Healthcare Analytics 3 (2023) 100183

Contents lists available at ScienceDirect

Healthcare Analytics

journal homepage: www.elsevier.com/locate/health

A systematic approach to enhance the explainability of artificial intelligence in healthcare with application to diagnosis of diabetes

Yu-Cheng Wang [a,*,1], Tin-Chih Toly Chen [b,1], Min-Chi Chiu [c,1]

[a] Department of Aeronautical Engineering, Chaoyang University of Technology, Taichung City, Taiwan
[b] Department of Industrial Engineering and Management, National Yang Ming Chiao Tung University, Hsinchu City, Taiwan
[c] Department of Industrial Engineering and Management, National Chin-Yi University of Technology, Taichung City, Taiwan

ARTICLE INFO

Keywords:
Explainable artificial intelligence
Healthcare
Local interpretable model-agnostic explanation
Diabetes diagnosis
Artificial intelligence

ABSTRACT

Explainable artificial intelligence (XAI) tools are used to enhance the applications of existing artificial intelligence (AI) technologies by explaining their execution processes and results. In most past research, XAI tools and techniques are typically applied to only the inference part of the AI application. This study proposes a systematic approach to enhance the explainability of AI applications in healthcare. Several AI applications for type 2 diabetes diagnosis are taken as examples to illustrate the applicability of the proposed methodology. According to experimental results, the XAI tools and technologies in the proposed methodology were more diverse than those in the past research. In addition, an artificial neural network was approximated to a simpler and more intuitive classification and regression tree (CART) using local interpretable model-agnostic explanation (LIME). The extracted rules were used to recommend actions to the users to restore their health.

1. Introduction

Artificial intelligence (AI) are a set of technologies that enable computers to imitate human behavior [1]. The computing speed, storage capacity, reliability and interconnectivity of computers combined with human reasoning capability give AI the ability to solve complex and large-scale problems, such as many problems in medicine and healthcare. So far, AI technologies have been widely applied in medicine and healthcare [2–5]. However, new AI technologies (such as deep learning and hybrid algorithms) are becoming more and more complex, making it difficult for users to understand. Difficulties in understanding AI technologies can lead to users not being willing to trust, use or recommend related applications [6]. To overcome this difficulty, the concept of explainable AI (XAI) has been proposed [7–12]. XAI is to enhance the practicality of an existing AI technology by explaining its execution process and result [7].

There are basically two types of XAI methods. One is to enhance the practicality of an existing AI technology by explaining its execution process and result [7]. The other is to improve existing AI technologies by incorporating easy-to-interpret tools such as heatmaps [13], which can be considered as a new stream of AI. This study belongs to the first type.

So far, many XAI techniques or tools have been proposed to better explain the applications of AI in healthcare. A comprehensive review of generic XAI refers to Adadi et al. [14]. The application of XAI in healthcare has been reviewed by Nazar et al. [15], Durán et al. [16], Yang et al. [17], and others [18]. Although these reviews provide a lot of information for reference, they still have the following shortcomings:

- The objects, purposes, degrees and requirements of AI technology applications in different domains are not necessarily the same. Therefore, the applicable XAI tools or techniques also vary. In this regard, the differences between healthcare and other domains should be highlighted. In addition, XAI tools and techniques should be tailored to specific AI applications in healthcare.
- AI applications in healthcare have a wide variety of stakeholders. Stakeholders have different roles, backgrounds, motivations and needs. XAI tools and techniques that are suitable for particular stakeholders should be selected.

In conclusion, the selection of XAI tools or techniques should make them suitable not only for specific AI applications in healthcare, but also for various stakeholders.

This study proposes a systematic approach to enhance the explainability/interpretability of AI applications in healthcare. Depending on the flow of information through an AI application, seven XAI tools and techniques are applied to various parts of the AI application: smart technologies, common expression, color management, local interpretable model-agnostic explanation (LIME), classification and regression trees (CART), donut charts, and graphical user interface (GUI).

[*] Corresponding author.
E-mail addresses: tony.cobra@msa.hinet.net (Y.-C. Wang), tcchen@nycu.edu.tw (T.-C.T. Chen), mcchiu@ncut.edu.tw (M.-C. Chiu).
[1] All authors contributed equally to the writing of this paper.

https://doi.org/10.1016/j.health.2023.100183
Received 27 February 2023; Received in revised form 4 April 2023; Accepted 20 April 2023

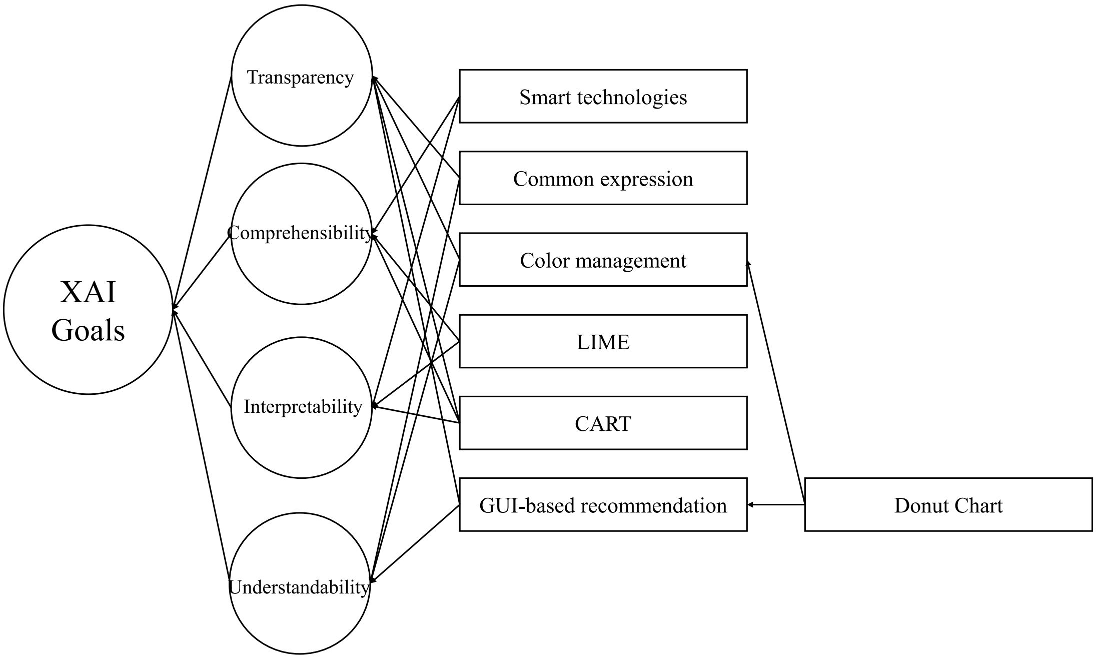


Fig. 1. XAI goals supported by the new techniques.

These XAI techniques were chosen because they are commonly used XAI techniques that have been applied in various fields to explain AI applications [6–8,10,15–20]. In contrast, in most of the past research, XAI tools and techniques were typically applied to a single part of an AI application (such as the inference part) [21–24]. The stakeholders targeted by each XAI application also varied. Similar systematic approaches have been rare in the past. The XAI tools and techniques applied in this study are also more diverse than those in past studies. The application of AI technologies to assist type 2 diabetes diagnosis is taken as an example to illustrate the applicability of the proposed methodology.

The remainder of this paper is organized as follows. Section 2 is dedicated to a literature review. Section 3 introduces the systematic approach proposed in this study. To illustrate the applicability of the proposed methodology, it has been used to better explain the process and/or outcome of diabetes diagnosis using AI technologies. Section 4 presents the experimental results and provides some discussions based on them. Finally, Section 5 concludes this study and provides some directions for future research.

2. Literature review

There are basically two ways to apply AI to diabetes diagnosis. One is to treat it as a binary classification problem [25–28]. Users are classified as yes (with diabetes) or no. The other builds a predictive model to predict the probability that a user has diabetes [29–34]. In addition, a number of studies highlighted the potential of new smart technologies for diabetes management [35–37]. For patients with diabetes, AI technologies can assist automated retinal screening, clinical decision support, predictive population risk stratification, and patient self-management [38–40].

Table 1 summarizes the AI technologies and XAI methods used in some AI applications in diabetes diagnosis [21–23,41].

Existing AI techniques for diabetes analysis and related XAI tools and techniques are summarized Table 2 [21–23,41].

3. Methodology

Like other information technology (IT) applications, AI applications can be roughly divided into three parts: input, processing, and output, according to the information flow in these AI applications. Among them, processing is the part focused by most AI applications, while input and output received less attention. In the proposed methodology, seven XAI tools and techniques are systematically applied to improve the explainability of AI applications for diabetes diagnosis: smart technologies, common expression, color management, LIME, CART, donut charts, and GUI. The XAI goals [42] supported by these new techniques are shown in Fig. 1. While the four XAI goals are important to all three parts of an AI application, transparency is critical to the input, while understandability adds value to the output. The comprehensibility and interpretability of processing are particularly important [20,42]. In this study, smart technology applications that facilitate users to input their data, understand the inference mechanism with simple decision rules, and explain possible treatments that can be employed to reduce the likelihood of developing diabetes increased their willingness to apply ANN/DNN applications. Furthermore, in this study, the explainability of ANN/DNN applications for diabetes diagnosis is mainly improved in terms of enhanced comprehensibility, interpretability, and understandability.

3.1. Enhancing the explainability of input

The explainability of inputs to an AI application for diabetes diagnosis can be enhanced in several ways:

- Smart technology applications to enhance users' willingness: A user interface based on the application of smart technologies, such as apps (smart phones) and humanoid robots, helps to enhance the willingness of a user to apply a diabetes prediction (or analysis) system [23,43], especially to reduce his/her tension and pressure when facing the doctor and enhance the autonomy of the user in diagnosis and treatment. For example, Shen et al. [23] designed an app for diabetes analysis, in which one screen let a user enter his/her attributes, and the analysis results were displayed on another screen.
- Transparency: Inputs to an AI application for diabetes diagnosis are usually the attributes (i.e., physical conditions and demographic data) of a user, and outputs are the probabilities that the user has and does not have diabetes. When collecting user

Table 1
AI and XAI methods for diabetes diagnosis.

Reference	AI technologies	XAI methods	Target stakeholder
Karan et al. [21]	• Artificial neural network (ANN)	• Inputs/outputs • Overall performance (accuracy) • Response surface modelling (RSM) • System snapshot • Textual description • User interface	• Developers • Users
Alian et al. [22]	• Decision rules	• Expert evaluation • Ontology • Overall performance: Accuracy, relevance, appropriateness • System architecture diagram • System snapshot • Textual description • Use cases • User interface	• Affected parties • Deployers • Developers • Users
Shen et al. [23]	• Support vector machine (SVM) • Random forest • Adaptive boosting (AdaBoost) • k-nearest neighbors (kNN) • Naïve Bayes (NB) • Extreme gradient boosting (XGBoost) • Gradient boosting decision tree (GBDT)	• Flow chart • Individual application • Overall performance: Accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), area under the curve (AUC) • Receiver operating characteristic (ROC) curve • System architecture diagram • Textual description • User interface	• Developers • Regulators • User
Tyler et al. [41]	• Decision support system • kNN	• Box-and-whisker plot • Expert evaluation • Flow chart • Line chart • Overall performance: Successful recommendation rate, % agreement, % disagreement, % additional, % not comparable • Textual description • User interface	• Deployers • Developers • Users

Table 2
Existing AI technologies for diabetes analysis and related XAI tools and techniques.

Existing AI technologies for diabetes analysis	Related XAI tools and methods
• AdaBoost • ANN • Decision rules, decision support system • XGBoost • GBDT • kNN • NB • Random forest • SVM	• Box-and-whisker plot • Expert evaluation • Flow chart • Line chart • Inputs/outputs • Ontology • Overall performance: Accuracy, relevance, appropriateness, sensitivity, specificity, PPV, NPV, AUC, successful recommendation rate, % agreement, % disagreement, % additional, % not comparable • RSM • ROC curve • System architecture diagram • System snapshot • Textual description • Use cases, individual application • User interface

attributes, first, it is better to provide users with access to information on why an attribute is needed. In addition, the collection of user attributes represents the storage, application and transfer of such information. In addition to obtaining a user's consent in advance, it shall comply with the provisions of government laws and regulations. Furthermore, it is better to let a user choose whether to enter these attributes. The prediction (or analysis) mechanism should be capable of reasoning based on only a subset of user attributes.

3.2. Enhancing the explainability of processing

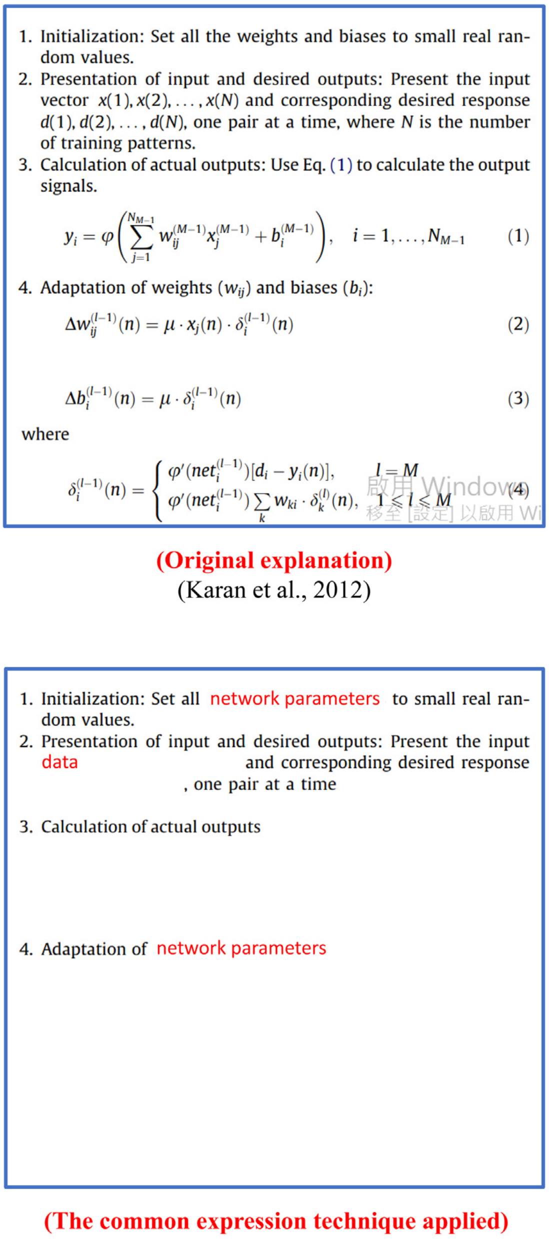
The explainability of processing in an AI application for diabetes diagnosis can be enhanced in several ways, as detailed in the following.

3.2.1. Explaining the reasoning process using common expressions

Textual description is the most common technique for explaining AI applications for diabetes diagnosis and is applicable to all stakeholders. However, it may not be necessary to explain the reasoning process to users. In this technique, text is used to describe the background, motivation, methodology and process of the AI application. It may be full of symbols and technical terms that are not easy to understand, but it is helpful if the reader needs to understand the logical structure of the formulas, algorithms, and systems that follow. Therefore, it is worth considering whether to apply XAI techniques, such as the common expression technique [26], to improve textual descriptions by replacing these symbols and technical terms with common expressions. An example is given in Fig. 2.

3.2.2. Improving the explainability of an ANN/DNN for diabetes diagnosis

ANNs or deep neural networks (DNNs) have been built in some past studies to predict the probability of a user having diabetes. Inputs to the ANN (or DNN) are the physical conditions or demographics of a user, including his/her age, physical activities (yes or no), weeks pregnant, diabetes in the family (yes or no), body mass index, skin



1. Initialization: Set all the weights and biases to small real random values.
2. Presentation of input and desired outputs: Present the input vector $x(1), x(2), \ldots, x(N)$ and corresponding desired response $d(1), d(2), \ldots, d(N)$, one pair at a time, where N is the number of training patterns.
3. Calculation of actual outputs: Use Eq. (1) to calculate the output signals.

$$y_i = \varphi\left(\sum_{j=1}^{N_{M-1}} w_{ij}^{(M-1)} x_j^{(M-1)} + b_i^{(M-1)}\right), \quad i = 1, \ldots, N_{M-1} \tag{1}$$

4. Adaptation of weights (w_{ij}) and biases (b_i):

$$\Delta w_{ij}^{(l-1)}(n) = \mu \cdot x_j(n) \cdot \delta_i^{(l-1)}(n) \tag{2}$$

$$\Delta b_i^{(l-1)}(n) = \mu \cdot \delta_i^{(l-1)}(n) \tag{3}$$

where

$$\delta_i^{(l-1)}(n) = \begin{cases} \varphi'(net_i^{(l-1)})[d_i - y_i(n)], & l = M \\ \varphi'(net_i^{(l-1)}) \sum_k w_{ki} \cdot \delta_k^{(l)}(n), & 1 \leqslant l \leqslant M \end{cases} \tag{4}$$

(Original explanation)
(Karan et al., 2012)

1. Initialization: Set all network parameters to small real random values.
2. Presentation of input and desired outputs: Present the input data and corresponding desired response, one pair at a time
3. Calculation of actual outputs
4. Adaptation of network parameters

(The common expression technique applied)

Fig. 2. Enhancing explainability using the common expression technique.

fold thickness, cholesterol, diastolic blood pressure, 2-h serum insulin, the pedigree of diabetes, plasma glucose concentration, etc. [21,44,45]. These inputs are transmitted through one to multiple hidden layers, some of which may have recurrent nodes. Finally, the network output is generated. If the transformation function on the output node is a sigmoid function [46]:

$$o = \frac{1}{1 + e^{-x}} \tag{1}$$

The output o is within [0, 1] and can be used to predict the probability of having (or not having) diabetes. However, whether o is greater than 0.5 can be considered that the user has diabetes is controversial. To solve this problem, the ANN constructed by Karan et al. [21] has two outputs. One predicts the probability of having diabetes and the other predicts the probability of not having it. Once the former is (significantly) greater than the latter, the user is deemed to have diabetes.

In most of the past studies, ANNs are usually explained using textual descriptions and network configuration (or system architecture) diagrams, as shown in Fig. 3. The operations in an ANN/DNN are very difficult to understand to some stakeholders. As a result, only some of them (e.g., developers, regulators and deployers) are interested in the operations. To overcome this difficulty, animation-based tools have been designed, such as ConvNetJS (for convolutional neural networks, CNNs), TensorFlow (for ANNs with multiple hidden layers, DNNs), Seq2Seq (for recurrent neural networks), MATLAB, etc., as illustrated in Fig. 4, to explain/animate the operations within the network.

The color management technique [47] is also applied in TensorFlow, where positive and negative connection weights are displayed in different colors, and the thickness of the line is proportional to the absolute value of the connection weight. However, it is a problem that the difference in the thickness of the lines is not sufficiently noticeable. In addition, symbols and technical terms in the network configuration can be replaced by common expressions. Further, nodes with different transformation functions should have different shapes. Also, in TensorFlow, connections between nodes have no directions. Stakeholders may not know that the signal flow in the network is from left to right. After considering these, a modified network configuration diagram is illustrated in Fig. 5

These existing animation-based XAI tools still have the following problems:

- These animation-based tools are useful for stakeholders who already have some ANN background knowledge.
- In other words, these tools will have more reference value for developers.
- To most stakeholders, these tools look interesting, but are still quite complicated.

3.2.3. Fitting a direct relationship between inputs and outputs using LIME

For stakeholders, if the relationship between inputs and outputs can be described in a simple way, the explainability of the prediction (or analysis) process can be greatly improved. That is a very challenging task for some deep learning-based AI technologies, but is still the direction of the efforts of many XAI researchers.

A stakeholder wants to know the relationship between his/her physical conditions or demographics and the probability of developing diabetes, which is difficult to describe explicitly using an ANN/DNN. To address this issue, decision/regression rules that approximate the reasoning mechanism of the ANN/DNN can be extracted using techniques such as LIME [48,49], which is to explain the classification (or regression) result using a machine learning model by identifying critical predictors and fitting a simple interpretable model. The procedure comprises the following steps:

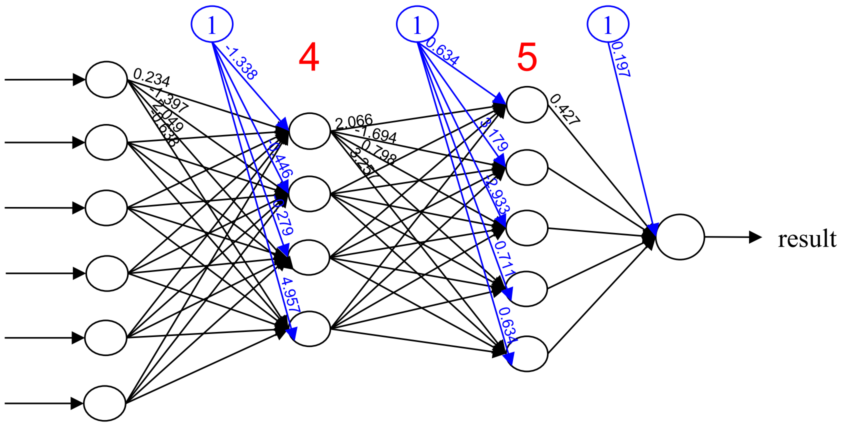
Step 1. Construct and train an ANN/DNN to predict the probability of having diabetes.
Step 2. Generate synthetic data from original data.
Step 3. Create a CART to fit the synthetic data.
Step 4. Explain the prediction mechanism of the ANN/DNN using the rules of the CART.

as illustrated in Fig. 6.

The collected data sometimes may not cover the entire sample space. To address this issue, synthetic data are generated from the original data as follows:

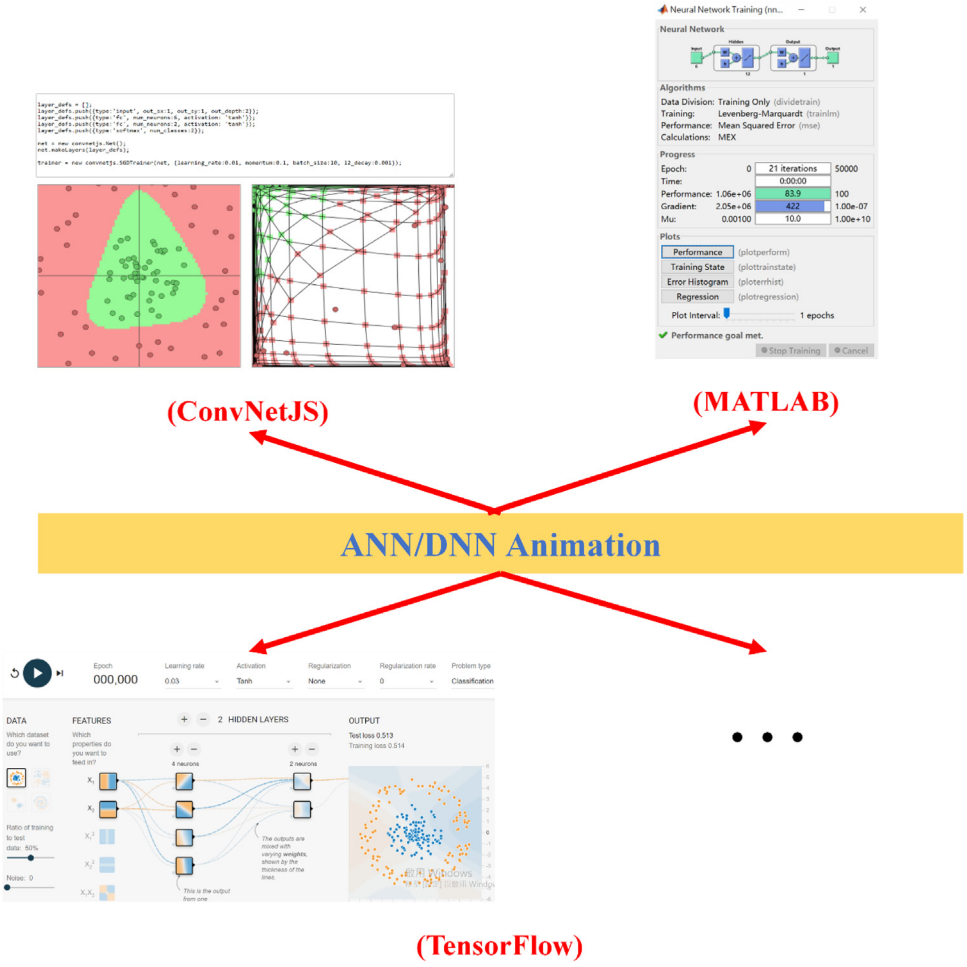
Step 1. Set the number of examples to T.
Step 2. $t = 1$.
Step 3. Generate a random number within [0, 1] for each user attribute.
Step 4. If attribute p is numeric, convert the random number to the value of the attribute as

$$\hat{x}_{tp} = \min_j x_{jp} + r_{tp}(\max_j x_{jp} - \min_j x_{jp}) \tag{2}$$



* Values of some network parameters are hidden for clarity.

Fig. 3. Network architecture diagram of a DNN using R language.



```
layer_defs = [];
layer_defs.push({type:'input', out_sx:1, out_sy:1, out_depth:2});
layer_defs.push({type:'fc', num_neurons:6, activation: 'tanh'});
layer_defs.push({type:'fc', num_neurons:2, activation: 'tanh'});
layer_defs.push({type:'softmax', num_classes:2});

net = new convnetjs.Net();
net.makeLayers(layer_defs);

trainer = new convnetjs.SGDTrainer(net, {learning_rate:0.01, momentum:0.1, batch_size:10, l2_decay:0.001});
```

Neural Network Training (nn...

Neural Network

Algorithms

Data Division: Training Only (dividetrain)
Training: Levenberg-Marquardt (trainlm)
Performance: Mean Squared Error (mse)
Calculations: MEX

Progress

Epoch:	0	21 iterations	50000
Time:		0:00:00	
Performance:	1.06e+06	83.9	100
Gradient:	2.05e+06	422	1.00e-07
Mu:	0.00100	10.0	1.00e+10

Plots

Plot Interval: 1 epochs

✓ Performance goal met.

Epoch 000,000
Learning rate 0.03
Activation Tanh
Regularization None
Regularization rate 0
Problem type Classification

DATA
Which dataset do you want to use?
Ratio of training to test data: 50%
Noise: 0

FEATURES
Which properties do you want to feed in?

2 HIDDEN LAYERS
4 neurons
2 neurons

OUTPUT
Test loss 0.513
Training loss 0.514

The outputs are mixed with varying **weights**, shown by the thickness of the lines.

This is the output from one

Fig. 4. Some existing animation-based tools for explaining an ANN/DNN.

where r_{tp} is the random number for attribute p at iteration t. $\hat{x}_{tp}$ is the generated synthetic sample. Otherwise, when attribute p is nominal

$$\hat{x}_{tp} = C_{p,l+1} if \sum_{q=1}^{l} f_{pq} \leq r_{tp} < \sum_{q=1}^{l+1} f_{pq} \tag{3}$$

$C_{p,l}$ is the lth nominal value of attribute p; $l = 1 \sim L$. f_{pq} is the probability of the qth nominal value of attribute p; $q = 1 \sim Q$.

Step 5. Feed the synthetic sample to the trained ANN/DNN to predict the probability of having diabetes.

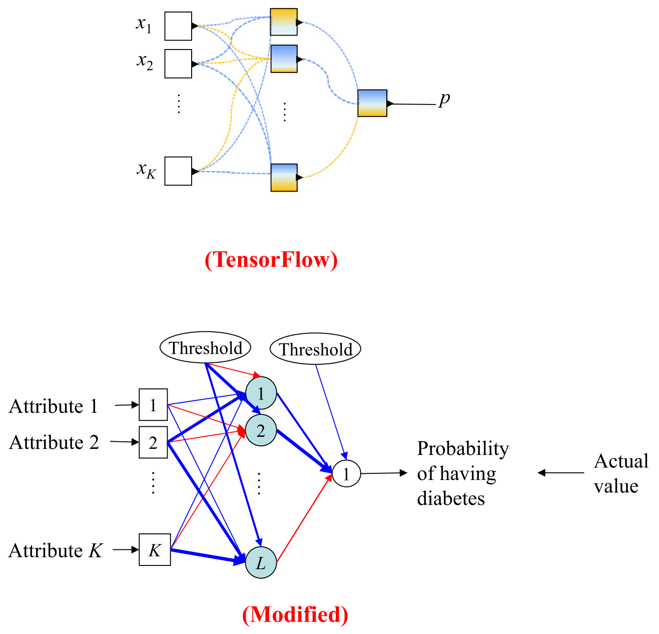


Fig. 5. Modified network configuration diagram.

Step 6. $t = t + 1$.
Step 7. If $t < T$, return to Step 3; otherwise, stop.

The original data and synthetic samples are merged as the synthetic data.

3.3. Enhancing the explainability of output

The explainability of outputs from an AI application for diabetes diagnosis can be enhanced in several ways:

- Better interpretation of the outputs: Some diagnostic systems only tell a user whether he or she has diabetes after analysis. Some systems also let a user know the probability of having diabetes. As a result, the probability of not having diabetes is one minus that. Other systems present both the probabilities of having diabetes and not having diabetes at the same time. For example, in Shen et al. [23], the prediction was presented in words (i.e., textual description), and the two probabilities are compared in a donut chart, as illustrated in Fig. 7. As such, the user is more certain to have diabetes if the probability of having diabetes is much greater than that of not having diabetes, and vice versa. Furthermore, the sum of these two probabilities does not necessarily have to equal one, as the factors that are weighted in assessing the presence and absence may be different.
- Enhancing users' interpretability via smart technology applications: Textual description is usually applied to provide recommendations based on the results of a diabetes analysis. This can be better explained by displaying recommendations in a GUI using smart technologies [22], as shown in Fig. 8.
- Clarifying the limitations of the application: Compared to explicitly telling a user whether or not he/she has diabetes, presenting the probability of yes (or no) allows the user to understand that the accuracy of the AI application is limited.
- Use case (recommendation) generation: The rules extracted can be used to generate use cases (or recommendations).

Table 3
Simulated data of 100 users.

User #	Attribute 1	Attribute 2	Attribute 3	Attribute 4	Having diabetes
1	Yes	54	No	26.7	Yes
2	Yes	52	No	29.3	Yes
3	No	31	Yes	17.6	No
…					
100	No	66	Yes	30.2	Yes

Table 4
Synthetic data.

User #	Attribute 1	Attribute 2	Attribute 3	Attribute 4	Having diabetes (actual/predicted value)
1	Yes	54	No	26.7	Yes
2	Yes	52	No	29.3	Yes
3	No	31	Yes	17.6	No
…					
150	No	42	Yes	29.5	No

4. Case study

4.1. Background

An example is given in Table 3, which involves the simulated data of 100 users.

An ANN with the following configuration is constructed to predict the probability of a user having diabetes.

- Inputs: the four attributes (physical conditions and demographics) of a user.
- Number of layers: There are three layers (the input layer, one hidden layer, and the output layer) in the ANN. The number of nodes in the hidden layer is two times the number of inputs.
- Output: the probability that the user has diabetes. If the network output is greater than 0.5, the user is predicted to have diabetes.
- Training algorithm: the Levenberg–Marquardt (LM) algorithm [50]. The first 3/4 of the data is used for training the ANN, and the rest is used for testing/evaluation.
- Learning rate: 0.2.
- Convergence conditions: The training process stops if mean squared error (MSE) $<$ 0.1 or 10000 epochs have been run.

The required MALTAB code is provided in Fig. 9. After applying the trained ANN to test data, the prediction results are summarized in Fig. 10.

In the previous example, 75 synthetic samples are generated, and their probabilities of having diabetes are predicted using the trained ANN. The results are then combined with training data as synthetic data, which have 150 examples, as shown in Table 4.

Subsequently, a CART was constructed to fit the synthetic data using MATLAB R2021a. The result is shown in Fig. 11. There are ten rules in the CART, as summarized in Table 5.

4.2. Results and discussion

According to the experimental results, the following discussion is made:

- The accuracy of diabetes diagnosis using the ANN application was 92%. For comparison, the classification accuracies achieved using multiple linear regression (MLR) and CART are 76% and 88%, respectively.
- Rules like MLR and CART were easier to explain than the ANN application.

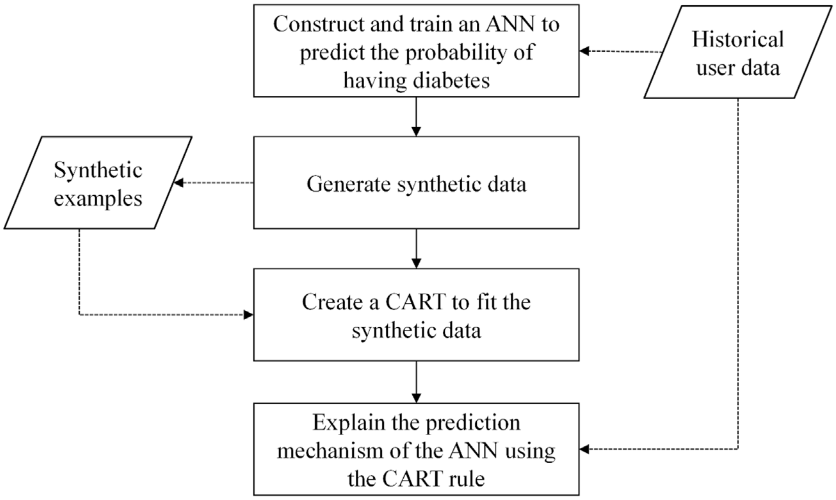

Construct and train an ANN to predict the probability of having diabetes
Historical user data
Synthetic examples
Generate synthetic data
Create a CART to fit the synthetic data
Explain the prediction mechanism of the ANN using the CART rule

Fig. 6. Procedure of LIME for enhancing the explainability of an ANN/DNN for diabetes diagnosis.

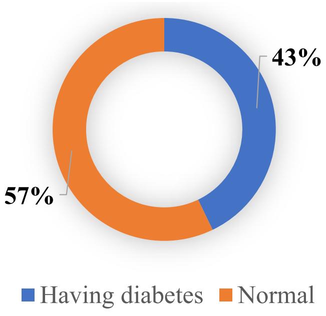
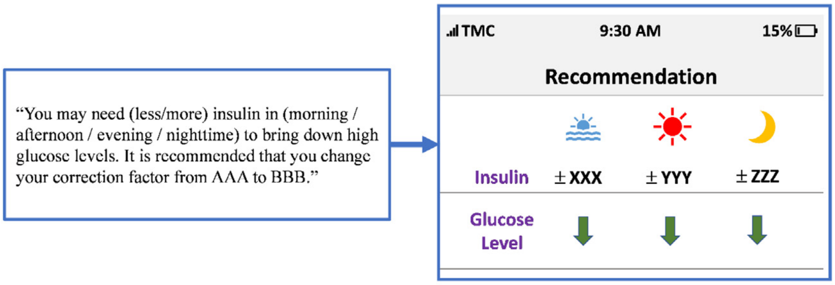


Fig. 7. Comparing the two probabilities using a donut chart.

Table 5
Fitted rules for diabetes prediction.

Rule #	Rule content
1	If Attribute 3 is ‘No’ and Attribute 2 is less than 53 and Attribute 4 is less than 23.9 Then Having Diabetes is ‘No’
2	If Attribute 3 is ‘No’ and Attribute 2 is less than 53 and 23.9 ≤ Attribute 4 < 24.25 Then Having Diabetes is ‘Yes’
3	If Attribute 3 is ‘No’ and Attribute 2 is less than 23.5 and 24.25 ≤ Attribute 4 < 27.85 Then Probability of Having Diabetes is 33.3%
4	If Attribute 3 is ‘No’ and 23.5 ≤ Attribute 2 < 53 and 24.25 ≤ Attribute 4 < 27.85 Then Having Diabetes is ‘No’
5	If Attribute 3 is ‘No’ and Attribute 2 is less than 53 and Attribute 4 is greater than or equal to 27.85 Then Having Diabetes is ‘Yes’
6	If Attribute 3 is ‘No’ and 53 ≤ Attribute 2 < 57 Then Probability of Having Diabetes is 75%
7	If Attribute 3 is ‘No’ and Attribute 2 is greater than 57 Then Having Diabetes is Yes
8	If Attribute 3 is ‘Yes’ and Attribute 4 is less than 29.3 Then Having Diabetes is ‘No’
9	If Attribute 1 is ‘No’ and Attribute 3 is ‘Yes’ and Attribute 4 is greater than or equal to 29.3 Then Having Diabetes is ‘No’
10	If Attribute 1 is ‘Yes’ and Attribute 3 is ‘Yes’ and Attribute 4 is greater than or equal to 29.3 Then Probability of Having Diabetes is 50%

- The prediction accuracy using the CART for synthetic data was 96%. Some rules do not use all attributes and are therefore still applicable when a user does not enter certain attributes.
- In Table 5, if a user triggers the fifth rule, it means that the user is diagnosed with diabetes because his/her attribute 3 is “no”, attribute 2 is less than 53, and attribute 4 is greater than or equal to 27.85. However, according to the fourth rule, if the user’s attribute 4 can be lowered below 27.85, he/she will be diagnosed as not having diabetes, resulting in the following recommendation:
  “Please take actions like... to lower your attribute 4 below 27.85”.
  A user may apply more than one rule. In addition, the rules applicable to different users may also be different.
- In summary, the XAI techniques applied to diabetes diagnosis in this study can be divided into animation-based illustration techniques, feature importance assessment and selection techniques, and diabetes probability (or risk) assessment techniques. Among the XAI tools and techniques proposed in this study, smart technologies, color management, donut charts and GUI-based recommendation belong to the category of animation-based illustration techniques. Other existing XAI tools and techniques in this category include ConvNetJS [51], TensorFlow [52], Seq2Seq [53] and MATLAB [54]. This study applied the color management technique and the common expression technique to improve the understandability of TensorFlow. LIME and CART falls in the feature importance assessment and selection category. Existing XAI tools and techniques in this category are numerous, e.g., partial derivation, odd ratio [55], out-of-bag (OOB) predictor importance [56], recursive feature elimination (RFE) [57], Shapely additive explanation value (SHAP) analysis [58]. These XAI tools and techniques are often combined with complex forecasting methods, whereas the combination of LIME and CART applied in this study is computationally simple and straightforward in interpreting the forecasting result. Donut charts, GUI-based recommendation and CART are diabetes probability (or risk) assessment techniques used in this study, which is basically based on decision trees. There are other existing XAI tools and techniques based on decision trees, such as random forests [10,59–61]. However, in such methods, multiple decision rules are applied to predict the probability of having diabetes, which confuses the patient and increases the difficulty of generating recommendations from these decision rules.
- The XAI tools and techniques presented in this study are general approaches that can be used to explain other AI applications in various fields.
- The XAI techniques proposed in this study are actually modifications of the existing XAI techniques or XAI techniques that are

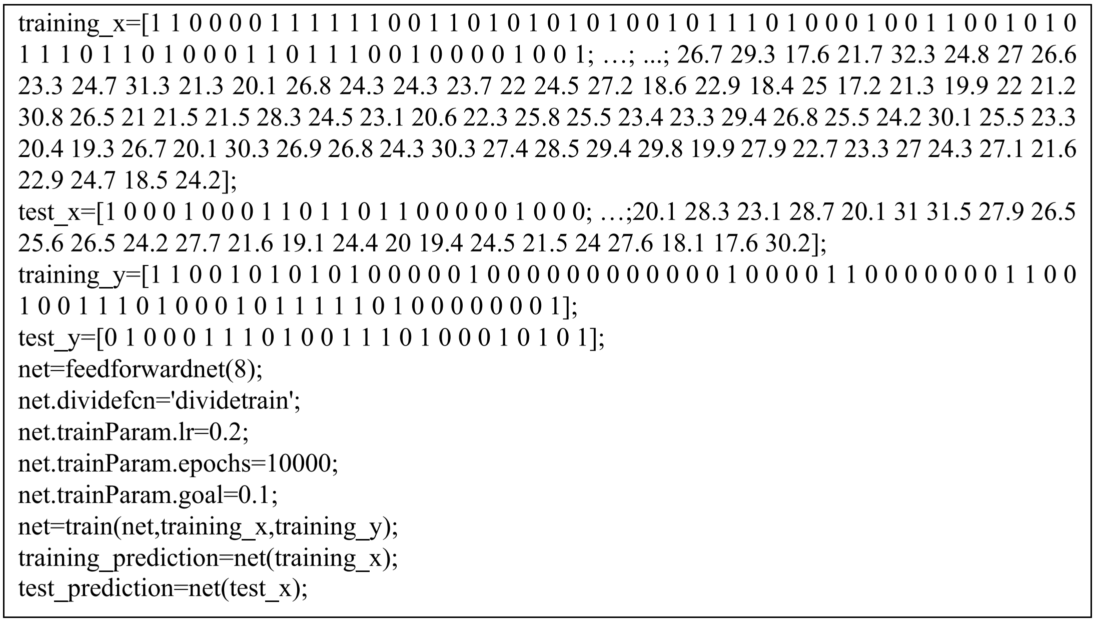


"You may need (less/more) insulin in (morning / afternoon / evening / nighttime) to bring down high glucose levels. It is recommended that you change your correction factor from AAA to BBB."

(Text-based recommendation)
(Tyler et al., 2020)

(GUI-based recommendation)

Fig. 8. Replacing text-based recommendation with GUI-based recommendation.



```
training_x=[1 1 0 0 0 0 1 1 1 1 1 1 0 0 1 1 0 1 0 1 0 1 0 1 0 0 1 0 1 1 1 0 1 0 0 0 1 0 0 1 1 0 0 1 0 1 0
1 1 1 0 1 1 0 1 0 0 0 1 1 0 1 1 1 0 0 1 0 0 0 0 1 0 0 1; …; ...; 26.7 29.3 17.6 21.7 32.3 24.8 27 26.6
23.3 24.7 31.3 21.3 20.1 26.8 24.3 24.3 23.7 22 24.5 27.2 18.6 22.9 18.4 25 17.2 21.3 19.9 22 21.2
30.8 26.5 21 21.5 21.5 28.3 24.5 23.1 20.6 22.3 25.8 25.5 23.4 23.3 29.4 26.8 25.5 24.2 30.1 25.5 23.3
20.4 19.3 26.7 20.1 30.3 26.9 26.8 24.3 30.3 27.4 28.5 29.4 29.8 19.9 27.9 22.7 23.3 27 24.3 27.1 21.6
22.9 24.7 18.5 24.2];
test_x=[1 0 0 0 1 0 0 0 1 1 0 1 1 0 1 1 0 0 0 0 0 1 0 0 0; …;20.1 28.3 23.1 28.7 20.1 31 31.5 27.9 26.5
25.6 26.5 24.2 27.7 21.6 19.1 24.4 20 19.4 24.5 21.5 24 27.6 18.1 17.6 30.2];
training_y=[1 1 0 0 1 0 1 0 1 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 1 1 0 0 0 0 0 0 0 1 1 0 0
1 0 0 1 1 1 0 1 0 0 0 1 0 1 1 1 1 1 0 1 0 0 0 0 0 0 0 1];
test_y=[0 1 0 0 0 1 1 1 0 1 0 0 1 1 1 0 1 0 0 0 1 0 1 0 1];
net=feedforwardnet(8);
net.dividefcn='dividetrain';
net.trainParam.lr=0.2;
net.trainParam.epochs=10000;
net.trainParam.goal=0.1;
net=train(net,training_x,training_y);
training_prediction=net(training_x);
test_prediction=net(test_x);
```

Fig. 9. MATLAB code for training the ANN.

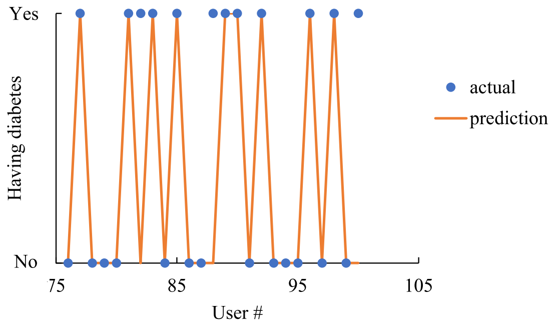


Fig. 10. Prediction results for test data.

prevalent in other domains, which ensures their effectiveness and reliability.

- Table 6 lists 6 requirements [62–66] to evaluate the effectiveness of various XAI techniques to explain diabetes prediction methods. From the comparison results, the advantages and disadvantages of these XAI techniques become apparent. Most existing XAI techniques can only satisfy at most four requirements. Obviously, there is no perfect XAI technique to explain the forecasting process and results.

Table 6
Comparison of existing XAI techniques for explaining diabetes prediction methods.

	I	II	III	IV	V	VI	VII
No background knowledge required		Yes	Yes				
Easy to understand		Yes	Yes			Yes	Yes
Easy to communicate		Yes	Yes			Yes	Yes
Can handle high-dimensional data	Yes			Yes	Yes		
Show the prediction process	Yes				Yes		
Evaluate the prediction performance				Yes	Yes		
Recommend improvement measures					Yes		

I: Smart technologies; II: Common expression; III: Color management; IV: LIME; V: CART; VI: Donut charts; VII: GUI-based recommendation.

5. Conclusions

As AI technologies are more and more widely used, the explainability of the process and result of an AI application has gradually been paid attention to. Compared with the applications of AI technologies in other fields, AI applications in healthcare pay more attention to individual/local benefits and acceptability. Therefore, XAI tools and techniques are continuously introduced and applied to improve the explainability of AI applications in this field. To this end, this study proposes a systematic approach that applies seven XAI tools and techniques to improve the explainability of AI applications for healthcare. These tools are arranged based on the flow of information through the

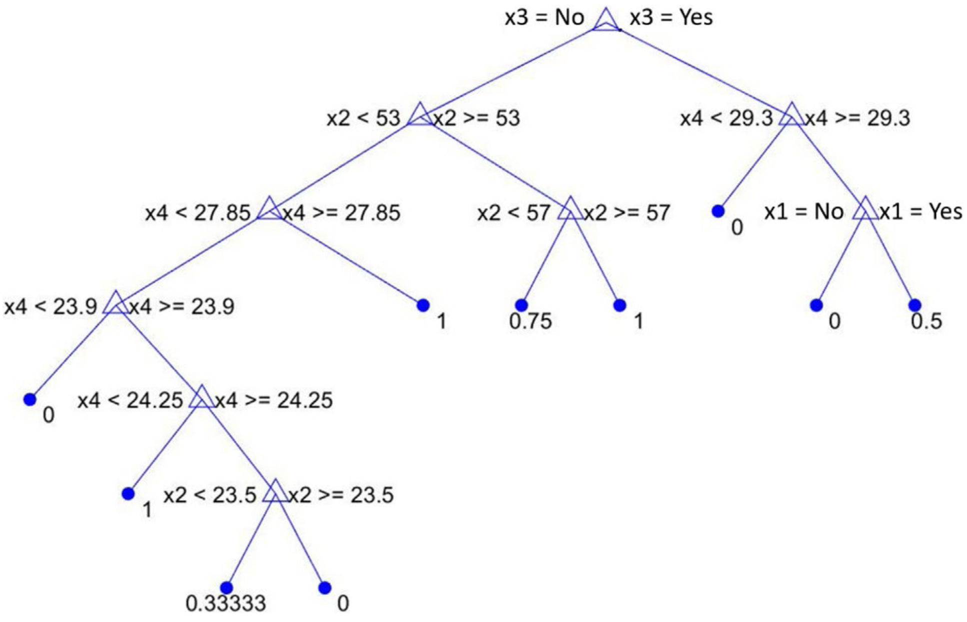


Fig. 11. Fitted CART for diabetes prediction.

AI application. To illustrate the applicability of the proposed methodology, it has been applied to several AI applications for diabetes diagnosis in the literature.

According to experimental results, the following conclusions are drawn:

- Compared with AI technologies, the tools and technologies required by XAI come from more diverse fields, such as statistics, psychology, human–computer interaction, natural language, system analysis and design, etc.
- The explanation of an AI application may not apply to all stakeholders. XAI tools and techniques must be selected and applied according to the needs of specific stakeholders.
- In the example used to illustrate the proposed methodology, the inference mechanism of ANN was successfully approximated by a simpler and more intuitive CART, which helped explain the application process and result. The extracted rules can also be used to recommend actions to the user to restore health.
- The explainability of AI applications in healthcare has indeed been enhanced by the systemic approach as transparency, comprehensibility, interpretability, understandability of ANN/DNN applications have been improved by using new XAI tools and techniques in the input (preparing synthesis data), processing (generating interpretable decision rules) and output (visualizing the probability of developing diabetes and providing follow-up recommendations).
- Another piece of evidence is that the systematic approach met the six requirements of XAI techniques in explaining AI applications for diabetes prediction, while most existing XAI techniques could only meet at most four requirements.

There are undoubtedly many ways of explaining the same AI application. This is a topic that future research can address and compare. In addition, in the process of explaining to relevant stakeholders, improving the effectiveness of the AI application based on their feedback is another direction worth exploring.

Funding

This research received no external funding.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

References

[1] D. Schultz, What's the difference between artificial intelligence and machine learning? 2020, https://blog.isa.org/whats-the-difference-between-artificial-intelligence-and-machine-learning.

[2] P. Hamet, J. Tremblay, Artificial intelligence in medicine, Metabolism 69 (2017) S36–S40.

[3] T.M. Maddox, J.S. Rumsfeld, P.R. Payne, Questions for artificial intelligence in health care, JAMA 321 (1) (2019) 31–32.

[4] M.C. Chiu, T.C.T. Chen, A ubiquitous healthcare system of 3D printing facilities for making dentures: Application of type-II fuzzy logic, Digital Health 8 (2022) 20552076221092540.

[5] K.C. Pai, W.C. Chao, Y.L. Huang, R.K. Sheu, L.C. Chen, M.S. Wang, S.H. Lin, Y.Y. Yu, C.L. Wu, M.C. Chan, Artificial intelligence–aided diagnosis model for acute respiratory distress syndrome combining clinical data and chest radiographs, Digital Health 8 (2022) 20552076221120317.

[6] C. Panigutti, A. Perotti, D. Pedreschi, Doctor XAI: An ontology-based approach to black-box sequential data classification explanations, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020, pp. 629–639.

[7] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, G.Z. Yang, XAI—Explainable artificial intelligence, Sci. Robot. 4 (37) (2019) eaay7120.

[8] T. Chen, M.-C. Chiu, Evaluating the sustainability of a smart technology application in healthcare after the COVID-19 pandemic: A hybridizing subjective and objective fuzzy group decision-making approach with XAI, Digital Health 8 (2022) 20552076221136381.

[9] M. Ghassemi, L. Oakden-Rayner, A.L. Beam, The false hope of current approaches to explainable artificial intelligence in health care, Lancet Digital Health 3 (11) (2021) e745–e750.

[10] T. Chen, Y.C. Wang, A two-stage explainable artificial intelligence approach for classification-based job cycle time prediction, Int. J. Adv. Manuf. Technol. 123 (2022) 2031–2042.

[11] T. Speith, A review of taxonomies of explainable artificial intelligence (XAI) methods, in: 2022 ACM Conference on Fairness, Accountability, and Transparency, 2022, pp. 2239–2250.

[12] T.C.T. Chen, Applications of XAI for decision making in the manufacturing domain, in: Explainable Artificial Intelligence (XAI) in Manufacturing: Methodology, Tools, and Applications, 2023, pp. 51–81.

[13] D. Kumar, A. Wong, G.W. Taylor, Explaining the unexplained: A class-enhanced attentive response (clear) approach to understanding deep neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 36–44.

[14] A. Adadi, M. Berrada, Peeking inside the black-box: A survey on explainable artificial intelligence (XAI), IEEE Access 6 (2018) 52138–52160.

[15] M. Nazar, M.M. Alam, E. Yafi, M.S. Mazliham, A systematic review of human–computer interaction and explainable artificial intelligence in healthcare with artificial intelligence techniques, IEEE Access 9 (2021) 153316-153348.

[16] J.M. Durán, Dissecting scientific explanation in AI (sXAI): A case for medicine and healthcare, Artificial Intelligence 297 (2021) 103498.

[17] G. Yang, Q. Ye, J. Xia, Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review two showcases and beyond, Inform. Fusion 77 (2022) 29–52.

[18] Y.C. Wang, T. Chen, M.-C. Chiu, An improved explainable artificial intelligence tool in healthcare for hospital recommendation, Healthcare Anal. 3 (2023) 100147.

[19] T.C.T. Chen, Explainable artificial intelligence (xai) in manufacturing, in: Explainable Artificial Intelligence (XAI) in Manufacturing: Methodology, Tools, and Applications, 2023, 1–11.

[20] T.C.T. Chen, Applications of XAI for forecasting in the manufacturing domain, in: Explainable Artificial Intelligence (XAI) in Manufacturing: Methodology, Tools, and Applications, 2023, pp. 13–50.

[21] O. Karan, C. Bayraktar, H. Gümüşkaya, B. Karlık, Diagnosing diabetes using neural networks on small mobile devices, Expert Syst. Appl. 39 (1) (2012) 54–60.

[22] S. Alian, J. Li, V. Pandey, A personalized recommendation system to support diabetes self-management for American Indians, IEEE Access 6 (2018) 73041–73051.

[23] J. Shen, J. Chen, Z. Zheng, J. Zheng, Z. Liu, J. Song, S.Y. Wong, X. Wang, M. Huang, P.-H. Fang, B. Jiang, W. Tsang, Z. He, T. Liu, B. Akinwunmi, C.C. Wang, C.J.P. Zhang, J. Huang, W.K. Ming, An innovative artificial intelligence–based app for the diagnosis of gestational diabetes mellitus (gdm-ai): development study, J. Med. Internet Res. 22 (9) (2020) e21573.

[24] L. Shinners, S. Grace, S. Smith, A. Stephens, C. Aggar, Exploring healthcare professionals' perceptions of artificial intelligence: Piloting the shinners artificial intelligence perception tool, Digital Health 8 (2022) 20552076221078110.

[25] I.A. Scott, S.M. Carter, E. Coiera, Exploring stakeholder attitudes towards AI in clinical practice, BMJ Health Care Inform. 28 (1) (2021) e100450.

[26] Y.-C. Lin, T. Chen, An intelligent system for assisting personalized COVID-19 vaccination location selection: Taiwan as an example, Digital Health 8 (2022) 20552076221109062.

[27] H.C. Wu, Y.C. Wang, T.C.T. Chen, Assessing and comparing COVID-19 intervention strategies using a varying partial consensus fuzzy collaborative intelligence approach, Mathematics 8 (10) (2020) 1725.

[28] C. Meske, E. Bunde, J. Schneider, M. Gersch, Explainable artificial intelligence: objectives stakeholders, and future research opportunities, Inform. Syst. Manag. 39 (1) (2022) 53–63.

[29] T. Chen, Y.C. Wang, Recommending suitable smart technology applications to support mobile healthcare after the COVID-19 pandemic using a fuzzy approach, Healthcare 9 (11) (2021) 1461.

[30] A. Preece, D. Harborne, D. Braines, R. Tomsett, S. Chakraborty, Stakeholders in explainable AI, 2018, arXiv preprint arXiv:1810.00184.

[31] T.C.T. Chen, Consensus measurement and enhancement, Adv. Fuzzy Group Decis. Mak. (2021) 55–72.

[32] H. Güngör, Creating value with artificial intelligence: A multi-stakeholder perspective, J. Creating Value 6 (1) (2020) 72–85.

[33] T.C.T. Chen, M.C. Chiu, Mining the preferences of patients for ubiquitous clinic recommendation, Health Care Manag. Sci. 23 (2) (2020) 173–184.

[34] M. Langer, D. Oster, T. Speith, H. Hermanns, L. Kästner, E. Schmidt, A. Sesing, K. Baum, What do we want from explainable artificial intelligence (XAI)?–A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research, Artificial Intelligence 296 (2021) 103473.

[35] M. Rigla, G. García-Sáez, B. Pons, M.E. Hernando, Artificial intelligence methodologies and their application to diabetes, J. Diabetes Sci. Technol. 12 (2) (2018) 303–310.

[36] G. Fagherazzi, P. Ravaud, Digital diabetes: Perspectives for diabetes prevention, management and research, Diabetes Metabolism 45 (4) (2019) 322–329.

[37] M. Neborachko, A. Pkhakadze, I. Vlasenko, Current trends of digital solutions for diabetes management, Diabetes Metabolic Syndrome Clin. Res. Rev. 13 (5) (2019) 2997–3003.

[38] I. Dankwa-Mullan, M. Rivo, M. Sepulveda, Y. Park, J. Snowdon, K. Rhee, Transforming diabetes care through artificial intelligence: The future is here, Popul. Health Manag. 22 (3) (2019) 229–242.

[39] S. Ellahham, Artificial intelligence: The future for diabetes care, Am. J. Med. 133 (8) (2020) 895–900.

[40] J. Chaki, S.T. Ganesh, S.K. Cidham, S.A. Theertan, Machine learning and artificial intelligence based diabetes mellitus detection and self-management: A systematic review, J. King Saud Univ.-Comput. Inform. Sci. 34 (6B) (2022) 3204–3225.

[41] N.S. Tyler, C.M. Mosquera-Lopez, L.M. Wilson, R.H. Dodier, D.L. Branigan, V.B. Gabo, F.H. Guillot, W.W. Hilts, J.E. Youssef, J.R. Castle, P.G. Jacobs, An artificial intelligence decision support system for the management of type 1 diabetes, Nat. Metabolism 2 (7) (2020) 612–619.

[42] U. Kamath, J. Liu, Explainable Artificial Intelligence: An Introduction to Interpretable Machine Learning, Springer, 2021.

[43] Y. Lau, D.G.H. Chee, X.P. Chow, S.H. Wong, L.J. Cheng, S.T. Lau, Humanoid robot-assisted interventions among children with diabetes: A systematic scoping review, Int. J. Nurs. Stud. 111 (2020) 103749.

[44] Ö. Deperlioğlu, U. Köse, Diagnogsis of diabete mellitus using deep neural network, in: 2018 Medical Technologies National Congress, 2018, pp. 1–4.

[45] H. Qteat, M. Awad, Using hybrid model of particle swarm optimization and multi-layer perceptron neural networks for classification of diabete, Int. J. Intell. Eng. Syst. 14 (2021) 11–22.

[46] H.-C. Wu, T.-C.T. Chen, M.-C. Chiu, Constructing a precise fuzzy feedforward neural network using an independent fuzzification approach, Axioms 10 (4) (2021) 282.

[47] Y.C. Lin, T.C.T. Chen, Type-II fuzzy approach with explainable artificial intelligence for nature-based leisure travel destination selection amid the COVID-19 pandemic, Digital Health 8 (2022) 20552076221106322.

[48] M.T. Ribeiro, S. Singh, C. Guestrin, Why should i trust you? Explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135–1144.

[49] T.C.T. Chen, Applications of XAI to job sequencing and scheduling in manufacturing, in: Explainable Artificial Intelligence (XAI) in Manufacturing: Methodology, Tools, and Applications, 2023, pp. 83–105.

[50] J. Nocedal, S. Wright, Numerical Optimization, Springer Science & Business Media, 2006.

[51] ConvNetJS, ConvnetJS demo: Toy 2d classification with 2-layer neural network, 2022, https://cs.stanford.edu/people/karpathy/convnetjs/demo/classify2d.html.

[52] GitHub, Tensorflow, 2022, https://github.com/tensorflow.

[53] Z. Li, J. Cai, S. He, H. Zhao, Seq2seq dependency parsing, in: Proceedings of the 27th International Conference on Computational Linguistics, 2018, pp. 3203–3214.

[54] L. Joss, E.A. Müller, Machine learning for fluid property correlations: Classroom examples with MATLAB, J. Chem. Educ. 96 (4) (2019) 697–703.

[55] M. Green, U. Ekelund, L. Edenbrandt, J. Björk, J.L. Forberg, M. Ohlsson, Exploring new possibilities for case-based explanation of artificial neural network ensembles, Neural Netw. 22 (1) (2009) 75–81.

[56] MathWorks, oobPermutedPredictorImportance, 2022, https://www.mathworks.com/help/stats/classificationbaggedensemble.oobpermutedpredictorimportance.html?searchHighlight=oobPermutedPredictorImportance&s_tid=srchtitle_oobPermutedPredictorImportance_1.

[57] K. Yan, D. Zhang, Feature selection and analysis on correlated gas sensor data with recursive feature elimination, Sensors Actuators B 212 (2015) 353–363.

[58] B.O. Kong, M.S. Kim, B.H. Kim, J.H. Lee, Prediction of creep life using an explainable artificial intelligence technique and alloy design based on the genetic algorithm in creep-strength-enhanced ferritic 9% Cr steel, Metals Mater. Int. (2022) 1–12.

[59] P. Palimkar, R.N. Shaw, A. Ghosh, Machine learning technique to prognosis diabetes disease: Random forest classifier approach, in: Advanced Computing and Intelligent Technologies: Proceedings of ICACIT 2021, 2022, pp. 219–244.

[60] T. Ooka, H. Johno, K. Nakamoto, Y. Yoda, H. Yokomichi, Z. Yamagata, Random forest approach for determining risk prediction and predictive factors of type 2 diabetes: large-scale health check-up data in Japan BMJ nutrition, Prevention & Health 4 (1) (2021) 140.

[61] A.A. Abokhzam, N.K. Gupta, D.K. Bose, Efficient diabetes mellitus prediction with grid based random forest classifier in association with natural language processing, Int. J. Speech Technol. 24 (3) (2021) 601–614.

[62] Y.-C. Wang, T.-C.T. Chen, M.-C. Chiu, An explainable deep-learning approach for job cycle time prediction, Decis. Anal. 6 (2023) 100153.

[63] J. Gerlings, M.S. Jensen, A. Shollo, Explainable ai but explainable to whom? an exploratory case study of xai in healthcare, in: Handbook of Artificial Intelligence in Healthcare: Vol 2: Practicalities and Prospects, 2022, pp. 169–198.

[64] O.I. Dauda, J.B. Awotunde, M. AbdulRaheem, S.A. Salihu, Basic issues and challenges on explainable artificial intelligence (XAI) in healthcare systems, in: Principles and Methods of Explainable Artificial Intelligence in Healthcare, 2022, pp. 248–271.

[65] A. Chaddad, J. Peng, J. Xu, A. Bouridane, Survey of explainable AI techniques in healthcare, Sensors 23 (2) (2023) 634.

[66] M.-C. Chiu, T. Chen, Assessing mobile and smart technology applications to active and healthy ageing using a fuzzy collaborative intelligence approach, Cogn. Comput. 13 (2) (2021) 431–446.