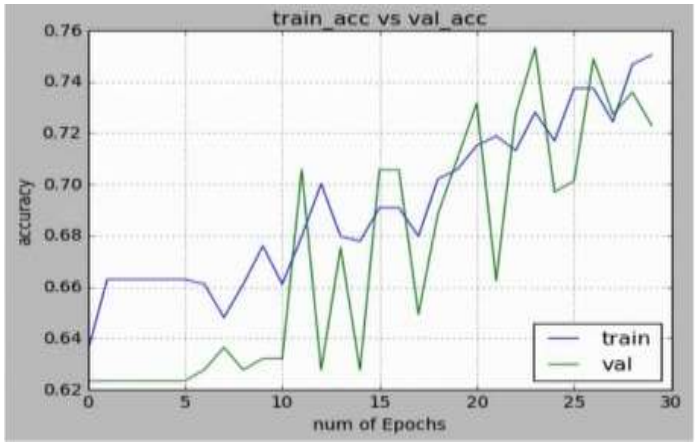

Fig. 6. Artificial Neural Network Model (ANN) accuracy visualization

Figure 7, represents the loss achieved using the ANN model. In this case, the maximum loss achieved is almost 0.6. Loss decreases with a rise in the number of Epochs. Figure 7 also shows the comparison of train-loss (training loss) and val-acc. As discussed earlier the best results were obtained by using the DLMNN, as the maximum accuracy achieved is almost 79% and the maximum loss obtained in this case is 0.1.

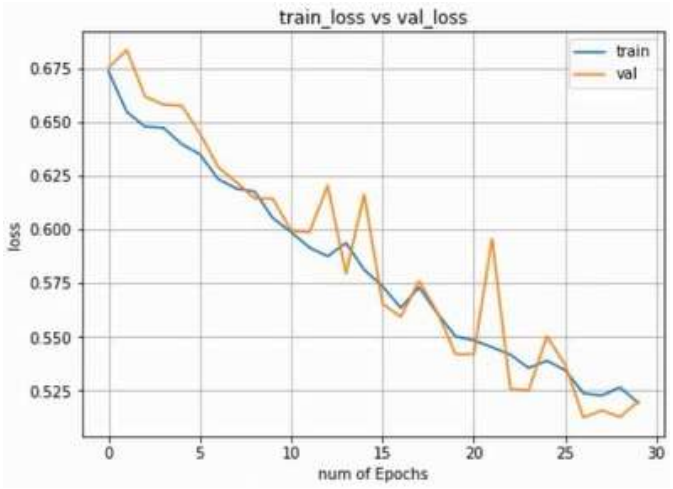

Fig. 7. Artificial Neural Network Model (ANN) loss obtained visualization

The table given presents a comparison of the performance of the proposed methodology with state-of-art methods. It shows that the accuracy of our proposed scheme is comparable to the previous studies. Also, the problem of missing data imputation present in previous studies was solved in this paper.

S.No	Model, Year	Accuracy	Classification Algorithm /Technique
1	[26], 2019	82.30%	Na¨ıve Bayesian
2	[27], 2016	-	Web Ontology Language 2
3	[28], 2019	-	Compositional pattern neural network
4	[29], 2017	-	Machine Learning Algorithms
5	[30], 2019	-	Data Mining
6	Proposed, 2020	79%	Entropy-Based DLMNN Classifier

CONCLUSION

In the proposed scheme, we have developed a system for the prediction of diabetes mellitus using an entropy-based DLMNN classifier. The dataset named “Pima Indians Diabetes Database” was used here. After getting the dataset, further steps in the preprocessing phase were done, such as sort instances by age, removal of duplication, missing data imputation, and normalization. After preprocessing, feature selection has been done for getting the best results by using the Modified Squirrel Algorithm. After selecting features, instances evaluation was done, and lastly, the

prediction of diabetes using Artificial Neural Network Algorithm, DLMNN Classifier, and Adam Optimizer was conducted. The maximum accuracy achieved is by using an entropy-based DLMNN classifier that is almost 79% by using the entropy-based DLMNN Classifier which is most suitable as compared to the schemes or systems proposed in the past. In addition, the problem of missing data imputation which was present in previous studies has been resolved.
We implemented the proposed scheme by using python the results prove that our proposed method is better than many existing techniques. Python is a common-purpose, interactive as well as high-level computer language for programming that can be used for generating an application whether a desktop or a web. It can also be used for making a GUI of any application and it can be compiled in Jupyter, SKULPT, Nutika, WinPython, Spyder, etc.

REFERENCES

1. Haldurai Lingaraj, Rajmohan Devadass, Vidya Gopi, and Kaliraj Palanisamy, "Prediction of diabetes mellitus using data mining techniques: a review", Journal of Bioinformatics Cheminformatics, vol.1, no. 1, pp. 1-3, 2015.
2. Messan Komi, Jun Li, Yongxin Zhai, and Xianguo Zhang, "Application of data mining methods in diabetes prediction", In 2017 2nd International Conference on Image, Vision and Computing (ICIVC), pp. 1006- 1010. IEEE, 2017.
3. J. Steff, Dr.R. Balasubramanian, and Mr.K. Aravind Kumar, "Predicting Diabetes Mellitus using Data Mining Techniques", International Journal of Engineering Development and Research, Vol. 6, pp. 2321-9939, 2018.
4. H. N. A. Pham and E. Triantaphyllou, "Prediction of diabetes by employing a new data mining approach which balances fitting and generalization," Computer and Information Science, vol. 131. Berlin, Germany: Springer, 2008, pp. 11–26.
5. S. Wild, G. Roglic, A. Green, R. Sicree, and H. King, "Global prevalence of diabetes: Estimates for the year 2000 and projections for 2030,"
6. Diabetes Care, vol. 27, no. 5, pp. 1047–1053, 2004
7. T. Szydło and M. Konieczny, "Mobile and wearable devices in an open and universal system for remote patient monitoring," Microprocessors Microsyst., vol. 46, pp. 44–54, Oct. 2016
8. G. Anderson and J. Horvath, "The growing burden of chronic disease in America," Public Health Rep., vol. 119, no. 21, pp. 263–270, 2004
9. J. Hartz, L. Yingling, and T. M. Powell-Wiley, "Use of mobile health technology in the prevention and management of diabetes mellitus," Current Cardiol. Rep., vol. 18, p. 130, Dec. 2016.
10. M. Hood, R. Wilson, J. Corsica, L. Bradley, D. Chirinos, and A. Vivo, "What do we know about mobile applications for diabetes self-management? A review of reviews," J. Behav. Med., vol. 39, pp. 981–994, Dec. 2016.
11. Center for Disease Control. (2017). National Diabetes Statistics Report. [Online].

Available: https://www.cdc.gov/diabetes/pdfs/data/ statistics/national-diabetes-statistics-report.pdf

12. Centers for Disease Control and Prevention, "Estimates of diabetes and its burden in the united states," National Diabetes Statistics Report, 2014. [Online]. Available: https://www.cdc.gov/diabetes/pdfs/data/2014- report-estimates-of-diabetes-and-its-burden-in-the-united-states.pdf
13. C. E. Koro, S. J. Bowlin, N. Bourgeois, and D. O. Fedder, "Glycemic control from 1988 to 2000 among U.S. adults diagnosed with type 2 diabetes: A preliminary report," Diabetes Care, vol. 27, no. 1, pp. 17–20, 2004.
14. X. Zhang et al., "Eye care in the United States: Do we deliver to high- risk people who can benefit most from it?" Arch Ophthalmol, vol. 125, no. 3, pp. 411–418, 2007.
15. P. C. Thirumal and N. Nagarajan, "Utilization of data mining techniques for diagnosis of diabetes mellitus—A case study," ARPN J. Eng. Appl. Sci., vol. 10, no. 1, pp. 8–13, 2015.
16. American Diabetes Association, "Diagnosis and classification of diabetes mellitus," Diabetes Care, vol. 33, pp. S62–S69, Jan. 2010.
17. X. Wang, D. Bi, and S. Wang, "Fault recognition with labeled multi-category support vector machine," in Proc. IEEE 3rd Int. Conf. Natural Comput. (ICNC), Aug. 2007, pp. 567–571.
18. B. Zhang, Z. Wei, J. Ren, Y. Cheng, and Z. Zheng, "An empirical study on predicting blood pressure using classification and regression trees," IEEE Access, vol. 6, pp. 21758–21768, 2018.
19. Vrushali Balpande, and Rakhi Wajgi, "Review on Prediction of Diabetes using Data Mining Technique", International Journal of Research and Scientific Innovation (IJRSI), no. 4, pp. 43-46, 2017.
20. Ioannis Kavakiotis, Olga Tsave, Athanasios Salifoglou, Nicos Maglaveras, Ioannis Vlahavas, and Ioanna Chouvarda, "Machine learning and data mining methods in diabetes research", Computational and structural biotechnology journal, no.15, pp. 104-116, 2017.
21. P. S. Kumar and V. Umatejaswi, "Diagnosing diabetes using data mining techniques," Int. J. Sci. Res. Publications, vol. 7, pp. 705–709, Jun. 2017.
    I. Kavakiotis, O. Tsave, and A. Salifoglou, "Machine learning and data mining methods in diabetes research," Comput. Struct. Biotechnol. J., vol. 15, no. 9, pp. 104–116, 2017.
22. M. Fatima and M. Pasha, "Survey of machine learning algorithms for disease diagnostic," J. Intell. Learn. Syst. Appl., vol. 9, no. 1, pp. 1–16, 2017.
23. R. Joshi and M. Alehegn, "Analysis and prediction of diabetes diseases using machine learning algorithm: Ensemble approach," Int. Res. J. Eng. Technol., vol. 4, no. 10, pp. 426–435, 2017.
24. S. Ravizza, T. Huschto, A. Adamov, L. Bo¨hm, A. Bu¨sser, F. F. Flo¨ther, R. Hinzmann,

H. Ko¨nig, S. M. McAhren, D. H. Robertson, T. Schleyer, B. Schneidinger, and W. Petrich, “Predicting the early risk of chronic kidney disease in patients with diabetes using real-world data,” Nature Med., vol. 25, pp. 57–59, Jan. 2019.

25. Han Wu, Shengqi Yang, Zhangqin Huang, Jian He, and Xiaoyi Wang, “Type 2 diabetes mellitus prediction model based on data mining”, Informatics in Medicine Unlocked, no. 10, pp. 100-107, 2018.
26. N. Sneha, and Tarun Gangil, “Analysis of diabetes mellitus for early prediction using optimal features selection”, Journal of Big Data, vol. 6, no. 1, 2019.
27. P. C. Sherimon, and Reshmy Krishnan, “OntoDiabetic: an ontology-based clinical decision support system for diabetic patients”, Arabian Journal for Science and Engineering, vol. 41, no. 3, pp. 1145-1160, 2016.
28. J. Jayashree, and S. Ananda Kumar, “Hybrid swarm intelligent redundancy relevance (RR) with convolution trained compositional pattern neural network expert system for diagnosis of diabetes”, Health and Technology, pp. 1-10, 2019.
29. N. Yuvaraj, and K. R. SriPreethaa, “Diabetes prediction in healthcare systems using machine learning algorithms on Hadoop cluster”, Cluster Computing, pp. 1-9, 2017.
30. Changsheng Zhu, Christian Uwa Idemudia, and Wenfang Feng, “Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques”, Informatics in Medicine Unlocked, vol. 100179, 2019.