Review on Diabetes Disease Using Data Mining Classification Techniques

Sweta[1], Pushpa Devi[2], Dr. Kishori Lal Bansal[3]
[1, 2] Dept of Computer Science
[3]Professor, Dept of Computer Science
[1, 2, 3] Himachal Pradesh University, Summerhill

Abstract- *Data mining is the process of extracting useful information from large datasets. Data mining has various types of techniques like classification, clustering, and association. Classification is one of the most popular techniques used in the medical field. Nowadays in the healthcare sector, a large amount of data is being generated every day. Data mining plays an important role to handle this huge amount of data and providing an accurate prediction of diseases. Diabetes is one of the most common diseases that affect millions of people all over the world. So, it is important to identify and detect diabetes at its early stage. This paper presents a detailed review and comparative study of existing data mining techniques used in the early prediction of diabetes disease. The comparative study of various existing classification techniques helps in deciding which technique is best for diabetes prediction purposes in the future. On comparing the accuracy of all the papers, the one technique which has outperformed all others in most of the studies taken into consideration is KNN with accuracy (99.0%).*

Keywords- Data Mining, Machine Learning, Classification, Diabetes Mellitus

I. INTRODUCTION

Data mining is a field of computer science used to find hidden patterns from large databases. The main aim of the data mining process is to extract useful information from a large amount of data and convert it into an understandable structure for future use[1]. Data mining also known as Knowledge discovery in the database is the process of finding useful information and other valuable knowledge from large datasets. In organizations and companies, data mining is used for transforming their raw data into useful knowledge. Today, the technology is advanced so million of data is being generated in the world. This data cannot be handled manually. So, there is a need for automatic data analysis that analyzes the data and extracts useful information. Data mining plays an important role in research, healthcare, banking, and bioinformatics. Diabetes Mellitus is one of the major health problems that affect millions of people all over the world. Diabetes Mellitus is a chronic and one of the dramatically increasing metabolic diseases in the world[2]. Diabetes is a common disease and it happens when the human body is not able to produce insulin properly. It causes very serious health problems like nerve damage, kidney ailment, circulatory strain, blindness and skin infections, etc[3]. There are three types of diabetes type1, type2, and gestational diabetes, which are described below.

Type1

Type1 diabetes occurs when the pancreas secretes a little or zero amount of insulin. It affects 10% of people and generally happens in childhood.

Type2

Type2 diabetes occurs when the pancreas does not secret sufficient insulin as per the human body. It generally affects 90% of people and happens in middle or old-age people.

Gestational

This type of diabetes occurs in women during the time of pregnancy due to high blood pressure and it can affect both mothers and babies[4].

In this paper, we present a detailed review and comparative study of existing data mining techniques used in the early prediction of diabetes. Selected papers published from 2015 to 2021 have been comparatively analyzed and conclusions were drawn. Various data mining classification techniques such as Naive Bayes (NB), Support vector machine (SVM), Random Forest (RF), K-nearest neighbor (KNN), Logistic Regression (LR), etc have been compared. The main aim of this paper is to compare these techniques and to conclude which is the best technique for achieving the highest accuracy in the prediction process. There are VI sections in this research paper. Section II describes data mining classification techniques. Section III describes literature reviews. Section IV gives a comparative study and discussion. Sections V and VI present the conclusion and future scope.

II. CLASSIFICATION TECHNIQUES

The classification technique of data mining is used to classify data in different classes. This technique is used to identify relevant information about data and metadata. The purpose of the classification technique is to predict the class of variable whose class label is known. The classification has two steps i.e Learning step and the classification step. There are various types of data mining classification techniques such as Naïve Bayes, Decision trees, support vector machine, random forest, logistic regression, etc have been applied for predicting diabetes. Some important techniques are discussed here.

(i) Naïve Bayes

Naive Bayes is the most rapid classification algorithm that is based on the Bayes theorem. The Bayes theorem is used to create a probabilistic classifier called Naive Bayes[5]. It is simple and easy to implement or widely used for text classification. It is used to predict the probability of different classes based on various attributes.

(ii)Decision Tree

A decision tree is a supervised learning method that is used for classification and regression. It is a tree-like structure that includes the root node, branches, and leaf nodes, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node holds a class label[6]. A decision tree is one of the most popular and important classification algorithms used in the health sector for predicting disease. The decision tree follows a top-down approach that starts with the root node. It can handle both numeric and categorical data. ID3, C4.5, and CART are the commonly used decision tree algorithms[7].

(iii)Support Vector machine

It is a supervised machine learning model that can be used for both classification and regression[8]. SVM algorithm is used to find the hyperplane in an N-dimensional space to classify the different data points. The dimension depends upon the number of features.

(iv)Random Forest

A random forest is an extension of a decision tree that can be used for classification and regression. It is the type of ensemble technique that combines multiple classifiers to solve a complex problem and improve the accuracy of the system. This algorithm is easy to implement and can manage a large number of input variables without overfitting [9].

(v) Logistic Regression

Logistic regression is a type of classification algorithm used to predict the probability of a target variable. It is a simple and more efficient method used to solve classification problems. It is used to describe the relationship between data. It can be either yes or no, or true or false, 0or1, etc.

III. LITERATURE REVIEW

The literature review has been done by studying various research papers. Many researchers are using classification techniques on the different datasets and analyzing their results. A detailed literature review of existing data mining techniques is given below.

In 2015, Nilesh Jagdish Vispute et al.[10] presented a comparative analysis of various classification techniques such as naïve Bayes, J48, SMO, REP tree, and Random forest by using the WEKA tool. In their work, the author used the diabetes dataset collected from Hospital Repository. After comparing the performances of all algorithms, they found that Naïve Bayes shows the best performing algorithm for classified accuracy of 76.3021%.

In 2015, Sadri Saadi et al.[11] presented a comparative analysis of various data mining algorithms such as Naïve Bayes, RBF network, and J48 for diagnosis of type-2 diabetes by using the WEKA tool. In their study, the author used the Pima Indians diabetes dataset which contains 768 samples and 9 attributes. After comparing the algorithms, they found that Naïve Bayes has the highest accuracy (76.95%).

In 2015, Masoum Farahmandian et al.[12] proposed a model using data mining techniques for diabetes prediction. They used six classification techniques such as SVM, Naïve Bayes, ID3, C4.5, C5.0, and CART. In their study, the author used the Pima Indian Diabetes Dataset. MATLAB is used for implementation purposes. According to their results, the SVM algorithm is more accurate as compared to other algorithms.

In 2015, Dr. D Ashok Kumar et al.[13] proposed a hybrid prediction model for the classification of diabetes patient datasets. They used five different techniques such as SVM, Regression, Bayes Net, Naïve Bayes, and Decision Table. In their study, the dataset is obtained from Pima Indians Diabetes Dataset(PIDD) which is sourced from UCI machine learning

Repository. After comparing the algorithms, they found that the Decision table algorithm outperformed all other techniques with the help of feature selection with an accuracy of 79.81%.

In 2015,Amit kumar Dewangan et al.[14] proposed a hybrid model using classification techniques like Multilayer Perceptron, and Bayesian Net for diabetes prediction. In their study, the author used Pima Indian diabetes data set which is collected from the UCI repository. WEKA tool is used for experiment purposes. After evaluating the algorithms, the author found that the hybrid model MLP and Bayesian network have achieved the highest accuracy of 81.89%.

In 2016, K.Saravananathan et al.[15] discussed the classification techniques such as J48, CART, SVM, and KNN to find the optimal solution for diabetes using the WEKA tool. In their study, the author used a diabetes dataset which is calculated from a Private Medical Diabetic Center. The 10-fold cross-validation method is used to calculate the accuracy. After analyzing the performance of all algorithms, the author found that the performance of J48 is significantly superior to the other three techniques.

In 2016,Sajida Perveen et al.[2]proposed a model using AdaBoost and bagging ensemble techniques such as J48 (c4.5) and standalone J48 for diabetic classification. In their study, the author used the diabetes dataset which is taken from the Canadian Primary Care Sentinel Surveillance Network (CPCSSN) database. For experiment purposes, they used the WEKA tool. After analyzing the performance, they found that the overall performance of the AdaBoost ensemble method is better than bagging as well as a standalone J48 decision tree.

In 2017, Deepika Verma et al.[16] presented a comparative analysis of various classification techniques such as Naïve Bayes, SMO, REP tree, and J48 by using the WEKA tool. The authors used the breast cancer and diabetes dataset which is taken from the UCI Machine Learning Repository. After analyzing the performances of all algorithms, the Authors found that SMO gives 76.80% accuracy on the diabetes dataset.

In 2017, Sonu Bala Garg et al.[17] presented a comparative analysis of different classification techniques such as Naive Bayes, Bayes Net, J48, SMO, and Random Forest by using the WEKA tool. In their study, the author used the Pima Indians Diabetes Database having 768 instances and 9 attributes. After comparing the algorithms, the author found that the best algorithm based on the Cross-validation is the SMO classifier with an accuracy of 77.34%, and the best algorithm based on the Percentage split is the Decision Table classifier with an accuracy of 81.99%.

In 2018, Deepti Sisodia et al.[18] presented the comparative analysis of three classification techniques such as Decision Tree, SVM, and Naive Bayes by using the WEKA tool. They used Pima Indians Diabetes Database (PIDD) which is sourced from the UCI machine learning repository. In their research work, the author used various measures like Precision, Accuracy, F-Measure, and Recall to evaluate the performance of three algorithms. After evaluating the performance, the author found that Naive Bayes outperforms with the highest accuracy of 76.30% compared to other algorithms.

In 2018, Fikirte Girma Woldemichael et al.[19]proposed a model to predict diabetes using four classification techniques like Backpropagation, J48, Naïve Bayes, and support vector machine (SVM) by using RStudio. The author used the PIMA Indian dataset which is taken from the UCI machine learning repository. According to their work, the result showed that the accuracy of the backpropagation algorithm is better than SVM, J48, and Naïve Bayes.

In 2018, Shuja Mirza et al.[20] designed a predictive model using data mining classification techniques such as Naïve Bayes and Support Vector Machine, KNN on the clinical dataset to find an optimum solution for diabetic prognosis. They used the weka tool for experiment purposes. The author found that SVM has the highest accuracy rate, highest TP rate, Precision, and F-measure.

In 2019, Abhilasha Nakra et al.[21]presented the comparative analysis of Bayes Net Classifier, Naïve classifier, and a combination of both classifiers by using Weka. The experiment is performed on the diabetes dataset available in the WEKA tool itself. After comparing algorithms, they found that the combination of Bayes Net and Naive Bayes provides a better result than using these classifiers individually.

In 2019, Sara A. Aboalnaser et al.[22] presented a comprehensive study of Diabetes Miletus prediction using Different Classification Algorithms such as Naïve Bayes, KNN, Artificial Neural Network, Decision Tree, Random Forest, Support Vector Machine, and Logistic Regression by using the orange tool. Pima Indians Diabetes dataset is used in their study. After analyzing performances, they found that KNN outperformed other classifiers, also both random forest and decision tree models did well.

In 2019 Changsheng Zhua et al.[23] proposed a model using data mining techniques for diabetes prediction on Pima Indian diabetes dataset. In their work, the author used PCA for dimensionality reduction, the K-mean algorithm for clustering, and logistic regression for classification, and python language

is used for experiment purposes. After analyzing performance, the author found that PCA enhanced the k-means clustering algorithm and logistic regression classifier accuracy.

In 2020, Kezban Alpan et al.[24] proposed a model to classify the accuracy of seven different algorithms like Bayes Network, Naïve Bayes, J48, Random Tree, Random Forest, K-NN, and SVM by using the weka tool. In their study, the dataset is taken from Sylhet Diabetes hospital which is publicly available on the UCI Machine Learning Repository having 520 instances and 17 attributes. After analyzing the performance, they found that the K-NN has the highest accuracy 98.07%.

In 2020, Subitha Sivakumar et al.[25]proposed a model to predict diabetes in the early stages by using classification algorithms such as Naïve Bayes, KStar, ZeroR, OneR, J48, and Random forest. The author used the diabetes dataset from the UCI Repository contains 768 instances and 9 attributes. In their study, the weka tool is used for experiment purposes. Their experimental results showed that the Naive Bayes has the best accuracy at 76.30%.

In 2020, Safae Sossi Alaoui et al.[26]presented a comparative analysis of various data mining classification techniques such as Naive Bayes, Neural Network, Support Vector Machine, and decision tree for diagnosing diabetes in women. The two different data mining tools WEKA and ORANGE are used for experiment purposes. In their study, the authors used PIDD (Pima Indians diabetes database) dataset from the UC Irvine Machine learning repository having 768 instances and 9 attributes. After comparing algorithms, they found that the Support Vector Machine classifier implemented in WEKA as SMO has the highest accuracy with 76.8%, sensitivity with 77.3%, and precision with 76.9%.

In 2021, Ritu Ratra et al.[27] presented a comparative analysis of different data mining techniques such as Naïve Bayes, SMO, J48, ANN, REP Tree, KNN, and Logistic Regression using the Weka tool. In their research work, the author used the Pima Indians Diabetes data set having 768 Instances and 9 attributes. After analyzing the performance, they found that some of the algorithms show high performance, and some of them show poor performance.

In 2021, Adetunji Olusogo Julius et al. [28]presented a comparative analysis of various classification techniques such as KNN, Support Vector Machine, and Functional Tree by using the WEKA tool. In their study, the author used the diabetes dataset taken from Irvine(UCI)Repository which contains 520 samples and 17 attributes. After comparing the algorithms, they found that the KNN has the highest accuracy, specificity, and precision of 98.08%, 99%, and 99.36% respectively.

In 2021, Raman Kumar Mondal et al.[29] implemented several machine learning classifiers such as Naïve Bayes, Random forest, ANN, KNN, SVM, and LR by using the weka tool. The author used the Pima Indian dataset (PIDD) obtained from UCI Repository having 768 instances and 9 attributes. For better accuracy, they used bootstrapping Resample technique which will increase the accuracy of almost all classifiers but the decision trees lead over others. Their result shows that the ANN classifier performed better with the highest accuracy 94.14%.

IV. COMPARATIVE STUDY AND DISCUSSION

This comparative study presents the complete analysis of what type of methodologies and datasets were used for conducting experiments. The following table1 shows the diabetes prediction accuracy of some selected papers achieved by the different researchers using various data mining classification techniques.

Table1 Comparison of diabetes prediction accuracy of existing Data Mining techniques

Ref.	Year	Dataset	Tool	Techniques	Prediction Results
[11]	2015	PIDD	Weka	NB, RBF Network, J48	NB (76.95%)
[12]	2015	PIDD	Matlab	SVM,KNN,NB,ID3,C4.5,C5.0,CART	SVM (81.77%)
[13]	2015	PIDD	Weka	SVM, Regression, BN,NB,DT	BN(78.25%)without feature selection DT(79.81%)after applying feature selection
[14]	2015	PIDD	Weka	C4.5,RF,MLP,Bayes Network	Hybrid model (MLP+Bayes Net) 81.89%
[15]	2016	Private Medical	Weka	J48,CART,SVM,KNN	J48(67.16)
[17]	2017	PIDD	Weka	DT,NB,BN,J48,MLP,RF,SMO	DT(81.99%)Percentage split, SMO(77.34%)10-cross validation
[18]	2018	PIDD	Weka	NB,SVM,DT	NB(76.30%)
[19]	2018	PIDD	R-Studio	Backpropagation,SVM,J48,NB	Back propagation (83.11%)
[20]	2018	Clinical dataset	Weka	NB,SVM,KNN	SVM(91.35%)
[21]	2019	Dataset available in weka	Weka	BN, NB	BN+NB(81.26%)
[22]	2019	PIDD	Orange	NB,KNN,ANN,DT,RF,SVM, LR	KNN (99.0%)
[23]	2019	PIDD	Python	PCA, k-means, logistic regression	PCA+K-means+LR=97.40% PCA+K-means=79.94%
[24]	2020	(Sylhet Diabetes Hospital)	Weka	BayesNetwork,NB,DT(J48),RT,RF,KNN,SVM	KNN (98.07%)
[26]	2020	PIDD	Weka, Orange	NB,DT,SVM,NN	SMO (76.8%), NN (76.7%)
[27]	2021	PIDD	Weka	NB, SMO,J48,ANN,REPTree, KNN,LR	SMO (77.34%), LR (77.21%)
[28]	2021	UCI	Weka	KNN,SVM,FT,RF	KNN (98.08%)
[29]	2021	PIDD	Weka	NB,RF,,ANN,SVM,LR,KNN	ANN (94.14%)

After a comparative study of diabetes prediction using different approaches of data mining techniques, it has been observed that some of the researchers have designed their methodologies using only individual techniques and some researchers have used ensemble techniques to verify the performance of diabetes prediction. Many researchers used PIMA Indian diabetes dataset and few researchers used the

clinical dataset. This paper shows that the Pima data set is widely used for the experiments but it is limited to only nine attributes. This comparative study reveals that the KNN classifier has the highest accuracy 99.0% on the orange tool as compared to other techniques.

V. CONCLUSION

In this paper, we presented a detailed review and comparative study of existing data mining techniques used in the early prediction of diabetes disease. After reviewing various data mining techniques used in diabetes prediction along with their experimental results some observations are highlighted. The selection of appropriate attributes improves the performance of the model and the accuracy of any model is dependent on the good dataset and algorithms. Data preprocessing also plays an important role in better prediction. The comparative study of various classification techniques helps in deciding which technique is best for diabetes prediction purposes in the future. On comparing the accuracy of all the papers, the one technique which has outperformed all others in most of the studies taken into consideration is KNN with accuracy (99.0%).

VI. FUTURE SCOPE

In the future, we can evaluate the performance of different classification techniques based on different parameters by using different data mining tools and also find the best classification algorithm for the diabetes dataset having a maximum accuracy rate.

REFERENCES

[1] S. Agarwal, "Data mining: Data mining concepts and techniques," *Proc. - 2013 Int. Conf. Mach. Intell. Res. Adv. ICMIRA 2013*, pp. 203–207, 2014, DOI: 10.1109/ICMIRA.2013.45.

[2] S. Perveen, M. Shahbaz, A. Guergachi, and K. Keshavjee, "Performance Analysis of Data Mining Classification Techniques to Predict Diabetes," *Procedia Comput. Sci.*, vol. 82, no. March, pp. 115–121, 2016, DOI: 10.1016/j.procs.2016.04.016.

[3] B. L. Shivakumar and S. Alby, "A survey on data-mining technologies for prediction and diagnosis of diabetes," *Proc. - 2014 Int. Conf. Intell. Comput. Appl. ICICA 2014*, pp. 167–173, 2014, DOI: 10.1109/ICICA.2014.44.

[4] M. Rout and A. Kaur, "Prediction of Diabetes Risk based on Machine Learning Techniques," *Proc. Int. Conf. Intell. Eng. Manag. ICIEM 2020*, pp. 246–251, 2020, DOI: 10.1109/ICIEM48762.2020.9160276.

[5] R. Fitria, D. Yulisda, and M. Ula, "Data Mining Classification Algorithms For Diabetes Dataset Using Weka Tool," pp. 117–124, 2017.

[6] G. G. Gokilam and K. Shanthi, "Performance Analysis of Various Data mining Classification Algorithms on Diabetes Heart dataset," *Compusoft*, vol. 5, no. 3, pp. 2074–2079, 2016.

[7] S. Singh and P. Gupta, "Comparative Study Id3, Cart and C4.5 Decision Tree Algorithm: a Survey," *Int. J. Adv. Inf. Sci. Technol. ISSN*, vol. 27, no. 27, pp. 97–103, 2014.

[8] M. Shuja, S. Mittal, and M. Zaman, "Effective Prediction of Type II Diabetes Mellitus Using Data Mining Classifiers and SMOTE," no. January, pp. 195–211, 2020, DOI: 10.1007/978-981-15-0222-4_17.

[9] P. O. Gislason, J. A. Benediktsson, and J. R. Sveinsson, "Random forests for land cover classification," *Pattern Recognit. Lett.*, vol. 27, no. 4, pp. 294–300, 2006, DOI: 10.1016/j.patrec.2005.08.011.

[10] N. Jagdish, D. Kumar, and A. Rajput, "An Empirical Comparison by Data Mining Classification Techniques for Diabetes Data Set," *Int. J. Comput. Appl.*, vol. 131, no. 2, pp. 6–11, 2015.

[11] S. Sa'di, A. Maleki, R. Hashemi, Z. Panbechi, and K. Chalabi, "Comparison of Data Mining Algorithms in the Diagnosis of Type Ii Diabetes," *Int. J. Comput. Sci. Appl.*, vol. 5, no. 5, pp. 1–12, 2015, DOI: 10.5121/ijcsa.2015.5501.

[12] M. Farahmandian, Y. Lotfi, and I. Maleki, "Data Mining Algorithms Application in Diabetes Diseases Diagnosis: A Case Study," *MAGNT Res. Rep.*, vol. 3, no. 1, pp. 989–997, 2015.

[13] D. A. Kumar and R. Govindasamy, "Performance and Evaluation of Classification Data Mining Techniques in Diabetes," *Int. J. Comput. Sci. Inf. Technol.*, vol. 6, no. 2, pp. 1312–1319, 2015, [Online]. Available: https://pdfs.semanticscholar.org/995c/954dccf57848c7393183687f9e95a24db0c4.pdf

[14] A. Kumar Dewangan and P. Agrawal, "file:///C:/Users/my/Desktop/data mining research paper/WEKA/2016 daibetes weka/15.Analyzing Diabetic Data using Classification algorithms in data mining.pdf," *Int. J. Comput. Eng. Appl. Sci.*, vol. 2, no. 5, pp. 145–148, 2015.

[15] K. Saravananathan and T. Velmurugan, "Analyzing Diabetic Data using Classification Algorithms in Data Mining," *Indian J. Sci. Technol.*, vol. 9, no. 43, 2016, DOI: 10.17485/ijst/2016/v9i43/93874.

[16] D. Verma and N. Mishra, "Analysis and Prediction of Breast cancer and Diabetes disease datasets using Data mining classification Techniques," *2017 Int. Conf. Intell. Sustain. Syst.*, no. Iciss, pp. 533–538, 2017.

[17] S. B. Garga, A. K. Mahajanb, and ..., "An approach for diabetes detection using data mining classification techniques," *Int. J. …*, vol. 6913, no. 63019, pp. 202–218, 2017, [Online]. Available: http://ijoes.vidyapublications.com/paper/Vol26/20-Vol26.pdf

[18] D. Sisodia and D. S. Sisodia, "Prediction of Diabetes using Classification Algorithms," *Procedia Comput. Sci.*, vol. 132, no. Iccids, pp. 1578–1585, 2018, DOI: 10.1016/j.procs.2018.05.122.

[19] F. G. Woldemichael and S. Menaria, "Prediction of Diabetes Using Data Mining Techniques," *Proc. 2nd Int. Conf. Trends Electron. Informatics, ICOEI 2018*, no. Icoei, pp. 414–418, 2018, DOI: 10.1109/ICOEI.2018.8553959.

[20] S. Mirza and S. Mittal, "Design and Implementation of Predictive Model for Prognosis of Diabetes Using Data Mining Techniques," *Int. J. Adv. Res. Comput. Sci.*, vol. 9, no. 2, pp. 393–398, 2018, doi: 10.26483/ijarcs.v9i2.5792.

[21] A. Nakra and M. Duhan, "Comparative Analysis of Bayes Net Classifier, Naive Bayes Classifier and Combination of both Classifiers using WEKA," *Int. J. Inf. Technol. Comput. Sci.*, vol. 11, no. 3, pp. 38–45, 2019, DOI: 10.5815/ijitcs.2019.03.04.

[22] S. A. Aboalnaser and H. R. Almohammadi, "Comprehensive study of diabetes miletus prediction using different classification algorithms," *Proc. - Int. Conf. Dev. eSystems Eng. DeSE*, vol. October-2019, pp. 128–133, 2019, DOI: 10.1109/DeSE.2019.00033.

[23] C. Zhu, C. U. Idemudia, and W. Feng, "Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques," *Informatics Med. Unlocked*, vol. 17, no. March, p. 100179, 2019, DOI: 10.1016/j.imu.2019.100179.

[24] K. Alpan and G. S. Ilgi, "Classification of Diabetes Dataset with Data Mining Techniques by Using WEKA Approach," *4th Int. Symp. Multidiscip. Stud. Innov. Technol. ISMSIT 2020 - Proc.*, 2020, DOI: 10.1109/ISMSIT50672.2020.9254720.

[25] S. Sivakumar, S. Venkataraman, and A. Bwatiramba, "Classification Algorithm in Predicting the Diabetes in Early Stages," *J. Comput. Sci.*, vol. 16, no. 10, pp. 1417–1422, 2020, DOI: 10.3844/jcssp.2020.1417.1422.

[26] S. Sossi Alaoui, B. Aksasse, and Y. Farhaoui, "Data Mining and Machine Learning Approaches and Technologies for Diagnosing Diabetes in Women," *Big data Networks Technol.*, vol. 81, pp. 59–72, 2020, DOI: 10.1007/978-3-030-23672-4_6.

[27] R. Ratra, P. Gulia, and N. S. Gill, "Performance Analysis of Classification Techniques in Data Mining using WEKA," *SSRN Electron. J.*, 2021, DOI: 10.2139/ssrn.3879610.

[28] O. Julius Adetunji, A. Ayeni Olusola, and I. Fasanya Olawale, "Early Diabetic Risk Prediction using Machine Learning Classification Techniques," *Int. J. Innov. Sci. Res. Technol.*, vol. 6, no. 9, pp. 502–507, 2021, [Online]. Available:https://www.researchgate.net/publication/353324389

[29] R. K. Mondal, R. Kumar, N. Mishra, S. Pandey, V. Kumar, and P. Kumar, "Evaluation and Implementation of a Diabetic Prediction Method based on Machine Learning," *Int. J. Adv. Eng. Manag.*, vol. 3, no. 7, pp. 1547–1562, 2021, DOI: 10.35629/5252-030715471562.