IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received September 19, 2021, accepted September 27, 2021, date of publication October 4, 2021, date of current version October 12, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3117269

Average Weighted Objective Distance-Based Method for Type 2 Diabetes Prediction

PRATYA NUANKAEW[1], SUPANSA CHAISING[2], AND PUNNARUMOL TEMDEE[3], (Member, IEEE)

[1]School of Information and Communication Technology, University of Phayao, Phayao 56000, Thailand
[2]Department of Information Technology, International College, Payap University, Chiang Mai 50000, Thailand
[3]Computer and Communication Engineering for Capacity Building Research Center, School of Information Technology, Mae Fah Luang University, Chiang Rai 57100, Thailand

Corresponding author: Punnarumol Temdee (punnarumol@mfu.ac.th)

This work was supported in part by the University of Phayao under Project FF64-UoE004, and in part by Mae Fah Luang University.

ABSTRACT Early detection of Type 2 diabetes is necessary for its prevention. The prediction models for detection systems normally employ common factors that may not properly fit all persons having different health conditions. Therefore, this study proposes a method for type 2 diabetes prediction with factors representing personal health conditions. More specifically, this study proposes a novel prediction method named Average Weighted Objective Distance (AWOD) based on the assumption that the individual has diverse health conditions resulting from different individual factors, a requirement for an effective prediction model. AWOD is a modification of Weighted Objective Distance (WOD) by applying information gain to reveal significant and insignificant individual factors having different priorities, which are represented by different weights. For AWOD, the data set is divided into a training set used to determine all relevant thresholds and constant values required for AWOD calculation and the testing set. In particular, AWOD is designed for binary classification problems with a relatively small dataset. To validate the proposed method, two datasets from open sources, Pima Indians Diabetes (Dataset 1) and Mendeley Data for Diabetes (Dataset 2) each containing 392 records, were studied. The prediction performance for both datasets is compared with the machine learning-based prediction methods, including K-Nearest Neighbors, Support Vector Machines, Random Forest, and Deep Learning. The comparison results showed that the proposed method provided 93.22% and 98.95% accuracy for Dataset 1 and Dataset 2, respectively, which are higher than those provided by other machine learning-based methods.

INDEX TERMS Objective distance, weighting factors, information gain, diabetes, prediction.

I. INTRODUCTION

Diabetes, formally called diabetes mellitus, is a group of abnormal metabolic and chronic diseases. It causes elevated blood glucose levels, which results in prolonged high blood sugar levels [1]. Elevated blood sugar levels can lead to increased urination, thirst, and hunger, especially for sweets. It also leads to severe damage to the blood vessels, heart, kidneys, eyes, and nerves. Without urgent treatment, diabetes mellitus can cause many other complications and serious negative side effects, up to and including diabetic ketoacidosis, nonketotic hyperosmolar, heart disease, stroke, kidney failure, foot ulcers, vision loss, and blindness [2], [3]. In addition, individuals with diabetes mellitus are more likely to be infected and are at a higher risk of complications and death from COVID-19 [4]. Recently, diabetes has become the leading cause of mortality and morbidity in the world. According to the International Diabetes Federation, approximately 463 million people had diabetes worldwide in 2019. This amount is expected to increase by 51% in the next 26 years with around 700 million people living with diabetes worldwide in 2045 [5]. Early detection and treatment of diabetes is a major step forward in necessary treatment for diabetic patients, which can reduce the risk of serious complications [6].

There are three main types of diabetes: type 1, type 2, and gestational diabetes. Type 1 diabetes develops when the body cannot produce insulin because cells in the pancreas responsible for that are destroyed. Type 2 diabetes develops when the body becomes resistant to insulin. Gestational diabetes

The associate editor coordinating the review of this manuscript and approving it for publication was Yiming Tang.

develops when insulin-blocking hormones are produced during pregnancy. In 2019, there were over 1.1 million children and adolescents living with type 1 diabetes. For gestational diabetes type, there were over 20 million live births affected by diabetes during pregnancy. However, type 2 diabetes is the most common type usually found in adults. Its prevalence has increased dramatically in most countries worldwide with approximately 374 million people at increased risk of developing it in 2019. Diabetes caused total mortality of about 4.2 million. Notably, it was found that 1 in 2 people with diabetes went undiagnosed, which increases the number of mortalities in the future. In general, many people with type 2 diabetes rarely show any symptoms, which results in increased risk factors generating complications [7]. People with this condition should undergo several tests to diagnose diabetes in advance. The increasing number of patients subjected to inadequate health care providers exacerbate the problem for diabetes diagnosis and care [8]. The primary objective of this study is to support health care providers with a prevention diabetes prediction method, particularly for early-stage type 2 diabetes.

For the early disease detection systems, the existing studies employ different machine learning algorithms, which normally go through large and diverse amounts of data, to predict the presence of type 2 diabetes. The machine learning-based methods can easily fail to categorize the diversity of individuals with a relatively small set of data. To address this issue, this study proposes a binary classification method based on distance measurement to predict the presence of type 2 diabetes. Furthermore, existing prediction methods normally employ a common set of factors for constructing the model. According to the health care professional principle in the diagnosing process, a wide swath of health conditions results in different disease diagnoses and treatment decisions [9]. Thus, this study proposes a novel prediction method to predict the presence of type 2 diabetes based on individual factors instead of using common factors as shown in the general prediction methods.

The proposed method, the Average Weighted Objective Distance (AWOD), is a modification of Weighted Objective Distance (WOD) [10]. Both methods are designed particularly for constructing prediction models from relatively small datasets because of infrequent and rare clinical data. To calculate the weight of each factor, WOD requires the pre-defined thresholds and constant values, which need to be assigned by a healthcare professional accordingly to the individual health diagnosis records. This process cannot be applied to any dataset without individual health diagnosis records. AWOD is thus designed to deal with this limitation of WOD. More specifically, AWOD derives all of them directly from the training dataset.

The organization of this paper is in the following way: Section II describes the literature review. Section III presents the research methodology. Section IV showcases experimentation with the proposed AWOD based method. Section V shows the results and discussion. Section VI gives the conclusion for this paper.

II. LITERATURE REVIEW

The literature reviews below discuss existing works in two different research areas, namely factors for prediction method and prediction method.

A. FACTORS FOR PREDICTION METHOD

The prediction of diabetes normally considers the common set of data for constructing the model. To take into account the diversity of individuals, the risk factors are considered. Novel feature extraction methods are proposed to select relevant factors required for effective prediction. The most significant risk factors are extracted from the whole dataset based on the attribute scores [11], [12]. There exist works that account for both risk factors and symptom-oriented variables for constructing the model [13]. Complex attributes containing several categorical attributes are also employed to provide better prediction performance [14]. However, the reduction of the number of input factors is also widely encouraged [15] to decrease the model complexity. From the previous studies regarding early diabetes prediction, the common set of factors are granted the most consideration for model construction. Conversely, this study bears in mind individual set of factors derived from different health conditions for each person.

B. PREDICTION METHOD

As mentioned before, machine learning algorithms have been widely used in early diabetes prediction for decades. Pieces of literature have shown they express an effective performance compared with the traditional statistical methods [16]. Several works for early diabetes detection using several machine learning techniques, such as K-Nearest Neighbor (K-NN), Support Vector Machine (SVM), Artificial Neural Network (ANN) [17], [18], and Random Forest (RF) [19] have demonstrated cost-effectiveness and time-saving for diabetic patients and doctors. Particularly, the algorithms were used to classify different types of datasets. For example, the RF classifier exhibited high accuracy for predicting type 2 diabetes of individuals based on their lifestyle and family background [19]. K-NN, SVM, Logistic Regression (LR), and ANN were applied for type 2 diabetes prediction by using long non-coding RNAs and demographic data [20]. ANN, RF, and Decision Trees were used to predict diabetes from physical examination data randomly selected for healthy people and diabetic patients [21]. For this set of clinical data, machine learning-based predictive models usually fail to characterize the diversity among individuals without enough numbers of data. Recent works have demonstrated the attempts to empower machine learning-based classifiers to deal with classification problems of small training sample size, such as a generalized mean distance-based K-NN classifier [22], a local mean representation-based K-NN classifier [23], and a locality constrained representation-based

K-NN classifier [24]. Similarly, this study thus proposes the binary classification method for a relatively small data set.

The binary classification method is based on distance measurements like Hamming, Euclidean, Manhattan distance, and Minkowski distance [25], [26]. Generally, a distance measure is an objective score describing the relative difference between two objects. Hamming distance calculates the distance between two binary vectors or binary strings. It is the number of bit positions in which the two bits are different. Euclidean and Manhattan distance determine the distance between two real-valued vectors. Euclidean distance is the straight-line distance between 2 data points in a plane, which can be calculated from the Cartesian coordinates of the points using the Pythagorean theorem. Manhattan distance is preferred for the vectors describing objects on a uniform grid, like a city block. The Manhattan distance thus is normally used to calculate the distance between two data points in a grid-like path. For Minkowski distance, it is a generalization of other distances having a parameter called the order to calculate different distance measures. More specifically, it gives control over the type of distance measure of real-valued vectors by using a hyperparameter that can be tuned. For example, when the order is ''1'', ''2'', and infinity, it is the Manhattan, Euclidean, and Chebychev distance respectively. For this study, the proposed measurement follows the principle of Euclidean distance which is to measure the distance between 2 real-valued vectors in the plane and have a scalar factor to change their lengths but does not change their directions.

The distance measurements have been using for solving various classification and prediction problems. For example, a combination of Jaccard and weighted Euclidean distances was presented for noise prediction [27]. A weighted Chebyshev distance method was proposed for the classification of hyperspectral imagery [28]. In personalized learning application domains, objective distance (OD) [29] was initially proposed to measure the distance between the current competency of a student and the expected level to attain learning expectations. Similarly, the OD was applied in the health care domain to classify a group of older people with hypertension by using the distance between current and expected health status of all risk factors developing hypertension [30]. It was also used to provide the appropriate recommendation individually based on each risk factor [31].

Later, WOD [10], which is the modification of the original OD, was proposed to improve the group classification performance of older people with hypertension by using prioritized individual factors instead of considering all of them. Following the principle of distance measurement, WOD has a scalar factor or weight for representing the priority. For WOD, priority is represented only for significant individual factors by different weights obtained from the information gain principle. For weight calculation, WOD requires predefined thresholds namely expected and acceptable levels of all factors and some constant values. Therefore, WOD cannot be obtained directly from any dataset with no clues of pre-designed thresholds and constant values. Then, AWOD is proposed in this study to be more generalized for different datasets and different diseases as two diabetes datasets in this study.

Information gain is still applied to prioritize factors for AWOD. Generally, information gain measures reductions in entropy [32] and determines irrelevant attributes of a dataset [33]–[36], including individual factors [37] by considering information gain levels after reducing entropy. For AWOD, the information gain is applied to determine both significant and insignificant factors for individuals. The former is defined as the factors the individual cannot control properly so that they have a noticeable effect on their health condition, while the latter are those controlled by the individual [10]. The limitation of WOD is that fewer significant individual factors decrease classification performance. For this study, the assumption is thus made that the significant and insignificant factors are different for each person and can be effectively used for constructing a model to provide higher prediction performance with proper priority settings.

Therefore, this study proposes the AWOD based method for type 2 diabetes prediction based on individual health conditions employing both significant and insignificant individual factors. Accordingly, the AWOD method is modified to be more generalized for a relatively small dataset for a binary classification problem. To represent priority, the weight is calculated from the obtained thresholds and constant values from the training dataset of both classes.

To validate the proposed prediction method, two open data sets, namely Pima Indians Diabetes (Dataset 1) [38] and Mendeley Data for Diabetes (Dataset 2) [39], are used in this study. As mentioned before, the proposed AWOD method cannot compare with WOD because there are no pre-defined thresholds, and constant values required for weight calculation of these datasets. Instead, the prediction results with them are compared with existing machine learning-based methods from the distance-based group including K-NN and SVM, implicit feature selection group as RF, and deep learning (DL) method, which is the state-of-the-art machine learning method.

K-NN calculates the distance from the interest data to every other in the dataset to find the closest data. The obtained distances are sorted to find the nearest neighbors, where the k number is defined as a minimum distance to predict the results. The principle of SVM for classification is that the algorithm creates a line or a hyperplane by separating the data into two classes. To perform classification, SVM finds the hyperplane that maximizes the margin between the two classes. RF is known as an ensemble method that is more effective than a single decision tree because it can reduce over-fitting by averaging the result. RF is a dimensionality reduction method because it identifies the most significant variables among input variables. DL can learn without human supervision by drawing from data that is both unstructured and unlabeled. Learning can be supervised, semi-supervised, or unsupervised. This algorithm is

essentially a neural network with three or more layers that attempt to mimic the human brain based on a combination of data inputs, weights, and biases.

III. RESEARCH METHODOLOGY

The research methodology of this study consists of two main procedures as shown in Figure 1, AWOD determination and evaluation of AWOD based prediction method. The details of each procedure are described in the following sections.

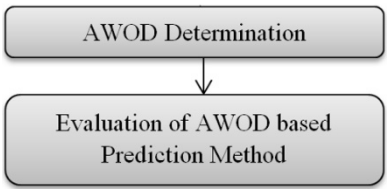

FIGURE 1. Research methodology.

A. AWOD DETERMINATION

1) AWOD CONCEPT

The principle underlying AWOD based method is based on the number of significant and insignificant factors representing real effects towards the prediction. This principle can represent the different individual health conditions and the general diagnostic procedures performed by health care professionals. The AWOD concept is illustrated in Figure 2.

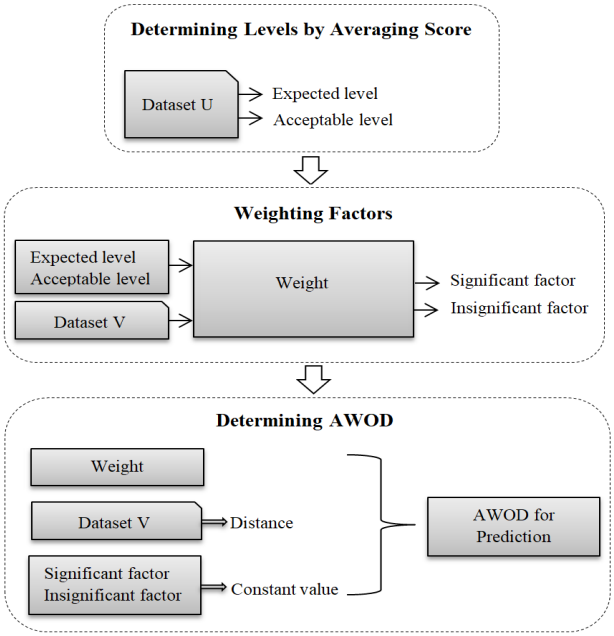

FIGURE 2. AWOD concept.

From Figure 2, three main steps are required for AWOD determination. The first step is to determine important levels for weight calculation namely an expected level and an acceptable level. The expected level is the health status level of each factor that the individual is recommended to have for an individual with no diabetes, while the acceptable level is the health status level for each factor that is acceptable for an individual with no diabetes. Next, the weights for both significant and insignificant factors are calculated. Finally, AWOD is determined for prediction.

To determine expected and acceptable levels used to calculate the current and the acceptable distance for each factor, the dataset will be split into two, dataset U and dataset V. The dataset U is the training set used for determining expected and acceptable levels of all factors by averaging score. Therefore, both levels can be representatives from those in the training set. Then, both levels are used to find weights for each factor. Dataset V is the testing set applied for weighting factors and determining AWOD. The value of weighting factors can represent significant and insignificant factors. These represent the real effects of each factor for each individual based on the set of significant and insignificant factors, which is denoted as a constant value. Next, Dataset V is used to calculate the distance between the acceptable distance and each factor's current distance. Then, the obtained distances, their associated weights, and a constant value are combined for deriving an individual AWOD. Since the values of each factor are on different scales, a min-max normalization is required for rescaling the range of attributes to scale the range in [0,1].

2) TARGET CLASS DETERMINATION WITH AWOD BASED METHOD

The algorithm for determining target class with AWOD based prediction method can be explained as the pseudocode as follows.

BEGIN
//Variable initialization
Step 0: Input all variables

u_{j+} ← value of factor j that is in the positive class for the U set
nu_{j+} ← total number of factor j that is in the positive class for the U set
u_{j-} ← value of factor j that is in the negative class for the U set
nu_{j-} ← total number of factor j that is in the negative class for the U set
Tn ← total number of factors
T ← total number of target classes
Z_j ← the current level of factor j
x_j ← value of the expected level of factor j
y_j ← value of the acceptable level of factor j
z_j ← value of the current level of factor j
lTn ← the minimum number of factors that can affect identifying the negative class
hTn ← the maximum number of factors that can affect identifying the negative class
$MAXnu_{j+}$ ← the factor with the maximum number of the positive class among all factors
$nND_{(v=0)}$ ← total number of factorswith the normalized average-based weighted objective

distance that is equal to 0
$ND_{(v=0)} \leftarrow$ the factor with the normalized average -based weighted objective distance that is equal to 0
$N \leftarrow$ total number of individuals
//Split data to Training Set (U set) and Testing Set (V set)
Step 1: Read Dataset
Random data for U set (70%)
Random data for V set (30%)
//Determine expected levels and acceptable levels of all //factors from U set.
//Stopping Condition is the number of factors that are //reached.
Step 2: WHILE Stopping Condition is False
Step 3: Calculate the expected level of each factor for all samples of positive class

$$X_j = \frac{\sum u_{j+}}{nu_{j+}}$$

Step 4: Calculate the average number of each factor for all samples of negative class

$$Y\,(a)_j = \frac{\sum u_{j-}}{nu_{j-}}$$

Step 5: Calculate the acceptable level of each factor for all samples

$$Y_j = \frac{X_j + Y\,(a)_j}{2}$$

ENDWHILE
//Determine the entropy of the target class with respect to //all factors in V set.
Step 6: Calculate equal probability for the positive target class (EP_+) and the negative target class (EP_-)

$$EP_+ = EP_- + \frac{Tn}{T}$$

Step 7: Calculate the positive-target class fraction (F_+) and the negative-target class fraction (F_-)

$$F_+ = \frac{EP_+}{Tn} = F_- = \frac{EP_-}{Tn}$$

Step 8: Calculate entropy of the target class with respect to all factors [$E(C)$]

$$E\,(C) = -F_+ * log_2\,(F_+) - F_- * log_2\,(F_-)$$

//Determine the entropy of each factor in V set.
//Stopping Condition is the number of factors that are //reached.
Step 9: WHILE Stopping Condition is False
Step 10: IF $x_j > y_j > z_j$ *or* $x_j < y_j < z_j$ THEN

$$Y_j = z_j$$

ELSE

$$Y_j = y_j$$

ENDIF
Step 11: Calculate acceptable distance for each factor that can identify as the positive class (dXY_j)

$$dXY_j = \sqrt{(X_j - Y_j)^2}$$

Step 12: Calculate the current distance that must be considered to identify as the positive class or the negative class (dXZ_j)
IF $Y_j = Z_j$ THEN

$$dXZ_j = 0$$

ELSE

$$dXZ_j = \sqrt{(X_j - Z_j)^2}$$

ENDIF
Step 13: Calculate the probability of the acceptable distance (pXY_{j+}) and the probability of the current distance (pXZ_{j-})

$$pXY_{j+} = \frac{dXY_j}{dXY_j + dXZ_j}$$
$$pXZ_{j-} = \frac{dXZ_j}{dXY_j + dXZ_j}$$

Step 14: IF $pXZ_{j-} = 0$ THEN $E\,(C_j) = 0$
ELSE

$$E\,(C_j) = -\frac{pXY_{j+}}{1} * log_2\left(\frac{pXY_{j+}}{1}\right) - \frac{pXZ_{j-}}{1} * log_2\left(\frac{pXZ_{j-}}{1}\right)$$

ENDIF
ENDWHILE
//Determine the entropy of all factors in V set.
//Stopping Condition is the number of factors that are //reached.
Step 15: WHILE Stopping Condition is False
Step 16: Calculate the entropy of all factors [$E(Ct)$]

$$E\,(Ct) = \sum_{j=1}^{Na}\left[E\,(C_j) * \left(\frac{pXY_{j+} + pXZ_{j-}}{Tn}\right)\right]$$

ENDWHILE
//Determine the information gain of the target class in V set.
Step 16: Determine the information gain of the target class with respect to all factors [$Gain\,(C, t)$]

$$Gain\,(C, t) = E\,(C) - E\,(Ct)$$

//Determine the weight of each factor in V set.
//Stopping Condition is the number of factors that are //reached.
Step 17: WHILE Stopping Condition is False
Step 18: Determine the significant gain for each factor (S_j)

$$S_j = \frac{E\,(C_j)}{Gain\,(C, t)}$$

Step 19: Determine the weight of each factor (W_j)

$$W_j = \frac{S_j}{\sum_{j=1}^{Na} S_j}$$

ENDWHILE

//Determining AWOD in V set.
//Stopping Condition is the number of factors that are
//reached.

Step 20: Determine the average-based weighted objective distance for each factor (D_j)

WHILE stopping condition is FALSE

$$D_j = W_j \times \left| \sqrt{(X_j - Y_j)^2} - \sqrt{(X_j - Z_j)^2} \right|$$

Step 21: Determine the maximum value of D_j ($\max D_j$) and the minimum value of D_j ($\min D_j$)

$$\max D_j = \text{Maximum } (D_j)$$
$$\min D_j = \text{Minimum } (D_j)$$

Step 22: Determine the normalized average-based weighted objective distance for each factor (ND_j)

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

Step 23: CAS *expression* OF

$nND_{(v=0)} > hTn$, *or* $lTn \leq nND_{(v=0)} \geq hTn$ *and* $ND_{(v=0)} \in MAXnu_{j+}$: $b = 0$

$nND_{(v=0)} < lTn$, *or* $lTn \leq nND_{(v=0)} \geq hTn$ *and* $ND_{(v=0)} \notin MAXnu_{j+}$: $b = 1$

ENDCASE

Step 24: Determine the average-based weighted objective distance for all factors for each individual ($AWOD_i$)

FOR i =1 to N

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

ENDFOR

ENDWHILE

// Identify the target class with AWOD condition

Step 25: FOR i = 1 to N

CASE expression OF

$0 < AWOD_i \leq 1$: Target Class == Negative Class

$AWOD_i = 0$: Target Class == Positive Class

ENDCASE

ENDFOR

END

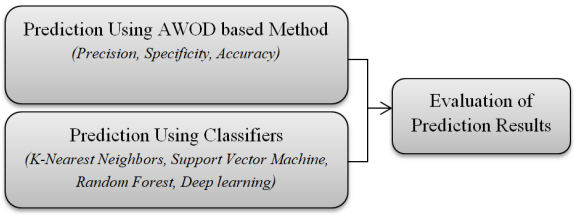

FIGURE 3. Prediction using AWOD based method.

B. EVALUATION OF AWOD BASED PREDICTION METHOD

Figure 3 shows the evaluation process of the proposed AWOD based method. In it, the predicted and the observed class were applied to evaluate the prediction performance using precision, specificity, and accuracy. The observed class refers to individuals' actual condition, specifically type 2 diabetes presence or absence. The predicted class refers to the prediction of either absence of type 2 diabetes (AD) or the presence of type 2 diabetes (PD), using the AWOD based method. In addition, K-NN, SVM, RF, and DL classifiers were employed with all original factors to compare their performance to the AWOD based prediction method. The classifiers used for evaluating results are described below. The details and results of prediction performance are explained in the next section.

IV. EXPERIMENT

A. DATA COLLECTION

Two datasets used for experimenting with type 2 diabetes prediction were designated Dataset 1 and 2. The former was collected from the Kaggle website, while latter from the Mendeley Data website. Dataset 1 is originally from the National Institute of Diabetes and Digestive and Kidney Diseases associated with all female patients at least 21 years old of Pima Indian heritage. The dataset contains 392 records after removing the missing value, and 8 factors including pregnancies, glucose, blood pressure, skin thickness, insulin, BMI, diabetes pedigree function, and age. Dataset 2 is originally from the Iraqi society, which was acquired from the laboratory of Medical City Hospital and the Specializes Center for Endocrinology and Diabetes-Al-Kindy Teaching Hospital. For this dataset, the data attribute used in this study includes 10 factors, which are age, urea, creatinine ratio, hemoglobin A1c (HBA1C), cholesterol, triglycerides, high-density lipoprotein (HDL), low-density lipoprotein (LDL), very-low-density lipoprotein (VLDL), and body mass index (BMI). Data with Diabetic and Non-Diabetic classes were only employed for predicting type 2 diabetes. In this dataset, 392 records were randomly selected for this study, which is the same as Dataset 1. Abbreviations of diagnostic factors for Dataset 1 and Dataset 2 are presented in Table 1 and Table 2 respectively. The abbreviation of diagnostic factors for Dataset 1 is used for a calculating demonstration to predict type 2 diabetes in the next section.

TABLE 1. Abbreviation of diagnostic factors for dataset 1.

Abbreviation	Factor	Detail
Pr	Pregnancies	Number of times pregnant
Gl	Glucose	Plasma glucose concentration a 2 hours in an oral 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

TABLE 2. Abbreviation of diagnostic factors for dataset 2.

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
Hd	HDL Cholesterol	High-density lipoprotein, which is good cholesterol
Ld	LDL Cholesterol	Low-density lipoprotein, which is bad cholesterol
Vl	VLDL	Very-low-density lipoprotein cholesterol produced in the liver
Bm	BMI	Body mass index

Examples of the gathered data for Dataset 1 and Dataset 2 based on factors associated with type 2 diabetes diagnoses are shown in Table 3 and Table 4 respectively. In this table, ''Yes'' and ''No'' for the diabetes factors (Di) represent presence and absence respectively.

B. DEMONSTRATION OF SAMPLE CALCULATION

This section presents the sample AWOD calculation to predict the presence or absence of type 2 diabetes using Dataset 1. The type 2 diabetes prediction using the proposed measurement method can be illustrated by applying data no. 1. In this study, the presence of type 2 diabetes was denoted as PD, whereas the absence as AD. The sample calculation performed to calculate the AWOD and predict which class the data no. 1 belongs to is shown below.

To determine levels by averaging score, the dataset of 392 samples as shown in Table 3 were divided into 2 sets,

TABLE 3. Examples of the collected data for dataset 1.

Data no.	Di	Factors							
		Pr	Gl	Bp	St	In	Bm	Dpf	Ag
1	Yes	7	150	78	29	126	35.2	0.692	54
2	Yes	0	121	66	30	165	34.3	0.203	33
3	Yes	8	100	74	40	215	39.4	0.661	43
4	Yes	5	187	76	27	207	43.6	1.034	53
5	No	1	116	78	29	180	36.1	0.496	25
6	No	2	56	56	28	45	24.2	0.332	22
7	No	1	112	80	45	132	34.8	0.217	24
8	Yes	7	97	76	32	91	40.9	0.871	32
9	No	2	87	58	16	52	32.7	0.166	25
10	Yes	8	167	106	46	231	37.6	0.165	43
…	…	…	…	…	…	…	…	…	…
392	No	3	102	44	20	94	30.8	0.400	26

TABLE 4. Examples of the collected data for dataset 2.

Data no.	Di	Factors									
		Ag	Ur	Cr	Hb	Ch	Tg	Hd	Ld	Vl	Bm
1	No	33	2.0	54	5.4	3.7	1.3	0.8	2.4	0.6	22
2	Yes	60	5.7	76	6.8	5.5	1.5	0.7	4.1	0.7	33
3	Yes	61	10.5	111	8.2	3.8	3.0	0.9	1.7	1.3	39
4	No	63	6.6	106	4.3	4.8	1.7	1.1	3.0	0.7	20
5	Yes	45	4.8	82	7.2	4.7	1.8	0.8	3.1	12.7	31
6	Yes	55	5.7	88	7.2	6.5	3.4	0.9	2.6	1.2	33
7	Yes	55	5.0	76	10.2	5.6	4.6	0.8	2.9	31.8	34
8	No	41	2.0	39	4.0	3.4	1.2	1.7	1.1	0.5	21
9	No	44	4.4	56	4.2	3.4	1.3	1.3	1.5	0.6	21
10	Yes	66	3.2	46	8.5	4.2	1.0	1.4	2.4	0.4	26
…	…	…	…	…	…	…	…	…	…	…	…
392	No	43	4.0	54	4.3	4.1	1.1	1.2	2.4	1.3	25

U and V, by splitting the data in 70:30 ratio due to their relatively small size. The proportion of the split ratio represents that 70% of the data used for determining the value of the expected and the acceptable level, which refers to the U set. Conversely, 30% of the data will be applied for weighting factors, which refers to the V set. The U set includes 274 samples, and the V set 118 samples. An example of the expected and acceptable level calculation for the Dpf factor is provided. X_j for the Dpf factor (X_{Dpf}) and Y_j for the Dpf factor (Y_{Dpf}) were determined, which applied the dataset from the U set. In this calculation sample, $\sum u_{Dpf+} = 89.017$, $nu_{Dpf+} = 184$, $\sum u_{Dpf-} = 55.508$, $nu_{Dpf-} = 90$, $Y(a)_j = 0.617(55.508/90)$. The values of X_{Dpf} and Y_{Dpf} were equal to 0.484 and 0.621, respectively, as follows:

$$X_{Dpf} = \frac{89.017}{184} = 0.484$$

$$Y_{Dpf} = \frac{0.484 + 0.617}{2} = 0.550$$

Thus, expected levels and acceptable levels of all factors for Dataset 1 are shown in Table 5.

TABLE 5. Expected levels and acceptable levels of all factors for dataset 1.

Factors	Expected Level	Variables (X_j)	Acceptable Level	Variables (Y_j)
Pr	3	x_{Pr}	4	y_{Pr}
Gl	113	x_{Gl}	127	y_{Gl}
Bp	70	x_{Bp}	71	y_{Bp}
St	27	x_{St}	30	y_{St}
In	135	x_{In}	170	y_{In}
Bm	31.9	x_{Bm}	33.7	y_{Bm}
Dpf	0.484	x_{Dpf}	0.550	y_{Dpf}
Ag	29	x_{Ag}	32	y_{Ag}

From above algorithm, the first step of determining weight factor is to find the entropy of the target class. The equal probability of the target class was initially determined. This refers to the equal probability between AD representing the positive target class (EP_+) and PD representing the negative target class (EP_-). The values of the AD (EP_+) and PD (EP_-) were equated to 4, as follows:

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

Next, the positive-target class fraction (F_+) and the negative-target class (F_+) with respect to all factors was determined. The fractions of the AD (F_+) and PD (F_-) with respect to all factors are shown as follows:

$$F_+ = \frac{4}{8} \quad F_- = \frac{4}{8}$$

The entropy of the target class with respect to all factors [$E(C)$] was thus determined. The value of $E\,(C)$ was equal to 1, as follows:

$$E\,(C) = -\frac{4}{8} \times log_2\left(\frac{4}{8}\right) - \frac{4}{8} \times log_2\left(\frac{4}{8}\right) = 1$$

The second step is to determine the entropy of each factor. An example of entropy calculation for the Dpf factor is provided. The value of acceptable distance (dXY_{Dpf}) and current distance (dXZ_{Dpf}) for Dpf factor were measured respectively. From Table 3, the current level of data no.1 for the Dpf factor (Z_{Dpf}) is 0.692, which belongs to the V set. From Table 5, $X_{Dpf} = 0.484$ and $Y_{Dpf} = 0.550$ calculated from the U set. The dXY_{Dpf} and dXZ_{Dpf} were equal to 0.066 and 0.208, respectively, as follows:

$$dXY_{Dpf} = \sqrt{(0.484 - 0.550)^2} = 0.066$$
$$dXZ_{Dpf} = \sqrt{(0.484 - 0.692)^2} = 0.208$$

Then, the probability of the acceptable distance for the positive class (pXY_{Dpf+}) and that of the current distance for the negative class (pXZ_{Dpf-}) for the Dpf factor were calculated, respectively. The pXY_{Dpf+} and pXZ_{Dpf-} were equal to 0.24 and 0.75, respectively, as follows:

$$pXY_{Dpf+} = \frac{0.066}{0.066 + 0.208} = 0.24$$
$$pXZ_{Dpf-} = \frac{0.208}{0.066 + 0.208} = 0.75$$

The entropy of Dpf factor [$E(C_{Dpf})$] was then computed. $E(C_{Dpf})$ was equal to 0.79, as follows:

$$E\left(C_{Dpf}\right) = -\frac{0.24}{1} \times log_2\left(\frac{0.24}{1}\right) - \frac{0.75}{1} \times log_2\left(\frac{0.75}{1}\right) = 0.79$$

The entropy calculated for each factor is shown in Table 6.

The third step is to determine the information gain of the target class with respect to all factors. Thus, the entropy of all factors [$E(Ct)$] was calculated. $E\,(Ct)$ was equal to 0.54, as follows:

$$\begin{aligned} E\,(Ct) = {} & 0.72 * \left(\frac{0.20 + 0.80}{8}\right) + 0.85 * \left(\frac{0.27 + 0.73}{8}\right) \\ & + 0.50 * \left(\frac{0.11 + 0.89}{8}\right) + 0 * \left(\frac{1 + 0}{8}\right) \\ & + 0 * \left(\frac{1 + 0}{8}\right) + 0.99 * \left(\frac{0.55 + 0.45}{8}\right) \\ & + 0.79 * \left(\frac{0.24 + 0.75}{8}\right) + 0.49 * \left(\frac{0.11 + 0.89}{8}\right) \\ = {} & 0.54 \end{aligned}$$

Then, the information gain of the target class with respect to all factors [$Gain\,(C, t)$] was determined. $Gain\,(C, t)$ was equal to 0.46, as follows:

$$Gain\,(C, t) = 1 - 0.54 = 0.46$$

The fourth step is to determine the weight of each factor. Dpf factor was used as an example of the determination of the weight for each factor. The significant gain for Dpf factor (S_{Dpf}) was calculated. S_{Dpf} was equal to 1.72 as follows:

$$S_{Dpf} = \frac{0.79}{0.46} = 1.72$$

The significant gain calculated for each factor is shown in Table 6.

TABLE 6. Values of each factor for the data no. 1.

	Factor							
	Pr	Gl	Bp	St	In	Bm	Dpf	Ag
$E(C_i)$	0.72	0.85	0.50	0	0	0.94	0.79	0.49
S_i	1.56	1.83	1.09	0	0	2.02	1.72	1.06
W_i	0.17	0.20	0.12	0	0	0.22	0.19	0.11
D_i	0.50	4.54	0.82	0	0	0.39	0.03	2.63
ND_i	0.11	1	0.18	0	0	0.09	0.01	0.58

TABLE 7. Examples of prediction results for dataset 1 using the proposed AWOD based method.

Data no.	Results											Observed Class	Matching Results
	Weight								Constant Value	AWOD	Predicted Class		
	Pr	Gl	Bp	St	In	Bm	Dpf	Ag					
1	0.17	0.20	0.12	0.00	0.00	0.22	0.19	0.11	1	0.25	PD	Yes	Y
2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0	0.00	AD	Yes	N
3	0.13	0.00	0.14	0.13	0.17	0.14	0.16	0.13	1	0.19	PD	Yes	Y
4	0.20	0.14	0.13	0.00	0.20	0.12	0.11	0.11	1	0.31	PD	Yes	Y
5	0.00	0.00	0.21	0.00	0.42	0.37	0.00	0.00	0	0.00	AD	No	Y
6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0	0.00	AD	No	Y
7	0.00	0.00	0.50	0.68	0.00	0.00	0.00	0.00	0	0.00	AD	No	Y
8	0.21	0.00	0.17	0.27	0.00	0.18	0.17	0.00	0	0.31	PD	Yes	Y
9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0	0.00	AD	No	Y
10	0.15	0.17	0.04	0.13	0.19	0.18	0.00	0.15	1	0.25	PD	Yes	Y
…	…	…	…	…	…	…	…	…	…	…	…	…	…
118	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0	0.00	AD	No	Y
												Total	Y = 110
													N = 8

The weight of the Dpf factor (W_{Dpf}) was determined. W_{Dpf} was equal to 0.18, as follows:

$$W_{Dpf} = \frac{1.72}{1.58 + 1.86 + 1.10 + 0 + 0 + 2.18 + 1.72 + 1.08} = 0.18$$

The weight of each factor is displayed in Table 6.

To determine the AWOD of all factors, the AWOD for the Dpf factor (D_{Dpf}) was first determined. D_{Dpf} was equal to 0.03, as follows:

$$D_{Dpf} = 0.18 \times \left| \sqrt{(0.484 - 0.550)^2} - \sqrt{(0.484 - 0.692)^2} \right| = 0.03$$

The average-based weighted objective distance of each factor is shown in Table 6.

Among all factors, the D_{max} was 4.54 and the D_{min} was 0. Then, the normalized average-based weighted objective distance for the Dpf factor (ND_{Dpf}) was determined. ND_{Dpf} was equal to 0.01, as follows:

$$ND_{Dpf} = \frac{0.03 - 0}{4.54 - 0} = 0.01$$

The normalized average-based weighted objective distance for each factor is demonstrated in Table 6.

The AWOD for all factors for data no.1 ($AWOD_1$) was determined. In this study, $lTn = 5$ and $hTn = 2$ were derived by observing the dataset. According to Step 23 in the AWOD algorithm, $b = 1 if\, nND_{(v=0)} < lTn$, which $nND_{(v=0)} = 2$ including St and In factors for the data no.1, so $b = 1$. $AWOD_1$ was equal to 0.25, as follows:

$$AWOD_1 = \frac{0.11 + 1 + 0.18 + 0 + 0 + 0.09 + 0.01 + 0.58}{8} \times 1 = 0.25$$

Different weights (W_i) representing significant and insignificant factors for data no. 1 in Table 6 include $W_{Pr} = 0.17$, $W_{Gl} = 0.20$, $W_{Bp} = 0.12$, $W_{St} = 0$, $W_{In} = 0$, $W_{Bm} = 0.22$, $W_{Dpf} = 0.19$, and $W_{Ag} = 0.11$. The weight with a value of 0 indicates an insignificant factor. In contrast, the weight with a value greater than 0 indicates a significant factor. Accordingly, the weights of Pr, Gl, Bp, Bm, Dpf, and Ag were deemed to be significant factors. The weights of St and In were indicated as insignificant factors. To identify the target class based on the obtained $AWOD_1$, the sample data no.1 was in the negative class because $AWOD_1 = 0.25$. Therefore, sample data no.1 can be predicted as an individual who has type 2 diabetes presence.

V. RESULTS AND DISCUSSION

The proposed AWOD based method for type 2 diabetes prediction uses Datasets 1 and 2 categorized into either AD ($AWOD = 0$) or PD ($0 < AWOD \leq 1$). To evaluate the prediction accuracy of the AWOD based method, the result was compared with the observed value for each data sample.

A. PREDICTION RESULTS OF THE AWOD BASED METHOD

Table 7 and Table 8 show examples of the type 2 diabetes prediction results for Dataset 1 and Dataset 2, respectively, using the proposed AWOD based method. In the tables, weights, constant values, AWOD values, and predicted classes are

TABLE 8. Examples of prediction results for dataset 2 using the proposed AWOD based method.

Data no.	Results													Observed Class	Matching Results
	Weight										Constant Value	AWOD	Predicted Class		
	Ag	Ur	Cr	Hb	Ch	Tg	Hd	Ld	Vl	Bm					
1	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0	0.00	AD	No	Y
2	0.14	0.13	0.16	0.17	0.12	0.00	0.06	0.09	0.00	0.14	1	0.29	PD	Yes	Y
3	0.19	0.07	0.13	0.21	0.00	0.13	0.11	0.00	0.00	0.17	1	0.20	PD	Yes	Y
4	0.17	0.13	0.13	0.00	0.21	0.00	0.15	0.20	0.00	0.00	1	0.15	PD	No	N
5	0.00	0.00	0.16	0.19	0.19	0.00	0.08	0.16	0.05	0.18	1	0.19	PD	Yes	Y
6	0.16	0.13	0.13	0.17	0.09	0.08	0.08	0.00	0.00	0.15	1	0.20	PD	Yes	Y
7	0.13	0.14	0.13	0.11	0.10	0.05	0.06	0.13	0.02	0.12	1	0.41	PD	Yes	Y
8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0	0.00	AD	No	Y
9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0	0.00	AD	No	Y
10	0.45	0.00	0.00	0.55	0.00	0.00	0.00	0.00	0.00	0.00	0	0.00	AD	Yes	N
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
118	0.19	0.11	0.20	0.15	0.00	0.10	0.09	0.00	0.03	0.13	1	0.26	PD	Yes	Y
														Total	Y = 116
															N = 2

showcased. ''Y'' means a class has been correctly predicted, and ''N'' means an incorrect prediction by matching between the predicted and the observed class. The predicted class as ''AD'' represents ''No'' for the observed class, whereas the predicted class as ''PD'' represents ''Yes'' for the observed class. The weight represents the factor that can affect the type 2 diabetes prediction. A constant value was used to calculate the AWOD value representing real effects towards the prediction based on the number of both significant and insignificant factors, and a factor with the maximum number of the AD class among all factors. For example, the significant factors for sample data no.1 include Pr, Gl, Bp, Dpf, and Ag. The insignificant factors are St, In, and Bm. According to Step 23 in the AWOD algorithm, the number of insignificant factors is less than the minimum number of factors affecting the prediction as the PD class, which represents significant factors, a constant value was equal to 1. Therefore, the AWOD value for sample data no.1 was equal to 0.25, which was predicted as the PD class ($0 < AWOD \leq 1$). In contrast, the significant factors for sample data no.5 include Bp, In, and Bm. The insignificant factors include Pr, Gl, St, Dpf, and Ag. According to Step 23 in the pseudocode for the proposed AWOD based method, the number of insignificant factors is more than the maximum number of factors that can affect identifying the negative class or the PD class, in other words, the number of insignificant factors is more than the number of significant factors so that a constant value was equal to 0. Therefore, the AWOD value for sample data no.5 was equal to 0.00, which was predicted as the AD class ($AWOD = 0$). Based on these two examples, both individuals have different sets of significant and insignificant factors that can be used to predict type 2 diabetes. In addition, each individual has specific health conditions influencing type 2 diabetes diagnosis, so a constant value is necessary for representing real effects towards the prediction to obtain the accurate class besides the weights.

B. AN EVALUATION OF PREDICTION PERFORMANCE FOR THE AWOD BASED METHOD

To evaluate prediction performance obtained from the AWOD based method, precision, specificity, and accuracy were used. Precision can measure how frequently the proposed AWOD based method correctly predicts true positive (TP) out of the total number of predicted positive classes. TP represents the individuals who were correctly predicted in the AD group. Specificity can measure the proportion of true negative (TN) that is correctly predicted out of the total number of negatives. TN represents the individuals who were correctly predicted in the PD class. Accuracy can measure the total prediction performance of the proposed AWOD based method, which indicates that both TP and TN are correctly predicted. To calculate precision, specificity, and accuracy, false positive (FP) and false negative (FN) are applied. FP means that the individuals were incorrectly predicted as the AD class, but the observed class is in the PD class. FN means that the individuals were incorrectly predicted as the PD class, but the observed class is in the AD class.

Table 9 demonstrates the performance of type 2 diabetes prediction for Dataset 1 and Dataset 2 using the AWOD based method. Each dataset used for type 2 diabetes prediction includes 392 records. The prediction performance for Dataset 1 indicated TP = 35, FP = 5, TN = 75, and FN = 3. For Dataset 2, the prediction performance provided TP = 88,

TABLE 9. AWOD Performance of type 2 diabetes prediction based on precision, specificity, and accuracy for dataset 1 and dataset 2.

	Type 2 diabetes prediction Performance	
	Dataset 1	Dataset 2
Precision	87.50%	98.88%
Specificity	93.75%	96.55%
Accuracy	93.22%	98.95%

FP = 1, TN = 28, and FN = 1. The precision was revealed at 87.50% for Dataset 1 and 98.88% for Dataset 2, indicating that the proposed method has the high potential for predicting an individual who has the type 2 diabetes absence. The specificity was revealed at 93.75% for Dataset 1 and 96.55% for Dataset 2, indicating that the proposed method has the ability in predicting an individual who has type 2 diabetes presence correctly. Particularly, the overall prediction accuracy revealed that the proposed AWOD based method provides high accuracy with 93.22% for Dataset 1 and 98.95% for Dataset 2.

After evaluating the prediction performance with precision, specificity, and accuracy, the AWOD based method has a high potential to predict the type 2 diabetes presence or absence. The determination of significant factors and insignificant factors applying information gain based on an average value of the acceptable level and the expected level, represented as weighting factors, can be used in the prediction process. The prioritization of those factors by different weights along with indicating a constant value affecting the actual class prediction appears to be a workable method for prediction.

C. COMPARATIVE PREDICTION RESULTS

The prediction accuracy obtained from the AWOD based method was compared against K-NN, SVM, RF, and DL classifiers, as shown in Table 10. The K-NN and SVM were employed in this study because both classifiers measure distances to obtain the prediction, which is similar to the proposed AWOD based method. For this study, the K-NN classifier performed the prediction with $K = 5$. The concept of AWOD based method is to consider significant individual factors and transform insignificant factors to be zero in the prediction process. Similarly, the RF classifier is widely used for feature extraction to identify the most significant features; therefore, the RF was employed to compare the prediction performance against the proposed method. Additionally, a state-of-the-art machine learning algorithm as DL was applied in this study to evaluate the prediction performance of the AWOD based method. A K-fold cross-validation technique was used for validating the prediction performance. This technique is appropriate for a limited dataset and yields minimum bias during the training process [42]. To obtain the prediction performance, the dataset was divided into 10 folds ($K = 10$) for employing in the training and testing process.

TABLE 10. Comparison between the AWOD based method and classifier results.

Method	Accuracy	
	Dataset 1	Dataset 2
K-Nearest Neighbors	71.68%	92.08%
Support Vector Machines	77.30%	93.45%
Random Forest	78.32%	98.20%
Deep Learning	74.74%	94.72%
AWOD based method	93.22%	98.95%

From Table 10, using the K-NN and SVM classifiers to predict type 2 diabetes presence or absence resulted in poor accuracy for Dataset 1 with 71.68% and 77.30% respectively. Although the K-NN and SVM classifiers provided good accuracy for Dataset 2 with 92.08% and 93.45% respectively, the prediction performance obtained from the proposed AWOD based method still provided better accuracy than those classifiers for both datasets, which are 93.22% for Dataset 1 and 98.95% for Dataset 2. It was caused by using all factors, significant and insignificant factors, to calculate the distance for all patients because some may not affect some individuals, but those factors were used in the prediction process. Moreover, the prediction accuracy obtained from DL provided 74.74% for Dataset1 and 94.72% for Dataset 2; however, the results obtained from the AWOD based method were still better. It caused from the DL requires a large training dataset for training the model based on a combination of data inputs, weights, and bias to derive better accuracy. Besides, the prediction results for both datasets using the RF classifier provided higher accuracy than those provided by other classifiers because this method chose only the most significant factors for prediction. According to the comparison results, the AWOD based method provided higher prediction performance than using other machine learning classifiers because this method works well with relatively small datasets, while larger datasets are required for those classifiers.

The proposed AWOD based method has the potential to predict the patients whether have type 2 diabetes presence or absence. Therefore, the assumption made by this study can be confirmed that the proposed AWOD based method can provide higher accuracy than those machine learning classifiers. In particular, the AWOD based method can determine significant factors and insignificant factors for the prediction process, which results in the high accuracy of prediction. It can be recognized that insignificant factors can affect type 2 diabetes prediction among individuals because patients have different health conditions. Some insignificant factors may represent the factors influencing the presence of type 2 diabetes for some individuals. Thus, those factors applied in the AWOD based method can enhance prediction performance.

However, the proposed AWOD based method still provided an approximate error of 6.78% for Dataset 1 and 1.05% for Dataset 2 for incorrect prediction. Among the incorrect

prediction cases, the individuals were predicted in the wrong class. Most cases may cause by an individual having specific health conditions. This condition resulted in the current level of those individuals being indicated in the improper range either below or above the acceptable level calculated by the average score. Additionally, inaccurate predictions may obtain from constant values. The minimum and maximum numbers of factors used for determining constant values may not be workable for predicting type 2 diabetes in those individuals. Therefore, determining the expected level, the acceptable level, and constant values will be considered for future studies by modifying AWOD based method or applying different analytical points of view for obtaining better prediction performance.

The proposed AWOD based method can benefit the diagnosis of any chronic diseases with a relatively small dataset that is hardly collected more often and many of them are low frequency of change. For the limitations, computational complexity should be further investigated for future work. Processing large datasets using AWOD based method may encounter computational complexity problems because there are several computational stages involved and requires many parameter settings. In addition, the proposed AWOD method requires complicated parameter settings in order to apply for multi-category classifications. It is also worth modifying the AWOD to be more generalized for other multi-category classifications for future work.

VI. CONCLUSION

This study proposes a novel prediction method, called average-based weighted objective distance (AWOD), for type 2 diabetes prediction. The AWOD based method is based on the principle of health care professionals that considers individual health conditions for diagnosis. The proposed method employed information gain based on average values of expected levels and acceptable levels to prioritize factors referred to as weighing factors. The prioritized factor indicates significant and insignificant factors for individuals. Those factors can represent real effects towards prediction based on diverse individual health conditions. The open data named Pima Indians Diabetes dataset and Mendeley Data for Diabetes dataset were studied for the experiment, which contains 392 records for each set. The prediction performance obtained from the AWOD based method was evaluated by precision, specificity, and accuracy. The comparison results for prediction performance revealed that the AWOD based method provided 93.22% and 98.95% accuracy for Dataset 1 and Dataset 2 respectively, which are more accurate than those of machine learning-based prediction methods including K-NN, SVM, RF, and DL.

APPENDIX

See Table 11.

TABLE 11. Table of symbols for the AWOD based method.

Symbol	Description	Symbol	Description
u_{j+}	Value of factor j that is in the positive class for the U set	pXZ_{j-}	Probability of the current distance for the negative class
nu_{j+}	Total number of factor j that is in the positive class for the U set	$E(C_j)$	Entropy of each factor
u_{j-}	Value of factor j that is in the negative class for the U set	$E(Ct)$	Entropy of all factors
nu_{j-}	Total number of factor j that is in the negative class for the U set	$Gain(C,t)$	Information gain of the target class with respect to all factors
$Y(a)_j$	An average number in the negative class for the U set	S_j	Significant gain for each factor
EP_+	Value of the equal probability for the positive class	W_j	Weight of each factor
EP_-	Value of the equal probability for the negative class	D_j	Average-based weighted objective distance for each factor
Tn	Total number of factors	ND_j	Normalized average-based weighted objective distance for each factor
T	Total number of target classes	D_{max}	Maximum value of D_j among all factors
F_+	Positive-target class fraction	D_{min}	Minimum value of D_j among all factors
F_-	Negative-target class fraction	$AWOD_i$	Average-based weighted objective distance for all factors for the i^{th} individual
$E(C)$	Entropy of the target class with respect to all factors	b	Constant value influencing the actual class
dXY_j	Acceptable distance of each factor	$nND_{(v=0)}$	Total number of factors with the normalized average-based weighted objective distance that is equal to 0
dXZ_j	Current distance of each factor	lTn	Minimum number of factors that can affect identifying the negative class

TABLE 11. *(Continued.)* **Table of symbols for the AWOD based method.**

X_j	Expected level of factor j	hTn	Maximum number of factors that can affect identifying the negative class
Y_j	Acceptable level of factor j	$ND_{(v=0)}$	Factor with the normalized average-based weighted objective distance that is equal to 0
Z_j	Current level of factor j	$MAXnu_{j+}$	Factor with the maximum number of the positive class among all factors
pXY_{j+}	Probability of the acceptable distance for the positive class		

REFERENCES

[1] N. Barakat, A. P. Bradley, and M. N. H. Barakat, "Intelligible support vector machines for diagnosis of diabetes mellitus," *IEEE Trans. Inf. Technol. Biomed.*, vol. 14, no. 4, pp. 1114–1120, Jul. 2010.

[2] K. Zarkogianni, M. Athanasiou, A. C. Thanopoulou, and K. S. Nikita, "Comparison of machine learning approaches toward assessing the risk of developing cardiovascular disease as a long-term diabetes complication," *IEEE J. Biomed. Health Informat.*, vol. 22, no. 5, pp. 1637–1647, Sep. 2018.

[3] A. E. Kitabchi, G. E. Umpierrez, J. M. Miles, and J. N. Fisher, "Hyperglycemic crises in adult patients with diabetes," *Diabetes Care*, vol. 32, no. 7, pp. 1335–1343, Jul. 2009.

[4] R. Muniyappa and S. Gubbi, "COVID-19 pandemic, coronaviruses, and diabetes mellitus," *Amer. J. Physiol.-Endocrinol. Metabolism*, vol. 318, no. 5, pp. E736–E741, May 2020.

[5] International Diabetes Federation. (2019). *Global Diabetes Data Report 2010–2045*. IDF Diabetes Atlas 9th Edition 2019. Accessed: Nov. 1, 2020. [Online]. Available: https://diabetesatlas.org/data/en/world/

[6] U.S. Department of Health and Human Services. *Diabetes Detection Initiative: Finding the Undiagnosed*. Accessed: Nov. 3, 2020. [Online]. Available: http://www.ndep.nih.gov/ddi/about/index.htm

[7] G. Roglic, "WHO Global report on diabetes: A summary," *Int. J. Noncommunicable Diseases*, vol. 1, no. 1, pp. 3–8, Jun. 2016.

[8] (2020). *International Diabetes Federation: Diabetes Facts & Figures*. IDF Diabetes Atlas 9th Edition 2019. Accessed: Nov. 1, 2020. [Online]. Available: https://www.idf.org/aboutdiabetes/what-is-diabetes/facts-figures.html

[9] S. Kjeldsen, R. D. Feldman, L. Lisheng, J.-J. Mourad, C.-E. Chiang, W. Zhang, Z. Wu, W. Li, and B. Williams, "Updated national and international hypertension guidelines: A review of current recommendations," *Drugs*, vol. 74, no. 17, pp. 2033–2051, Oct. 2014.

[10] S. Chaising, P. Temdee, and R. Prasad, "Weighted objective distance for the classification of elderly people with hypertension," *Knowl.-Based Syst.*, vol. 210, Dec. 2020, Art. no. 106441.

[11] M. S. Islam, M. K. Qaraqe, S. B. Belhaouari, and M. A. Abdul-Ghani, "Advanced techniques for predicting the future progression of type 2 diabetes," *IEEE Access*, vol. 8, pp. 120537–120547, 2020.

[12] N. L. Fitriyani, M. Syafrudin, G. Alfian, and J. Rhee, "Development of disease prediction model based on ensemble learning approach for diabetes and hypertension," *IEEE Access*, vol. 7, pp. 144777–144789, 2019.

[13] F. Aofa, P. S. Sasongko, Sutikno, Suhartono, and W. A. Adzani, "Early detection system of diabetes mellitus disease using artificial neural network backpropagation with adaptive learning rate and particle swarm optimization," in *Proc. 2nd Int. Conf. Informat. Comput. Sci. (ICICoS)*, Semarang, Indonesia, Oct. 2018, pp. 1–5.

[14] M. T. Mira Kania Sabariah, S. T. Aini Hanifa, and M. T. Siti Sa'adah, "Early detection of type II Diabetes Mellitus with random forest and classification and regression tree (CART)," in *Proc. Int. Conf. Adv. Inform.: Concept, Theory Appl. (ICAICTA)*, Bandung, Indonesia, Aug. 2014, pp. 238–242.

[15] T. M. Le, T. M. Vo, T. N. Pham, and S. V. T. Dao, "A novel wrapper–based feature selection for early diabetes prediction enhanced with a metaheuristic," *IEEE Access*, vol. 9, pp. 7869–7884, 2020.

[16] B. G. Choi, S. W. Rha, S. W. Kim, J. H. Kang, J. Y. Park, and Y. K. Noh, "Machine learning for the prediction of new-onset diabetes mellitus during 5-year follow-up in non-diabetic patients with cardiovascular risks," *Yonsei Med. J.*, vol. 60, no. 2, pp. 191–199, Feb. 2019.

[17] K. Saxena, Z. Khan, and S. Singh, "Diagnosis of diabetes mellitus using K nearest neighbor algorithm," *Int. J. Comput. Sci. Trends Technol.*, vol. 2, no. 4, pp. 36–43, Jul./Aug. 2014.

[18] T. N. Joshi and P. P. M. Chawan, "Diabetes prediction using machine learning techniques," *J. Eng. Res. Appl.*, vol. 8, no. 1, pp. 9–13, Jan. 2018.

[19] N. P. Tigga and S. Garg, "Prediction of type 2 diabetes using machine learning classification methods," *Proc. Comput. Sci.*, vol. 167, pp. 706–716, Mar. 2020.

[20] F. Kazerouni, A. Bayani, F. Asadi, L. Saeidi, N. Parvizi, and Z. Mansoori, "Type2 diabetes mellitus prediction using data mining algorithms based on the long-noncoding RNAs expression: A comparison of four data mining approaches," *BMC Bioinf.*, vol. 21, no. 1, pp. 1–13, Aug. 2020.

[21] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, and H. Tang, "Predicting diabetes mellitus with machine learning techniques," *Frontiers Genet.*, vol. 9, p. 515, Nov. 2018.

[22] J. Gou, H. Ma, W. Ou, S. Zeng, Y. Rao, and H. Yang, "A generalized mean distance-based K-nearest neighbor classifier," *Expert Syst. Appl.*, vol. 115, pp. 356–372, Jan. 2019.

[23] J. Gou, W. Qiu, Z. Yi, Y. Xu, Q. Mao, and Y. Zhan, "A local mean representation-based K-nearest neighbor classifier," *ACM Trans. Intell. Syst. Technol.*, vol. 10, no. 3, p. 1–25, 2019.

[24] J. Gou, W. Qiu, Z. Yi, X. Shen, Y. Zhan, and W. Ou, "Locality constrained representation-based K-nearest neighbor classification," *Knowl.-Based Syst.*, vol. 167, pp. 38–52, Mar. 2019.

[25] A. Dhar, N. Dash, and K. Roy, "Classification of text documents through distance measurement: An experiment with multi-domain Bangla text documents," in *Proc. 3rd Int. Conf. Adv. Comput., Commun. Autom. (ICACCA) (Fall)*, Dehradun, India, Sep. 2017, pp. 1–6.

[26] G. Latifa, M. Jazouli, N. Es-Sbai, A. Majda, and A. Zarghili, "Comparison between Euclidean and Manhattan distance measure for facial expressions classification," in *Proc. Int. Conf. Wireless Technol., Embedded Intell. Syst. (WITS)*, Fez, Morocco, Apr. 2017, pp. 1–4.

[27] N. Kwon, J. Lee, M. Park, I. Yoon, and Y. Ahn, "Performance evaluation of distance measurement methods for construction noise prediction using case-based reasoning," *Sustainability*, vol. 11, no. 3, p. 871, Feb. 2019.

[28] S. Demirci, I. Erer, and O. Ersoy, "Weighted Chebyshev distance classification method for hyperspectral imaging," *Proc. SPIE*, vol. 9482, Jun. 2015, Art. no. 948218.

[29] S. Chaising and P. Temdee, "Determining recommendations for preventing elderly people from cardiovascular disease complication using objective distance," in *Proc. Global Wireless Summit (GWS)*, Chiang rai, Thailand, Nov. 2018, pp. 151–155.

[30] S. Chaising, R. Prasad, and P. Temdee, "Personalized recommendation method for preventing elderly people from cardiovascular disease complication using integrated objective distance," *Wireless Pers. Commun.*, vol. 117, pp. 215–233, Aug. 2019.

[31] L. H. Patil and M. Atique, "A novel feature selection based on information gain using WordNet," in *Proc. Sci. Inf. Conf.*, London, U.K., Oct. 2013, pp. 625–629.

[32] T. A. Alhaj, M. M. Siraj, A. Zainal, H. T. Elshoush, and F. Elhaj, "Feature selection using information gain for improved structural-based alert correlation," *PLoS ONE*, vol. 11, no. 11, Nov. 2016, Art. no. e0166017.

[33] R. B. Pereira, A. P. D. Carvalho, B. Zadrozny, and L. H. D. C. Merschmann, "Information gain feature selection for multi-label classification," *J. Inf. Data Manage.*, vol. 6, no. 1, pp. 48–58, Feb. 2015.

[34] B. Azhagusundari and A. S. Thanamani, "Feature selection based on information gain," *Int. J. Innov. Technol. Exploring Eng.*, vol. 2, no. 2, pp. 18–21, Jan. 2013.

[35] M. Gupta, "Dynamic k-NN with attribute weighting for automatic web page classification (Dk-NNwAW)," *Int. J. Comput. Appl.*, vol. 58, no. 10, pp. 34–40, Nov. 2012.

[36] S. Chaising, P. Temdee, and R. Prasad, "Individual attribute selection using information gain based distance for group classification of elderly people with hypertension," *IEEE Access*, vol. 9, pp. 82713–82725, 2021, doi: 10.1109/ACCESS.2021.3084623.

[37] UCI Machine Learning. *Pima Indians Diabetes Database*. Kaggle. Accessed: May 10, 2020. [Online]. Available: https://www.kaggle.com/uciml/pima-indians-diabetes-database
[38] A. Rashid. *Diabetes Dataset*. Mendeley Data. Accessed: May 10, 2020. [Online]. Available: https://data.mendeley.com/datasets/wj9rwkp9c2/1
[39] Cochrane U.K. *How Did They Determine Diagnostic Thresholds?*. Accessed: Dec. 30, 2020. [Online]. Available: https://uk.cochrane.org/news/how-did-they-determine-diagnostic-thresholds
[40] A. Indrayan. *Diagnostic Thresholds of Medical Measurements*. Medical Biostatistics & Research. Accessed: Dec. 30, 2020. [Online]. Available: http://www.medicalbiostatistics.com/Diagnostic%20Threshold.pdf
[41] J. Brownlee. *A Gentle Introduction to K-Fold Cross-Validation. Mach. Learn. Mastery (2018)*. Accessed: Aug. 6, 2020. [Online]. Available: https://machinelearningmastery.com/k-fold-crossvalidation/

PRATYA NUANKAEW received the bachelor's degree in educational technology and the master's degree in information technology from Naresuan University, Thailand, in 2001 and 2008, respectively, and the doctorate's degree in computer engineering from Mae Fah Luang University, Thailand, in 2018. His research interests include applied informatics technology, educational data mining, educational engineering, educational technology, learning analytics, and learning strategies for lifelong learning.

SUPANSA CHAISING received the bachelor's degree in accounting, the master's degree, and the Ph.D. degree in computer engineering from Mae Fah Luang University, Thailand. She is currently a Lecturer with the Department of Information Technology, International College, Payap University, Thailand. Her research interests include artificial intelligence, machine learning, and data analysis.

PUNNARUMOL TEMDEE (Member, IEEE) received the bachelor's degree in electronic and telecommunication engineering, the master's degree in electrical engineering, and the Ph.D. degree in electrical and computer engineering from King Mongkut's University of Technology Thonburi, Thailand. She is currently an Associate Professor with Mae Fah Luang University, Chiang Rai, Thailand. Her research interests include social network analysis, artificial intelligence, software agent, context-aware computing, and ubiquitous computing.