International Journal of Research in Engineering, Science and Management
Volume 5, Issue 9, September 2022
https://www.ijresm.com | ISSN (Online): 2581-5792

Prediction Analysis using Weighted Product Method to Compare Machine learning Algorithms for Diabetes Disease

Jerry Malapane[1*], W. Doorsamy[2], B. S. Paul[3]
[1]Department of Electrical and Electronic Engineering, University of Johannesburg, Gauteng, South Africa
[2,3]Institute for Intelligent Systems, University of Johannesburg, Gauteng, South Africa

***Abstract*: Diabetes is one of the well-known and most serious chronic diseases in the world, causing a person to suffer from a raised level of blood sugar due to body resistance to producing the essential volume of insulin. People suffering from diabetes may have complications such as cardiovascular diseases, blindness, and kidney diseases. Early prediction and diagnosis of diabetes can save many lives by alarming people to require medical attention. One possible solution for the diagnosis and prediction of diabetes disease is the use of machine learning approaches. In this paper, prediction analysis of diabetes disease is performed using various machine learning techniques and comparative analysis based on the Multi-Criteria Decision-Making method to vote for the best algorithm. Few Machine learning techniques such as Support Vector Machine, K-NN, Random Forest Classifier, Naïve Bayes, Extreme Gradient Boosting, Adaptive Boosting, Multilayer Perceptron, and hybridized K-mean Random Forest. In our study experiment, we used PIMA Indian Dataset retrieved from UCI Repository. The results show that all evaluation criteria are valid and calculated using Weighted Product Methods to provide the best algorithm. Our experimental results show that the Extreme Gradient Boosting technique achieved the highest ranking when relating to the other seven machine learning algorithms.**

***Keywords*: MCDM, Diabetes, Machine learning, Dataset, WPM.**

1. Introduction

Diabetes is a class of chronic diseases that occurs when the blood sugar level is very high. Increased blood sugar levels can lead to dysfunction and failures of different body organs such as the heart, kidneys, and blood vessels. According to International Diabetes Federation, it's estimated that in 2021, 536 million people had diabetes globally with diabetes-related fatalities of 32% in people under the age of 60. While it is predicted that by 2045, 783 million people will be diabetic [1]. There are three main categories of diabetes: Type 1, Type 2, and Gestational diabetes. Type 1 diabetes, also recognized as Juvenile-onset diabetes and is common among young people mostly less than 30 years of age. Type 1 diabetes is triggered by an autoimmune response in which the body's immune system attacks insulin due to insufficient cells in the pancreas [2]. People with type 1 diabetes need regular injections of insulin to sustain blood sugar levels in the suitable range. Type 2 diabetes, is known as Adult-onset diabetes. It is mostly presented in older adults with similar symptoms to those of type 1 diabetes. Type 2 diabetes happens when the human body is not able to produce insulin. Gestational diabetes develops during women's pregnancy. This type of diabetes causes blood glucose levels to increase which can affect pregnancy and the health of a baby [3]. Diabetic patients usually require continuous care to avoid life-threatening complications. For early diabetes disease, the existing studies proposed and deployed machine learning algorithms for the prediction of patients with diabetes.

In this paper, we analyze different types of machine learning algorithms based on their performance evaluation by using the Weighted Product Method (WPM) to select the best algorithm. Machine learning is a method that is used to train machines explicitly and provide efficient results in the collection of information by structuring various classification models from a collected dataset. The experimental tests for diabetes prediction are accomplished using Pima Indian Diabetes Database (PIDD) [4]. six different machine learning techniques used in this study are Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Random Forest, Naïve Bayes, XGBoost, Adaptive Boosting, Multilayer Perceptron, and hybridized K-mean Random Forest. we conducted a comparative analysis to identify the most precise and efficient algorithm using the Weighted Product Method (WPM) as a Multi-Criteria Decision-Making technique by using performance evaluations such as accuracy, sensitivity, precision, and specificity.

The rest of this paper is prepared as follows: Section 2 literature review, Section 3 Methodology description, and discussion of the Dataset used in our experiment and Evaluation method. Section 4 experimental procedure and results. Section 5 discusses and concludes the paper.

2. Literature Review

The most recent researchers are involved in developing various prediction machine learning algorithms which are helpful in the different healthcare sectors. With available and observed datasets few studies have been conducted with suggested proposals to predict the results of some diseases. Sisodia and Sisodia et al. [5] applied NB, DT, and SVM

*Corresponding author: 216088717@student.uj.ac.za

machine learning algorithms to identify diabetes at an early stage. The authors used the PIDD dataset to perform experiment and evaluate parameters like F-Measure, Precision, Recall, and Accuracy. Naïve Bayes was regarded as the best with 76.30% accuracy compared to DT and SVM algorithms. Gupta et al. [6] used only two ML techniques: Support Vector Machine and Naïve Bayes for diabetes prediction based on PIMA India Diabetes dataset. K-fold cross-validation and feature selection-based approach methods were used by the authors to improve the accuracy of the algorithm. After test performances, the Support Vector Machine was showing improved results better than the Naïve Bayes algorithm.

Aishwarya et al. [7] used the PIMA dataset to predict diabetes with a performance estimation on five machine learning algorithms which are Decision Tree, Support Vector Machine, Random Forest, K-Nearest Neighbor, and Logistic Regression. Based on the authors Logistic Regression show the highest accuracy of 77.6% and an Area Under the Curve (AUC) of 73.6%. Similarly, Jakka et al. [8] performed diabetes prediction by evaluating machine learning techniques such as DT, SVM, RF, LR, and KNN. LR achieved the best accuracy compared to other algorithms. Choubey et al. [9] used Adaboost, KNN, and ANN algorithms for diabetes classification. The authors used the PIMA Indian dataset captured from the UCI Repository. For feature selection, Principle component analysis (PCA) and linear discriminant analysis (LDA) are used for removing unwanted features and improving accuracy. Zou et. al. [10] used minimum redundancy maximum relevance (mRMR) and PCA to reduce the dimensionality, and used machine learning algorithms such as Decision Tree, Neural Network, and Random Forest, to predict diabetes. The dataset contains 14 attributes obtained from the hospital physical examination data in Luzhou, China. The results show that when all attributes are been used, Random Forest provides the highest accuracy of 80.84% compared to other algorithms.

Based on previous research, it has been demonstrated that the classification process needs to be improved. Most machine learning algorithms are selected based on accuracy value, while other evaluation parameters such as Recall, F-Measure, Precision, and Receiver Operating Curve (ROC) are not considered. In our work, we introduce the Weighted Product Method (WPM) which is the Multi-Criteria Decision Making (MCDM) technique. WPM combines all used evaluation parameters and calculates the best algorithm.

3. Methodology

The main goal of this paper is to use MCDM to select the best ML algorithm for diabetes prediction. Figure 1 demonstrates the framework for the proposed predictive model. The framework contains the following important stages: (1) collected PIMA Indian diabetes dataset from the UCI machine learning repository. (2) The data is pre-processed and cleaned to remove irrelevant features. (3) Dataset is split into two, 80% for the training set and 20% for the test set (4) eight machine learning algorithms are used to analyze and predict diabetes in the training stage. (5) The performance evaluation criteria results are used by WPM to finalize and determine the best algorithm.

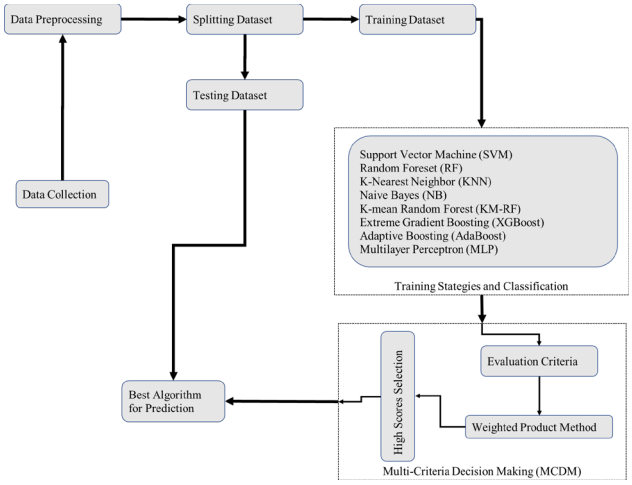

Fig. 1. Proposed diabetes prediction framework

A. Dataset

In this paper, we used PIMA Indian Diabetes Dataset from the University of California (UCI), Irvine repository [11]. The dataset consists of medical records of 768 instances of women patients from the age of 26 years old. As shown in Table 1, the dataset contains eight attributes which are: Number of pregnancies, glucose tolerance, blood pressure, skin thickness, body mass index, age, Diabetes Pedigree Function and insulin, and one target variable that indicates 0 for a patient without diabetes and 1 for a patient with diabetes.

Table 1
PIMA dataset description

Attributes	Description	Attributes type
Pregnancies	Number of pregnancies	Numeric
Glucose	Plasma glucose concentration for 2 hours of oral glucose tolerance test	Numeric
BP	Blood pressure in mm Hg	Numeric
Skin thickness	Triceps skinfold thickness in mm	Numeric
Insulin	2 hours serum insulin in µIU/mL	Numeric
BMI	Body mass index in kg/(m*m)	Numeric
DPF	Diabetes pedigree function	Numeric
Age	Age in years	Numeric
Outcome	Diabetes diagnoses outcomes (1 = positive; 0 = negative)	Nominal

The available dataset consists of 500 people who are healthy with an outcome of 0 and 268 people who are diabetic, with an outcome of 1. For experiment purposes, the dataset is divided into 80% for training and 20% for testing with ML algorithms.

B. Brief Description of Used Machine Learning Algorithms

After data is preprocessed and split, it becomes ready to be used by applying different ML techniques, for diabetes prediction. The main objective is to apply machine learning algorithms to clean the dataset and analyze the performance of each algorithm. In the following, we briefly describe each of the machine learning algorithms evaluated in our experiment.

1) Support Vector Machine (SVM)

SVM is one of the used and known machine learning algorithms, used in both classification and regression problems. The SVM classifier's references are based on the Structural Risk Minimisation (SRM) theory, by maximizing the margins on the training dataset of a classifier may result in a lower generalization error [12].

2) Random Forest (RF)

RF refers to the collection of decision trees which are commonly used in classification problems but also can be used in regression tasks. The main purpose of RF is for each decision tree model to predict results and vote on the majority results [13]. The advantage of RF is that can easily handle large datasets compared to other machine learning algorithms.

3) K-Nearest Neighbor (KNN)

KNN is also a supervised machine learning technique and is used to solve both regression and classification tasks. KNN is also known as a lazy, instance-based, and non-parametric prediction algorithm. KNN classifier is based on Euclidean distance and classifies an instance by finding nearest neighbors [14].

4) Naïve Bayes (NB)

NB is one of the best classification machine learning algorithms based on the Naïve Bayes Theorem. NB classifies data into determined categories using conditional probability [15]. NB is observed as a descriptive and also predictive technique. The probabilities are descriptive and are deployed to predict the categories of untrained data.

5) K-Mean Random Forest (KM-RF)

This is a developed hybrid stacking model which combines K-Mean and Random Forest algorithms to improve the accuracy and efficiency of the prediction technique. In this combined method, K-means is only executed and transform data into smaller and more appropriate sets and RF performs classification.

6) Extreme Gradient Boosting (XGBoost)

XGBoost is an improved supervised model based on a decision tree algorithm. XGBoost is mostly used by data scientists to solve both regression and classification problems as it has high accuracy and high execution speed [17]. This technique is considered a significant and more advanced algorithm that can reduce overfitting and solve data irregularities.

7) Adaptive Boosting (AdaBoost)

AdaBoost is one of the simplest boosting techniques, it modifies weak algorithms by assigning higher weights to certain results based on weak rules to improve the accuracy of the misclassified models [18]. AdaBoost works on a combination of other algorithms which failed to reach the required accuracy and uses boosting methods to increase accuracy.

8) Multilayer Perceptron (MLP)

MLP is a part of Neural Network algorithms that consist of one input layer, one output layer, one or more hidden layers of nodes, and an output layer. MLP is used to solve nonlinear and complex problems compare to Single-Layer Perceptron (SLP) which solves only linearly separable problems [19]. The perceptron is a linear technique that uses multiple inputs and produces one output. MLP generates pattern classification of input patterns and multiplies them with weights and adds a bias term to predict the output.

4. Evaluation Metrics

The evaluation metrics are used to evaluate the performance of the algorithms. Evaluation metrics are calculated based on four factors which are True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). We use eight evaluation metrics such as; Accuracy, Precision, Specificity, Recall, F1-score, False Positive Rate (FPR), False Negative Rate (FNR), and Receiver Operating Characteristic – Area Under Curve (AUC). The performance metrics are calculated as follows:

$$\text{Accuracy} = (TP + TN) / (TP + FP + TN + FN) \quad (1)$$

$$\text{Specificity} = TN / (FN + TN) \quad (2)$$

$$\text{Precision} = TP / (TP + FP) \quad (3)$$

$$\text{Recall} = TP / (TP + FN) \quad (4)$$

$$\text{F1-Score} = 2^* (\text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (5)$$

$$\text{False Positive Rate (FPR)} = FP / (TN + FP) \quad (6)$$

$$\text{False Negative Rate (FNR)} = FN / (FN + TP) \quad (7)$$

A. Model Selection based on Weighted Product Method (WPM)

WPM is one of the most popular multi-criteria decision-making methods which determine the best alternative based on multiplication criteria. WPM is also known as dimensionless analysis because of its mathematical structure that abolishes any units of measure [20]. The steps used to calculate WPM are as follows:

1. Determine types of criteria based on category and weight.
2. Calculate the weight values of all criteria using the below formula 1.

$$W_j = \frac{w_j}{\Sigma_{j=1}^{N} w_j} \quad (8)$$

3. Calculate the value of Vector S using the following formula 9.

$$S = (W_{ij}^{Awj}.w) \,.\, (W_{in}^{Awn}.w) \quad (9)$$

4. Calculate the relative preference value Vector V using formula 10.

$$V_{jn} = \frac{s_i}{\Sigma_{i=1}^{N} s_i} \quad (10)$$

5. The maximum value of *V* presents a better alternative. *Where: W* is a Weight of criteria; *S* is Alternative preference analogous; *V* is Alternative preferences; *j* is criteria; *i* is Alternative; *n* is the number of criteria.

5. Experiments and Results

After data is processed is divided into the training set and test set, 80% of the data is for the training set and 20% of the dataset is for testing. All experiments and analyses were performed using Python (Version 3) and Figure 2, shows the correlation

coefficient matrix which demonstrates a linear correlation between various parameters and diabetes.

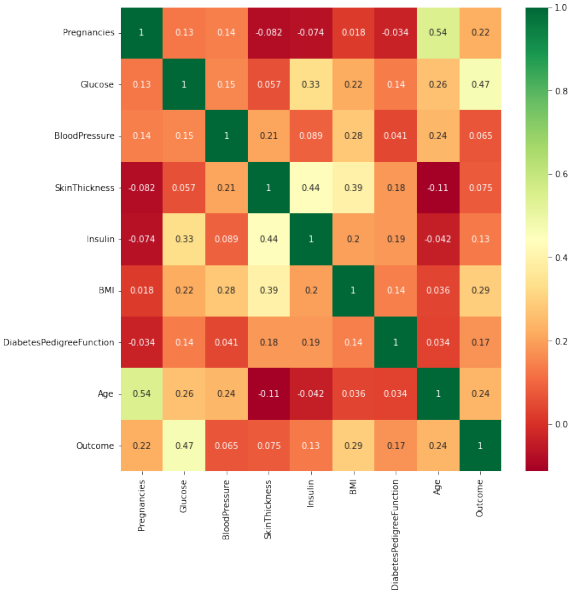

Fig. 2. Correlation coefficient matrix

Table 2 demonstrates the application of performance results of all eight machine learning algorithms and their performances are examined by using different evaluation metrics as specified in equations (1) to (7).

Table 2
Performance results

Model	Accuracy[%]	F1-score[%]	FPR[%]	FNR[%]	Precision[%]	Specificity[%]	Recall[%]	AUC[%]
SVM	78.99	65.96	19.82	24.24	58.41	90.00	75.76	79.0
RF	86.64	79.30	13.20	13.74	73.36	93.75	86.26	80.0
KNN	79.97	65.93	20.34	19.05	55.61	93.00	80.95	76.0
NB	75.57	63.60	19.95	33.84	61.21	83.25	66.16	81.0
XGBoost	98.53	97.88	1.49	1.42	97.20	99.25	98.58	79.0
AdaBoost	83.71	76.08	13.41	22.06	74.30	88.75	77.94	80.0
MLP	85.18	77.86	12.95	18.78	74.77	90.75	81.22	78.0
KM-RF	87.46	80.41	12.87	11.73	73.82	94.75	88.27	81.0

Figure 3, shows the Receiver operating characteristic (ROC) curve which is used to visualize the performance of algorithms by plotting the True Positive Rate against the False Positive Rate. It is found that the AUC values of NB and KM-RF are 0.81. NB and AdaBoost with AUC values of 0.80 while SVM, XGBoost, KNN, and MLP values are 0.79, 0.78, and 0.76. AUC values are included in Table 2 as part of the overall evaluation criteria.

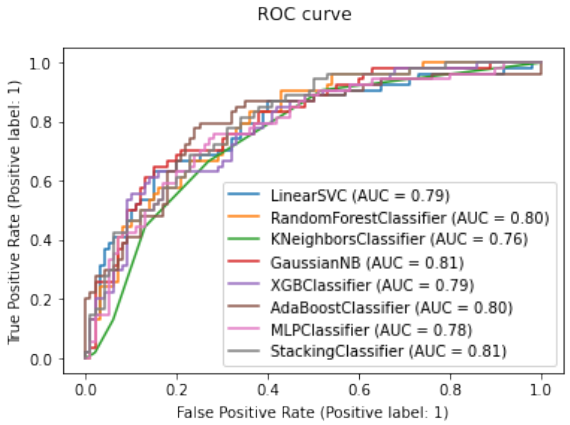

Fig. 3. ROC curve

Measuring the performance of predictive algorithms based only on one criterion such as accuracy may provide imbalanced results.in other to integrate WPM, all evaluation criteria are converted to entropy weights based and normalized as shown in Table 3. All categories are calculated and summed up and WPM is executed to provide high ranking predictive algorithm.

Table 3
Entropy weights

Model	Accuracy[▲ 0.125]	F1-score[▲ 0.125]	FPR[▲ 0.125]	FNR[▲ 0.125]	Precision[▲ 0.125]	Specificity[▲ 0.125]	Recall[▲ 0.125]	AUC[▲ 0.125]
SVM	0.132970	0.141290	0.044803	0.038053	0.147922	0.127044	0.133496	0.125346
RF	0.121229	0.117522	0.067272	0.067132	0.117777	0.121962	0.117246	0.123779
KNN	0.131341	0.141354	0.043658	0.048420	0.155370	0.122945	0.124937	0.130294
NB	0.138988	0.146532	0.044511	0.027258	0.141155	0.137344	0.152867	0.122251
XGBoost	0.106600	0.095213	0.595969	0.649574	0.088890	0.115203	0.102593	0.125346
AdaBoost	0.125473	0.122495	0.066219	0.041813	0.116287	0.128833	0.129762	0.123779
MLP	0.123307	0.119695	0.068571	0.049116	0.115556	0.125994	0.124522	0.126953
KM-RF	0.120093	0.115899	0.068997	0.078636	0.117043	0.120675	0.114576	0.122251

Table 4 presents the ranking results after integrating WPM, XGBoost shows a higher ranking, and KM-RF ranks second compared to other machine learning techniques. When comparing only supervised ML algorithms, RF shows better results when compared to SVM, KNN, and NB.

Table 4
Ranking results after integrating WPM

Algorithms	Voting Rank
SVM	7
RF	3
KNN	6
NB	8
XGBoost	1
AdaBoost	5
MLP	4
KM-RF	2

After using the Weighted Product Method as one of the MCDM techniques, we can conclude that the XGBoost algorithm performed better compared to the other seven algorithms. Not all the evaluation parameters of XGBoost are high, such as an AUC value of 79% while NB and KM-RF have the highest AUC values of 81% by using the Weighted Product Method we can use all the evaluation parameters to calculate and vote for the highest performing algorithm. Besides XGBoost and KM-RF algorithms as enhanced and modified techniques, RF is also performing better compared to the other five stated algorithms. In this paper, we conclude that the most suitable classification algorithm for diabetes prediction based on PIMA Indian Diabetes Dataset is the XGBoost technique.

6. Conclusion

Early prediction and detection of patients with diabetes are becoming more challenging in the healthcare system. The main purpose of this paper is to design and implement the method that analyzes the performance criteria and highest performing algorithm using the Weighted Product Method. We used eight input attributes and one output feature from the PIMA dataset and used eight different machine learning algorithms, including SVM, RF, KNN, NB, XGBoost, AdaBoost, MLP, and KM-RF to predict diabetes. Different evaluation parameters such as Accuracy, F1-score, Recall, Precision, Specificity, and AUC were in the Multi-Criteria Decision method (WPM) to determine the best performing algorithm. based on our proposed method of using WPM to rank the techniques using

evaluation criteria, it shows that the XGBoost classifier ranks the best model compared to the other seven algorithms. therefore, we can conclude that the implementation of machine learning systems in healthcare can assist in the early prediction and detection of diseases such as diabetes.

References

[1] International Diabetes Federation. (2021). *Global Diabetes Data Report 20002045*.IDF Diabetes Atlas 10th Edition 2021. Data last updated: Nov.08,2021:https://diabetesatlas.org/data/en/world/

[2] Martinsson, J., Schliep, A., Eliasson, B., and Mogren, O. (2020). Blood glucose prediction with variance estimation using recurrent neural networks. *J. Healthc. Inform. Res*. 4, 1–18.

[3] M. F. Zohora, M. H. Tania, M. S. Kaiser, and M. Mahmud, "Forecasting the risk of type ii diabetes using reinforcement learning," in *2020 Joint 9th International Conference on Informatics, Electronics Vision (ICIEV)*.

[4] Hassan AS, Malaserene I, Leema AA. Diabetes Mellitus Prediction using Classification Techniques". *International Journal of Innovative Technology and Exploring*. 2020.

[5] D. Sisodia, D.S. Sisodia, Prediction of diabetes using classification algorithms, *Procedia Comput. Sci*. 132 (2018) 1578-1585.

[6] S. Gupta, H. K. Verma, and D. Bhardwaj, "Classification of diabetes using naive Bayes and support vector machine as a technique," *Lecture Notes on Multidisciplinary Industrial Engineering*, Springer, Singapore, pp. 365–376, 2021.

[7] Aiswarya, L., Jeyalatha, S. and Sumbaly R. "Diagnosis of Diabetes using Classification Mining Techniques", in *International Journal of Data Mining & Knowledge Management Process*, vol. 5, no. 1, pp. 1-14, 2015.

[8] A. Jakka and R. J. Vakula, "Performance evaluation of machine learning models for diabetes prediction," *International Journal of Innovative Technology and Exploring Engineering*, vol. 8, no. 11, pp. 1976–1980, 2019.

[9] D. K. Choubey, M. Kumar, V. Shukla, S. Tripathi, and V. K. Dhandhania, "Comparative analysis of classification methods with PCA and LDA for diabetes," *Current Diabetes Reviews*, vol. 16, no. 8, pp. 833–850, 2020.

[10] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, and H. Tang, "Predicting diabetes mellitus with machine learning techniques," *Frontiers in Genetics*, vol. 9, Nov. 2018.

[11] Lichman, M., 2013, "UCI Machine Learning Repository" [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[12] V. N. Vapnik. An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, 10(5):988–999, 1999.

[13] N. Yuvaraj and K. R. SriPreethaa, "Diabetes prediction in healthcare systems using machine learning algorithms on hadoop cluster," *Cluster Computing*, vol. 22, no. 1, pp. 1–9,2019.

[14] K. Saxena, Z. Khan, and S. Singh, "Diagnosis of Diabetes Mellitus using K Nearest Neighbor Algorithm," in *International Journal of Computer Science Trends and Technology*, vol. 2, no. 4, pp. 36–43, 2014.

[15] Pandiangan N, Buono MLC, Loppies SHD (2020) Implementation of decision tree and Naı¨ve Bayes classification method for predicting study period. *J Phys Conf Ser* 1569:022022.

[16] T. Chen and C. Guestrin, "XGBoost: a scalable tree boosting system," in *Proceedings of the 22nd ACM international conference on knowledge discovery and data mining*, pp. 785–794, ACM, San Francisco California USA, 13 August 2016.

[17] S. Perveen, M. Shahbaz, A. Guergachi, and K. Keshavjee, "Performance analysis of data mining classification techniques to predict diabetes," *Procedia Computer Science*, vol. 82, pp. 115–121, 2016.

[18] Tzeng, G.-H., & Huang, J.-J., "Multiple attribute decision making: methods and applications," 1st ed., Chapman and Hall/CRC, 2011.

[19] J. S. Sonawane, and D. R. Patil, "Prediction of heart disease using multilayer perceptron neural network", in *Information Communication and Embedded Systems (ICICES)*, pp. 1–6, IEEE, February 2014.