Journal of Computing & Biomedical Informatics
ISSN: 2710 - 1606

Volume 04 Issue 01
2022

Research Article
https://doi.org/10.56979/401/2022/110

Type-II Diabetes Prediction by using Classification and Novel based Method (AWOD)

Aleena Farooq[1,*], Muhammad Kamran Abid[1], Wasif Akbar[1], Hafiz Humza[2], and Naeem Aslam[1]

[1]Department of Computer Science, NFC Institute of Engineering and Technology, Multan, Pakistan.
[2]Department of Computer Science, Air University, Multan, Pakistan.
[*]Corresponding Author: Aleena Farooq. Email: 2k19mscs212@nfciet.edu.pk.

Received: September 29, 2022, **Accepted:** November 22, 2022, **Published:** December 29, 2022.

Abstract: Type II diabetes is the deadliest disease. It must be identified early to be cured. Prediction models for detection systems typically use common parameters that might not be suitable for all individuals with various health conditions. As a result, this study suggests a way for diabetes type II prediction using variables that reflect individual health issues. More specifically, this work proposes a unique prediction method called Average Weighted Objective Distance (AWOD) based on the idea that the person has a variety of health states coming from a variety of individual characteristics, which is a need for an efficient prediction model. Using information gain to expose significant and inconsequential individual components with varying priorities, denoted by distinct weights (AWOD), modifies Weighted Objective Distance (WOD). To calculate AWOD, the data set is split into a training set and a testing set. The training set is utilized to establish all significant thresholds and constant values. In particular, binary classification issues with a small dataset are developed by AWOD. Two open-source datasets, Pima Indians Diabetes (First Dataset) and Mendeley Data for Diabetes (Second dataset) are examined to validate the suggested methodology. With machine learning-based prediction techniques like (K-Nearest Neighbors, Support Vector Machines, and Random Forest), the prediction performance for both datasets is compared, including statistical measures: Accuracy, Sensitivity, and specificity. The AUC-ROC curve graph reveals how well the model can differentiate across classes for all ML classifiers. The comparison findings demonstrated that the ML classifiers resulted in poor accuracy for the first Dataset. Although better for the second Dataset, the proposed method had greater accuracy than other machine learning-based methods, with 95.26 percent and 99.01 percent for Datasets I and II, respectively.

Keywords: Diabetes; Disease prediction; Machine learning; Objective distance; Weighted factors.

1. Introduction

Diabetes Mellitus (DM) is a group of metabolic diseases characterized by insufficient insulin production and action. Diabetes mellitus (DM) is one of the most common endocrine diseases, affecting approximately 200 million people worldwide. Diabetes is expected to become far more prevalent in the coming years. The evolution of diabetes is tightly linked to several problems, the majority of which are caused by high blood glucose levels. It is well understood that DM encompasses a broad range of pathophysiology disorders (Kavakiotis et al., 2017).

DM is one of the most common diseases among the elderly in the country (World Health Organization, 20202). In 2017, 451 million people worldwide had diabetes, according to the International Diabetes Federation. This figure is expected to rise to 693 million people over the next 26 years. Although the precise cause of diabetes mellitus is unknown, researchers believe that environmental and genetic factors are involved. While it is incurable, it can be controlled with medications and treatments. Diabetes patients face additional health risks, such as cardiac arrest and organ damage. Early detection and treatment of diabetes can also help to avoid complications and lower the risk of serious health issues (Chaki et al., 2020).

If you don't take medicine for diabetes, your body won't create as much insulin as it should. A condition known as high blood sugar occurs when the body retains a large amount of glucose. It may cause serious or life-threatening health problems. There are various types of diabetes, including prediabetes, Type-I diabetes, Type-II diabetes, and gestational diabetes. Type 1 diabetes is an autoimmune disorder. Non-insulin-dependent diabetes, often known as adult-onset hyperglycemia, is type 2 diabetes. However, it's become increasingly common in kids and adolescents in the last twenty years, owing to an increase in fat or obese young people. Type 2 affects around 90 percent of sufferers. Another essential fact is that diabetes is the seventh leading cause of death. To our knowledge, type 1 diabetes cannot be prevented. Type-2 diabetes, on the other hand, can be discovered early and properly managed. Type-2 diabetes is mentioned in a lot of research studies (Chaki et al., 2020; Sethi, 2020).

One in ten persons, or 537 million people between the ages of 20 and 79, have diabetes. By 2030, this number is projected to reach 643 million, and by 2045, it will reach 783 million. Diabetes will be responsible for 6.7 million deaths by 2021, or one every five seconds. Diabetes has cost the United States economy at least USD 966 billion in healthcare costs over the last 15 years, representing a 316 percent increase. Impaired Glucose Tolerance (IGT) affects 541 million adults and puts them at risk of developing type 2 diabetes (*IDF Diabetes Atlas | Tenth Edition*, n.d.).

Artificial intelligence (AI) advancements, particularly in machine learning and computer vision, make it feasible to create apps that automate processes requiring intelligent human behavior, knowledge, and adaptability, thus bringing answers to real-world problems like diabetes control (Sowah et al., 2020).

Although blood sugar levels have historically been the primary focus of diabetes therapy, various indicators, including non-esterified fatty acids (NEFA), have been linked to type 2 diabetes (T2D). Patients with T2D are known to have high levels of NEFA, which is linked to increased insulin sensitivity and poor glucose clearance rates.

The presence of oleic acid (OA) and glucose were then determined. Because these two examples (OA-CoA and OA) are two of the most commonly used markers in clinical tests, they were chosen as NEFA of interest (Şahin et al., 2018).

The frequency of hypertension is expanding worldwide, with 600 million cases in 1980 and 1 billion in 2008. Despite frequently going unnoticed, hypertension is one of the major risk factors for cardiovascular disease (CVD). Fifty-one percent of deaths from stroke and 45 percent from heart disease are attributed to hypertension. According to the World Health Organization (WHO), complications from hypertension will cause the deaths of 9.4 million people annually, and if no action is taken, this number will increase. With a short amount of data, machine learning-based systems can soon fail to classify various people. This paper provides a binary classification technique based on distance measurement to forecast the occurrence of type 2 to address this issue. A wide range of health conditions leads to various health diagnoses and treatment choices, according to the diagnostic process theory used by healthcare professionals (Kjeldsen et al., 2014).

WOD (weighted objective distance) is a novel measurement method that prioritizes significant factors only on an individual basis that help predicts chronic diseases and control hypertension. AWOD (average weighted objective) is the moderation of WOD.AWOD method involves significant and insignificant factors supporting the prediction of type 2 diabetes depending on an individual's health condition. Because clinical data is uncommon and rare, both techniques are intended specifically for building prediction models from limited datasets. WOD requires pre-defined thresholds and constant values, which a healthcare expert must provide to the individual health diagnosis data to determine each factor's weight. This technology cannot be implemented in any dataset without personal health diagnosis records. As a result, AWOD was created to address WOD's shortcomings. AWOD, in particular, generates them all straight from the training dataset (Chaising et al., 2020; Nuankaew et al., 2021).

Thus this study contributes to the improvement of the accuracy rate of the AWOD-based method for type 2 prediction applied to Pima Indians Diabetes (Dataset 1) and Mendeley Data for Diabetes (Dataset 2), as previous work had an approximate error of 6.78% for Dataset 1 and 1.05% for Dataset 2 for inaccurate prediction. We need to modify AWOD's accuracy or performance for better prediction results. After implementing AWOD, we will implement different machine-meaning techniques on the same Dataset to compare the results, including KNN, SVM, RF, and DL (ANN). So other machine learning techniques are also applied, including the statistical metrics: Accuracy, Sensitivity, and Specificity. AUC-ROC curve graphs reveal how well the model can differentiate across classes for all ML classifiers. The main focus is on improving the AWOD-based method by determining the acceptable level, expected level, and constant values that cause inaccurate prediction or error. The individuals were predicted in the wrong class.

The next part of this paper describes the literature review in section 2. Section 3 tells about the research methodology. Section 4 shows experimentation with a modified AWOD-based method. Section 5 shows the results and discussions. Section 6 gives the conclusion of this paper.

2. Literature Review

In most cases, while creating a model for diabetes prediction, the common set of data is used. The risk factors are considered to account for the variability of individuals. To choose relevant factors required for good prediction, novel extracting features approaches are proposed. The most significant risk variables are extracted from the entire Dataset based on the attribute ratings. (Fitriyani et al., 2019; Islam et al., 2020). For building the model, studies exist that consider factors and symptom-oriented variables (Aofa et al., 2018).To improve prediction results, complex characteristics containing numerous category attributes are also used (Sabariah et al., 2015). However, to reduce model complexity, it is usually recommended that the number of input factors is reduced (Le et al., 2021). The common collection of elements is given the most consideration for model construction in prior studies involving early diabetes prediction. In contrast, this study considers an individual's unique collection of factors resulting from various health conditions.

The binary classification approach used distance metrics such as Hamming, Euclidean, Manhattan, and Minkowski distances. In conclusion, an iteration looping design procedure is necessary while constructing neural networks to determine the best possible erection for the network. When designing a neural network, no set rule can be followed too precisely determine the number of hidden neurons or layers that should be used when an insufficient or excessive number of hidden neurons, either underfitting or overfitting, is possible [15], [16]. Both the Euclidean and Manhattan distances contribute to the calculation of the gap that exists between two real-valued vectors. This distance can be calculated using the Pythagorean Theorem in conjunction with the Cartesian coordinates of the locations in question.

In conclusion, the model's applicability was validated using two distinct approaches: first, specific comparisons were made using the results of arbitrarily chosen test cases, and second, the LOOCV of all acquired cases was analyzed [17]. A weighted Chebyshev distance approach was developed to categorize hyperspectral imaging. This paper improves the MS-VT method, and a new way for the supervised classification of HSI called Weighted Chebyshev Distance (WCD) is presented. Both of these methods are spectral similarity-based techniques. According to the optimization results, the optimum performance rates for AVIRIS, BOTS, and KSC data are obtained with p-values of 0.60, 0.55, and 0.59, respectively. As a result, the value of p can be chosen as 0.6, which is nearly the best value for the three data sets examined (Demirci et al., 2015). The objective distance (OD) was first suggested in the customized learning application areas to assess the distance between a student's current competency and the required level to meet learning goals. So research was done to establish appropriate suggestions for seniors at high risk of CVD complications, considering their ability to fulfill personal goals. High blood pressure denotes the current situation, whereas lowered or stable blood pressure denotes the fulfilled objective level after following the prescribed treatment (Chaising & Temdee, 2018). Other research studies used hypertension to determine useful ideas for the elderly. The suggestions help older people regulate their lifestyles to protect against CVD complications (Chaising et al., 2021).

Many strategies have been developed to deal with feature selection approaches. Still, there are issues with attribute reduction, so a hybrid technique for feature selection was created to increase speed. The study focuses on document preparation, which includes selecting the most essential feature using information gain based on entropy and calculating the class probability to minimize the feature's high dimensionality (Patil & Atique, 2013). WOD (Chaising et al., 2020), A variant of the original objective distance, is created to make better class categorization of aged adults escorted by high Bp alongside prioritizing single criteria. WOD expresses priority using variable quantity or weight based on distance. The first concern is expressed by weights determined from the information gain theory for important weight objective distance components. AI and ml methods had widely employed in the untimely prediction of diabetes for ten years, as previously indicated. Literature has demonstrated that when compared to established statistical procedures, they produce beneficial results (Choi et al., 2019). Decision tree, Naïve Bayes, and SVM were used as models to predict diabetes with maximum accuracy. Naïve Bayes performs better than other algorithms, with maximum accuracy of 76.50 % (Sisodia & Sisodia, 2018). In another study on predicting diabetes, the basic goal of this paper was to apply three machine learning algorithms. KNN, SVM, and random forest for diabetes prediction with comparing their performance. The random forest algorithm gives the highest accuracy, with 74.47 % of others. The algorithms with normalization give better prediction results (Raut et al., 2021). The deep neural network was used to diagnose diabetes by training its attributes in a fivefold and tenfold cross-validation fashion. A confusion matrix was made for both five-fold and ten-fold cross-validation. Five-fold validation gave 98.35 % accuracy, which shows a better result than ten-fold. This system supports medical staff and general human beings(Islam Ayon & Milon Islam, 2019). In prior work, techniques for K-nn, random forest, SVM, naive Bayes, logistic regression, and gradient booting were used to create machine-learning models.

Random forest predictive learning-based model achieves 88.76% accuracy; however, the gradient boosting predictive learning-based model and random Forest predictive learning-based model each achieve 86.28% predictive ability [26]. To predict blood glucose, LTSM and Bi-LTSM layer were used [27]. When it comes to precision, the comparison between the deep learning model and the rough set theory shows that the deep learning model is significantly superior to the rough set theory [28]. It was an ABC-DNN-based diagnostic model that was used. ABC stands for an artificial bee colony, and DNN stands for a deep neural network. Compared to the DNN technique, the ABC-DNN algorithm achieved an accuracy of 94.74%. In addition, the tenfold method was utilized here for diabetes prediction (Srivastava et al., 2021).

To get around problems, diabetes patients' risk levels are classified using DLMNN and naive Bayes (NB). The disease prediction model DLMNN gave the highest accuracy regarding a recall, F-measure, and precision or computational analysis. Also, the NB classifier examined the risk analysis phase, which provides proficiency in disease prediction and analysis(Appavu alias Balamurugan & Salomi, 2020). Another previous work used hidden layers of DNN and dropout regularization (Zhou et al., 2020). Using machine learning algorithms and the PIMA dataset, a study predicted diabetes. The study found that DL works best with promising extracted features. DL's 98.07 percent accuracy can improve the automatic prognosis tool. Including omics data for illness and onset prediction can improve DL accuracy [32]. The experimental analysis used the UC Irvine machine learning repository's Pima Indian Diabetes Database (PIDD). Random Forest (RF) outperformed other categorization algorithms with 87.66% accuracy [33]. Prior research developed a DSS for diabetes prediction using ML. Deep learning, support vector machine, and random forest had an accuracy of 76.81%, 65.38%, and 83.67%, respectively. RF was more successful for diabetes prediction than deep learning and SVM [34].

This paper suggests generalizing AWOD to multiple datasets and disorders, including two diabetes datasets. AWOD prioritizes using information gained. Information gain quantifies entropy decreases [35] and identifies irrelevant data [36]–[39], including individual factors (*Pima Indians Diabetes Database | Kaggle*, n.d.). The information gain method is used in AWOD to identify significant and minor aspects for individuals. The former is characterized as elements that a single person can’t check well and have a discernible impact on their well-being, while the later are those that individuals can regulate (Chaising et al., 2020). The constraint of WOD is fewer major individual characteristics harm that classification performance. This study assumes that significant and insignificant factors differ for each individual and can be used to build a model that provides better prediction performance with suitable priority settings. As a result, using both major and insignificant individual characteristics, this study suggests a Method based on AWOD for predicting type II diabetes, including the patient's medical history and conditions.

This investigation utilized two publicly available data sets, namely Pima Indians Diabetes (Dataset I) [39] and Mendeley Data for Diabetes (Dataset II) [41], to Validate the prediction method. The suggested AWOD approach can't be compared to the WOD method because these datasets don't need already defined thresholds or unchanging values for mass calculation. Instead, these approaches are compared to pre-existing machine learning-based methods, including K-nearest neighbors, Support vector machines, Random Forest, and Deep learning (DL). K-nearest neighbors calculate the space between the target data and all other data. This helps it find nearby data. Intervals are arranged to find the closest neighbors, and k defines a minimum distance needed to predict results. SVM for classification divides data into two groups. The algorithm builds a line or hyperplane from these two groups. SVM finds the best-balanced hyperplane to classify data. RF is a more effective ensemble method than a decision tree. By averaging output, it can reduce overfitting. RF can reduce a problem's dimension by determining the most important input variables. DL doesn't need human supervision because it uses unstructured, unlabeled data. Learning can be supervised, semi-supervised, or unsupervised. This technique uses an ANN architecture with two or more layers, weights, and biases to imitate the cerebrum. It does this by making human-like decisions.

3. Methodology

The first figure below shows: AWOD calculation, assessment of the average weighted objective distance prediction method, and comparison with other ML techniques are the two primary components that comprise this study's research methodology. In the following paragraphs, the particulars of each process will be discussed.

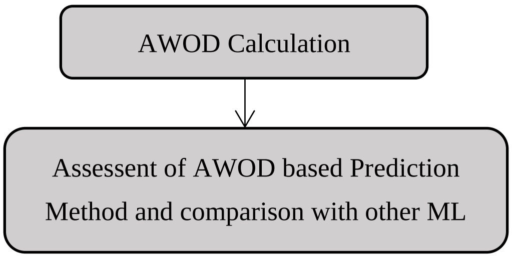
AWOD Calculation
Assessent of AWOD based Prediction Method and comparison with other ML

Figure 1. Methodology diagram

3.1. Attack's traffic AWOD Calculation

The average weighted objective distance technique's idea is rooted in the number of important and unimportant components representing significant influences on the prediction. This principle can be applied to a variety of particular health conditions as well as general diagnostic processes carried out by healthcare providers. Figure 2 illustrates the AWOD principle.

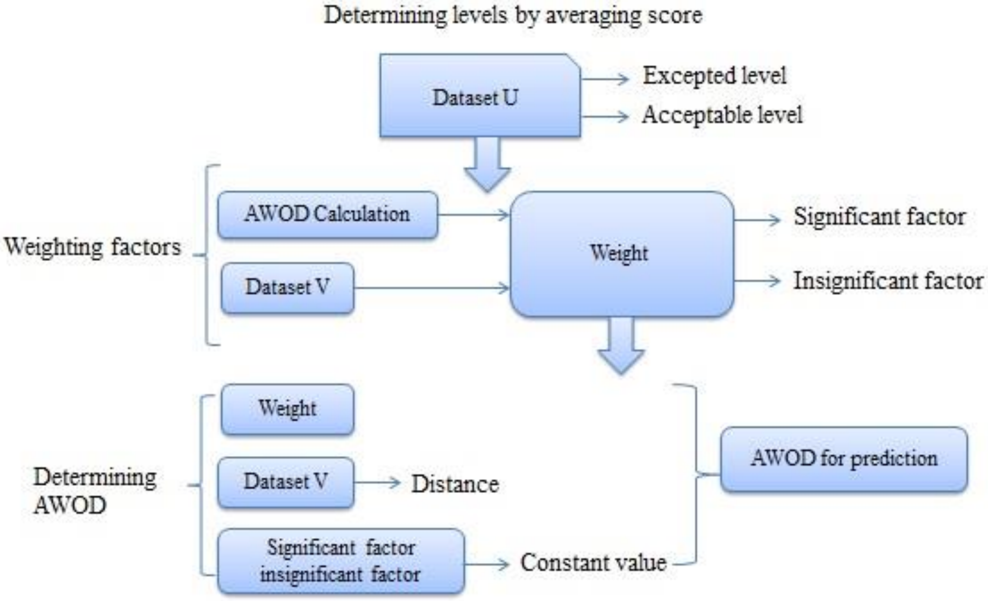

Figure 2. AWOD concept.

3.2 Significant Factors

Significant factors are those factors that are helping or important in the prediction of type 2 diabetes.

3.3 Insignificant Factors

Insignificant factors are those that are unimportant in the Dataset, including and not having much effect in predicting disease.

Figure 2 shows the three steps needed to determine AWOD. Before moving on, determine expected and acceptable weight-calculating levels. The expected level is the health status a person without diabetes should have for each component, while the acceptable level is for each factor. This step calculates both significant and unimportant factor weights. Finally, AWOD is computed to make predictions. The Dataset will be halved, creating datasets U and V. These datasets will be used to resolve expected and acceptable levels for every factor's current and acceptable distance. Dataset U will be used to determine expected, and acceptable levels of each attribute by averaging scores. These two measures can therefore be considered representative of the training set participants. Therefore, these two measures determine the weights of each factor. AWOD and weight factors were calculated using Dataset V. The weighting factor value can represent important and unimportant factors. These are the actual aftermath of every component of each person based on a constant value representing significant and insignificant factors. Same value. Then, we use Dataset V to determine each factor's acceptable and current distances. After that, distances, weights, and a constant value are combined to create a user's AWOD. Because each factor's results are on separate scales, rescaling attributes in [0, 1] requires min-max normalization.

3.4 Pseudocode for Target Class Calculation

The following pseudocode explains the algorithm for determining the target class using the proposed prediction method.

The parameters used in the algorithms are:

N = "total number of patients"

A_j = "value of anticipated level"

B_j = "value of accepted level"

C_j = "current level or distance"

T_n = "total number of attributes"

U = "training set"

V = "testing set"

Step 1: The first step is to read the Dataset and split it into two sets

- Choose 70% of the training set's data randomly
- Choose 30% of the test set's data randomly
- Set a Condition to Stop parameters according to the training set

Step 2: Stop when the parameter reaches an acceptable level, or the condition is False

Step 3: Determine the expected level of each parameter for all samples of positive class

$$A_j = \frac{\sum u_{j+}}{n\, u_{j+}} \tag{1}$$

Step 4: Calculate the average number of every attribute for all negative class samples.

$$B\left(a\right)_{j}=\frac{\sum u_{j-}}{n\,u_{j-}} \tag{2}$$

Step 5: Calculate the acceptable level of every attribute.

$$B_{j}=\frac{A_{j}+B_{(a)_{j}}}{2} \tag{3}$$

End (WHILE)

//Calculate the entropy of the target class for every attribute in the V set.

Step 13: Determine the probability of acceptable (pABj+) and current distance (pabj_)

$$pAB_{j+}=\frac{dAB_{j}}{dAB_{j}+dAC_{j}} \tag{4}$$

$$pAC_{j-}=\frac{dAC_{j}}{dAB_{j}+dAC_{j}} \tag{5}$$

Step 14: IF pAC_{j-} = 0 THEN $E\left(C_{j}\right)$ =0

ELSE

$$E\left(C_{j}\right)=-\frac{pAB_{j+}}{1}*\log_{2}\left(\frac{pAB_{j+}}{1}\right)-\frac{pAC_{j-}}{1}*\log_{2}\left(\frac{pAC_{j-}}{1}\right) \tag{6}$$

End If

End While

//Calculate the entropy of each attribute in the V set.

//Stopping condition is the number of attributes reached.

Step 15: (WHILE) stopping condition is false

Step 16: Calculate the entropy of each attribute $E\left(Ct\right)$

$$E\left(Ct\right)=\sum_{j=1}^{Na}\left[E(C_{j})*\left(\frac{pAB_{j+}+pAC_{j-}}{Tn}\right)\right] \tag{7}$$

End (WHILE)

//Calculate the information gain of the target class in the V set.

Step 18: Calculate the significant progress for every attribute (Sg_{j})

$$Sg_j = \frac{E(C_j)}{Gain(C,\mathrm{t})} \tag{8}$$

Step 19: Calculate the mass of all attributes (W_j)

$$W_j = \frac{Sg_j}{\sum_{j=1}^{Na} Sg_j} \tag{9}$$

Step 20: Calculate the average weighted objective distance (AWOD) for each factor (Dj)

$$D_j = W_j \times | \sqrt{(A_j - B_j)^2} - \sqrt{(A_j - C_j)^2} \, | \tag{10}$$

Step 22: Calculate the normalized (AWOD) for each factor.

$$ND_j = \frac{D_j - D_{\min}}{D_{\max} - D_{\min}} \tag{11}$$

Step 23: CAS expression OF

Step 24: Calculate the average weighted objective distance for every attribute for each individual.

$$AWOD_i = \frac{\sum_{j=1}^{Tn} ND_j}{Tn} \times b \tag{12}$$

3.5 Comparison with ML-based Prediction Methods

The assessment procedure for the suggested technique is depicted in Figure 3. In it, the anticipated and observed classes were used to compare the performance of the predictions in terms of accuracy, sensitivity, and specificity.

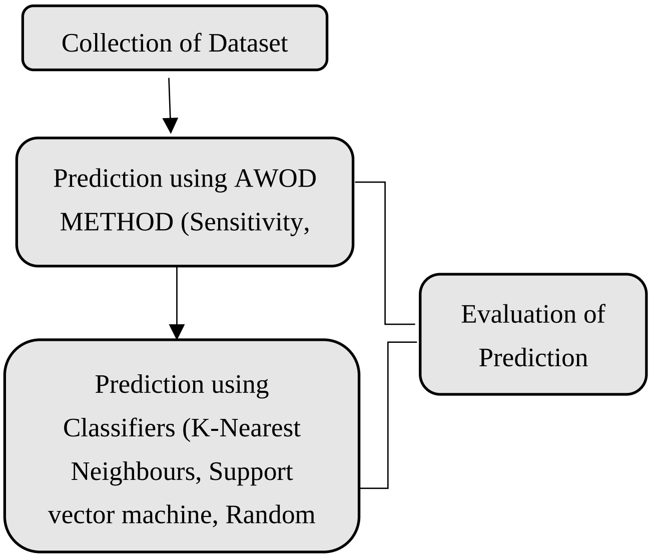
Collection of Dataset
Prediction using AWOD METHOD (Sensitivity,
Prediction using Classifiers (K-Nearest Neighbours, Support vector machine, Random
Evaluation of Prediction

Figure 3. Prediction using proposed method and classifiers.

The actual condition of individuals, specifically the presence or absence of type 2 diabetes, is referred to as the observed class. Using AWOD as a basis, one can predict either the absence (AD) or the presence

(PD) of diabetes type 2, and this is what is meant by the term "predicted class." In addition, the K-nearest neighbour, Support vector machine, Random forest, and Deep learning classifiers are used with all of the initial attributes to evaluate and contrast their performance with that of the proposed technique. The following paragraphs will describe the classifiers that were utilized in the evaluation of the results. The next section will explain the particulars as well as the outcomes of the prediction performance.

3.5.1 K-NN(K-Nearest Neighbor)

The K-Nearest Neighbor algorithm, which uses the Supervised Learning method, is one of the most straightforward examples of machine learning. The K-Nearest Neighbors algorithm assumes that the new case or data is similar to existing cases and places the new case into the category that is the most similar to the already available categories.

3.5.2 Support Vector Machine (SVM)

The Support Vector Machine, or SVM as it is more frequently known, is a well-known Supervised Learning technique that may be used to solve issues requiring classification and regression. But for classification-related problems, machine learning is where it is most commonly used. The Support Vector Machine (SVM) algorithm aims to find the best boundary for classification in an n-dimensional space. This will simplify us to classify freshly obtained data points in the future.

3.5.3 Random Forest (RF)

Random Forest is a well-known machine learning algorithm derived from the supervised learning method of instruction. In machine learning, it can be used to solve problems relating to classification and regression in that order. It is based on ensemble learning, a technique that combines several different classifiers to solve a challenging problem and improve the model's overall performance.

3.5.4 Deep Learning

Deep learning is an automatic learning technique that uses artificial neural networks in conjunction with representation learning. This technique is also known as deep structured learning. Three levels of supervision can be applied to the educational process: full supervision, semi-supervision, or no supervision. ANN(artificial neural network) classification algorithm is used in this work that contains three layers: Input layer, hidden layer, and output layer.

3.6 Feature scaling technique

Feature scaling technique is applied on all classifiers to get better accuracy results. The process of normalizing the many different features is called "feature scaling," It is performed. Real-world datasets almost always contain features that range in size, scope, and the units they are measured in. Because of this, feature scaling needs to be done for machine learning models to comprehend these features on the same scale.

3.6.1 Normalisation

The normalization procedure, also known as min-max scaling, involves rearranging the numbers in a column so that they are contained within a predetermined range that is between 0 and 1. The Scikit-learn function responsible for normalization is referred to as MinMaxScaler.

3.6.2 Standardisation

A different scaling method, standardization or Z-score normalization, rescales the values in a column to exhibit the characteristics of a conventional Gaussian distribution, which has a mean of 0 and a variance of 1.

3.7 Data Collection

Datasets I and II were chosen for the type 2 diabetes prediction study. First from Kaggle and second from Mendeley Data. The first Dataset includes all female Pima Indians over 21 with diabetes from the National Institute of Diabetes and Digestive and Kidney Diseases. The Dataset has 8 factors, including Dpf, Ag, Gl, Bp, St, In, Bm, and Pr. Dataset II was collected in Iraq's Medical City Hospital and Al-Kindy Teaching Hospital. Iraqi hospitals. Ag, Ur, Cr, Hb, Ch, Tg, Hd, Ld, VI, and Bm were used. Only diabetic and non-diabetic data were used to predict diabetes type 2. This Dataset was arbitrarily chosen for this investigation. Tables 1 and 2 lists the diagnostic factor abbreviations for Datasets I and II. We'll abbreviate Dataset I's diagnostic factors in the next calculation demonstration. This demo predicts type 2 diabetes.

Table 1. Description of dataset-1.

Abbreviation	Factor	Detail
Pr	Pregnancies	Number of times pregnant
Gl	Glucose	Plasma glucose concentration a 2 hours in an aral glucose tolerance test
Bp	Blood Pressure	Diastolic blood pressure
St	Skin Thickness	Triceps skinfold thickness
In	Insulin	2-Hour serum insulin
Bm	BMI	Body mass index
Dpf	Diabetes pedigree Function	Diabetes pedigree function
Ag	Age	Patient ages

So, table 2 explains the description of dataset 2, including 10 factors: Age, Urea, Creatinine Ratio, HBA1C, Cholesterol, Triglycerides, HDL Cholesterol, LDL Cholesterol, VLDL, and BMI of patients. Give details of abbreviations used in Dataset 2.

Table 2. Description of dataset-II.

Abbreviation	Factor	Detail
Ag	Age	Patient ages
Ur	Urea	A diamine, chief nitrogenous waste product in humans
Cr	Creatinine Ratio	Parameter to assess Kidneys
Hb	HBA1C	Average blood glucose (sugar) Levels
Ch	Cholesterol	A fatty, waxy substance produced by the liver
Tg	Triglycerides	A type of fat in the blood used for energy
Ld	LDL Cholestrol	Low-density lipoprotein, which is bad cholesterol
VI	VLDL	Very-low-density lipoprotein cholesterol produced in the liver
Bm	BMI	Body mass index

Table 3. Illustration of Dataset I.

	Factors								
Index	Pr	Gl	Bp	St	In	Bm	Dpf	Ag	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1
5	5	116	74	0	0	25.6	0.201	30	0

6	3	78	50	32	88	31	0.248	26	1
7	10	115	0	0	0	35.3	0.134	29	0
8	2	197	70	45	543	30.3	0.158	53	1
…	…	…	…	…	…	…	…	…	…
767	1	93	70	31	0	30.4	0.135	23	0

Only data consisting of diabetic and non-diabetic classes were considered for predicting type 2 diabetes.

Table 4. Illustration of Dataset II.

					Factors						
Index	Ag	Ur	Cr	Hb	Ch	Tg	Hd	Ld	Vl	Bm	Class
0	50	4.7	46	4.9	4.2	0.9	2.4	1.4	0.5	24	0
1	26	4.5	62	4.9	3.7	1.4	1.7	2.1	0.6	23	0
2	50	4.7	46	4.9	4.2	0.9	2.4	1.4	0.5	24	0
3	50	4.7	46	4.9	4.2	0.9	2.4	1.4	0.5	24	0
4	33	7.1	46	4.9	4.9	1	0.8	2	0.4	21	0
5	45	2.3	24	4	2.9	1	1	1.5	0.4	21	0
6	50	2	50	4	3.6	1.3	0.9	2.1	0.6	24	0

For this investigation, 946 records from this Dataset were chosen, including both diabetic and non-diabetic classes. Therefore, an example of the data collected for datasets 1 and 2 is presented in the table above 3 and 4.

The example AWOD computation used with Dataset 1 to determine whether type 2 diabetes exists or not is shown in this section. Applying data no. 1 will exhibit the diabetes mellitus type 2 prediction using the suggested measuring technique. In this study, diabetes mellitus type 2 was identified as PD when present and AD when absent. Below is an example of a computation that was done to get the average weighted objective distance and determine which class the data point number 1 belongs to the random samples taken in the Dataset, By dividing the data in a ratio of 70:30 because of their relatively tiny sizes, U and V. The split ratio percentage denote the fact that the U set accounts for 70% of the data needed to calculate the expected and acceptable levels.

In contrast, weighting factors will be applied to 30% of the data, which refers to the V set. The computation of the expected and permissible amount for the Dpf attribute is shown as an example in Table 3, which was split into two sets to identify levels by average score. A_j(Adpf) and B_j for the dpf element (B_{dpf}) are calculated, which was applied on "U" set from the Dataset. Following the decision, we observe according to our training dataset:

$$\begin{aligned}&\sum u_{Dpf+} = 87.015, n\ u_{j+} = 184,\\&\sum u_{Dpf-} = 58.508, n\ u_{j-} = 90, B\\&\left(a_j\right) = 0.65\left(58.610 / 90\right)\end{aligned} \tag{13}$$

$$\begin{aligned}A_{Dpf} &= \frac{87.015}{184} = 0.472\\B_{Dpf} &= \frac{0.472 + 0.65}{2} = 0.56\end{aligned} \tag{14}$$

As a result, Table 5 displays each element's acceptable and expected amounts for Dataset. Acceptable and expected values of each factor were calculated by the value of the expected level of element (Aj) and the acceptable level of element (Bj). So these two acceptable and expected amounts are important for each factor that will distinguish the significant and insignificant factors in the end. Significant factors are those having value more than zero, and insignificant factors include zero value. The expected level is the health condition level of every element, including the person with no diabetes, and on the other hand, the acceptable level is the health condition degree which is acceptable for a person with no diabetes.

Finding the entropy of the target class is the first step in the procedure mentioned above for determining the weight factor. The target class's equal probability was initially established. The likelihood that AD will represent the positive target class (EP+) and PD will represent the negative target class (EP-) is equal. The (EP+) and (EP-) values were equaled to 4, shown below:

$$EP_{+} = EP_{-} = \frac{8}{2} = 4 \tag{15}$$

$$EP_{+} = EP_{-} = \frac{8}{2} = 4 \tag{16}$$

The fractions of the positive target group (F+) or negative target group (F+) were then calculated regarding each parameter. Following are the fractions of AD (F+) and PD (F-) about each parameter:

$$PF_{+} = \frac{4}{8} NF_{-} = \frac{4}{8} \tag{17}$$

Thus, it was possible to calculate the target class's [E(C)] entropy concerning all factors. E (C) had a value of 1, as seen below:

$$E\left(C\right) = -\frac{4}{8} \times log2\left(\frac{4}{8}\right) - \frac{4}{8} \times log2 \times \left(\frac{4}{8}\right) = 1 \tag{18}$$

Finding each factor's entropy is the second step. The Dpf factor's entropy computation is demonstrated. Both the current distance (dACdpf) and the acceptable distance (dABdpf) for the dpf parameter are calculated. According to third table, V set's current status of the first data for the dpf parameter (Cdpf) is 0.695. Adpf = 0.472 and Bdpf = 0.56 were calculated from the U set in Table 5.1 . The dABdpf and dACdpf, respectively, were equivalent to 0.089 and 0.223.

The next step is calculating the target class's information gain across all variables. Thus, [E(Ct)] was used to calculate the entropy of every factor. E (Ct) had a value of 0.61 as shown below. The weight of each component is decided in the fourth phase. The Dpf factor was used to illustrate how the weights for the various factors were determined.

$$E\left(Ct\right) = 0.74 * \left(\frac{0.18 + 0.78}{8}\right) + 0.79 * \left(\frac{0.30 + 0.75}{8}\right) + 0.52 \tag{19}$$

The target class's information gain [Gain (C, t)] with regard to all factors was then calculated. Gain (C, t) equaled 0.39 as seen below:

$$\text{Gain}\left(C,t\right) = 1 - 0.61 = 0.39 \tag{20}$$

what had significant gain for dpf parameter (SDpf). SDpf was 2.05 after calculation as stated below:

$$Sg_{Dpf} = \frac{0.80}{0.39} = 2.05 \tag{21}$$

The significant gain calculated for all parameters is shown in table 5. The Dpf factor's weight (WDpf) was established. WDpf was set at 0.21, as shown below:

$$W_{Dpf} = \frac{2.05}{1.26+1.72+1.20+0+0+2.25+2.05+1.15} \qquad (22)$$
$$= 0.21$$

The average weighted objective distance for the Dpf attribute (DDpf) was first established, and then the AWOD for all factors was calculated. DDpf was set at 0.05, as shown below:

$$D_{Dpf} = 0.21\times \left| \sqrt{\left(0.472-0.56\right)^2} - \sqrt{\left(0.472-0.712\right)^2} \right| = 0.05 \qquad (23)$$

Dmax and Dmin for all factors were 4.54 and 0, respectively. The weighted objective distance for the Dpf factor (NDDpf) Dmax and Dmin for all factors were 4.54 and 0, respectively. The weighted objective distance for the Dpf factor (NDDpf) was then calculated using a normalized average basis. NDDpf was set at 0.01, as shown below:

$$ND_{Dpf} = \frac{0.05-0}{4.60-0} = 0.01 \qquad (24)$$

It was decided what the AWOD for data no. 1 for all factors (AWOD1) would be. lTn = 5 and hTn = 2 were calculated in this study by looking at the Dataset. The AWOD algorithm's Step 23 states that b = 1 if ND (v=0) < lTn, and since n ND (v=0) = 2 when St and In parameters are taken into account for Data No. 1, b = 1. AWOD1 was set at 0.29, as shown below:

$$AWOD_1 = \frac{0.12+1.2+0.20+0+0+0.11+0.02+0.69}{8} \times 1 = 0.29 \qquad (25)$$

Table 5. Expected and acceptable measure of each parameter for the first Dataset.

Factors	Expected	Variable	Acceptable	Variable
	level	(Aj)	Level	(Bj)
Pr	4	A_{pr}	5	B_{pr}
Gl	114	A_{Gl}	128	B_{Gl}
Bp	71	A_{BP}	72	B_{BP}
St	26	A_{St}	31	B_{St}
In	136	A_{In}	171	B_{In}
Bm	32.9	A_{Bm}	33.6	B_{Bm}
Dpf	0.494	A_{Dpf}	0.853	B_{Dpf}
Ag	28	A_{Ag}	31	B_{Ag}

Table 5 will list each factor’s relative weight, including factors: E (Ci), Si, Wi, Di, and NDi, concerning each factor’s weight which was calculated so the remaining detail was in the table described below.

Table 6. Lists each factor's relative weight.

Factor	Pr	Gl	Bp	St	In	Bm	Dpf	Ag
E(Ci)	0.74	0.79	0.52	0	0	0.96	0.80	0.47
Si	1.29	1.72	1.20	0	0	2.25	2.05	1.15
Wi	0.22	0.31	0.08	0	0	0.17	0.19	0.09
Di	0.65	4.39	0.76	0	0	0.45	0.03	2.58
NDi	0.12	1.2	0.20	0	0	0.11	0.02	0.69

For data no. 1 in Table 5, various weights (Wi) corresponding to significant and insignificant factors are WPr = 0.22, WGl = 0.31, WBp = 0.08, WSt = 0, WIn = 0, WBm = 0.17, WDpf = 0.19, and WAg = 0.09. A weight with a value of 0 indicates an insignificant factor. The weight with a value greater than 0 denotes a major parameter in comparison. As a result, it was decided that the mass of Pr, Gl, Bp, Bm, Dpf, and Ag were important variables. St In's weights were identified as unimportant variables. The first data representative was in the negative class because AWOD1 = 0.29, which was used to determine the target class. As a result, type 2 diabetes can be expected to be present in sample data no 1.

4. Result and Discussion

Both datasets are divided into AD (AWOD = 0) and PD (0< AWOD <1) for the proposed average weighted objective technique for the prediction of diabetes mellitus type 2. For each data sample, the output was differentiated with the perceived value to assess the detection accuracy of the average weighted objective distance technique.

Some AWOD-based type 2 diabetes prediction results are shown in Tables 5 and 6 for Datasets 1 and 2. The tables include expected class values, weights, constants, and AWODs. Type 2 diabetes is predicted by weight. The AWOD value—which represents actual effects on prediction based on significant and insignificant variables—was computed using a constant value. This number represents actual forecasts based on important and irrelevant factors. The data from sample no. 1 depend on Pr, Gl, Bp, Dpf, and Ag. The minors are St, In, and Bm. The PD group, which denotes major parameters, has a value of 1, according to Step 23 of the AWOD method. This means that unimportant attributes are less than the minimum number of attributes affecting the prediction. Sample five's three most crucial variables are Bp, In, and Bm.

The variables Pr, Gl, St, Dpf, and Ag are unimportant. Because there are more unimportant factors than important ones, the constant has a value of 0. The unimportant variables outweigh the number that can influence determining the negative or PD class. This information is contained in the AWOD-based pseudocode step 23. Sample data no. 5, therefore, had an AWOD of 0.00, indicating that it belonged to the AD class (AWOD = 0). We can infer from these two cases that every individual has a distinct set of major and minor traits that can predict diabetes mellitus type 2. So the next table includes the sample of prognosis outcome for the second Dataset using the proposed average weighted objective distance method. Accuracy, sensitivity, and specificity were utilized to gauge how well the AWOD-based technique performed predictions. In terms of the total number of projected positive classes, sensitivity can be defined as the frequency with which the suggested AWOD-based technique properly predicts true positives (TP).

TP represents AD members whose predictions came true. Specificity is the proportion of true negatives among correctly predicted negatives. PD correctly predicted TN residents. The AWOD-based method's accuracy can be evaluated. This evaluation verifies TP and TN predictions. Calculating sensitivity, accuracy, and specificity involves false positives and false negatives. FP means the predicted class is AD, but the perceived group is PD. A false negative indicates that the observed class is AD despite the number of individuals. Table 6 shows the AWOD-based method's accuracy for Datasets 1 and 2. Each type 2 diabetes dataset has 392 records taken. The parameters were TP = 40, FP = 3, TN = 86, and FN = 2. For Dataset 2, TP=90, FP=1, TN=30, and FN=1. The sensitivity for Datasets 1 and 2 was 76.22 and 92.36 percent, respectively, indicating that the suggested method accurately predicts a person's lack of type 2 diabetes.

Datasets 1 and 2 had 96.55 and 99.42% specificities, respectively, demonstrating that the proposed technique accurately predicts type 2 diabetes.

According to total prediction accuracy, the suggested AWOD-based technique has 95.26% for the first and 99.01% for the second Dataset. The average weighted objective distance technique has great potential to predict diabetes mellitus type 2 based on its sensitivity, accuracy, and specificity. In the prediction method, important factors and irrelevant components can be identified using an average of the permitted level and the predicted level as weighting factors. Differentiating these criteria and predicting a constant PD class value are viable prediction strategies.

The accuracy obtained from the average weighted objective distance method is compared against ML classifier algorithms, including K-nearest neighbours, Support vector machine, Random forest, and Deep learning, as given in table 6.

KNN, which stands for "k- nearest neighbours," SVM, which stands for "support vector machine," RF, which stands for "random forest," and ANN, which stands for "artificial neural network," are all used for diabetes prediction in comparison with the AWOD method that was proposed. Only supervised learning was done, which included the use of classification algorithms.

The accuracy of the prediction results was examined using the K-fold cross-validation method. This approach produces only a small amount of bias during training, making it suitable for use with small datasets. The Dataset was folded ten times, giving it a K value of ten, so that it could be used in the training and testing procedure to achieve accurate prediction performance. Import the Dataset first. Next, take the mean of the data to fill in missing values in the Dataset for scaling their units for better performance in the results of the classifiers.

Table 7. KNN Outcomes of Diabetes type II prediction performance

KNN	Dataset I	Dataset II
Accuracy	78.57%	98.94%
Sensitivity	86.27%	93.33%
Specificity	63.46%	99.42%

Table 8. SVM Outcomes of Diabetes type II prediction performance

SVM	Dataset 1	Dataset 2
Accuracy	80.51%	97.36%
Sensitivity	82.92%	77.77%
Specificity	72.97%	99.41%

Table 7 shows the prediction performance of random forest on datasets 1 and 2, including accuracy, sensitivity, and specificity. Results show that it gave much better accuracy on dataset 2 and poorly on dataset 1 for diabetes type 2 prediction.

Table 9. RF Outcomes of Diabetes type II prediction performance

RF	Dataset 1	Dataset 2
Accuracy	79.87%	98.94%
Sensitivity	83.92%	93.33%
Specificity	69.04%	99.42%

Finally, generate the confusion matrix after data are split into training and testing sets and after getting the prediction results of x and y train. Accuracy, sensitivity, and specificity were some of the statistical measures after the cm results were gathered. When making a prediction, the idea behind the AWOD-based strategy is to consider individual factors that are significant to the overall prediction while disregarding those that are insignificant. Because it is frequently used for feature extraction to determine which features are the most important, the RF classifier was chosen to evaluate the performance of the suggested methodology in comparison to the performance of the predictions.

On dataset 1, the k-nearest neighbour algorithm was utilised. First, a variety of libraries are incorporated, and then the Dataset is read from where it is stored, if one is present. The following step involved using scikit learn to separate the Dataset into a training and testing set. After making the x and y train predictions, a confusion matrix was created with 88 correct positive answers, 19 incorrect positive answers, 14 incorrect negative answers, and 33 correct negative answers. Accuracy, sensitivity, and specificity were all produced due to this cf. The correlation between sensitivity and specificity is inverse, meaning that as sensitivity increases, specificity typically decreases and vice versa. Tests with a high level of specificity will demonstrate that patients who do not have a finding do not have an illness. In contrast, tests with a high level of sensitivity will result in positive findings for patients with a condition.

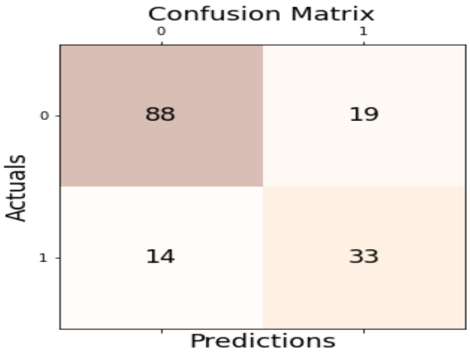

Figure 4. Confusion matrix of KNN

K-nearest neighbour gave an accuracy of 78.57 percent on dataset 1 with a sensitivity 86.27 percent and specificity 63.46 percent. Sensitivity informs us of the percentage of positives class members who were accurately identified. Finding out what percentage of the real sick persons the model properly identified would be an easy example. So these are calculated, and the results of these measures are given below:

We must evaluate the ML model and determine how effective (or ineffective) it is. In this situation, the AUC-ROC curve is useful. AUC-ROC curves mean the Area Under the Curve (AUC) of Receiver Characteristic Operator (ROC). We can see how effectively our machine learning classifier works using the AUC-ROC curve. AUC-ROC curve graph of KNN shows that AUC = 0.76 means that, if we were to take two data points from different classes, there is a 76% chance that the model would be able to separate them or rank them correctly and that auc value was acceptable or fine. Therefore, the better a classifier differentiates between positive and negative classifications, the greater its AUC score.

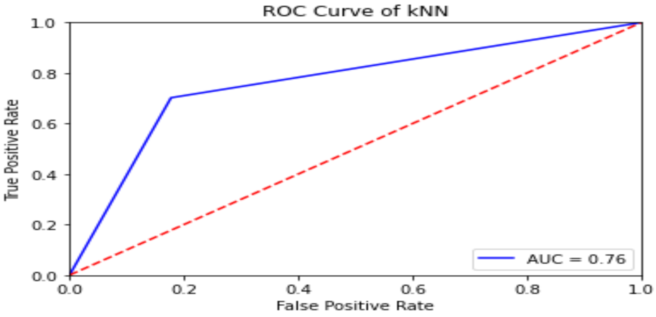

Figure 5. AUC-ROC graph of KNN

Random forest was applied on dataset 1. Different libraries are included first, then read the Dataset from the location where it exist. The next step was to split the Dataset into training and testing sets through scikit learn. After predicting the x and y train, the confusion matrix was generated with 94 TP, 13 FP, 18 FN and 29 TP. Accuracy, sensitivity and specificity came through this confusion matrix generated. Sensitivity and specificity are negatively correlated; as sensitivity rises, specificity tends to fall, and vice versa.

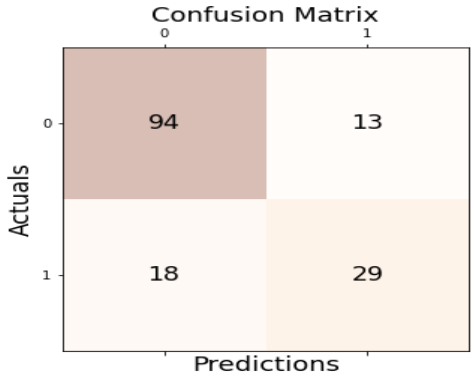

Figure 6. Confusion matrix of RF

We need to evaluate the ML model and determine how effective (or ineffective) it is. In this situation, the AUC-ROC curve is useful. AUC-ROC curves mean Area Under the Curve (AUC) of the Receiver Characteristic Operator (ROC). We need to see how effectively our machine learning classifier is working using the AUC-ROC curve. AUC-ROC curve graph of RF shows that AUC = 0.75 means that, if we were to take two data points from different classes, there is a 75% chance that the model would be able to separate them or rank them correctly and that auc value was acceptable or fine. Therefore, the better a classifier differentiates between positive and negative classifications, the greater its AUC score.

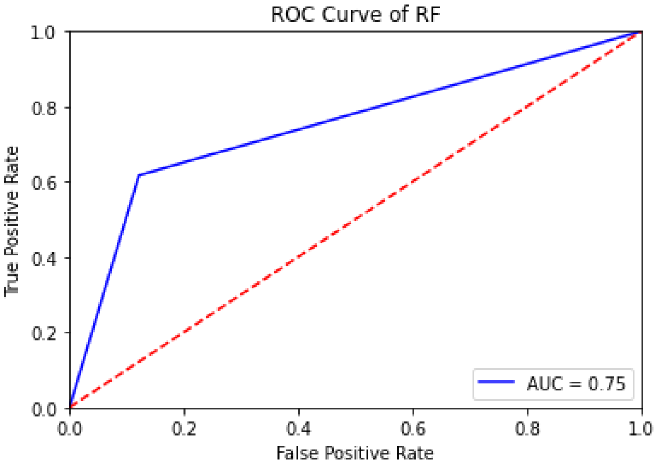

Figure 7. AUC-ROC graph of RF

On dataset 1, a support vector machine analysis was performed. First, a variety of libraries are incorporated, and then the Dataset is read from where it is stored, if one is present. The technique of feature scaling was applied to it. The following step involved using scikit learn to separate the Dataset into a training and testing set. After making the x and y train predictions, a confusion matrix was created with 97 correct positive answers, 10 incorrect positive answers, 20 incorrect negative answers, and 27 correct negative answers.

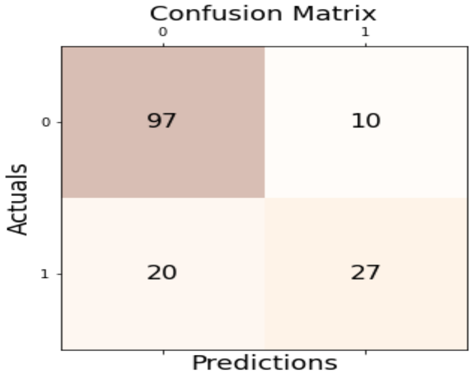

Figure 8. Confusion matrix of SVM

AUC-ROC curves mean Area Under the Curve (AUC) of the Receiver Characteristic Operator (ROC). We can see how effectively our machine learning classifier works using the AUC-ROC curve. AUC-ROC curve graph of SVM shows that AUC = 0.74 means that, if we were to take two data points from different classes, there is a 74% chance that the model would be able to separate them or rank them correctly and that auc value was acceptable or fine.

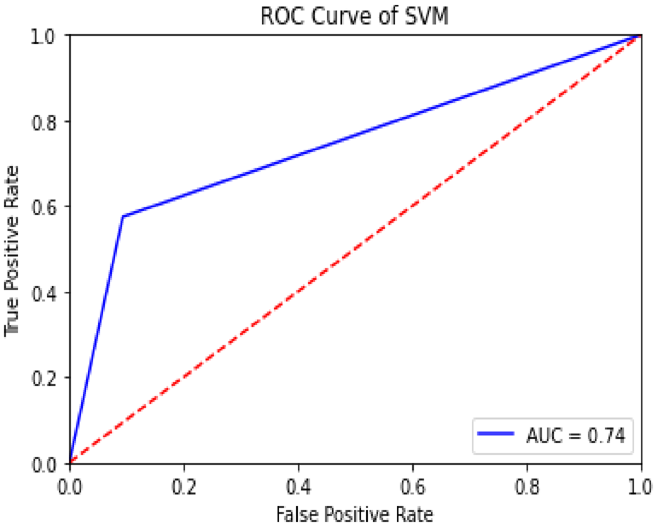

Figure 9. AUC-ROC graph of SVM

The artificial neural network was applied on datasets 1 and 2 with hidden layers giving an accuracy of 80.46 on dataset 1 and 95.85 percent on dataset 2. One to a hundred epochs are gone with 80.46 accuracies and a loss of 0.4062.

K-NN and SVM classifiers performed poorly when predicting diabetes type II for the first Dataset, with success rates of 78.57 and 80.51 percent, respectively. The K-nearest neighbours and Support vector machine classifiers have good accuracy for second Dataset (98.94% and 97.36%), but the

Table 10. Differentiation between the Proposed method and other ML techniques.

Method	Dataset 1	Dataset 2
K-Nearest Neighbors	78.57%	98.94%
Support Vector Machines	80.51%	97.36%
Random Forest	79.87%	98.94%
Deep Learning	80.46%	95.85%
AWOD based method	95.26%	99.01%

AWOD-based method has better accuracy for both datasets, with 95.26% for first and 99.01% for second. It was caused by computing the distance for all patients using both significant and insignificant factors, which may not have affected all patients. DL's (ANN) prediction accuracy was 80.46 percent for Dataset 1 and 95.85 percent for Dataset 2, but AWOD-based results were better. In addition, the RF classifier selected only the most significant factors to use for prediction, which resulted in the prediction results for both datasets being more accurate than the results provided by other classifiers. The RF classifier's predictions for both datasets were more accurate than those obtained from other classifiers. This is because the RF classifier only used the most important criteria when making its predictions. The comparison findings show that the average weighted objective distance technique offered a superior prognosis outcome than utilizing those classifiers. The potential for predicting the presence or absence of diabetes type II in patients using the suggested AWOD-based technique. As a consequence, the hypothesis of the study that the suggested AWOD-based technique can offer higher accuracy than other Machine Learning classifiers can be proven to be true.

The Proposed technique nevertheless offered an approximation error for inaccurate prediction of 4.74 percent for Dataset 1 and 0.99 percent for Dataset 2, respectively. The people were placed in the incorrect class in some inaccurate prediction scenarios. In most cases, a person's medical condition may be to blame. Due to this circumstance, those people's present level was shown to be in the incorrect range, either lower than or above the permitted level determined by the averaging them. Constant values can result in incorrect predictions as well. It's possible that the minimum and maximum numbers of components utilized to arrive at constant values won't be useful in detecting diabetes type II in those people. To improve prediction performance, future studies will consider establishing the expected level, acceptable level, and constant values by adjusting the proposed technique or employing various analytical vantage points.

Table 11. Comparison of the results with our proposed technique with state-of-the-art methods.

Authors	**Method**	**Accuracy**
A. Yahyaoui and M. Yesiltepe	SVM	65.38
P. Nuankaew and P. Temdee	KNN	71.68
J. R. Raut	RF	74.47
J. R. Raut	DL	74.74
D. Sisodia and D. S. Sisodia	Naïve Bayes	76.30
A. Yahyaoui and M. Yesiltepe	Deep Learning	76.81
P. Nuankaew and P. Temdee	SVM	77.70
P. Nuankaew and P. Temdee	RF	78.32
Proposed method (KNN)	Feature scaling (KNN)	**78.57**
Proposed method (RF)	Feature scaling (RF)	**79.87**
Proposed method (SVM)	Feature scaling (SVM)	**80.51**
Proposed method (DL(ANN))	Deep learning (ANN)	**80.71**
Proposed method (AWOD)	Average weighted objective distance method	**95.26**

Table 11 compares the results with our proposed technique with state-of-the-art methods. So from the above table, it is clear that our proposed practices gives higher accuracy than the previous or existing methods.

5. Conclusions

This study suggests that the proposed method (AWOD) can predict diabetes type II. According to the proposed technique, doctors should consider a patient's unique medical issues when diagnosing them. The proposed strategy used average predicted and acceptable levels to prioritize weighing factors. Prioritized factor lists important and unimportant human aspects. Depending on the patient's health, such factors may affect the prognosis. The experiment used the Pima Indians Diabetes dataset and the Mendeley Data for Diabetes dataset. AWOD's sensitivity, specificity, and accuracy were evaluated as a predictive tool. KNN, SVM, and RF classifiers had poor accuracy in predicting type 2 disease for Dataset no 1, with 78.57% for KNN, 80.51% for SVM, and 79.87% for RF.

Similarly, KNN, SVM, and RF gave much better results than Dataset 1, including 98.94% accuracy for KNN and RF 97.36% for SVM. Still, the AWOD-based approach has a prediction accuracy of 95.26% for Dataset 1 and 99.01% for Dataset 2, which is better than ML prediction techniques, including K-nearest neighbours, Support vector machine, and Random forest. Deep learning (ANN) had an accuracy of 75.26% for Dataset I and 95.85% for Dataset II, but the proposed technique was better. For future research purposes, it is essential to modify the AWOD to make it more applicable to other types of multi-category groups. Work should also be done to improve the performance of ML classifiers by performing parameter tuning and employing other preprocessing methods that contribute to improved prediction performance.

References

1. Aofa, F., Sasongko, P. S., Sutikno, Suhartono, & Adzani, W. A. (2018). Early Detection System of Diabetes Mellitus Disease Using Artificial Neural Network Backpropagation with Adaptive Learning Rate and Particle Swarm Optimization. 2018 2nd International Conference on Informatics and Computational Sciences, ICICoS 2018, 191–195. https://doi.org/10.1109/ICICOS.2018.8621683
2. Appavu alias Balamurugan, S., & Salomi, M. (2020). A predictive risk level classification of diabetic patients using deep learning modified neural network. Journal of Ambient Intelligence and Humanized Computing, 0123456789. https://doi.org/10.1007/s12652-020-02490-1
3. Alhaj, T. A., Siraj, M. M., Zainal, A., Elshoush, H. T., & Elhaj, F. (2016). Feature selection using information gain for improved structural-based alert correlation. PloS one, 11(11), e0166017.doi: 10.1371/journal.pone.0166017.
4. Azhagusundari.B and Thanamani A. S., Feature Selection based on Information Gain, Int. J. Innov. Technol. Explor. Eng., vol. 2, no. 2, pp. 18–21, 2013, doi: 10.1016/2008-6005.
5. Chaising, S., Prasad, R., & Temdee, P. (2021). Personalized Recommendation Method for Preventing Elderly People from Cardiovascular Disease Complication Using Integrated Objective Distance. Wireless Personal Communications, 117(1), 215–233. https://doi.org/10.1007/s11277-019-06639-w
6. Chaising, S., & Temdee, P. (2018). Determining Recommendations for Preventing Elderly People from Cardiovascular Disease Complication Using Objective Distance. 6th Global Wireless Summit, GWS 2018, 151–155. https://doi.org/10.1109/GWS.2018.8686527
7. Chaising, S., Temdee, P., & Prasad, R. (2020). Weighted objective distance for the classification of elderly people with hypertension. Knowledge-Based Systems, 210, 106441. https://doi.org/10.1016/j.knosys.2020.106441
8. Chaki, J., Thillai Ganesh, S., Cidham, S. K., & Ananda Theertan, S. (2020). Machine learning and artificial intelligence based Diabetes Mellitus detection and self-management: A systematic review. Journal of King Saud University - Computer and Information Sciences. https://doi.org/10.1016/J.JKSUCI.2020.06.013
9. Choi, B. G., Rha, S. W., Kim, S. W., Kang, J. H., Park, J. Y., & Noh, Y. K. (2019). Machine Learning for the Prediction of New-Onset Diabetes Mellitus during 5-Year Follow-up in Non-Diabetic Patients with Cardiovascular Risks. Yonsei Medical Journal, 60(2), 191–199. https://doi.org/10.3349/YMJ.2019.60.2.191
10. Demirci, S., Erer, I., & Ersoy, O. (2015). Weighted Chebyshev distance classification method for hyperspectral imaging. Next-Generation Spectroscopic Technologies VIII, 9482, 948218. https://doi.org/10.1117/12.2181914
11. Dhar, A., Dash, N., & Roy, K. (2017, September). Classification of text documents through distance measurement: An experiment with multi-domain Bangla text documents. In 2017 3rd International Conference on Advances in Computing, Communication & Automation (ICACCA)(Fall) (pp. 1-6) doi: 10.1109/ICACCAF.2017.8344721.
12. Fitriyani, N. L., Syafrudin, M., Alfian, G., & Rhee, J. (2019). Development of Disease Prediction Model Based on Ensemble Learning Approach for Diabetes and Hypertension. IEEE Access, 7, 144777–144789. https://doi.org/10.1109/ACCESS.2019.2945129
13. Greche, L., Jazouli, M., Es-Sbai, N., Majda, A., & Zarghili, A. (2017, April). Comparison between Euclidean and Manhattan distance measure for facial expressions classification. In 2017 International conference on wireless technologies, embedded and intelligent systems (WITS) (pp. 1-4). doi: 10.1109/WITS.2017.7934618.
14. Gupta, M. (2012). Dynamic k-NN with attribute weighting for automatic web page classification (Dk-NNwAW). International Journal of Computer Applications, 58(10).doi: 10.5120/9321-3554. IDF Diabetes Atlas | Tenth Edition. (n.d.). Retrieved January 25, 2022, from https://diabetesatlas.org/
15. Islam Ayon, S., & Milon Islam, Md. (2019). Diabetes Prediction: A Deep Learning Approach. International Journal of Information Engineering and Electronic Business, 11(2), 21–27. https://doi.org/10.5815/ijieeb.2019.02.03
16. Islam, M. S., Qaraqe, M. K., Belhaouari, S. B., & Abdul-Ghani, M. A. (2020). Advanced Techniques for Predicting the Future Progression of Type 2 Diabetes. IEEE Access, 8, 120537–120547. https://doi.org/10.1109/ACCESS.2020.3005540 IDF Diabetes Atlas | Tenth Edition. https://diabetesatlas.org/ (accessed Jan. 25, 2022).

17. Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine Learning and Data Mining Methods in Diabetes Research. Computational and Structural Biotechnology Journal, 15, 104–116. https://doi.org/10.1016/j.csbj.2016.12.005
18. Kjeldsen, S., Feldman, R. D., Lisheng, L., Mourad, J. J., Chiang, C. E., Zhang, W., Wu, Z., Li, W., & Williams, B. (2014). Updated national and international hypertension guidelines: A review of current recommendations. Drugs, 74(17), 2033–2051. https://doi.org/10.1007/s40265-014-0306-5
19. Kwon, N., Lee, J., Park, M., Yoon, I., & Ahn, Y. (2019). Performance evaluation of distance measurement methods for construction noise prediction using case-based reasoning. Sustainability, 11(3), 871. doi: 10.3390/su11030871.
20. Le, T. M., Vo, T. M., Pham, T. N., & Dao, S. V. T. (2021). A Novel Wrapper-Based Feature Selection for Early Diabetes Prediction Enhanced with a Metaheuristic. IEEE Access, 9, 7869–7884. https://doi.org/10.1109/ACCESS.2020.3047942
21. Muhammad, L. J., Algehyne, E. A., & Usman, S. S. (2020). Predictive supervised machine learning models for diabetes mellitus. SN Computer Science, 1(5), 1-10.doi: 10.1007/s42979-020-00250-8.
22. Nuankaew, P., Chaising, S., & Temdee, P. (2021). Average Weighted Objective Distance-Based Method for Type 2 Diabetes Prediction. IEEE Access, 9, 137015–137028. https://doi.org/10.1109/ACCESS.2021.3117269
23. Naz, H., & Ahuja, S. (2020). Deep learning approach for diabetes prediction using PIMA Indian dataset. Journal of Diabetes & Metabolic Disorders, 19(1), 391-403.doi: 10.1007/s40200-020-00520-5.
24. Patil, L. H., & Atique, M. (2013). A novel feature selection based on information gain using WordNet. 2013 Science and Information Conference, 625–629.
25. Pima Indians Diabetes Database | Kaggle. (n.d.). Retrieved May 19, 2022, from https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database
26. Pereira, R. B., Plastino, A., Zadrozny, B., & Merschmann, L. H. (2015). Information gain feature selection for multi-label classification. Journal of Information and Data Management, 6(1), 48-48.https://sol.sbc.org.br/journals/index.php/jidm/article/view/1555.
27. Raut, J. R., Sharma, Y., & Shinde, V. D. (2021). Performance Evaluation of Various Supervised Machine Learning Algorithms for Diabetes Prediction. European Journal of Molecular & Clinical Medicine, 7(8), 4921–4925.
28. Ramesh, S., Balaji, H., Iyengar, N. C. S., & Caytiles, R. D. (2017). Optimal predictive analytics of pima diabetics using deep learning. International Journal of Database Theory and Application, 10(9), 47-62. doi: 10.14257/ijdta.2017.10.9.05.
29. Rajput, M. R., & Khedgikar, S. S. Diabetes prediction and analysis using medical attributes: A machine learning approach. Journal of Xian University of Architecture and Technology.doi: 10.17632/WJ9RWKP9C2.1.
30. Sabariah, M. K., Hanifa, A., & Sa'adah, S. (2015). Early detection of type II Diabetes Mellitus with random forest and classification and regression tree (CART). Proceedings - 2014 International Conference on Advanced Informatics: Concept, Theory and Application, ICAICTA 2014, 238–242. https://doi.org/10.1109/ICAICTA.2014.7005947
31. Şahin, S., Merotra, J., Kang, J., Trenell, M., Catt, M., & Yu, E. H. (2018). Simultaneous Electrochemical Detection of Glucose and Non-Esterified Fatty Acids (NEFAs) for Diabetes Management. IEEE Sensors Journal, 18(22), 9075–9080. https://doi.org/10.1109/JSEN.2018.2870071
32. Sethi, N. (2020). a Comprehensive Analysis of Machine Learning Techniques for Incessant. January. https://doi.org/10.33832/ijgdc.2020.13.1.01
33. Sisodia, D., & Sisodia, D. S. (2018). Prediction of Diabetes using Classification Algorithms. Procedia Computer Science, 132(Iccids), 1578–1585. https://doi.org/10.1016/j.procs.2018.05.122
34. Sowah, R. A., Bampoe-Addo, A. A., Armoo, S. K., Saalia, F. K., Gatsi, F., & Sarkodie-Mensah, B. (2020). Design and Development of Diabetes Management System Using Machine Learning. International Journal of Telemedicine and Applications, 2020. https://doi.org/10.1155/2020/8870141

35. Srivastava, A. K., Kumar, Y., & Singh, P. K. (2021). Artificial Bee Colony and Deep Neural Network-Based Diagnostic Model for Improving the Prediction Accuracy of Diabetes. International Journal of E-Health and Medical Communications, 12(2), 32–50. https://doi.org/10.4018/ijehmc.2021030102

36. Sun, Q., Jankovic, M. V., Bally, L., & Mougiakakou, S. G. (2018, November). Predicting blood glucose with an lstm and bi-lstm based deep neural network. In 2018 14th symposium on neural networks and applications (NEUREL) (pp. 1-5).doi: 10.1109/NEUREL.2018.8586990.

37. Tripathi, G., & Kumar, R. (2020, June). Early prediction of diabetes mellitus using machine learning. In 2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO) (pp. 1009-1014). doi: 10.1109/ICRITO48877.2020.9197832.

38. Yahyaoui, A., Jamil, A., Rasheed, J., & Yesiltepe, M. (2019, November). A decision support system for diabetes prediction using machine learning and deep learning techniques. In 2019 1st International Informatics and Software Engineering Conference (UBMYK) (pp. 1-4). doi: 10.1109/UBMYK48245.2019.8965556.

39. Zhou, H., Myrzashova, R., & Zheng, R. (2020). Diabetes prediction model based on an enhanced deep neural network. Eurasip Journal on Wireless Communications and Networking, 2020(1). https://doi.org/10.1186/s13638-020-01765-7