Early Prediction of Diabetes Disease Based on Data Mining Techniques

Salma N. Elsadek, Lama S. Alshehri, Rawan A. Alqhatani, Zainah A. Algarni, Linda O. Elbadry, Eyman A. Alyahyan

Early Prediction of Diabetes Disease Based on Data Mining Techniques

Salma N. Elsadek[0000-0003-3670-6104], Lama S. Alshehri, Rawan A. Alqhatani, Zainah A. Algarni, Eyman A. Alyahyan[1] [0000-0002-9272-6129] and Linda O. Elbadry [0000-0003-0704-4164]

Computer Science Department, College of Science and Humanities, Imam Abdulrahman Bin Faisal University, P.O.Box 31961, Jubail, Kingdom of Saudi Arabia
[1]eaalyahyan@iau.edu.sa

Abstract. As a common yet chronic disease**,** Diabetes Mellitus (DM) affects millions of people all over the world. Some groups are more vulnerable to diabetes in comparison to others, such as people with a family history, and those suffering from obesity. Early detection of such people, in conjunction with preventive measures, can go a long way in saving the lives of people who are most likely to be infected with these diseases and avoid suffering. Consequently, Prediction solutions have been found using data mining techniques which help to discover hidden information about the disease and supports decision-making. This study aims to ensure that diabetes is predicated at the initial stages using two algorithms of machine learning: Random Forest and Multi-Layer Perceptron (MLP) using the WEKA environment to estimate the accuracy. The experiments were applied using a dataset obtained from the Machine Learning Repository of UCI. The dataset comprises 16 attributes and 520 instances collected via questionnaires from patients at Sylhet-based Sylhet Diabetes Hospital (Bangladesh). According to the findings, The RF method has been found to yield better results than MLP when it comes to enabling early prediction of diabetes with a high accuracy of 97.88%.

Keywords: Data mining, Diabetes, Random Forest, MLP.

1 Introduction

Diabetes mellitus is one of the most chronic and frequent diseases, affecting422 million people globally; by 2030, over 552 million people around the world are likely to experience diabetes [1]. World Health Organization (WHO), suggests that1.6 million people die due to diabetes annually. It is caused when the pancreas have trouble producing and building glucose in the blood, which leads to an increase in blood sugar concentration [2]. Glucose, which is blood sugar, comes from different types of foods. Insulin t helps the glucose get into the body cells to give them energy. Diabetes is of three types, the most dangerous of which is type1, and can cause serious issues. For example, it can damage kidneys, eyes, and other organs. Type 2, and Gestational Diabetes are not as dangerous as type 1, and can be controlled by organizing meals and making exercises. Diabetes may lead to serious health complications or in extreme cases, lead to death if it is not discovered in a timely manner [3].

Diabetes is the biggest and most important epidemic confronting global economy. Despite allocating large financial and medical resources to confront it in developed countries, decisive economic considerations have not been reached to arrest or reduce its costly complications that are now increasingly affecting the quality of healthcare. For example the United States of America spent 327 billion dollar in diagnosing diabetes in 2017, which includes 237 billion in the form of direct medical costs and reduced productivity worth 90 billion[4] This requires us to adopt a long-term outlook to reduce diabetes and lower the economic cost, while ensuring early diagnosis and attention to avoid complications such as kidney failure, retinopathy, loss of limbs, arterial stiffness, stroke and heart disease are the most important means to reduce the economic burden of diabetes.

The process of finding meaningful knowledge from hidden patterns is called data mining, which assumes great significance in healthcare, especially in forecasting diseases like diabetes and cancer. It can also facilitate the diagnosis for doctors in making their clinical decisions [5] by extracting knowledge from copious amounts of data to identify patterns as well as to establish relationships that help solve problems. Applying data mining techniques will facilitate the early prediction of diabetes which improves treatment [6]. Then it can be used to create an efficient decision-making process in the medical field. Therefore, the accurate right decision planning of predictive data mining is a highly creative methodology that the WHO) would do well to take seriously [7].

In the majority of studies conducted so far, machine learning algorithms have been used more than deep learning. The Pima Indians Diabetes Database (PIDD) was the most widely used. Different algorithms were applied to this database, therefore, the result differed from one study to another regarding the best algorithm to apply. In our research, we used a data set recently collected from the patients of Sylhet (Bangladesh)-based Diabetes Hospital. Thus far, only two similar studies have been carried out using the same dataset [8], [9]. In [8], the authors applied Naive Bayes (NB), J48 Decision Trees, Logistic Regression, and Random Forest (RF), while in [9], the authors investigated the disease on the basis of Naïve Bayes, Random Tree, SVM, K-NN, Bayes Network, J48, and Random Forest.

This study aims to build two data mining models using classification methods Random Forest (RF) and Multilayer Perceptron (MLP) using WEKA. Since no study thus far on the Sylhet dataset has compared MLP with RF, we're seeking a comparison between them to determine the best performance to help forecast the onset of diabetes using different measurements: accuracy and Roc curve. Also, MLP and Random Forest are essential data mining methods, and they can diagnose diseases at an early stage and obtain high performance compared with other classifiers. Thus, these methods are adopted in this study. In addition, this study attempts to determine the most salient issues (predictive) that can impact diabetes; the findings can also be used to issue early warnings about the disease.

The organization of this paper is done in the following manner. Section 2 reviews associated literature work. Section 3 contains a description of the techniques proposed in this study. Section 4 contains dataset description, experimental setup, whereas section 5 elaborates on Optimization Strategy. Results and discussion is presented in section 6 whereas section 7 contains the conclusion.

2 Related Work

In [8], the authors aimed to discover the best algorithm for forecasting the risk factors for diagnosis. The dataset comprises 520 instances as well as 16 attributes gathered via Sylhet Diabetes Hospital. The dataset was assessed using J48 Decision Trees, RF and Logistic Regression (LR) algorithms. As per the findings, the accuracy of RF was the highest at 97.4% as compared to others.

In [9], the researchers proposed a classifier that classifies diabetes using data mining techniques. They study seven algorithms: Naïve Bayes, Random Tree, SVM, K-NN, Bayes Network, J48, and Random Forest. This study attempted to assess the algorithms' execution for a diabetes dataset using the WEKA program and applied to Sylhet Diabetes Hospital. As per this study, the highest accuracy was attained by k-NN (98.07%), and also helped classify diabetes within the dataset.

In [10], the authors discussed a method for determining the incidence of diabetes using machine learning algorithms. This paper aimed to discover diabetes in its early stages and design a model that can accurately predict the prospect of getting diagnosed with this disease. Three machine learning algorithms DT, SVM, and NB were utilized and the evaluation of the algorithm took place based on accuracy, F-scaling, and other measures, and to determine the correct classification of cases. The results showed that the NB algorithm achieved higher accuracy by 76.30%.

In [11], the authors aimed to early prediction of diabetes using various techniques of data mining. The dataset comprised 768 instances and the dataset was assessed using WEKA and MATLAB tool with Naive Bayes, MLP, Bayesian Network, PLS-LDA, Homogeneity-Based, ANN, C4.5, Amalgam KNN, ANFIS, and Modified J48. The results showed that Modified J48 had the highest accuracy of 99.87% for predicting the disease, while the ANN algorithm had the worst accuracy of 73.44%.

In [12], the authors presented a novel model to forecast type 2 diabetes using the techniques mentioned above to not only enhance the prediction model's accuracy, but also to increase its suitability to more than a single dataset. The main technique applied was the enhanced algorithm of LR and K-means cluster. According to the findings, the accuracy of this model 3.04%higher as compared to other findings, which revealed 95.42%. Also, evaluating the model using two datasets, the model revealed an accuracy of 90.7% and 94, respectively, providing that the performance of the model was good

In [13], the problem of early prediction of gestational diabetes was discussed. The authors aimed to compare three data mining algorithms depending on some attributes to provide the best algorithm. The main technique applied to the dataset was NB, KNN, and DT. The results showed that the DT's accuracy was the highest with 75.65%, while the accuracy of KNN was the lowest (65.16%).

In [14], the authors studied the same problems. The dataset used in this research comprised nine attributes with 768 instances. The relationship between the attributes was identified by the researchers. The main techniques applied in this dataset included RF, ANN, as well as K-means clustering. The accuracy of ANN was the highest at 75.7%. The results also revealed a tight linkage between diabetes, glucose and Body Mass Index (BMI).

In [15], the authors have aimed to build a model of prediction for three complications related to diabetes in Indonesia and to ascertain their relationship to characteristics, including age and blood glucose level. The database was collected from three sources, consists of 158 medical records. Three data mining techniques used included C4.5 decision tree, NB, and k-mean clustering for data analysis purposes. The authors assessed performance through clustering and classification techniques. The classification technique gives better information and performance compared to the clustering technique. The researchers used accuracy as a metric for evaluating this model's accuracy. The accuracy of Diabetes Complication Prediction Model was revealed to be 68%. The authors recommended the use of other algorithms like RF, LM, RF, and RT, for gauging their reliability and accuracy.

Table 1. Related Work

Ref	Year	Proposed Method	dataset	Best method
8	2020	NB, J48 D, Logistic Regression, and Random Forest.	520 instances from Diabetes symptom Dataset.	Random Forest 97.4%.
9	2021	NB, Random Tree, SVM, k-NN, Bayes Network, J48, and Random Forest	520 instances from Diabetes symptom Dataset.	k-NN 98.07%.
10	2018	DT, SVM and NB	768 instances from Pima Indians Diabetes Database (PIDD).	NB 76.30%.
11	2018	NB, MLP, Bayesian Network, PLS-LDA, Homogeneity-Based, ANN, C4.5, Amalgam KNN, ANFIS, and Modified J48.	768 instances from PIMA Indian Dataset	Modified J48 99.87%
12	2018	The improved K-means and Logistic regression algorithm.	768 instances from Pima Indians Diabetes Database (PIDD).	Proposed model 95.42%.
13	2018	DT, NB, and K-NN	768 instances from Pima Indians Diabetes Database (PIDD).	DT 75.65%.
14	2019	ANN, Random Forest, and K-means	768 instances from the National Institute of Diabetes and Digestive and Kidney Diseases.	ANN 75.7%.
15	2019	NB Tree , C4.5 and k means	158 instances from three sources (Sri Pamela Hospital and Kumpulan Pane Hospital	proposed model 68%.

3 Description of The Proposed Techniques

3.1 ANN

ANN or Artificial Neural Network is a class of algorithms that uses artificial intelligence technology. ANN is an interconnected set of virtual neurons created by computer programs that aims to collect knowledge through training and then storing it via the neurons' connecting forces referred to as synaptic weights [16]. ANN aims to simulate the intelligence of the human brain. The idea of neural networks goes back to two researchers at the University of Chicago, Warren McCullough, a neuroscientist, and Walter Bates, a mathematician, in 1943 [17].

The essential elements in ANN are nodes called neurons and connections between them. As mentioned before, the training process determines the value of connection weights and is paramount to discern its learning ability. The three layers of ANN are Output, Hidden, and Input [18] Multilayer Perceptron (MLP) is one of ANN algorithms, characterized by adding hidden layers to simple perception [19]. Fig.1 shows a schematic graph of MLP [20]

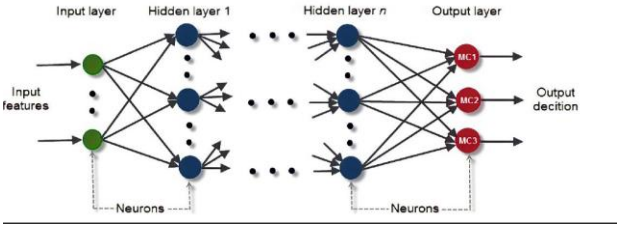

Fig. 1. A Schematic View of an MLP Neural Network [20]

3.2 Random Forest

This form of machine learning technology is an integrated and multi-purpose tool that uses a set of data and variables to build more than one decision tree. It is then combined to obtain more stable and accurate forecasts. RF can be expedited on large datasets and is used in regression as well as classification problems. Leo Breiman had proposed RF in 2001 [21] [22]. RF needs two parameters to be tuned: total number of variables and trees [23]. Fig. 2 shows a pictorial representation of RF [22].

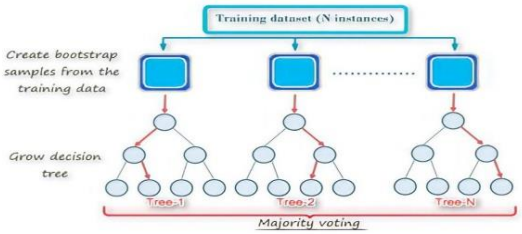

Fig. 2. Pictorial Representation of RF [22]

4 Contains Dataset Description and Setup

4.1 Description of the Dataset

The dataset was extracted from the UCI repository, and collected using a questionnaire shared with the patients of Sylhet Diabetes Hospital. It contains 17 attributes, and 520 instances, the attributes are all nominal except age which is numeric. The range of age of the participants was 20- 65 and above. Fig. 3 summarizes the data preparation's approach.

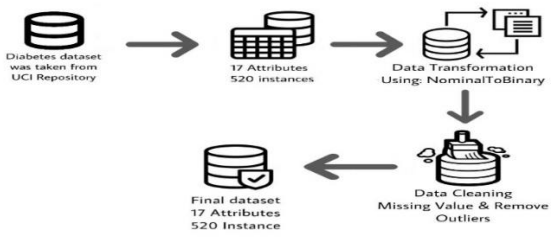

Fig. 3. Process of Dataset Preparation

4.2 Experimental Setup

The experiment was applied using WEKA, WEKA is a workspace that provides an excellent mechanism to examine common problems of data mining like clustering, classification and it facilitates comparison of different machine learning techniques [24].

Initially, the data set was preprocessed in preparation for the experiment. The nominal data were transformed to numeric binary data using a supervised attribute filter (NominalToBinary). No missing data was observed. Also, an unsupervised attribute filter (InterquartileRange) was used to handle outliers in the data set in WEKA.

Additionally, the optimized parameters of RF, and MLP were determined by adjusting the number of the tree, and the seed parameter was resetting. In addition to adjusting the seed, the learning rate, and the hidden layers parameter for the MLP. As per the findings, the accuracy of RF and MLF was 97.88% and 96.34%, respectively.

Besides, to rank the selected features' attributes, correlation coefficients and information gain between the class variable and each attribute were determined. Tables 5 and 6 show the results.

Subsequently, we investigated the performance by choosing attributes and determining significant attributes to help the prediction of Diabetes during the infancy stage. In this process, correlation-based feature selection and InfoGain method were utilized. For all classifiers, we then tested these models on these attributes via optical parameters as well as 10-fold cross-validation. Table 6 as well as Table 7 depict the results.

Finally, from the aforementioned experimentation, we opted for the optimal choice encompassing all attributes for undertaking the model's development and attaining the optimal outcome in terms of cross-validation

5 Strategy of Optimization

Intending to optimize the results of classification, we used the CVParameterSelection, a search methodology for undertaking optimal parameters for each classifier using WEKA [25]. Our attempt was to yield the best criteria of performance based on accuracy using 10 cross-validations. Table 2 depicts both default parameters as well as optimal parameters.

Table 2. Parameters for all classifiers: Default and Optimal

Model	Parameters Values		
RF	**Parameters**	**Default**	**Optimal**
	Num Iterations	100	34
	Seed	1	34
MLP	Seed	0	0
	Hidden Layers	a	a
	Learning Rate	0.3	0.3

Table 3 shows the comparison of findings associated with optimal parameters and that of their default counterparts. It can be seen that the RF algorithm's accuracy increases, but that of MLP remains unchanged.

Table 3. Classifiers performance: Default and Optimal parameters

Model	Accuracy of Performance	
	Default	**Optimal**
RF	97.11%	97.88%
MLP	96.34%	96.34%

6 Results and Discussion

6.1 Effect of Feature Selection on the Dataset

The information gain and correlation-based selection of feature methods were used to determine the best performing subset along with the most salient attributes to ensure early prediction of Diabetes. The correlation coefficient was used to rank the features based on the Pearson values, from the highest to the lowest variable relationship with the class variable (output), as shown in Table 4. Besides, InfoGain was applied to classify the features based on the class information gain measure, as shown in Table 5.
Table 6 and Table 7 clearly point out that the implementation of all features (16 in totality) helped obtain the best performance. Moreover, the results have been shown that decreasing the number of features leads to a decrease in the percent-age of accuracy.

Table 4. The correlation of each attribute and the target

Sequence	Attribute name	Correlation
1	Polyuria	0.6659
2	Polydipsia	0.6487
3	Gender	0.4492
4	sudden weight loss	0.4366
5	partial paresis	0.4323
6	Polyphagia	0.3425
7	Irritability	0.2995
8	Alopecia	0.2675
9	visual blurring	0.2513
10	weakness	0.2433
11	muscle stiffness	0.1225
12	Genital thrush	0.1103
13	Age	0.1087
14	Obesity	0.0722
15	delayed healing	0.047
16	Itching	0.0134

Table 5. The information gain of each attribute and the target

Sequence	Attribute name	InfoGain
1	Polyuria	0.36225
2	Polydipsia	0.35906
3	Gender	0.16342
4	sudden weight loss	0.14877
5	partial paresis	0.14465
6	Polyphagia	0.14465
7	Irritability	0.07287
8	Alopecia	0.05116
9	visual blurring	0.04661
10	weakness	0.04267
11	Age	0.02239
12	muscle stiffness	0.01097
13	Genital thrush	0.00905
14	Itching	0
15	delayed healing	0
16	Obesity	0

Table 6. InfoGain Feature Selection Results

Number of features	Features	RF	MLP	AVG
Using features: All (16)	All	97.88%	96.34%	97.09%
Using 8 features	Polydipsia, Polyuria, Sex Sudden loss of weight, Partial paresis, Irritability, Alopecia, Polyphagia	94.03%	93.65%	93.94%
Using 4 features	Polydipsia, Polyuria, Sex sudeen loss of weight	88.84%	89.42%	89.13%
Using 2 features	Polyuria ,Polydipsia	86.92%	86.92%	86.92%

Table 7. Correlation-Based Feature Selection Results

Number of features	Features	RF	MLP	AVG
Using features: All (16)	All	97.88%	96.34%	97.09%
Using 8 features	Polydipsia, Polyuria, Sex Sudeen loss of weight, Partial paresis, Alopecia, Irritability, Polyphagia	94.03%	93.65%	93.94%
Using 4 features	Polydipsia, Polyuria, Sex sudeen loss of weight	88.84%	89.42%	89.13%
Using 2 features	Polyuria ,Polydipsia	86.92%	86.92%	86.92%

6.2 Further Discussion

Table 8 and fig 4 show the model that was used to ensure early prediction of diabetes using 16 features involving optimal parameters. The technique of 10-fold cross-validation was used to attain optimal findings of MLP and RF. RF was found to outperform MLP in terms of forecasting Diabetes. Its accuracy was found to be 97.88%. Meanwhile, the accuracy was negatively impacted when the number of features was lowered and enhanced after using all features. Based on these findings, it can be inferred that it is necessary to incorporate all features to increase the accuracy of forecasting diabetes at an early stage. The Receiver Operating Characteristic (ROC) curve is another indicator of how the model of classification performs. Figures 5 and 6 illustrate that the proximity and placement of this curve to the left-hand side (on the top) indicate that the experiment's accuracy is high. Overall, the area under the curve for each classifier shows that the most suitable classifier is determined to be RF compared to MLP.

Table 8. Findings of Best Options based on Optimal Parameters

Techniques	RF	MLP
10-fold validation	97.88%	96.34%

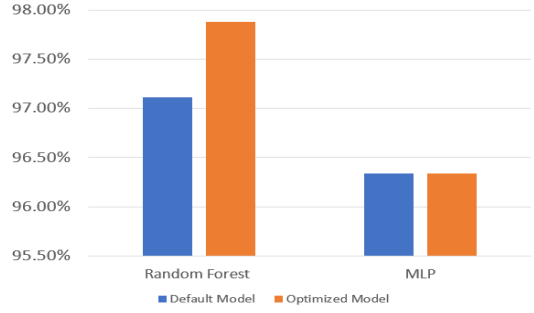

Fig. 4. Default and Optimized Models: A Comparison

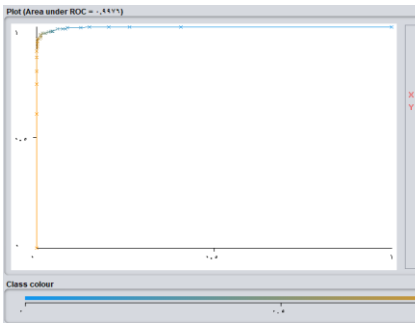

Fig. 5. Positive Class RF ROC curve

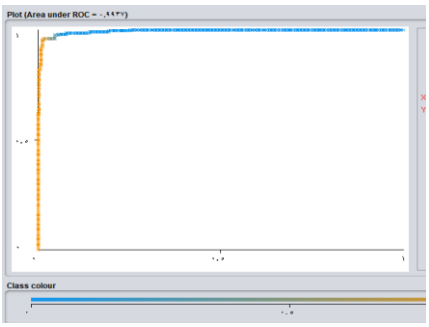

Fig. 6. Positive Class MLP ROC curve

7 Conclusion

Diabetes is characterized by a high blood sugar level due to the inability of the pancreas to generate sufficient insulin. It can also be caused when the cells of the human body are unable to respond adequately to the insulin produced in the body. It prevents the body from ensuring proper utilization of energy. Early prediction of diabetes can go a long way in helping people avert the negative economic consequences that emerge from the constant rise in the number of cases involving this serious disease that affects people from all over the world. In this context, data mining has emerged as an important tool that facilitates the discovery of hidden information.This study implemented two classifiers, random forest and MLP, in order to obtain the best accuracy rates for estimating diabetes. Accordingly, the random forest method was found to yield better results than MLP when it comes to enabling early prediction of diabetes with high accuracy of 97.88%. The experiment revealed that both algorithms work better when all features are utilized; therefore, each of these features was used to achieve the highest possible accuracy ratio. In the aftermath of the optimization strategy, the random forest was also found to outperform the results derived in [8], which was found to have an accuracy of 97.4%.

References

1. International Diabetes Federation. The diabetes atlas. 5th ed. Brussels: International Diabetes Federation; 2011.
2. Diabetes,World Health Organization (WHO): https://www.who.int/news-room/fact-sheets/detail/diabetes .
3. Alassaf, R. A., Alsulaim, K. A., Alroomi, N. Y., Alsharif, N. S., Aljubeir, M. F., Olatunji, S. O., ... & Alturayeif, N. S. (2018, April). Preemptive Diagnosis of Diabetes Mellitus Using Machine Learning. In 2018 21st Saudi Computer Society National Computer Conference (NCC) (pp. 1-5). IEEE.

4. Bommer, C., Sagalova, V., Heesemann, E., Manne-Goehler, J., Atun, R., Bärnighausen, T., ... & Vollmer, S. (2018). Global economic burden of diabetes in adults: projections from 2015 to 2030. Diabetes care, 41(5), 963-970.
5. Jothi, N., & Husain, W. (2015). Data mining in healthcare–a review. Procedia computer science, 72, 306-313.
6. Agrawal, P., & Dewangan, A. (2015). A brief survey on the techniques used for the diagnosis of diabetes-mellitus. Int. Res. J. of Eng. and Tech. IRJET, 2, 1039-1043.
7. Cheung, J. Y. (2012). Data mining: Concepts and techniques. Middletown: American Library Association dba CHOICE.
8. Islam, M. F., Ferdousi, R., Rahman, S., & Bushra, H. Y. (2020). Likelihood prediction of diabetes at early stage using data mining techniques. In Computer Vision and Machine Intelligence in Medical Image Analysis (pp. 113-125). Springer, Singapore.
9. Alpan, K., & İlgi, G. S. (2020, October). Classification of Diabetes Dataset with Data Mining Techniques by Using WEKA Approach. In 2020 4th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT) (pp. 1-7). IEEE
10. Sisodia, D., & Sisodia, D. S. (2018). Prediction of diabetes using classification algorithms. Procedia computer science, 132, 1578-1585.
11. Sengamuthu, M. R., Abirami, M. R., & Karthik, M. D. (2018). Various Data Mining Techniques Analysis to Predict Diabetes Mellitus. Int Res J Eng Technol (IRJET), 5(5), 676-9.
12. Wu, H., Yang, S., Huang, Z., He, J., & Wang, X. (2018). Type 2 diabetes mellitus prediction model based on data mining. Informatics in Medicine Unlocked, 10, 100-107.
13. Azrar, A., Ali, Y., Awais, M., & Zaheer, K. (2018). Data mining models comparison for diabetes prediction. Int. J. Adv. Comput. Sci. Appl.(IJACSA), 9, 320-323.
14. Alam, T. M., Iqbal, M. A., Ali, Y., Wahab, A., Ijaz, S., Baig, T. I., ... & Abbas, Z. (2019). A model for early prediction of diabetes. Informatics in Medicine Unlocked, 16, 100204.
15. Fiarni, C., Sipayung, E. M., & Maemunah, S. (2019). Analysis and prediction of diabetes complication disease using data mining algorithm. Procedia Computer Science, 161, 44
16. Jain, A., & Kumar, A. (2006). An evaluation of artificial neural network technique for the determination of infiltration model parameters. Applied Soft Computing, 6(3), 272-282.
17. Yaqub, F. A Study on Artificial Neural Network.
18. Grossi, E., & Buscema, M. (2007). Introduction to artificial neural networks. *European journal of gastroenterology & hepatology*, *19*(12), 1046-1054.
19. Azeez, O. S., Pradhan, B., Shafri, H. Z., Shukla, N., Lee, C. W., & Rizeei, H. M. (2019). Modeling of CO emissions from traffic vehicles using artificial neural networks. *Applied Sciences*, *9*(2), 313.
20. Ebrahimi, E., Mollazade, K., & Arefi, A. (2012). An expert system for classification of potato tubers using image processing and artificial neural networks. International Journal of Food Engineering, 8(4).
21. Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.
22. Al Amrani, Y., Lazaar, M., & El Kadiri, K. E. (2018). Random forest and support vector machine based hybrid approach to sentiment analysis. Procedia Computer Science, 127, 511-520.
23. Naghibi, S. A., Ahmadi, K., & Daneshi, A. (2017). Application of support vector machine, random forest, and genetic algorithm optimized random forest models in groundwater potential mapping. Water Resources Management, 31(9), 2761-2775.
24. Frank, E., Hall, M., Trigg, L., Holmes, G., & Witten, I. H. (2004). Data mining in bioinformatics using Weka. Bioinformatics, 20(15), 2479-2481.
25. https://weka.sourceforge.io/doc.dev/weka/classifiers/meta/CVParameterSelection.html.