

Soft computing techniques for early diabetes prediction

Sabah Anwer Abdulkareem¹, Hussien Yossif Radhi¹, Yousra Ahmed Fadil², Hussain Falih Mahdi¹

¹Department of Computer Engineering, College of Engineering, University of Diyala, Diyala, Iraq

²College of Law and Political Sciences University of Diyala, Diyala, Iraq

Article Info

Article history:

Received Jul 31, 2021

Revised Nov 9, 2021

Accepted Dec 1, 2021

Keywords:

Artificial neural networks

Diabetes

Multi criteria decision making

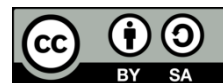
fuzzy analytical hierarchy processes

Support vector machines

ABSTRACT

Diabetes mellitus is a chronic, life-threatening, and complicated condition. Around 1.5 million deaths due to diabetes have been documented, according to a World Health Organization (WHO) estimation in 2019. In the world of medicine, predicting diabetes risk is a difficult and time-consuming task. Many past studies have been conducted to investigate and clarify diabetes symptoms and variables. To solve these persisting issues, however, more critical clinical criteria must be considered. A comparative analysis based on three soft computing strategies for diabetes prediction has been carried out and achieved in this work. Among the computational intelligence methods used in this study are fuzzy analytical hierarchy processes (FAHP), support vector machine (SVM), and artificial neural networks (ANNs). The techniques reveal promising performance in predicting diabetes reliably and effectively in terms of several classification evaluation metrics, according to experimental analysis and assessment conducted on 520 participants using a publicly available dataset.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Sabah Anwer Abdulkareem

Department of Computer Engineering, College of Engineering, University of Diyala

Diyala, Iraq

Email: sbh_anwar@uodiyala.edu.iq

1. INTRODUCTION

One of the most prevalent endocrine diseases is diabetes mellitus. That requires ongoing medical care with several strategies to reduce the external risk of glycemic control. An imbalance occurs in the person's nutritional metabolism, leading to many complications and long-term effects, including the heart, kidneys, eyes, nerves, and blood vessels. To diagnose symptomatic diabetes by doctors, the patient shows many signs and symptoms resulting from the osmotic separation causing high blood sugar [1]. The diseases of diabetes complications can be divided into two combinations according to their damage, including macrovascular damage (the arteries) and microvascular damage (small blood vessels). Accelerated cardiovascular disease, which presents as strokes and other catastrophic illnesses, is the most devastating macrovascular consequence. Microvascular illnesses such as retinopathy in the eye, nephropathy in the kidney, and neuropathy in the nervous system are examples of organ-specific disorders [2], [3].

According to the World Health Organization (WHO), diabetes affects 422 million people worldwide in 2018 (WHO). Type 1 and type 2 diabetes are the two forms of diabetes. Type 1 diabetes mellitus is an autoimmune disease that can strike anyone at any age, but it strikes children and adolescents more frequently. The immune system incorrectly destroys pancreatic beta cells, resulting in total insulin insufficiency, a small amount of insulin released into the body, or even no insulin released into the body. Mellitus is a Type 2 diabetes complication that arises when the body loses b-cell insulin secretion over time, generates insufficient insulin, or stays insulin-resistant. Gestational diabetes, on the other hand, is caused by hormonal changes that only occur during pregnancy. Type 1 diabetes, Type 2 diabetes, gestational diabetes, and other types of

diabetes caused by various factors are classified by some researchers and the American diabetes association (ADA) [4]. Gestational diabetes mellitus (GDM) is a kind of diabetes that only affects pregnant women and usually develops between the 24th and 28th week of pregnancy (except for women who already have chronic diabetes). When blood glucose levels rise over normal during pregnancy, it is detected. After the baby is born, the majority of mothers will not get diabetes. After childbirth, however, some women will continue to have high blood glucose levels [5].

According to diabetes Australia, diabetes can be present for up to seven years before clinical diagnosis. During this time, a person may acquire potentially fatal conditions such as blindness from eye damage, foot ulcers that may need amputation of the affected limbs, renal failure, and heart attacks [6]–[10]. Fiarni and others coined the term "silent killer" to describe it for these reasons [3]. With regular examinations, early detection, and treatment initiation, these repercussions can be prevented, managed, or even eliminated in some persons, saving roughly 1415 US dollars [5].

Artificial intelligence technologies are routinely employed to detect and diagnose diseases automatically. The authors in [11] recommended the use of a subset evaluator (CSE) as a method for identifying the most important risk variables for diabetes prevalence in the body. Based on the Pima Indian Diabetes dataset, the authors combined CSE and decision tree (DT) to create a classifier subset evaluator decision tree (CSE-DT) (PIDD). Moreover, Shuja *et al.* [12] developed a two-stage approach for diabetic prediction based on data mining categorization techniques: For data preprocessing, the initial stage is synthetic minority oversampling technique (SMOTE). Five machine learning classifiers are used in the second stage: simple logistic, decision tree, bagging, artificial neural networks (ANN), and support vector machine (SVM). Swapna *et al.* [13] a method for distinguishing normal heart rate variability (HRV) signals and diabetes was proposed utilizing deep learning architectures. To construct an integrated system, the authors merged a convolutional neural network (CNN) with long short-term memory (LSTM). Multiple criteria decision making (MCDM), a sub-field and key branch of operations research (OR) have been widely employed in many scientific domains for problem-solving and decision-making in addition to artificial intelligence algorithms [14], [15]. The MCDA or MCDM problem normally comprises four phases: formulation of options, criterion selection, criterion weighting, and decision making [16]. By merging MCDM approaches with artificial intelligence techniques, particularly soft computing technologies, hybrid methodologies can be constructed. The fuzzy analytic hierarchy process (FAHP) is a multi-criterion decision-making technique analytic hierarchy process (AHP) that incorporates fuzzy theories (a branch of soft computing) [17].

In the literature, there have been many methods developed for diabetes prediction. logistic regression is one of the classification techniques used to predict diabetes [18]. The authors employed seven factors as attributes in their data analysis, resulting in a prediction probability of 78.5565. Another work, Morgan *et al.* [19] used the world health survey plus (WHS+), which was performed with WHO support across five Gulf cooperation council (GCC) nations in 2008 and 2009, including the UAE, Kuwait, Saudi Arabia, and Oman. The sample sizes for the authors were UAE (n=2569), Kuwait (n=3828), Saudi Arabia (n=8629), and Oman (n=4717). According to the findings, Oman has the lowest standardized prevalence of diabetes at 8.5% (7.4%–9.8%), followed by Saudi Arabia at 10.5% (9.6%–11.4%), the United Arab Emirates at 13.2% (11.4%–15.2%), and Kuwait at 15.3% (13.9%–16.8%). Singh *et al.* [20] The Pima Indians diabetes dataset was also used to develop machine learning approaches for diabetes diagnosis, including likelihood-based naive bayes (NB), decision tree-based random forest (RF), and multi-layered function-based random forest (RF) (MLP). They demonstrated that data pre-processing can improve the performance of machine learning algorithms.

Vidhya and Shanmugalakshmi [21] considered many risk factors to predict diabetes mellitus including, the patients' body mass index level, bad eating habits, poor exercise, smoking, nature of work, and other factors regardless of the age or gender of the patient. They found the existing techniques, ANN and SVM, achieved an accuracy of 57.41% and 62.81%. In comparison, deep belief network (DBN) achieved an accuracy of 80.99%, achieving better results than the other machine learning methods. Ding *et al.* [22] proposed a novel approach for predicting diabetic complications based on the similarity enhanced latent Dirichlet assignment (seLDA) model. After preprocessing the data, they computed the similarity between each pair based on medical records, then used the similarity estimates as constraints in seLDA-based diabetes complications mining. The proposed approach (SVM-seLDA) consistently beat the traditional and seLDA-based approaches by 22.49% in estimating similarity and predicting diabetic complications, according to the experimental data. Liu *et al.* [23] the disease type 2 diabetes mellitus (T2DM) was used as a case study, with the focus on issues that occur after the first diagnosis. Modeling some risk complications and analyzing the linkages between risk factor selection patterns are the foundations of their strategy. They concluded that the Bayesian hierarchical framework outperformed the most recent models.

FAHP has been used in many applications and different fields, such as weapon selection [24], personnel selection [25], job selection [26], energy alternatives selection [27], and performance evaluation systems [28], [29]. Chamodrakas *et al.* [30] used a fuzzy AHP method for supplier selection in electronic marketplaces. Similarly, Kilincci *et al.* [31] used the fuzzy AHP approach in the washing machine company for selection purposes. Shaw *et al.* [32] proposed combining fuzzy AHP and fuzzy objective linear programming to better select a supplier for developing a low-carbon supply chain. Furthermore, Arikan [33] to handle multiple objective supplier selection challenges, we created an interactive solution using Fuzzy AHP. Their proposed strategy had three objectives: reduce total financial costs, improve overall quality, and improve customer service.

Data mining techniques were used in the most recently established diabetes prediction algorithms [34]. Islam *et al.* [34] a dataset of 520 occurrences was examined using three machine learning techniques: the Naive Bayes method, the logistic regression algorithm, and the random forest algorithm. Random forest algorithms produced the most accurate results, according to their research. To predict diabetes García-Ordás *et al.* [35] used state-of-the-art deep learning approaches. On the Pima Indians diabetes dataset, they used a variational autoencoder (VAE) to increase data, a sparse autoencoder (SAE) to increase features, and a CNN for classification, using a variational autoencoder (VAE) to increase data, a sparse autoencoder (SAE) to increase features, and a CNN to increase classification.

Despite the researchers' best efforts, more research and comparison of diabetes prediction approaches are still needed and open for investigation. For early diabetes prediction, this paper presents a comparative study involving three different soft computing techniques and a multi-criteria decision analysis method. For diabetes prediction and classification, the researchers used the fuzzy analytical hierarchy process (FAHP), artificial neural network (ANN), and support vector machine (SVM) methodologies. Selecting these three soft computing approaches in our study is based on the diversity of concepts and strength/weaknesses of those methods. This is the first study conducted to compare the performance of a multi-criteria decision-making method for diabetes prediction to the other computational intelligence methods. On publicly available datasets gathered from 520 patient and control subjects, the methods were tested.

The remaining sections of this work are presented as in: the approaches were provided and explained in section 2. The experimental data and comments are presented in section 3. Finally, section 4 brings the process to a close.

2. RESEARCH METHOD

To conduct our comparative study, three methods have been considered. The description of these methods is presented briefly in this section as in:

2.1. Fuzzy analytical hierarchy process (FAHP)

Saaty [17] a multi-criteria decision-making technique called the FAHP was devised. To build FAHP, the fuzzy theory is incorporated into the fundamental AHP. The AHP approach, for example, has been used in the past to forecast sickness [36]. As a result, FAHP is used to predict the occurrence of diabetes in this study. FAHP is a decision-making procedure that is frequently employed in cases involving conflicting criteria. The pairwise comparisons of the criteria and alternatives are implemented in FAHP by using triangular numbers to represent the variables [37]. The steps of the FAHP method can be described as:

- Step 1: By using the language concepts in Table 1, the decision-maker compares the criteria and options. As a result, the comparison matrix can be created. In terms of the fuzzy triangular scales that these linguistic concepts correspond to, for example, if the decision-maker declares that Criterion 1 (C1) is less important than Criterion 2 (C2), it uses the fuzzy triangle scale as: (2)-(4). In the pairwise contribution matrix of the criteria, the comparison of C2 to C1 will use the fuzzy triangle scale (1/4, 1/3, 1/2). The pairwise contribution matrix is shown in (1), where signifies the *zth* decision maker's preference for the *xth* criterion over the *yth* criterion, using fuzzy triangular scales. "Tilde" indicates to the triangular scale description, for the instance, the first decision maker's preference of criterion1 over criterion 2, equals to $m_{12}^1 = (2,3,4)$. *M* matrix of pairwise comparison is formed (1), where \tilde{m}_{xy}^z denotes that the *zth* makes the selection of the *xth* dimension over the *yth* dimension.

$$\tilde{M}^z = \begin{bmatrix} \tilde{m}_{11}^z & m_{12}^z & \dots & \tilde{m}_{1n}^z \\ \tilde{m}_{21}^z & \dots & \dots & \tilde{m}_{2n}^z \\ \dots & \dots & \dots & \dots \\ \tilde{m}_{n1}^z & \tilde{m}_{n2}^z & \dots & \tilde{m}_{nn}^z \end{bmatrix} \tag{1}$$

- Step 2: Priority of each defendant (\tilde{m}_{xy}^z) is collected \tilde{m}_{xy} using (2):

$$\tilde{m}_{xy} = \frac{\sum_{z=1}^z \tilde{m}_{xy}^z}{z} \tag{2}$$

Table 1. Linguistic terms and the corresponding fuzzy triangular scales

Integrate Scale	Linguistic Description	Fuzzy Triangular Scale
1/9	Extremely less important	(1/9, 1/9, 1/9)
1/8	The intermediate values between two adjacent scales	(1/9, 1/8, 1/7)
1/7	Very strongly less important	(1/8, 1/7, 1/6)
1/6	The intermediate values between two adjacent scales	(1/7, 1/6, 1/5)
1/5	strongly less important	(1/6, 1/5, 1/4)
1/4	The intermediate values between two adjacent scales	(1/5, 1/4, 1/3)
1/3	Moderately less important	(1/4, 1/3, 1/2)
1/2	The intermediate values between two adjacent scales	(1/3, 1/2, 1/1)
1	Equal Important	(1, 1, 1)
2	The intermediate values between two adjacent scales	(1, 2, 3)
3	Moderately more important	(2, 3, 4)
4	The intermediate values between two adjacent scales	(3, 4, 5)
5	Strongly more important	(4, 5, 6)
6	The intermediate values between two adjacent scales	(5, 6, 7)
7	Very strongly more important	(6, 7, 8)
8	The intermediate values between two adjacent scales	(7, 8, 9)
9	Extremely more important	(9, 9, 9)

- Step 3: The matrix of pairwise comparison is updated based on the average of responses.

$$\tilde{M} = \begin{bmatrix} \tilde{m}_{11} & \dots & \tilde{m}_{1n} \\ \vdots & \ddots & \vdots \\ \tilde{m}_{n1} & \dots & \tilde{m}_{nn} \end{bmatrix} \tag{3}$$

- Step 4: Calculating the geometric mean of a fuzzy valuation matrix for each dimension.

$$\tilde{r}_x = \left(\prod_{y=1}^n \tilde{m}_{xy} \right)^{1/n}, x = 1, 2 \dots \dots n \tag{4}$$

- Step 5: Merged 3 steps together and calculating the fuzzy weights of each criterion as shown in:

- Step 5.1: Calculate the vector summation of \tilde{r}_x .

$$\tilde{r}_x = \sum_{x=1}^n \tilde{r}_x \tag{5}$$

- Step 5.2: Calculate the power of negative one of summation vector.

$$(\tilde{r}_x)^{-1} = \left(\sum_{x=1}^n \tilde{r}_x \right)^{-1} \tag{6}$$

- Step 5.3: Calculated the fuzzy weights of each criterion by multiply with reverse each \tilde{r}_x of summation vector:

$$\tilde{w}_x = \tilde{r}_x \otimes (\tilde{r}_1 \oplus \tilde{r}_2 \oplus \dots \dots \tilde{r}_n)^{-1} = (lw_x, mw_x, uw_x) \tag{7}$$

- Step 6: Using the area center method, the non-fuzzy value is calculated by using (8).

$$A_x = \frac{lw_x + mw_x + uw_x}{3} \tag{8}$$

- Step 7: Non -Fuzzy values is normalized, by using (9):

$$N_x = \frac{A_x}{\sum_{x=1}^n A_x} \tag{9}$$

2.2. Artificial neural networks (ANNs)

A multi-layer perceptron architecture called an ANN [38] is used to train and classify input patterns to produce the required output. We used a three-layer network in our ANN model, with 14 neurons in the

input layer, 16 neurons in the hidden layer, and two neurons in the output layer. The number of neurons in hidden layer was chosen empirically, providing the best performance. Backpropagation was used as an optimization tool to adjust the network's weights with slope decline after our ANN model was trained on the data. To improve the performance of the ANN model, we use the cross-entropy loss function to alter the weights by minimizing the error at each stage. The soft-max activation function is employed in the output layer to generate the final prediction. The fourteen criteria serve as the network's input. At the same time, the output layer distinguishes between diabetes presence (1) and diabetes absence (0).

The equation of the cross-entropy loss function can be defined as (10):

$$L = - \sum_{i=1}^2 t_i \text{Log}(p_i) \tag{10}$$

where t refers to the ground truth value of sample i and p represents the probability of the sample resulted from soft-max activation function.

2.3. Support vector machines (SVMs)

SVMs are a type of machine learning and artificial intelligence method that is often employed in supervised learning. One of the most powerful prediction algorithms is SVMs. During the training process, SVMs develop a model for the dataset to predict point labels. SVMs learn a linear decision boundary to distinguish between the two classes based on a set of binaries labeled training vectors. The model is evaluated using the derived linear classification rule to categorize additional test instances [39]. A solid margin classifier, the simplest sort of SVM, is used to discover the linear classifier rule with the greatest geometric margin. Many optimization problems can be solved with linear SVM. The SVM static margin develops the hard hyperplane in the linearly separable state to obtain all data accurately sorted and increase the distance to the nearest training data points. In real-world data, datasets are frequently not linearly separable, necessitating the adjustment of the SVM. By using the soft margin principle, this modification is required to achieve a trade-off between maximizing geometric margin and decreasing classification error on the points of training data. The soft margin creates a hyperplane, allowing incorrect classification of difficult cases to increase the distance between them and the next entirely separated data samples [39], [40]. Suppose the training set is given as $(x_1, l_1), (x_2, l_2), \dots, (x_n, l_n)$, where x is the feature set and l_i are the labels with R classes, $l_i \in \{1, 2, \dots, R\}$. The primal problem for SVM is given as (11):

$$\begin{aligned} \min_{w,b,p} & \frac{1}{2} \sum_{m \in R} \|W_m\|^2 + C \sum_{i=1}^N \sum_{m \neq y_i} P_i^m \\ \text{s. t. : } & W_{y_i}^T \cdot X_i + b_{l_i} - (W_m^T \cdot X_i + b_m) \geq 1 - P_i^m \\ & P_i^m \geq 0, i = 1, \dots, n; m \in \{1, \dots, R\} \end{aligned} \tag{11}$$

where P_i represents the distance of the point that is classified in the wrong class from the margin, W represents the weights, b is bias, and C is a constant coefficient that its value reflects the weight of the penalty.

3. EXPERMENAL RESULTS AND AYSISNTAL

To carry out the experiments, we used a dataset collected by Islam *et al.* [34]. The number of subjects in the provided dataset is 520 persons with fourteen attributes represent symptoms that may cause diabetes. In addition, two more attributes (age and gender) representing socio-demographic are included in the dataset. This study only focuses on symptom attributes of diabetes disease, including genital thrush, alopecia, weakness, obesity, muscle stiffness, delayed healing, polydipsia, polyuria, polyphagia, and visual blurring, irritability, sudden weight loss, partial paresis, and itching. This dataset has been collected by a conducting survey using questionnaires targeting people who have recently got diabetic or are still nondiabetic but have some symptoms. There are 404 female and 116 male persons, their age between (16-90). The allocation of symptoms among persons is illustrated in Figure 1.

The pairwise comparison (PWC) matrix of the FAHP should be prepared initially to establish the FAHP diagnosis technique depicted in Figure 2. The importance of symptoms was rated using the doctor's judgment, as indicated in Table 2. The symptoms are abbreviated as: GT: genital thrush, AC: Alopecia, WN: weakness, OS: obesity, MS: muscle stiffness, DH: delayed healing, PD: Polydipsia, PR: Polyuria, PG: Polyphagia, VB: visual blurring, IA: irritability, SWL: sudden weight loss, PP: partial paresis, and LI: itching. The consistent comparison matrix, which is valid for experiments, was established after multiple revisions in the pairwise comparison matrix with the doctor's assistance. The achieved consistency ratio of the pairwise comparison matrix is 0.09, which is less than 0.1.

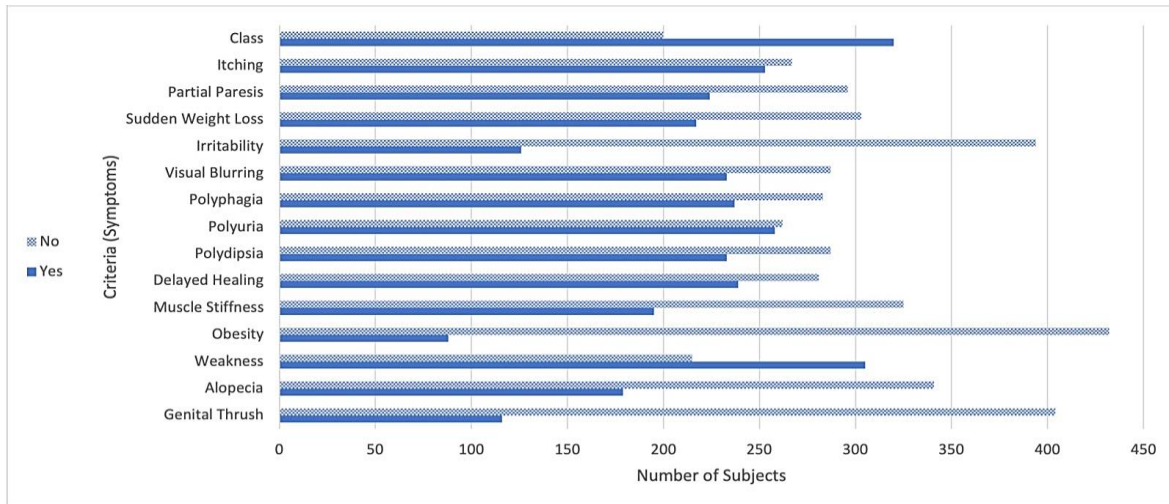


Figure 1. Symptoms distribution among 520 persons. Yes: indicates the symptoms presence and No: indicates the symptom absence

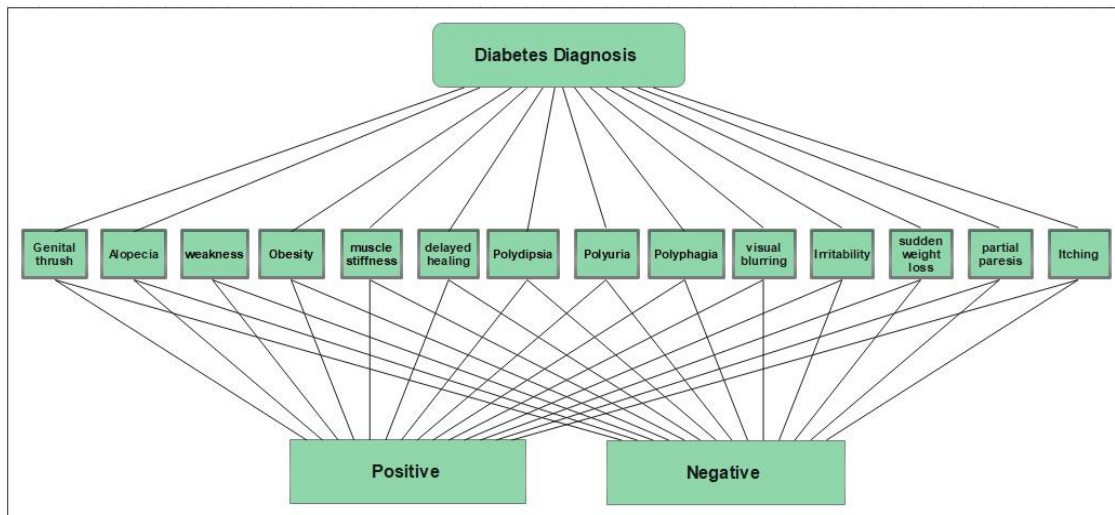


Figure 2. Conceptual level of FAHP for diabetes prediction

The value of the geometric mean of the fuzzy matrix is obtained as indicated in Table 3 after generating pairwise a comparison matrix. For example, according to (4), the product of the fourteen vectors yields the value of the geometric mean of fuzzy (r_x) for the first criterion. As shown in:

$$\tilde{r}_1 = \left(\prod_{y=1}^{14} \tilde{m}_{xy} \right)^{\frac{1}{14}} = [(1 * 2 * 1 * 4 * 3 * 4 * 0.250 * 0.2 * 1 * 1 * 1 * 1 * 1 * 1 * 1)^{\frac{1}{14}};$$

$$(1 * 3 * 1 * 5 * 4 * 5 * 0.333 * 0.250 * 1 * 1 * 1 * 1 * 1 * 1 * 1)^{\frac{1}{14}};$$

$$(1 * 4 * 1 * 6 * 5 * 6 * 0.500 * 0.333 * 1 * 1 * 1 * 1 * 1 * 1 * 1)^{\frac{1}{14}}] = [1.408; 1.258; 1.119]$$

thus, the total values are found by the sum of the fourteen criteria of r_x . The reverse values of total sum P (-1), shown in Table 3, are found by (total sum) $^{-1}$, $(14.227)^{-1} = 0.070$. In addition, Increasing Order of P (-1) is obtained by exchange for the first column with for the third column as shown in the last row (INCR) of Table 3.

Tabel 2. Pairwise comparison matrix (PWC) matrix

CRI	GT	AC	WN	OS	MS	DH	PD	PR	PG	VB	IA	SWL	PP	LI
GT	(1,1,1)	(2,3,4)	(1,1,1)	(4,5,6)	(3,4,5)	(4,5,6)	$(\frac{1}{4}, \frac{1}{3}, \frac{1}{2})$	$(\frac{1}{5}, \frac{1}{4}, \frac{1}{3})$	(1,1,1)	(1,1,1)	(1,1,1)	(1,1,1)	(1,1,1)	(1,1,1)
AC	$(\frac{1}{4}, \frac{1}{3}, \frac{1}{2})$	(1,1,1)	$(\frac{1}{4}, \frac{1}{3}, \frac{1}{2})$	(1,2,3)	(2,3,4)	(2,3,4)	$(\frac{1}{6}, \frac{1}{5}, \frac{1}{4})$	$(\frac{1}{9}, \frac{1}{8}, \frac{1}{7})$	(1,1,1)	(1,1,1)	(1,1,1)	(1,1,1)	(1,1,1)	(1,1,1)
WN	(1,1,1)	(2,3,4)	(1,1,1)	(4,5,6)	(6,7,8)	(7,8,9)	(1,1,1)	(1,1,1)	(1,1,1)	(1,1,1)	(1,1,1)	(1,1,1)	(1,1,1)	(1,1,1)
OS	$(\frac{1}{6}, \frac{1}{5}, \frac{1}{4})$	$(\frac{1}{3}, \frac{1}{2}, \frac{1}{1})$	$(\frac{1}{6}, \frac{1}{5}, \frac{1}{4})$	(1,1,1)	$(\frac{1}{4}, \frac{1}{3}, \frac{1}{2})$	$(\frac{1}{5}, \frac{1}{4}, \frac{1}{3})$	$(\frac{1}{8}, \frac{1}{7}, \frac{1}{6})$	$(\frac{1}{9}, \frac{1}{8}, \frac{1}{7})$	$(\frac{1}{9}, \frac{1}{9}, \frac{1}{9})$	$(\frac{1}{7}, \frac{1}{6}, \frac{1}{5})$	$(\frac{1}{4}, \frac{1}{3}, \frac{1}{2})$	$(\frac{1}{5}, \frac{1}{4}, \frac{1}{3})$	$(\frac{1}{3}, \frac{1}{2}, \frac{1}{1})$	$(\frac{1}{9}, \frac{1}{8}, \frac{1}{7})$
MS	$(\frac{1}{5}, \frac{1}{4}, \frac{1}{3})$	$(\frac{1}{4}, \frac{1}{3}, \frac{1}{2})$	$(\frac{1}{8}, \frac{1}{7}, \frac{1}{6})$	(2,3,4)	(1,1,1)	$(\frac{1}{3}, \frac{1}{2}, \frac{1}{1})$	$(\frac{1}{6}, \frac{1}{5}, \frac{1}{4})$	$(\frac{1}{7}, \frac{1}{6}, \frac{1}{5})$	$(\frac{1}{4}, \frac{1}{3}, \frac{1}{2})$	$(\frac{1}{3}, \frac{1}{2}, \frac{1}{1})$	(1,1,1)	$(\frac{1}{4}, \frac{1}{3}, \frac{1}{2})$	$(\frac{1}{3}, \frac{1}{2}, \frac{1}{1})$	$(\frac{1}{5}, \frac{1}{4}, \frac{1}{3})$
DH	$(\frac{1}{6}, \frac{1}{5}, \frac{1}{4})$	$(\frac{1}{4}, \frac{1}{3}, \frac{1}{2})$	$(\frac{1}{9}, \frac{1}{8}, \frac{1}{7})$	(3,4,5)	(1,2,3)	(1,1,1)	$(\frac{1}{9}, \frac{1}{8}, \frac{1}{7})$	$(\frac{1}{7}, \frac{1}{6}, \frac{1}{5})$	(1,1,1)	(1,1,1)	(1,1,1)	$(\frac{1}{7}, \frac{1}{6}, \frac{1}{5})$	$(\frac{1}{5}, \frac{1}{4}, \frac{1}{3})$	$(\frac{1}{4}, \frac{1}{3}, \frac{1}{2})$
PD	(2,3,4)	(4,5,6)	(1,1,1)	(6,7,8)	(4,5,6)	(7,8,9)	(1,1,1)	$(\frac{1}{3}, \frac{1}{2}, \frac{1}{1})$	(1,1,1)	(1,1,1)	(1,1,1)	(1,1,1)	(1,1,1)	(1,1,1)
PR	(3,4,5)	(9,9,9)	(1,1,1)	(7,8,9)	(5,6,7)	(5,6,7)	(1,2,3)	(1,1,1)	(1,1,1)	(1,1,1)	(1,1,1)	(1,1,1)	(1,1,1)	(1,1,1)
PG	(1,1,1)	(1,1,1)	(1,1,1)	(9,9,9)	(2,3,4)	(1,1,1)	(1,1,1)	(1,1,1)	(1,1,1)	(1,1,1)	(2,3,4)	$(\frac{1}{3}, \frac{1}{2}, \frac{1}{1})$	$(\frac{1}{3}, \frac{1}{2}, \frac{1}{1})$	(1,1,1)
VB	(1,1,1)	(1,1,1)	(1,1,1)	(5,6,7)	(1,2,3)	(1,1,1)	(1,1,1)	(1,1,1)	(1,1,1)	(1,1,1)	(1,1,1)	(1,1,1)	(1,1,1)	(1,1,1)
IA	(1,1,1)	(1,1,1)	(1,1,1)	(2,3,4)	(1,1,1)	(1,1,1)	(1,1,1)	(1,1,1)	$(\frac{1}{4}, \frac{1}{3}, \frac{1}{2})$	(1,1,1)	(1,1,1)	(1,1,1)	(1,1,1)	(1,1,1)
SWL	(1,1,1)	(1,1,1)	(1,1,1)	(3,4,5)	(2,3,4)	(5,6,7)	(1,1,1)	(1,1,1)	(1,2,3)	(1,1,1)	(1,1,1)	(1,1,1)	(3,4,5)	(1,1,1)
PP	(1,1,1)	(1,1,1)	(1,1,1)	(1,2,3)	(1,2,3)	(3,4,5)	(1,1,1)	(1,1,1)	(1,2,3)	(1,1,1)	(1,1,1)	$(\frac{1}{5}, \frac{1}{4}, \frac{1}{3})$	(1,1,1)	(1,1,1)
LI	(1,1,1)	(1,1,1)	(1,1,1)	(7,8,9)	(3,4,5)	(2,3,4)	(1,1,1)	(1,1,1)	(1,1,1)	(1,1,1)	(1,1,1)	(1,1,1)	(1,1,1)	(1,1,1)

To calculate the fuzzy weights of each criterion (W_x), in (7) is then applied as follows: $W1 = [1.119*0.055; 1.258*0.063; 1.408* 0.070] = [0.062; 0.079; 0.099]$. In the last step, the non-fuzzy weight value of each criterion (A_x) is found by taking the average of fuzzy weight values for each criterion using (8) as follows: $A_1 = [(0.062+0.079+0.099)/3] = 0.080$. Thus, the calculation of the total of A_x is obtained by summing of A_x values for each criterion. Finally, the normalization of non –Fuzzy values (N_x) is calculated by using (9). For example, $N_1 = A_1 / \text{total of } N1 = 0.080 / 1.008 = 0.079$. The weight calculation for all criteria is presented in Table 3.

Table 3. The geometric mean of fuzzy (r_x), with fuzzy weight (w_x), with averaged weight criterion (A_x), and normalized weight criterion (N_i)

CRI	r_x	W_x	A_x	N_x
C1	1.408	1.119	0.062	0.079
C2	0.925	0.801	0.037	0.050
C3	1.703	1.618	0.083	0.102
C4	0.322	0.245	0.000	0.015
C5	0.578	0.000	0.324	0.018
C6	0.627	0.461	0.385	0.021
C7	1.936	1.727	1.547	0.085
C8	2.193	2.034	1.830	0.101
C9	1.426	1.240	1.104	0.061
C10	1.243	1.194	1.122	0.062
C11	1.051	1.000	0.952	0.052
C12	1.727	1.575	1.379	0.076
C13	1.312	1.160	0.964	0.053
C14	1.449	1.385	1.306	0.072
Total	14.227	15.698	17.900	
P (-1)	0.070	0.064	0.056	
INCR	0.056	0.064	0.070	

The highest weights were assigned among the symptoms to Polyuria (PR) followed by Polydipsia symptom (PD). The lowest weight is assigned obesity symptom (OS) followed by delayed healing symptom (DH). After calculating the weight of the criteria, the data entry of each criterion in the dataset is then multiplied by the assigned weight. The summation value of each subject is then calculated by summing the weighted criteria. Hence, each subject will have a certain value resulted from the total sum of the weighted criteria values. The mean of these weighted criteria is then computed and taken as the threshold value. The threshold is then compared to each value in the weighted sum connected to a specific criterion. If the weighted sum value is equal to or more than the threshold, the result is (1), indicating that the patient has diabetes. If the weighted sum value is less than the threshold, the result is (0), indicating a negative diabetes diagnosis. The value of the threshold obtained from our experiment was 2.43.

To achieve the purpose of our comparison study, we also trained and tested ANN and SVM models. To process label imbalance, the dataset's minority class (negative classes) was oversampled substantially during training. The oversampling doubles the size of the negative examples and subsequently, balances the

two classes, producing better data representation during the training phase. The data was split into three categories: 60% for training, 15% for validation, and 25% for testing. By comparing the label of anticipated diabetes with the actual value provided along with the dataset, the performance for diabetes prediction is evaluated using various evaluation criteria. accuracy, sensitivity, specificity, precision, F-measure, and G-mean are among the evaluation metrics, which are defined as:

$$ACCURACY(AC) = \frac{TP+TN}{TP+TN+FP+FN} \quad (12)$$

$$SENSITIVITY(SEN) = \frac{TP}{TP+FN} \quad (13)$$

$$SPECIFICITY(SP) = \frac{TN}{TN+FP} \quad (14)$$

$$PRECISION(PR) = \frac{TP}{TP+FP} \quad (15)$$

$$F - MEASURE = \frac{2 \times \text{sensitivity} \times \text{precision}}{\text{sensitivity} + \text{precision}} \quad (16)$$

$$G - MEAN = \sqrt{\text{Sensitivity} \times \text{Specificity}} \quad (17)$$

where TP, FP, TN, and FN represent true positive, false positive, true negative, and false negative, respectively. The results obtained from the three diabetes prediction models including, FAHP, ANN, and SVM, using the six-evaluation metrics have been depicted in Table 4.

Table 4. The comparison among the three diabetes prediction models

Model	Accuracy	Sensitivity	Specificity	Precision	F-Measure	G-Mean
FAHP	0.7654	0.7312	0.82	0.8667	0.7932	0.7744
ANN	0.8385	0.747	1	1	0.8552	0.8643
SVM	0.8923	0.8793	0.9028	0.8793	0.8793	0.891

The findings show that the FAHP model is an excellent tool for diagnosing medical disorders based on many criteria, where the relative importance (priority) of each criterion to the others is not well defined. The reported sensitivity shows 0.7312, 0.747, and 0.8793 from FAHP, ANN, and SVM, respectively. These values indicate that the methods could be used in clinical practice as a computer-assistant diagnosis tool and as a second observer. Yet, it will not replace the decision taken by an expert physician. It's worth noting that the assessment metrics for FAHP are slightly lower than those published for ANN and SVM since it was tested on the entire dataset with no oversampling, unlike machine learning models. Overall, the three diabetes prediction models produce competitive findings and good performance, demonstrating their possibility of detecting diabetes early.

4. CONCLUSION

In this paper, a comparison study has been conducting harnessing soft computing techniques to early predict diabetes from publicly available data. The generated models, which included FAHP, ANN, and SVM, were successful in detecting diabetes in a group of people. The FAHP approach was used to create, implement, and evaluate the interest of weighting criteria from diabetic symptoms. Furthermore, in terms of accuracy, sensitivity, specificity, precision, F-measure, and F-mean, the findings obtained from applying the proposed approaches demonstrated that it is promising in accurately and effectively predicting diabetes. The proposed diabetes prediction algorithms can be readily and smoothly applied to models for other diseases. For future research, we suggest studying the performance of these methods based on an ensemble learning paradigm.

ACKNOWLEDGEMENTS

The authors would like to express their special thanks to Dr. Abdullateef Al-Bayati from Al-Mustansiriyah University, College of Medicine for providing the ground truth judgment of the comparison matrix used in this study. We also would like to thank Dr. Baidaa Al-Bander for her advice through conducting this work.




REFERENCES

- [1] American Diabetes Association, "Standards of Medical Care in Diabetes-2018 Abridged for Primary Care Providers," *Clinical Diabetes*, vol. 36, no. 1, pp. 14–37, Jan. 2018, doi: 10.2337/cd17-0119.
- [2] M. Wegmuller, J. P. Weid, P. Oberson, and N. Gisin, "High resolution fiber distributed measurements with coherent OFDR," *ECOC'00*, vol. 11, no. 4, p. 109, 2000.
- [3] C. Fiarni, E. M. Sipayung, and S. Maemunah, "Analysis and Prediction of Diabetes Complication Disease using Data Mining Algorithm," *Procedia Computer Science*, vol. 161, pp. 449–457, 2019, doi: 10.1016/j.procs.2019.11.144.
- [4] American Diabetes Association, "2. Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes—2018," *Diabetes Care*, vol. 41, no. Supplement_1, pp. S13–S27, Jan. 2018, doi: 10.2337/dc18-S002.
- [5] Diabetes Australia, "Failure to detect type 2 diabetes early costing \$700 million per year," *Diabetes Australia*. <https://www.diabetesaustralia.com.au/news/failure-to-detect-type-2-diabetes-early-costing-700-million-per-year/> (accessed Jul. 30, 2021).
- [6] M. I. Harris, R. Klein, T. A. Welborn, and M. W. Knuiman, "Onset of NIDDM occurs at Least 4–7 yr Before Clinical Diagnosis," *Diabetes Care*, vol. 15, no. 7, pp. 815–819, Jul. 1992, doi: 10.2337/diacare.15.7.815.
- [7] J. M. Forbes and M. E. Cooper, "Mechanisms of Diabetic Complications," *Physiological Reviews*, vol. 93, no. 1, pp. 137–188, Jan. 2013, doi: 10.1152/physrev.00045.2011.
- [8] Centers for Disease Control and Prevention, "National diabetes statistics report: estimates of diabetes and its burden in the United States, 2014," Atlanta, 2014.
- [9] American Diabetes Association, "Economic Costs of Diabetes in the U.S. in 2017," *Diabetes Care*, vol. 41, no. 5, pp. 917–928, May 2018, doi: 10.2337/dci18-0007.
- [10] B. Zhou *et al.*, "Worldwide trends in diabetes since 1980: a pooled analysis of 751 population-based studies with 4.4 million participants," *The Lancet*, vol. 387, no. 10027, pp. 1513–1530, Apr. 2016, doi: 10.1016/S0140-6736(16)00618-8.
- [11] M. T. Ogedengbe and C. O. Egbunu, "CSE-DT Features selection technique for Diabetes classification," *Applications of Modelling and Simulation*, vol. 4, pp. 101–109, 2020.
- [12] M. Shuja, S. Mittal, and M. Zaman, "Effective Prediction of Type II Diabetes Mellitus Using Data Mining Classifiers and SMOTE," 2020, pp. 195–211.
- [13] G. Swapna, R. Vinayakumar, and K. P. Soman, "Diabetes detection using deep learning algorithms," *ICT Express*, vol. 4, no. 4, pp. 243–246, Dec. 2018, doi: 10.1016/j.icte.2018.10.005.
- [14] T. Hanne, "On the classification of MCDM literature," in *Proceedings of the 5th Workshop of the DGOR-Working Group. Multicriteria Optimization and Decision Theory*, 1995, pp. 113–120.
- [15] A. Mardani, E. K. Zavadskas, Z. Khalifah, A. Jusoh, and K. M. Nor, "Multiple criteria decision-making techniques in transportation systems: A systematic review of the state of the art literature," *TRANSPORT*, vol. 31, no. 3, pp. 359–385, Dec. 2015, doi: 10.3846/16484142.2015.1121517.
- [16] P. Adhikary and S. Kundu, "MCDA or MCDM based selection of transmission line conductor: Small hydropower project planning and development," *International Journal of Engineering Research and Applications*, vol. 4, no. 2, pp. 357–361, 2014.
- [17] T. L. Saaty, "Optimization by the Analytic Hierarchy Process," Jan. 1979. doi: 10.21236/ADA214804.
- [18] V. Mishra, C. Samuel, and S. S.K., "Use of Machine Learning to Predict the Onset of Diabetes," *International Journal of Recent advances in Mechanical Engineering*, vol. 4, no. 2, pp. 9–14, May 2015, doi: 10.14810/ijmech.2015.4202.
- [19] S. A. Morgan *et al.*, "Prevalence and correlates of diabetes and its comorbidities in four Gulf Cooperation Council countries: evidence from the World Health Survey Plus," *Journal of Epidemiology and Community Health*, vol. 73, no. 7, pp. 630–636, Jul. 2019, doi: 10.1136/jech-2018-211187.
- [20] D. A. A. G. Singh, E. J. Leavline, and B. S. Baig, "Diabetes prediction using medical data," *Journal of Computational Intelligence in Bioinformatics*, vol. 10, no. 1, pp. 1–8, 2017.
- [21] K. Vidhya and R. Shanmugalakshmi, "Deep learning based big medical data analytic model for diabetes complication prediction," *Journal of Ambient Intelligence and Humanized Computing*, vol. 11, no. 11, pp. 5691–5702, Nov. 2020, doi: 10.1007/s12652-020-01930-2.
- [22] S. Ding, Z. Li, X. Liu, H. Huang, and S. Yang, "Diabetic complication prediction using a similarity-enhanced latent Dirichlet allocation model," *Information Sciences*, vol. 499, pp. 12–24, Oct. 2019, doi: 10.1016/j.ins.2019.05.037.
- [23] B. Liu, Y. Li, S. Ghosh, Z. Sun, K. Ng, and J. Hu, "Complication Risk Profiling in Diabetes Care: A Bayesian Multi-Task and Feature Relationship Learning Approach," *IEEE Transactions on Knowledge and Data Engineering*, vol. 32, no. 7, pp. 1276–1289, Jul. 2020, doi: 10.1109/TKDE.2019.2904060.
- [24] M. Dağdeviren, S. Yavuz, and N. Kılıç, "Weapon selection using the AHP and TOPSIS methods under fuzzy environment," *Expert Systems with Applications*, vol. 36, no. 4, pp. 8143–8151, May 2009, doi: 10.1016/j.eswa.2008.10.016.
- [25] Z. Güngör, G. Serhadlıoğlu, and S. E. Kesen, "A fuzzy AHP approach to personnel selection problem," *Applied Soft Computing*, vol. 9, no. 2, pp. 641–646, Mar. 2009, doi: 10.1016/j.asoc.2008.09.003.
- [26] H. S. Kılıç and E. Çevikcan, "Job selection based on fuzzy AHP: an investigation including the students of Istanbul Technical University Management Faculty," *International journal of business and management studies*, vol. 3, no. 1, pp. 173–182, 2011.
- [27] C. Kahraman and İ. Kaya, "A fuzzy multicriteria methodology for selection among energy alternatives," *Expert Systems with Applications*, vol. 37, no. 9, pp. 6270–6281, Sep. 2010, doi: 10.1016/j.eswa.2010.02.095.
- [28] H. S. Kılıç, "A fuzzy AHP based performance assessment system for the strategic plan of Turkish Municipalities," *International Journal of Business and Management Studies*, vol. 3, no. 2, pp. 77–86, 2011.
- [29] H. S. Kılıç and E. Çevikcan, "A Hybrid Weighting Methodology for Performance Assessment in Turkish Municipalities," 2012, pp. 354–363.
- [30] I. Chamodrakas, D. Batis, and D. Martakos, "Supplier selection in electronic marketplaces using satisficing and fuzzy AHP," *Expert Systems with Applications*, vol. 37, no. 1, pp. 490–498, Jan. 2010, doi: 10.1016/j.eswa.2009.05.043.
- [31] O. Kilincci and S. A. Onal, "Fuzzy AHP approach for supplier selection in a washing machine company," *Expert Systems with Applications*, vol. 38, no. 8, pp. 9656–9664, Aug. 2011, doi: 10.1016/j.eswa.2011.01.159.
- [32] K. Shaw, R. Shankar, S. S. Yadav, and L. S. Thakur, "Supplier selection using fuzzy AHP and fuzzy multi-objective linear programming for developing low carbon supply chain," *Expert Systems with Applications*, vol. 39, no. 9, pp. 8182–8192, Jul. 2012, doi: 10.1016/j.eswa.2012.01.149.
- [33] F. Arikani, "An interactive solution approach for multiple objective supplier selection problem with fuzzy parameters," *Journal of Intelligent Manufacturing*, vol. 26, no. 5, pp. 989–998, Oct. 2015, doi: 10.1007/s10845-013-0782-6.
- [34] M. M. F. Islam, R. Ferdousi, S. Rahman, and H. Y. Bushra, "Likelihood Prediction of Diabetes at Early Stage Using Data Mining Techniques," 2020, pp. 113–125.




- [35] M. T. García-Ordás, C. Benavides, J. A. Benítez-Andrades, H. Alaiz-Moretón, and I. García-Rodríguez, "Diabetes detection using deep learning techniques with oversampling and feature augmentation," *Computer Methods and Programs in Biomedicine*, vol. 202, p. 105968, Apr. 2021, doi: 10.1016/j.cmpb.2021.105968.
- [36] B. Al-Bander, Y. A. Fadil, and H. Mahdi, "Multi-Criteria Decision Support System for Lung Cancer Prediction," *IOP Conference Series: Materials Science and Engineering*, vol. 1076, no. 1, p. 012036, Feb. 2021, doi: 10.1088/1757-899X/1076/1/012036.
- [37] P. J. M. van Laarhoven and W. Pedrycz, "A fuzzy extension of Saaty's priority theory," *Fuzzy Sets and Systems*, vol. 11, no. 1–3, pp. 229–241, 1983, doi: 10.1016/S0165-0114(83)80082-7.
- [38] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, Oct. 1986, doi: 10.1038/323533a0.
- [39] V. N. Vapnik, *Statistical Learning Theory*, 1st ed. New York: Wiley, 1998.
- [40] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, 2000.

BIOGRAPHIES OF AUTHORS






Sabah Anwer Abdulkareem    received the B.Sc. degree in computer and software engineering from the the college of engineering, University of Diyala, Iraq, the M.Sc. degree in software Engineering from Chongqing University China. Her research interests include soft computing, and intelligent systems. She can be contacted at email: sbh_anwar@uodiyala.edu.iq.






Hussien Yossif Radhi    He got the BSc degree in Electronic Engineering from College of Engineering, University of Diyala, Iraq. He also got the MSc degree in Electronic and Communication from College of Engineering, Mustansiriya University, Iraq. He is working as a lecturer at Computer Engineering Department. Also, he is appointed as media department and web site manager for the college of engineering, diyala university since 2014. His research interests are: (Cryptography, Wireless sensor security, Image processing, and intelligent systems). He can be contacted at email: hussien.yossif19820@gmail.com.



Dr. Yousra Ahmed Fadil    has received her BSc in computer science from Computer science Department, University of Technology, Iraq in 1997. Then she completed her MSc in Computer Science from Information institute for postgraduate studies, Iraq in 2005, and she worked at University of Diyala, and she received a Ph.D degree in Artificial Intelligence from the University of Besançon, France in 2017. During that time, she published many papers in International Conferences and Journals. Currently, she continues to work at University of Diyala. She can be contacted at email: Yousra.comp@uodiyala.edu.iq.



Dr. Hussain Falih Mahdi    received the PhD from university of Kebangsaan Malaysia and Master of Science from University of Technology, Bagdad, Iraq. He is IEEE Region 10 Young Professional Committee South-East Asia coordinator (2017-2019), IEEE Region 10 Humanitarian activities committee (2017-2020), IEEE PES Young Professional Committee academic lead (2017-2020), IEEE IAS Chapters Area Chair, R10 Southeast Asia, Australia, and Pacific (2018-2019), and IEEE Region 10 PES students Chapters Chair (2019-2020), IEEE PES Day 2019 Global Chair, and IEEE HAC Event committee member 2019-2020. He can be contacted at email: Hussain.mahdi@ieee.org.