DOI: 10.21917/ijdsml.2025.0185

PROBABILITY BASED WEIGHTED DATA TO PREDICT DIABETES MELLITUS

A. Suriya Priyanka and T. Kathirvalavakumar
Department of Computer Science, Virudhunagar Hindu Nadar's Senthikumara Nadar College, India

Abstract

In the recent modern world, most of the human beings are affected by diabetes. Diabetes mellitus is a silent killer which may destroy the organs in the body without being noticed. Now many researchers are developing model to predict diabetes at early stage to prevent from other health complications caused by diabetes. Most of the existing works focused on weighted KNN rather than normal KNN to obtain better performance. In this work, instead of using normal KNN or weighted KNN for doing classification, weighted dataset is computed, significant features are identified and KNN is used for predicting type 2 diabetes patients. The weight is calculated for each feature value by calculating probability from its neighbourhood. N2PS pruning algorithm is used for identifying significant features. The oversampling technique SMOTE is applied to balance the dataset when it is imbalanced. Prediction accuracy of the intended method is found as better when weighted dataset is used for prediction.

Keywords:

Prediction, Network Pruning, KNN, Weighted Data, IQR, SMOTE

1. INTRODUCTION

In the present era, all humans are getting afraid of the word "Diabetes" since it leads to acute health complications like heart ailments, kidney disease etc. Preventing and prediction of diabetes is a challenging task. This disease can be controlled but cannot be cured. So, most of the researchers are giving attention to evolve model for predicting diabetes using machine learning approaches.

Jun Li et al. [1] have established a non-invasive diabetes risk prediction model by combining the features of ML and deep learning algorithm based on tongue features fusion [1]. Hafiz et al. [2] have used electronic health records of Saudi hospitals to predict diabetic patients. They have accustomed Logistic Regression (LR), Support Vector Machine (SVM), Decision Tree, Random Forest (RF) and ensemble Majority Voting (EMV) for classification. To increase the prediction accuracy, they have used feature permutation and hierarchical clustering to remove the undesirable features. [2]. Jobeda et al. [3] have applied different ML algorithms and neural network model on PIDD for classification. They selected features using Pearson's correlation approach and ascertained that the NN model yielded finer classification precision [3]. Jayroop et al. [4] classified the PIDD dataset using KNN, LR, Gaussian NB, and SVM-RBF. They have put out a comprehensive remote monitoring framework that uses smartphones, wearable technology, and personal health gadgets to automatically predict diabetes risk. They have deduced that SVM produced better result only after feature scaling, imputation, feature selection and augmentation using SMOTE [4]. Haohui et al. [5] combined ML classifiers with a patient network model to create a prognosis model using data from the Australian CBHS health funds company. A patient network was created by the authors using a bipartite graph, and they used patient characteristics and network properties to make predictions. [5]. Umair et al. [6] have put forward a MLP based model for classification and deep learning based Long Short-Term Memory (LSTM) for prediction. The writers have concluded that fine-tuned MLP and LSTM produced higher performance [6]. Tawfik et al. [7] have developed an competent medical decision system for diabetes prediction postulated on Deep Neural Network (DNN) over the dataset collected from Frankfurt Hospital, Germany. The authors have used various ML models over the dataset and they have found that DNN produced superior performance [7]. Victor et al. [8] have designed an e-diagnosis system, Internet of Medical Things (IoMT), based on ML algorithms. PID dataset is used for the model. They have employed k-means clustering, PCA and importance ranking methods for feature selection. They have concluded that Navie Bayes works well on fine-tuned selection of features, while RF works satisfactorily with more features [8]. Chollette et al. [9] have created a robust machine learning framework called the twice-Growth Deep Neural Network (2GDNN) to forecast diabetes mellitus using the PIDD and dataset from the Medical City Hospital's (LMCH) laboratory. For feature selection and imputation, they have embraced polynomial regression and spearman correlation [9]. Shamreen et al. [10] have used Logistic Regression, Gradient Boosting, Decision Trees, Extra Trees, RF and light gradient boosting machine (LGBM) on PIDD and have obtained better classification accuracy in the LGBM classifier [10]. Umm et al. [11] have produced an ensemble predictive model for diabetes mellitus using data on diabetes and 17 variables gathered from the UCI repository. High grade classification accuracy was accomplished by the authors using Ada Boost, Bagging, and Random Forest ensemble approaches to forecast diabetes mellitus. [11]. Koushik et al. [12] have employed a ML technique to predict diabetes mellitus over PID dataset. The authors have identified various feature subsets of PID using different feature selection methods and applied ten different classifiers on PIDD along with its sub-datasets. They came upon that the Generalized Additive Model utilizing LOESS (G-AMLOESS) was the finest classifier and that the most salient characteristics for improved classification were age, BMI, DBF, and glucose. [12]. Salliah et al. [13] have predicted diabetes mellitus for the clinical dataset gathered from Bandipora during the years April 2021 – Feb 2022. They have used various ML models and have concluded that RF classifier brought about greater accuracy than other classifiers such as Multi-layer Perceptron, Support Vector Machine, Gradient Boost, Decision Tree and Logistic Regression [13]. Yifan et al. [14] have applied five different ML models to do diabetics prediction for the dataset with lifestyle-related variables. For the examination, the National Health and Nutrition Examination Survey (NHANES) database is utilized. The AIC forward propagation algorithm is used to extract features, while SMOTE-NC is used to balance the data samples. The authors have found that Cat Boost model has produced good result [14]. Juyoung et al. [15] have developed a high-performance diabetes prediction model using XGBSE (XGBoost

Survival Embedding) algorithm with threshold adjustment over the electronic medical records collected during the period July 2009 to April 2019 at Seoul, Korea. They have compared the outcomes procured from the classification models LR, DT, RG, XGBoost and Cox regression [15].

This study develops a prediction model for type2 diabetes. The experiment uses the Pima Indians Diabetes (PID) dataset, the diabetes type dataset, and the diabetes categorization dataset. Pre-processed data helps the prediction model perform better. For feature selection, the N2PS pruning [16] method is employed. The weighted data for the chosen features is determined by the data's likelihood of occurring. KNN is used for classifying the weighted data. The remaining section of the work is organized as follows: Materials and Methods are discussed in Section 2, Experimental resultants are bandied in section 3 and conclusion of the proposed work is described in section 4.

2. MATERIALS AND METHODS

The PID dataset is amassed from UCI machine learning library, and diabetes type and diabetes classification datasets are collected from dataworld. The feature details of each dataset are shown in the Table.1-Table.3. Variations in the range of values in the feature and missing values in the dataset may impact the prediction accuracy. So missing values in the diabetic patients data and non-diabetic patients data in the dataset are replaced separately by the mean value of the corresponding feature. Data is normalized by dividing the current maximum value of the relevant patient type attribute.

A feedforward neural network with a single hidden layer identifies important elements in the dataset. For network training, the backpropagation algorithm is employed. N2PS pruning algorithm is used to identify the significant traits of the dataset from the trained network. The dataset is transmuted into a weighted dataset by assigning a weight to every important feature depending on how likely it is that the data will occur in its immediate vicinity. KNN classification algorithm is habituated to presage the diabetic patient from the weighted dataset.

Table.1. PID dataset

Feature	Description of feature
Pregnancies	The number of pregnancies
Glucose	The amount of plasma glucose in an oral glucose forbearance test
Blood Pressure	Diastolic Blood Pressure(mm Hg)
Skin Thickness	Skin fold thickness of the triceps (mm)
Insulin	2h serum insulin(mu U/ml)
BMI	Weight in kilograms/(height in meters)^2 is the body mass index.
Diabetes Pedigree Function	Diabetes Pedigree Function
Age	Age
Outcome	Class (0 or 1)

Table.2. Diabetes type dataset

Feature	Description of feature
Age	Age of Patient
BS_Fast	Blood sugar while fasting
BS_PP	Blood sugar 90 minutes after a meal
Plasma_R	Plasma glucose test randomly taken at anytime.
Plasma_F	Plasma glucose test usually taken in the morning or 8 hours after a meal
HbA1C	
Type	Type of Diabetes
Class	Boolen Yes/No

Table.3. Diabetes-classification dataset

Feature	Description of feature
Patient Number	Identity of the patient
Cholesterol	
Glucose	
HDLChol	
Chol-HDL ratio	
Age	
Gender	
Height	
Weight BMI Systolic_bp Diastolic_bp Waist Hip Waist_Hip_ratio Diabetes	Body Mass Index (weight in kg/(height in m)^2) Class (0 or 1)

2.1 MISSING VALUES REPLACEMENT

Researchers employ standard datasets from well-known sources to support or validate their suggested methodologies, although other researchers gather data for their work independently. There are certain datasets in the repositories that include missing values. The data sample with missing values may be eliminated if the dataset is very large and the majority of the samples reflect the same features because it has no bearing on the dataset's outcome. But samples available in the diabetic dataset are less in its size. So it is mandatory to restore the missing values in the samples.

This work replaces the corresponding missing values by calculating the mean value of each feature independently for each class.

2.2 NORMALIZATION

Normalization is done for the diabetic and non-diabetic patients separately by dividing each feature value by the utmost value of the corresponding feature of the class.

2.3 OUTLIER DETECTION

An outlier is typically defined as a sample that is not like the rest. The model carryout better when the outlier is found and eliminated. Here, Interquartile Range (IQR) outlier detection method is used for all the three datasets. Q1 is the value which marks the first 25% of the distribution ends and Q3 is the value which marks the 75% of the distribution ends. IQR is the range of values between the value Q1 and the value Q3: IQR = Q3 – Q1. The data indicates which fall below Q1 – 1.5 * IQR or above Q3 + 1.5 * IQR are outliers.

2.4 FEATURE EXTRACTION

significant features are found using the N2PS [16] pruning procedure after a single hidden layer has been trained using the conventional backpropagation technique.

2.5 N2PS

A feedforward neural network with one hidden layer is taken into consideration. In this architecture, every neuron's output in the input and hidden layers is coupled to every other neuron in the output and hidden layers, respectively. The sum of inputs for each neuron is calculated. The weighted sum (s_i) for the i^{th} neuron in the input layer is calculated as follows.

$$p = \sum_{i=1}^{m} | f(tx_{ip}) + w_{ij} | \tag{1}$$

where f is a linear function as $f(x)=x$ and $tx_{ip} = \sum_{p=1}^{n_p} x_{ip}$, i represents input neuron, m represents number of hidden neurons, n_p represents number of samples in the dataset, x_{ip} represents input value for the i^{th} neuron of the p^{th} sample, w_{ij} represents the weights of the i^{th} input neurons connected with the j^{th} hidden neuron.

The significance of the input neuron is decided by Eq.(2) and the insignificant input neurons are pruned [16].

$$u_i = \begin{cases} \text{insignificant}, & \text{if } s_i \le \alpha \\ \text{significant}, & \text{otherwise} \end{cases} \tag{2}$$

where u_i represents i^{th} input neuron, n represents number of input neurons.

2.6 WEIGHTED VALUE

All values under every extracted significant feature are sorted individually. Weight for every data under the significant features is calculated based on its k-neighbor values of the corresponding feature. Value for k can be selected by trial. If k is odd then ⌜k/2⌝ data are considered above the data in the column otherwise same number of data is selected from above and below the data.

$$P(x_i) = \frac{\text{No. of neighbors belonging to the same class as } x_i}{\text{Size of neighbors}} \tag{3}$$

$$\text{Weighted value of } x_i = x_i \times P(x_i) \tag{4}$$

where x_i represents the i^{th} value of a feature and probability $P(x_i)$ represents the weight of i^{th} value of the feature.

2.7 IMBALANCE RATIO

A dataset is called imbalanced if the data of each class in the dataset is distributed unevenly. Using the following formula, the dataset's imbalance ratio (IR) is determined.

$$IR = \frac{\text{Number of patterns in the majority class}}{\text{Number of patterns in the minority class}} \tag{5}$$

2.8 OVERSAMPLING

A KNN based Synthetic Minority Oversampling Technique SMOTE is used for oversampling the imbalanced data. This SMOTE working principle is presented below.

1. Select the minority class set A from the given dataset.
2. k = No. of samples of majority class / No. of samples of minority class.
3. For each sample $x \in A$
4. Find Euclidean distances between x and every other sample in the set A.
5. Select k-nearest neighbors of x, namely A_1
6. For each sample y in A_1, generate k synthetic samples
7. Every synthetic sample x' is generated by

$$x' = x + \text{rand}(0,1) \times | x - y_s | \text{ , where } s=\{1,2,\ldots k\}$$

2.9 ACCURACY

The prediction model is assessed using the confusion matrix shown in Table.4. Accuracy of the model is calculated using the following formula.

$$\text{Accuracy} = \frac{TN + TP}{TN + TP + FP + FN} \tag{6}$$

Table.4. Confusion matrix

	Positive(1)	**Negative(0)**
Positive(1)	True Positive (TP)	False Negative (FN)
Negative(0)	False Positive (FP)	True Negative (TN)

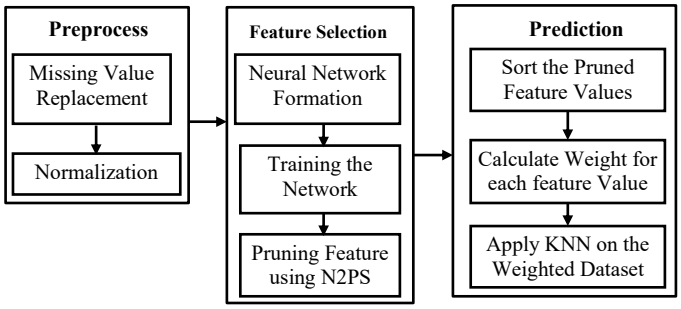

Fig.1. Procedure of the Proposed Work

3. EXPERIMENT RESULTS

The mean value of the relevant feature with the same class is used to restore the missing values of the PID dataset's glucose, blood pressure, skin thickness, insulin and BMI, as well as the BS_Fast of the Diabetes type dataset and is shown in Table.5 and Table.6. There is no missing value in diabetes_classification

dataset. After replacement the values under all features are normalized.

Table.5. Missing value replacement - PID dataset

Feature name	Non-diabetic data		Diabetic data	
	No. of missing	Replacement value	No. of missing	Replacement value
Glucose	3	109.98	2	141.26
Blood pressure	19	68.18	16	70.82
Skin thickness	139	19.66	88	22.16
Insulin	236	68.79	138	100.34
BMI	9	30.30	2	35.14
Diabetes pedigree function	0	-	0	-
Age	0	-	0	-

Table.6. Missing value replacement – Diabetes type dataset

Feature Name	Non-diabetic data		Diabetic data	
	No. of Missing	Replacement Value	No. of Missing	Replacement value
Age	0	-	0	-
BS_Fast	134	16.30	0	-
BS_PP	0	-	0	-
Plasma_R	0	-	0	-
Plasma_F	0	-	0	-
HbA1C	0	-	0	-
Type	0	-	0	-

Outliers of the datasets are detected using IQR method. The outlier detection is done for diabetic and non-diabetic patient separately. In PID dataset, 1 outlier is detected for non-diabetic patient and 4 outliers are detected for diabetic patients. In Diabetes type dataset, 1 outlier is detected for non-diabetic patients. No outlier is spotted in diabetes classification dataset for both diabetic and non-diabetic patient. The imbalance ratio of the dataset is shown in the Table.7.

The above Table.7 shows that the IR value of diabetes classification dataset is high. It leads to apply SMOTE oversampling technique on the diabetes classification dataset to overcome the class imbalance problem.

The considered architecture of the neural network is 9-11-1 for PID Dataset and diabetes classification dataset and 7-11-1 for diabetes type dataset. In the input layer, one bias neuron is included. The number of hidden neuron is selected based on trial and error. The output and hidden layers make use of a sigmoidal activation function. Random integers from the range [-1 1] are used to initialize the weights between the input layer and hidden layer as well as between the hidden layer and output layer. The network is trained upto 10,000 epochs and λ value is set by trial and error as 0.09 for diabetes classification and diabetes type datasets and 0.01 for PID dataset. After 25 different trials, network training error is 0.000998723 for diabetes classification and diabetes type datasets and 0.04298 for PID dataset. The measured error is a mean squared error.

Next, the N2PS pruning algorithm is applied on the trained network for extracting the significance features. Table.8, 9 and 10 shows the obtained s_i value of the features during pruning.

Table.7. Obtained imbalance ratio

Dataset	IR value
PIMA	1.86567
Diabetes_classification	5.5
Diabetes type	1.66

Table.8. Significant value - PID dataset

Feature	s_i Value
No. of times pregnant	2262.393715079818
Glucose	6130.979441303143
Blood pressure	5934.897467085981
Skin thickness	2637.893695983289
Insulin	1328.917737959378
Body mass index	4821.534577048848
Diabetes Pedigree Function	1955.4308894864764
Age	4121.510150165891

Table.9. Significant value - Diabetes_classification dataset

Feature	s_i Value
Glucose	3007.694632995436
Chol-HDL ratio	2578.049737067305
Age	3892.8094836464034
Gender	2930.731910455088
BMI	4177.942208035695
Systolic_bp	4543.610849560891
Diastolic_bp	4877.070845108763
Waist_Hip_ratio	5638.5427319443015

Table.10. Significant value – diabetes type dataset

Feature	s_i Value
Age	4934.574662610504
BS_Fast	6398.637432562369
BS_PP	9205.577579133847
Plasma_R	10091.998821228106
Plasma_F	8490.572698673968
HbA1C	8427.313416180785

The calculated threshold alpha for PID, diabetes-classification and Diabetes type datasets are 2966.0061050572103, 3516.9689266663136 and 6792.667801484225 respectively. Table.8 shows that the significant value s_i of glucose, blood pressure, BMI and age are greater than alpha. Table.9 shows that Age, BMI, Systolic_bp, Diastolic_bp and Waist_Hip_ratio are

greater than the threshold value. Table.10 shows that BS_PP, Plasma_R, Plasma_F and HbA1C are greater than the alpha. Only these features are taken for further processing. The probability for each data in PID, diabetes classification and Diabetes type datasets are calculated by considering size of the neighbor as 11, 4 and 7 respectively. Here neighbor(k) means they occur nearer by their index in the sorted data of the feature and not by their distance.

The different k values are considered and the results obtained are tabulated in Table.11 and Table.12. Optimal k value is taken for probability calculation. Sensitivity tells the correct disease status of the patient. An example for probability calculation is given in the Table.13. The Table.14 shows that when number of neighbors for calculating weight of the dataset is decreased, the classification accuracy of the dataset is getting increased.

This study shows that the classification accuracy of the oversampled dataset is higher than the weighted data of the diabetes classification dataset. But before oversampling, only 84.61539% of accuracy is obtained for the dataset.

Table.11. Test Ability Measurement - PID dataset

k	Sensitivity	Specificity
7	98.68421	1.3157895
9	98.68421	1.3157895
11	99.34211	0.65789473
13	96.05263	3.9473681

Table.12. Test ability measurement - Diabetes classification dataset

k	Sensitivity	Specificity
4	99.206345	0.79365087
5	98.4127	1.5873017
7	97.61904	2.3809524
8	95.2381	4.7619047

Table.13. Sample weight calculation

Feature Name	S. No	Feature Value x_i	Target Class Value	Neighbors	# of data having same diagnosis category of x_i	Probability value	Weighted feature value
Glucose	1	0.417085	0	2,3,4,5,6,7,8	6	0.857143	0.357501
	2	0.422111	0	1,3,4,5,6,7,8	6	0.857143	0.361809
	3	0.497487	0	1,2,4,5,6,7,8	6	0.857143	0.426417
	4	0.502513	0	1,2,3,5,6,7,8	6	0.857143	0.430725
	5	0.592965	1	2,3,4,6,7,8,9	1	0.142857	0.084709
	6	0.59799	0	3,4,5,7,8,9,10	4	0.571429	0.341709
	7	0.688442	0	3,4,5,6,8,9,10	4	0.571429	0.393395
	8	0.693467	0	3,4,5,6,7,9,10	4	0.571429	0.396267
	9	0.959799	1	3,4,5,6,7,8,10	2	0.285714	0.274228
	10	0.969849	1	3,4,5,6,7,8,9	2	0.285714	0.2771

Table.14. Diabetes _classification

	Diabetic	Non-Diabetic
Diabetic	53	0
Non-Diabetic	1	98

Table.15. Obtained sensitivity for different k values

k	Sensitivity	
	PIMA	Diabetes_classification
7	98.68421	98.4127
8	98.68421	98.4127
9	98.68421	99.206345
10	98.68421	99.206345
11	99.34211	99.206345
12	99.34211	99.206345
13	99.34211	99.206345
14	98.02631	99.206345
15	99.34211	99.206345
16	98.02631	98.4127
17	98.02631	99.206345
18	98.02631	99.206345
19	98.02631	99.206345
20	95.39474	98.4127

Table 16. Confusion Matrix - PID dataset

# of neighbors for calculating weight of the data	KNN Classifier k- value	Classification accuracy
5	11	98.4127%
7	25	97.61904%
8	17	95.2381%
4	9	99.206345%

Table17. Confusion Matrix - Diabetes_classification dataset

	Diabetic	Non-Diabetic
Diabetic	60	0
Non-Diabetic	1	65

Table18. Confusion Matrix - Diabetes type dataset

	Diabetic	Non-Diabetic
Diabetic	76	0
Non-Diabetic	2	124

Table19. Analysis of classification accuracy

Authors	Dataset	Handling Missing Values	Feature Selection	Classifier	Accuracy
[3]	PIMA	Mean value of the feature.	Pearson's correlation matrix	LR KNN SVM NB DT RF AB	76.82% 75.10% 76.82% 75.53% 74.24% 74.96% 73.96%
[4]	PIMA	MICE approach with the decision tree regression	Any one of Chisquared test, Extra Trees, Lasso	KNN LR NB SVM - RBF	79.80% 73.30% 73.10% 83.20%
[6]	PIMA	Not Used	Not Used	LR RF Fine-tuned MLP	73.05% 77.4% 86.083%
[8]	PIMA	median value of the feature	PCA with KMeans	J48 DT RF NB	74.78% 79.57% 78.67%
[10]	PIMA,LMCH	Polynomial Regression	1) Samples with missing values are removed. 2) Applied Spearman correlation	2GDNN	97.931%
[17]	PIMA		N2PS Pruning Algorithm	KNN	95.92%
Proposed Model with weighted data	PIMA, Diabetes type	Mean value of the feature of the corresponding class	N2PS Pruning Algorithm	KNN	99.34% 99.21%

4. CONCLUSION

The proposed method predicts Type 2 diabetes. The model encompasses missing value replacement, normalization, extracting significant features, weight calculation based on probability for each feature value and prediction. Significant features are identified and extracted using N2PS pruning algorithm. Weighted values are calculated for the extracted features based on the occurrence of the same class in its neighbour. KNN classifier is used for prediction. Experimental results show that the proposed model using weighted data produces better accuracy in predicting diabetes than the weighted KNN.

REFERENCES

[1] J. Li, P. Yuan H. Fu and J. Xu, "A Tongue Features Fusion Approach to Predicting Prediabetes and Diabetes with Machine Learning", *Journal on Biomedical and Information*, Vol. 115, pp. 103693-103705, 2020.

[2] H.F. Ahmad and A. Alhumam, "Investigating Health-Related Features and their Impact on the Prediction of Diabetes using Machine Learning", *Applied Science*, Vol. 11, pp. 1-18, 2021.

[3] J.J. Khanam "A Comparison of Machine Learning Algorithms for Diabetes Prediction", *ICT Express*, Vol. 7, pp. 432-439, 2021.

[4] J. Ramesh and A. Sagahyroon, "A Remote Healthcare Monitoring Framework for Diabetes Prediction using Machine Learning", *Healthcare Technology Letters*, Vol. 8, pp. 45-57, 2021.

[5] H. Lu, S. Uddin, F. Hajati and M. Khushi, "A Patient Network-based Machine Learning Model for Disease Prediction, The Case of Type 2 Diabetes Mellitus", *Applied Intelligence*, Vol. 52, pp. 2411-2422, 2022.

[6] U.M. Butt, S. Letchmunan and A. Baqir, "Machine Learning Based Diabetes Classification and Prediction for Healthcare Applications", *Journal on Healthcare Engineering*, Vol. 2021, pp. 1-17, 2021.

[7] T. Beghriche, M. Djerioui, "An Efficient Prediction System for Diabetes Disease Based on Deep Neural Network", *Complexity*, Vol. 2021, pp. 1-14, 2021.

[8] V. Chang, J. Bailey and Z. Sun, "Pima Indians Diabetes Mellitus Classification based on Machine Learning (ML) Algorithms", *Neural Computer Applications*, Vol. 35, pp. 16157-16173, 2023.

[9] C.C. Olisah, L. Smith, "Diabetes Mellitus Prediction and Diagnosis from a Data Preprocessing and Machine Learning Perspective", *Computer Methods and Programs in Biomedicine*, Vol. 220, pp. 1-12, 2022.

[10] B.S. Ahamed, M.S. Arya and V. Nancy, "Prediction of Type-2 Diabetes Mellitus Disease using Machine Learning Classifiers and Techniques", *Frontiers Computer Science*, Vol. 4, No, 1 pp. 1-5, 2022.

[11] U.E. Laila, F. Khan and W. Taekeun, "An Ensemble Approach to Predict Early-Stage Diabetes Risk using Machine Learning: An Empirical Study", *Sensors*, Vol. 22, No. 2, pp. 1-15, 2022.

[12] K.C. Howlader and Mohammed S. Islam, "Machine Learning Models for Classification and Identification of Significant Attributes to Detect Type 2 Diabetes", *Health Informatics, Information Science, and Systems*, Vol. 10, No. 3, pp. 1-13, 2022.

[13] S.S. Bhat, V. Selvam, G.A. Ansari and M.H. Rahman, "Prevalence and Early Prediction of Diabetes using Machine Learning in North Kashmir: A Case Study of District Bandipora", *Computational Intelligence and Neuroscience*, Vol. 2022, pp. 1-23, 2022.

[14] Y. Qin, J. Wu, W. Xiao, K. Wang and Z. Ren, "Machine Learning Models for Data-Driven Prediction of Diabetes by Lifestyle Type", *International Journal of Environmental Research and Public Health*, Vol. 19, No. 1, pp. 1-13, 2022.

[15] J. Shin, J. Lee and H.S. Kim, "Improving Machine Learning Diabetes Prediction Models for the Utmost Clinical Effectiveness", *Journal of Personalized Medicine*, Vol. 12, pp. 1899-1897, 2022. 2022.

[16] M. Gethsiyal Augasta and T. Kathirvalavakumar, "A Novel Pruning Algorithm for Optimizing Feed Forward Neural Network of Classification Problems", *Neural Processing Letters*, Vol. 34, pp. 241-258, 2011.

[17] A. Suriya Priyanka, T. Kathirvalavakumar and Rajendra Prasad, "Type 2 Diabetes Prediction from the Weighted Data", *MIKE Lecture Notes in Computer Science*, Vol. 2022, pp. 1-12, 2022.