A probabilistic modeling approach for interpretable data inference and classification

Shuaiyu Yao[a,*], Jian-Bo Yang[b], Dong-Ling Xu[b]

[a]School of Naval Architecture, Ocean and Civil Engineering, Shanghai Jiao Tong University, Shanghai 200240, China
[b]Alliance Manchester Business School, The University of Manchester, Manchester M15 6PB, United Kingdom

Abstract

In this paper, we propose a new probabilistic modeling approach for interpretable inference and classification using the maximum likelihood evidential reasoning (MAKER) framework. This approach integrates statistical analysis, hybrid evidence combination and belief rule-based (BRB) inference, and machine learning. Statistical analysis is used to acquire evidence from data. The BRB inference is applied to analyze the relationship between system inputs and outputs. An interdependence index is used to quantify the interdependence between input variables. An adapted genetic algorithm is applied to train the models. The model established by the approach features a unique strong interpretability, which is reflected in three aspects: (1) interpretable evidence acquisition, (2) interpretable inference mechanism, and (3) interpretable parameters determination. The MAKER-based model is shown to be a competitive classifier for the *Banana*, *Haberman's survival*, and *Iris* data set, and generally performs better than other interpretable classifiers, e.g., complex tree, logistic regression, and naive Bayes.

Keywords: Probabilistic modeling; Interpretable inference and classification; Maximum likelihood evidential reasoning (MAKER) framework; Belief rule-base; Machine learning.

*Corresponding author. E-mail address: shuaiyuyao@sjtu.edu.cn

1. Introduction

Machine learning has attracted great attention, for its astounding capability to accurately predict a wide range of complicated phenomena [1]. Despite these successes, some black box machine learning models also have limitations and drawbacks [2]. One of the most prominent issues is insufficient transparency in models' decision behaviors, which leaves human observers with very limited understanding and knowledge of how certain decisions are generated [2].

Broadly speaking, the term "interpretability" describes the ability to explain or present model predictions to human observers in understandable terms [3-4]. In low-risk environments, some machine learning models may not require interpretability, whether because mistakes will not have serious consequences (e.g., movie recommender systems) or because the methods used have been extensively studied and assessed (e.g., optical character recognition) [5]. In other environments, however, a lack of interpretability can have more damaging consequences [6]. For example, a driverless car equipped with black box machine learning models, which does not brake when confronted with a stationary fire truck while driving at highway speeds [2]. This incorrect decision can lead the car into serious consequences (e.g., crashing into the fire truck) [2]. Such decisions may be related to bias in the models' training set, which can lead models to discriminate against a feature to maximize prediction accuracy [5]. Additionally, the black box nature of certain models makes it more challenging to identify the factors and logic leading to a wrong decision, which can be used to prevent future problems. To take another example, a black box neural network outperforms other candidate models in prediction accuracy when predicting risk of death among pneumonia patients [6,7]. The neural network predicts that pneumonia patients who also have asthma will have a lower probability of death, when in reality patients with both pneumonia and asthma have a higher risk of death [6,7]. Yet, more timely and aggressive treatments are provided to such patients who consequently have better survival prospects than non-asthmatic counterparts [6]. This type of data leakage can feed misinformation to models or artificially inflate test performance [8]. When such problems surface, interpretation is necessary to interrogate and rectify models, so that they can learn sensible features rather than spurious and misleading correlations [6,9].

Interpretability is thus necessary in many contexts, due to incomplete problem formalization which creates a major barrier to optimization and evaluation [10]. Prediction accuracy alone cannot justify a model's validity [11], and problems require both correct prediction and explicit interpretation [10]. Interpretability bridges the gap between domain knowledge and data science [12]. It facilitates a system's learning, verification, and improvement [13], and reinforces human trust in a system [14]. Understanding model predictions and how information is coded in models helps humans understand why models fail, avoiding undesirable trials and flawed development procedures [15]. In domains such as medicine, justice, and education, model interpretability helps decision makers criticize, refine, and trust in a model, based on their expert knowledge [16].

A growing body of research has focused on interpretability [1–4,10,11,14]. However, there is no consensus on the either the definition of "interpretability" in the machine learning context or how interpretability can be evaluated for benchmarking [10]. This study proposes a new probabilistic modeling approach to define "interpretability" in the context of machine learning. Evidence is first acquired from data using statistical analysis according to the maximum likelihood evidential reasoning (MAKER) framework. The MAKER framework is used for data-driven inferential modeling and decision making under different types of uncertainty [17]. It consists of state space model and evidence space model, which are driven by the data reflecting the input-output relationship. The reliability of evidence and interdependence between a pair of evidence can be explicitly measured in the MAKER framework. In such a framework, various types of uncertainty can be considered for inferential modeling, probabilistic prediction and decision making [17]. Based on the acquired evidence, the belief rule base (BRB), which features a unique strong interpretability, is established between the inputs and outputs of a numerical system. A machine learning algorithm is used to train the parameters of the interpretable model. The mechanism of the model's inference and classification is briefly analyzed and graphically presented. The model is then validated through three representative data sets, and its performance is compared with that of other models.

In the remaining part of this paper, Section 2 reviews related studies, and highlights the major innovations of this modeling approach. Section 3 discusses the

underlying methodologies of the new approach using the *Iris* data set. In Section 4, we analyze the new proposed approach, and discuss the interpretability achieved by the approach. In Section 5, the models constructed by the new approach and other alternatives are compared in terms of their prediction performance on the *Banana*, *Haberman's survival*, and *Iris* data set. Finally, Section 6 offers concluding remarks and suggestions for further research.

2. Related research

One of the simplest methods of achieving interpretability is creating interpretable models. Rule-based models, which are represented as a set of IF-THEN rules, are among the most interpretable models. The BRB proposed by Yang et al. [18] was developed by adding a belief structure to the conventional IF-THEN rule base [19]. In a BRB system, information is integrated using the evidential reasoning (ER) rule to implement inference [20,21]. ER can handle both qualitative and quantitative information under uncertainty and incompleteness [18], based on Dempster-Shafer (D-S) theory [22–24]. In the D-S theory, a frame of discernment (FoD) is used to contain pre-assigned classes to which basic probabilities are assigned to generate a belief distribution (BD). Basic probabilities are used to measure the extent to which observations of input variables point to different classes or their subsets. The BD for each observation of an input variable is referred to as a piece of evidence. While Dempster's rule can be used to combine multiple pieces of evidence, it can be applied only if certain conditions are satisfied, including that any evidence is assumed to be fully reliable. However, this assumption is impractical and can lead to counter-intuitive results when Dempster's rule is used to combine highly or completely conflicting evidence [25].

Built on the basic concepts of the D-S theory, the ER rule [26] eliminates this assumption by taking into account the reliability and relative importance of evidence, while still preserving the desirable features of Dempster's rule. One of the most important features of the ER rule is that it constitutes a unique probabilistic inference process for conjunctive combination of independent evidence. The ER rule is used to deal with discrete probabilistic inference problems where both input and output variables are assumed to take discrete numerical or categorical values. In reality,

however, inference and classification problems can have both discrete and continuous variables. As such, it is necessary to develop new methods to help solve such problems.

The new modeling approach proposed in this paper is developed to address these issues in the classification context. Several researchers have applied the BRB inference methodology and the ER approach to classification problems. Chang et al. [27] proposed a new rule activation and weight calculation procedure to construct a BRB classifier. Jiao et al. [28] developed a BRB classification system to deal with incomplete or imprecise information. Xu et al. [29] proposed a new classification method based on the ER approach. Yang et al. [21] proposed an ensemble BRB modeling method to handle classification problems. In the field of healthcare classification problems, it has been used for risk stratification of patients with cardiac chest pain [19], trauma outcome prediction [20], and diagnosis of lymph node metastasis in gastric cancer [30,31].

In addition, researchers have made attempts to use other methods to interpret the results of computational intelligence methods. Du et al. [32] propose a guided feature inversion framework which not only determines each input variable's contribution but also provides insights into the decision-making process of deep neural network models. In the study of Brian and McKeown [33], evidence is defined as the intersection of a feature's actual and expected contribution. Based on this definition, the study categorize features which are important to prediction. Long Short-Term Memory (LSTM) caption generation model [34] which has a loss function encouraging class discriminative information is used to generate justifications for image classification of a convolutional neural network. Non-iterative supervised learning models [35,36] provide a fast solution to classification and regression models with increased accuracy. They are based on the use of Ito decomposition (Kolmogorov–Gabor polynomial) and the neural-like structure of the successive geometric transformations model (SGTM). Simple interpretation of the results of the regression or classification tasks can be established on the basis of the transition from neural-like structure to the solution of the task, which is in the form of a linear polynomial.

Compared with these studies, this paper's major innovations are as follows. (1) Continuous data from an input space are discretized based on referential values at which evidence is generated through statistical analysis. (2) Acquired evidence is combined in the MAKER framework [17], which captures the interdependence between any two

pieces of evidence. (3) Based on the MAKER framework, combined evidence is used to generate belief rules, so as to formulate a BRB [18] for interpretable inference that deduces an explicit probabilistic relationship between input and output variables. (4) The parameters of the MAKER-based models such as referential values and weights are trained using machine learning algorithms. In sum, the model built by the new approach is characterized by a unique strong interpretability. It provides a specific definition of "interpretability" in the machine learning context.

3. The proposed probabilistic modeling approach and its application to the *Iris* data set

This section provides a step-by-step description of how the proposed probabilistic modeling approach is used to establish a model for a classification problem. Suppose that a training input-output data set that has N instances is indicated by $x=\left\{x_n \middle| n=1,\ldots,N\right\}$. Each instance featuring M continuous input variables ($x_{(l)}=\left\{x_{(l)} \middle| l=1,\ldots,M\right\}$) is denoted by $x_n=\left\{x_{n,l} \middle| n=1,\ldots,N; l=1,\ldots,M\right\}$. These instances are classified in a nominal output variable: $y=\left\{k \middle| k=1,\ldots,K\right\}$ where an integer represents a class of output variable. y_n indicates the specific class of output variable for each instance. For example, in the *Iris* data set, the input variables include *sepal length*, *sepal width*, *petal length*, and *petal width*, which are denoted by $x_{(1)}$, $x_{(2)}$, $x_{(3)}$, and $x_{(4)}$, respectively. The output variable contains three classes: *Iris Setosa*, *Iris Versicolor*, and *Iris Virginica*, which are signified by "1", "2", and "3", respectively. The output variable is hence represented as $y=\left\{1,2,3\right\}$.

In this section, a training set of a fold of the *Iris* data set (hereinafter referred to as "training set") is used as an example to demonstrate how a MAKER-based classifier is constructed using the training set based on the proposed approach for the classification of the *Iris* data set. The associated trained referential values (displayed in Table 1) and weights are used to develop a MAKER-based classifier for the "training set" (the complete training set is provided in Table S1 of the supplementary materials). The rest of this section is organized into five subsections: statistical evidence acquisition,

evidence independence analysis, belief rule-base inference, rules combination for classification, and training of model parameters. Fig. 1 shows the flow diagram describing the probabilistic modeling based on the training data set, adapted genetic algorithm, classification based on BRB for test data set, and their interrelationship.

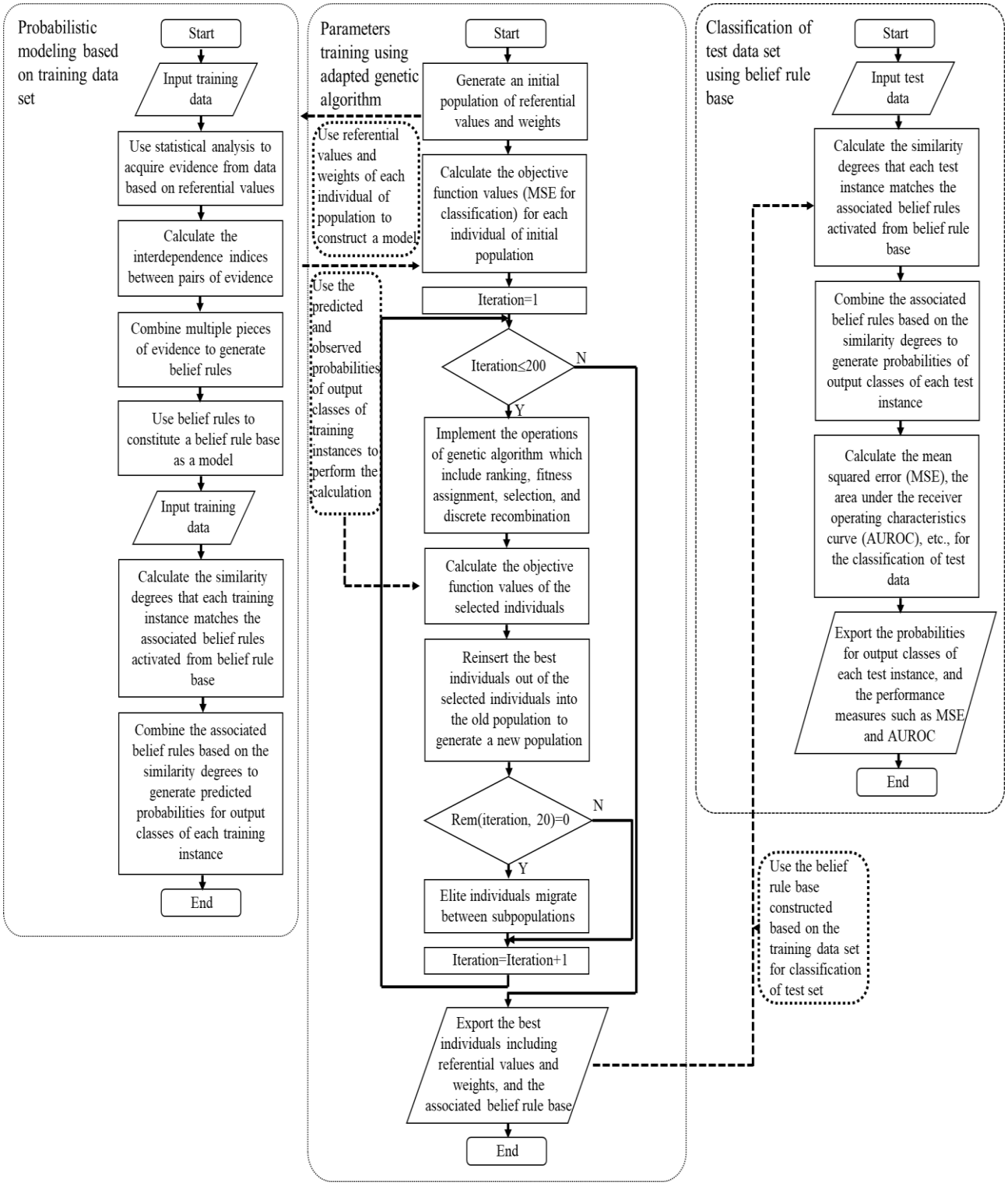

Fig. 1. Probabilistic modeling, training of model parameters, and classification

Table 1

Referential values of input variables of the example training set used for demonstration

Input variables	$x_{(1)}$ (*Sepal length*)	$x_{(2)}$ (*Sepal width*)	$x_{(3)}$ (*Petal length*)	$x_{(4)}$ (*Petal width*)
Boundary referential values (minima)	4.3000	2.0000	1.0000	0.1000
Trained referential values	4.9991	2.8969	4.4044	1.3389
Boundary referential values (maxima)	7.7000	4.4000	6.7000	2.5000

3.1. Statistical evidence acquisition

Since input variables are continuous, referential values of each variable can be used for probabilistic modeling, while any other observations in between a pair of adjacent reference values can be simulated by the probabilistic interpolation [37]. The matching degrees are generated to construct a probability distribution that is equivalent to the observation to be interpolated using certain principles. Referential values can be initially assigned through statistical analysis and subsequently fine-tuned by optimal training [29,38].

Using the referential values, we can transform an observation $x_{n,l}$ into a belief distribution of a referential value $A_{i,l}$, which is shown in Eq. (1).

$$S\left(x_{n,l}\right)=\left\{\left(A_{i,l},\alpha_{n,i,l}^{(k)}\right)\middle|n=1,\ldots,N;l=1,\ldots,M;i=1,\ldots,T_l;k=1,\ldots,K\right\}$$

where

$$\alpha_{n,i,l}^{(k)}=\frac{A_{i+1,l}-x_{n,l}}{A_{i+1,l}-A_{i,l}}\ \textit{and}\ \alpha_{n,i+1,l}^{(k)}=1-\alpha_{n,i,l}^{(k)},\ \textit{if}\ A_{i,l}\le x_{n,l}\le A_{i+1,l};$$
$$\alpha_{n,i',l}^{(k)}=0,\ \textit{for}\ i'=1,\ldots,T_l\ \textit{and}\ i'\ne i,i+1. \quad (1)$$

In Eq. (1), $\alpha_{n,i,l}^{(k)}$ is the matching degree for the n^{th} observation of the l^{th} input variable (indicated by '$x_{n,l}$') matching the i^{th} referential value of the l^{th} input variable (denoted by '$A_{i,l}$') which points to a class (represented by "k") of output variable. For instance, an observation of *sepal length* is 5.8000. According to Table 1, 5.8000 is between two adjacent referential values of *sepal length*: 4.9991 and 7.7000. The matching degrees that 5.8000 matches 4.9991 and 7.7000 are generated as $\frac{7.7000-5.8000}{7.7000-4.9991}\approx 0.7035$ and $1-\frac{7.7000-5.8000}{7.7000-4.9991}\approx 0.2965$, respectively. As 5.8000 is not between 4.3000 and 4.9991, the matching degree of 5.8000 to 4.3000 is 0. The associated belief distribution of an observation: 5.8000 over referential values: 4.3000, 4.9991, and 7.7000 is hence presented as $\left(0.0000,\ 0.7035,\ 0.2965\right)$.

Matching degrees for each referential value are aggregated for each class to generate an associated total matching degree for each class, as shown in Eq. (2). The total matching degree of a referential value for a class is then treated as the frequency for the referential value matching the class. Using the referential values shown in Table 1, we can generate all the frequencies of the referential values for an input variable. Table 2 displays the frequencies of the referential values: 4.3000, 4.9991, and 7.7000 matching classes of the input variable: *sepal length*.

$$\alpha_{i,l}^{(k)}=\sum_{n=1}^{N}\alpha_{n,i,l}^{(k)} \quad (2)$$

Table 2

The frequencies of referential values for input variable of *sepal length*

$y_n=k\setminus A_{i,1}$	4.3000	4.9991	7.7000
1 (*Iris setosa*)	7.5588	30.3599	2.0812

2 (*Iris versicolor*)	0.0000	26.2507	13.7493
3 (*Iris virginica*)	0.1417	17.0017	22.8566

For each row of a frequency table such as Table 1, we can generate a sum of frequency values ($\delta_l^{(k)}$), using $\delta_l^{(k)} = \sum_{i=1}^{T_l} \alpha_{i,l}^{(k)}$. For example, the sum of frequency values for the third row of Table 1 is generated by $\sum_{i=1}^{3} \alpha_{i,2}^{(3)} = 0.1417 + 17.0017 + 22.8566 = 40.0000$. Let $c_{n,i,l}^{(k)}$ be the likelihood to which the i^{th} referential value of the l^{th} input variable is observed given that the k^{th} class of output variable is known. The likelihood can be calculated as shown in Eq. (3) [17,39]. This is exemplified by the likelihood of observing the referential value of sepal length: 4.9991 given that the output class is *Iris virginica* (indicated by "3"), which is generated by $\frac{17.0017}{40.0000} \approx 0.4250$. Table 3 presents the associated likelihoods of referential values of *sepal length* being different species, which are calculated from the frequencies in Table 2.

$$c_{i,l}^{(k)} = \begin{cases} \dfrac{\alpha_{i,l}^{(k)}}{\sum_{i=1}^{T_l} \alpha_{i,l}^{(k)}}, \ if \ \sum_{i=1}^{T_l} \alpha_{i,l}^{(k)} \neq 0 \\ 0, \ if \ \sum_{i=1}^{T_l} \alpha_{i,l}^{(k)} = 0 \end{cases} \tag{3}$$

Table 3

The likelihoods of referential values of input variable: *sepal length* matching classes of output variable

$y_n = k \setminus A_{i,1}$	4.3000	4.9991	7.7000
1	0.1890	0.7590	0.0520
2	0.0000	0.6563	0.3437
3	0.0035	0.4250	0.5714

Based on $\eta_{i,l} = \sum_{k=1}^{K} c_{i,l}^{(k)}$, the sum of the likelihoods (signified by '$\eta_{i,l}$') can be produced for each column of a table of likelihoods such as Table 3. An example is that the sum of the likelihoods for the column beginning with "4.9991" is obtained by

$\eta_{2,1} = 0.7590 + 0.6563 + 0.4250 = 1.8403$. With these likelihoods, the probability of a referential value of an input variable pointing to a class of output variable ($p_{i,l}^{(k)}$) can be generated by Eq. (4). For example, the probability that the referential value of sepal length: 4.9991 points to the class: *Iris versicolor* (indicated by "2") is given by $p_{2,1}^{(2)} = \frac{c_{2,1}^{(2)}}{\eta_{2,1}} = \frac{0.6563}{1.8403} \approx 0.3566$. Tables 4 and 5 show the probabilities of referential values of *sepal length* and *sepal width*, respectively, pointing to classes of output variable, which are calculated through the normalization of likelihoods in tables such as Table 3 by Eq. (4). Further, we can acquire a piece of evidence at each referential value of Table 3 under the framework of the ER rule, which is defined as a belief distribution [17,39] shown in Eq. (5).

$$p_{i,l}^{(k)} = \begin{cases} \dfrac{c_{i,l}^{(k)}}{\sum_{k=1}^{K} c_{i,l}^{(k)}} \ \ if \ \ \sum_{k=1}^{K} c_{i,l}^{(k)} \neq 0 \\ 0 \ , if \ \ \sum_{k=1}^{K} c_{i,l}^{(k)} = 0 \end{cases} \tag{4}$$

Table 4

The probabilities of referential values of input variable: *sepal length* matching classes of output variable

$y_n = k \setminus A_{i,1}$	4.3000	4.9991	7.7000
1	0.9816	0.4124	0.0538
2	0.0000	0.3566	0.3554
3	0.0184	0.2310	0.5908

Table 5

The probabilities of referential values of input variable: *sepal width* matching classes of output variable

$y_n = k \setminus A_{j,2}$	2.0000	2.8969	4.4000
1	0.0000	0.3019	0.7055
2	0.7011	0.3324	0.0847

3	0.2989	0.3657	0.2098

$$e_j = \left\{ \left(\theta, p_{\theta,j} \right), \forall \theta \subseteq \Theta, \sum_{\theta \subseteq \Theta} p_{\theta,j} = 1 \right\} \quad (5)$$

In Eq. (5), $\Theta = \left\{ h_1, \ldots, h_N \right\}$ is referred to as a frame of discernment, which denotes a set of mutually exclusive and collectively exhaustive hypotheses [39]. $\left(\theta, p_{\theta,j} \right)$ is an element of evidence e_j that points to proposition θ, which can be any subset of Θ other than the empty set, with a probability: $p_{\theta,j}$ [39]. For instance, the classes of output variable: *Iris setosa*, *Iris versicolor*, and *Iris virginica* can be signified by "1", "2", and "3", respectively. The frame of discernment is hence denoted as $\Theta = \{1, 2, 3\}$. As shown in Table 4, the probabilities: 0.9816, 0.0000, and 0.0184, of a referential value: 4.3000, indicate that if *sepal length* is 4.3000 cm, the probability of this flower being *Iris setosa* (class: "1") is 0.9816, and that of this flower being *Iris versicolor* (class: "2") is 0.0000, and that of this flower being *Iris virginica* (class: "3") is 0.0184. Thus, we can define a piece of evidence at *sepal length* of 4.3000 cm, where it points to class: "1", "2", and "3" with a probability of 0.9816, 0.0000, and 0.0184, respectively.

3.2. Evidence interdependence analysis

To achieve greater predictive power, it is necessary to combine multiple pieces of evidence to generate predicted probabilities for certain classes. In the ER rule, independence is assumed between a pair of evidence. Under the MAKER framework, the degree of interdependence between a pair of evidence is measured by an interdependence index "α" generated by marginal and joint likelihood function [17]. To generate the interdependence index between a pair of evidence, we need first to estimate the joint probabilities for a pair of evidence. Let $x_{n,l}$ and $x_{n,m}$ be the n^{th} observation of the l^{th} and m^{th} input variable, respectively. Given that the simultaneous observations of $x_{n,l}$ and $x_{n,m}$ are characterized by a probability distribution, we can use Eq. (6) to generate $\alpha_{n,il,jm}^{(k)}$, which stands for the joint matching degree for these two observations matching the combination of referential values:

$\{A_{i,l}, A_{j,m}\}$ that points to a class of output variable (indicated by "k"). An instance: $\{5.0000,\ 2.3000\}$ is cited as an example. The observations: 5.0000 and 2.3000 in the instance can activate two sets of adjacent referential values: $\{4.9991,\ 7.7000\}$ and $\{2.0000,\ 2.8969\}$, respectively. The matching degree of 5.0000 to 4.9991 is generated by $\frac{7.7000-5.0000}{7.7000-4.9991}\approx 0.9997$, and that of 2.3000 to 2.0000 is obtained by $\frac{2.8969-2.3000}{2.8969-2.0000}\approx 0.6655$. Hence, the joint matching degree that $\{5.0000,\ 2.3000\}$ matches the combination of two referential values: $\{7.7000,\ 2.0000\}$ is obtained by $0.9997\times 0.6655\approx 0.6653$.

$$\alpha_{n,il,jm}^{(k)} = \alpha_{n,i,l}^{(k)}\alpha_{n,j,m}^{(k)} \tag{6}$$

Similarly, we can generate joint matching degrees for each instance in a data set matching associated combinations of referential values which point to different classes of output variable. These joint matching degrees can be further aggregated to generate frequency values for associated referential values combinations pointing to classes of output variable ($\alpha_{il,jm}^{(k)}$), using Eq. (7).

$$\alpha_{il,jm}^{(k)} = \sum_{n=1}^{N}\alpha_{n,il,jm}^{(k)} \tag{7}$$

Based on these frequency values, the joint likelihoods for referential values combinations pointing to classes of output variable ($c_{il,jm}^{(k)}$) are generated by Eqs. (8) and (9).

$$\delta_{lm}^{(k)} = \sum_{i=1}^{T_l}\sum_{j=1}^{T_m}\alpha_{il,jm}^{(k)} \tag{8}$$

$$c_{il,jm}^{(k)} = \frac{\alpha_{il,jm}^{(k)}}{\delta_{lm}^{(k)}} \tag{9}$$

With joint likelihoods, Eq. (10) is used to obtain the joint probabilities for referential values combinations pointing to classes of output variable ($p_{il,jm}^{(k)}$). Table 6 exhibits the joint probabilities of referential values combinations of sepal length and sepal width pointing to classes of output variable.

$$p_{il,jm}^{(k)}=\begin{cases}\dfrac{c_{il,jm}^{(k)}}{\sum_{k=1}^{K}c_{il,jm}^{(k)}},\ if\ \ \sum_{k=1}^{K}c_{il,jm}^{(k)}\neq 0;\\ 0,\ if\ \ \sum_{k=1}^{K}c_{il,jm}^{(k)}=0.\end{cases} \tag{10}$$

Table 6

The joint probabilities of referential values combinations of input variables: *sepal length* and *sepal width* pointing to different classes of output variable (a referential value of *sepal length* and that of *sepal width* are represented by $A_{i,1}$ and $A_{j,2}$, respectively)

$A_{i,1}$	$A_{j,2}$	$y_n=1$	$y_n=2$	$y_n=3$
4.3000	2.0000	0.0000	0.0000	1.0000
4.3000	2.8969	0.9876	0.0000	0.0124
4.3000	4.4000	1.0000	0.0000	0.0000
4.9991	2.0000	0.0000	0.7462	0.2538
4.9991	2.8969	0.3777	0.3602	0.2621
4.9991	4.4000	0.8406	0.0649	0.0945
7.7000	2.0000	0.0000	0.6003	0.3997
7.7000	2.8969	0.0284	0.3556	0.6160
7.7000	4.4000	0.2558	0.1615	0.5826

Using Eqs. (4) and (10), we can generate $\alpha_{A,B,i,j}$ representing interdependence indices between a pair of evidential elements indicated by $e_{i,l}(A)$ and $e_{j,m}(B)$, which is shown in Eq. (11).

$$\alpha_{A,B,i,j}=\begin{cases}0,\ if\ \ p_{A,i,l}=0\ or\ \ p_{B,j,m}=0\\ \dfrac{p_{A,B,il,jm}}{p_{A,i,l}p_{B,j,m}},\ otherwise\end{cases} \tag{11}$$

In Eq. (11), $p_{A,i,l}$ and $p_{B,j,m}$ are the basic probabilities for single input variables: x_l at $x_{i,l}$ and x_m at $x_{j,m}$, which point to propositions A and B, respectively. $p_{A,B,il,jm}$ represents the joint basic probability for both $x_{i,l}$ and $x_{j,m}$ being observed for the proposition $\theta\left(\theta=A\cap B,\theta\subseteq\Theta\right)$, which is generated using Eq. (10) based on the

joint table for $p_{A,i,l}$ and $p_{B,j,m}$. When $e_{i,l}(A)$ and $e_{j,m}(B)$ are disjoint, $\alpha_{A,B,i,j}$ is given as 0. If $e_{i,l}(A)$ is independent from $e_{j,m}(B)$, 1 is assigned to $\alpha_{A,B,i,j}$ [17]. An example of this is that the interdependence index between a piece of evidence from *sepal length* at 4.9991 and another piece from sepal width at 2.8969, which points to class: *Iris versicolor* (indicated by "2"), is obtained by $\frac{0.3602}{0.3566 \times 0.3324} \approx 3.0388$ (0.3566, 0.3324, and 0.3602 are from Tables 4, 5, and 6, respectively). In a general sense, the larger the interdependence index is, the more independent two evidential elements are from each other. Based on the probabilities shown in Tables 4, 5, and 6, Eq. (11) is used to generate the interdependence indices between a piece of evidence from *sepal length* and another piece from *sepal width*, which are displayed in Table 7.

Table 7

The interdependence indices between a piece of evidence from *sepal length* ($e_{i,1}$) and another piece from *sepal width* ($e_{j,2}$)

$e_{i,1}$ *at* $A_{i,1}$	$e_{j,2}$ *at* $A_{j,2}$	1	2	3
4.3000	2.0000	0.0000	0.0000	181.8097
4.3000	2.8969	3.3328	0.0000	1.8426
4.3000	4.4000	1.4441	0.0000	0.0000
4.9991	2.0000	0.0000	2.9846	3.6764
4.9991	2.8969	3.0340	3.0388	3.1024
4.9991	4.4000	2.8891	2.1496	1.9491
7.7000	2.0000	0.0000	2.4091	2.2636
7.7000	2.8969	1.7463	3.0103	2.8510
7.7000	4.4000	6.7411	5.3646	4.7003

From Table 7, it is clear that the majority of interdependence indices are moderate, as they are between 1.0000 and 7.0000. For example, the interdependence between *sepal length*: 4.9991 and *sepal width*: 4.4000 is considered to be moderate, as the associated indices matching *Iris setosa* (class "1"), *Iris versicolor* (class "2"), and *Iris virginica* (class "3") are 2.8891, 2.1496, and 1.9491, respectively. An exceptional

example is that *sepal length*: 4.3000 and *sepal width*: 2.0000 are highly independent from each other, as the associated interdependence index is 181.8097.

3.3. Belief rule-base inference

Based on the acquired evidence and interdependence analysis, we are now in a position to construct a BRB to infer the likelihood of a class for an instance in a data set. A BRB can capture incomplete, fuzzy, and ignorant information, along with nonlinear causal relationship between antecedent attributes and consequents [18]. It consists of a finite number of belief rules, which are defined as follows [18].

$$\begin{aligned} &R_k : if\ \left(x_1\ is\ A_1^k\right) \wedge \left(x_2\ is\ A_2^k\right) \wedge \cdots \wedge \left(x_{M_k}\ is\ A_{M_k}^k\right), \\ &then \left\{\left(D_1, \beta_{1,k}\right), \left(D_2, \beta_{2,k}\right), \ldots, \left(D_N, \beta_{N,k}\right)\right\}, \\ &with\ a\ rule\ weight\ \theta_k\ and\ attribute\ weights\ \delta_1, \delta_2, \ldots, \delta_{M_k}. \end{aligned} \tag{12}$$

In (12), R_k denotes the $k^{th}\left(k=1,\ldots,L\right)$ belief rule. M_k represents the number of antecedent attributes in the k^{th} rule, and $x_m\left(m=1,\ldots,M_k\right)$ indicates the m^{th} antecedent attribute. $A_j^k\left(j=1,\ldots,M_k; k=1,\ldots,L\right)$ is the referential value of the j^{th} antecedent attribute in the k^{th} rule. "$\wedge$" signifies a logical connective which denotes the relationship of "AND". $\beta_{n,k}\left(n=1,\ldots,N; k=1,\ldots,L\right)$ implies the belief degree for a consequence D_n which can be initially provided by experts. Given that $\sum_{n=1}^{N}\beta_{n,k}=1$, the k^{th} rule is complete. Otherwise, it is incomplete. θ_k and $\delta_m\left(m=1,\ldots,M_k\right)$ represent the relative weight of the k^{th} rule and the m^{th} antecedent attribute in the k^{th} rule, respectively.

As per the belief rule described in (12), the antecedent of a belief rule, which is represented in the form of "$if\ \left(x_1\ is\ A_1^k\right) \wedge \left(x_2\ is\ A_2^k\right) \wedge \ldots \wedge \left(x_2\ is\ A_{M_k}^k\right)$", in an example of classification problem should be understood as "if the observations of an instance are just equal to their respective associated referential values". It is noted that as a single piece of evidence is defined at a referential value, the antecedent of a belief rule can be considered as the combination of multiple pieces of evidence. The associated

385 consequent part of a belief rule, which is expressed in the form of
386 " *then* $\{(D_1,\beta_{1,k}),(D_2,\beta_{2,k}),\ldots,(D_N,\beta_{N,k})\}$ ", should then be interpreted as "the
387 consequences that an instance is classified as different classes have respective
388 probabilities".

389 To obtain the probabilities for the consequences, multiple pieces of evidence from
390 input variables are combined using the conjunctive MAKER rule. To combine evidence,
391 it is necessary to consider the reliability of evidence to measure the degree of its support
392 for proposition θ, as evidence is seldom fully reliable. Let $r_{\theta,i,l}$ be the reliability of
393 evidence $e_{i,l}$ pointing to proposition θ. As displayed in Eq. (13), $r_{\theta,i,l}$, which
394 essentially measures the quality of $e_{i,l}$, is defined as a conditional probability for θ
395 being true given that $e_{i,l}$ points to θ [17]. $r_{\theta,i,l}$ is related to how data are generated
396 and how $e_{i,l}$ is acquired from data [17].

397 $$r_{\theta,i,l} = p\left(\theta \middle| e_{i,l}(\theta)\right) \tag{13}$$

398 Using Eq. (13), we can further obtain the reliability of a piece of evidence $e_{i,l}$,
399 which is displayed in Eq. (14).

400 $$r_{i,l} = \sum_{\theta\subseteq\Theta} r_{\theta,i,l} p_{\theta,i,l} \tag{14}$$

401 To consider reliability in the process of combining multiple pieces of evidence, it
402 is necessary to use the probability mass to combine evidence and reliability. There are
403 two possible scenarios about generating the probability mass for the proposition θ
404 being supported by $e_{i,l}$, in terms of whether $e_{i,l}$ and other pieces of evidence are
405 acquired from the same data source. These scenarios are shown as follows.

406 **Scenario 1:** If $e_{i,l}$ and other pieces of evidence are acquired from the same data
407 source, the probability mass for the proposition θ being supported by $e_{i,l}$ is generated
408 by Eq. (15).

409 $$m_{\theta,i,l} = p\left(\theta \middle| e_{i,l}(\theta)\right) p\left(e_{i,l}(\theta)\right) = r_{\theta,i,l} p\left(e_{i,l}(\theta)\right) \tag{15}$$

Scenario 2: If $e_{i,l}$ featuring a probability function p_l, is acquired from a data source that is different from other pieces of evidence, the probability mass for the proposition θ is produced by Eq. (16).

$$m_{\theta,i,l} = w_{\theta,i,l}\, p_l\left(e_{i,l}\left(\theta\right)\right) \tag{16}$$

In Eq. (16), $w_{\theta,i,l} = \omega_{i,l}\, p_l\left(\theta \middle| e_{i,l}\left(\theta\right)\right)$ denotes the weight of an evidential element $e_{i,l}\left(\theta\right)$. $w_{\theta,i,l}$ is in proportion to the conditional probability for θ being true provided that $e_{i,l}$ points to θ [17]. The conditional probability is measured by a probability function p_l built from data for x_l only [17]. Of note is that if $p = p_l$, $w_{\theta,i,l} = r_{\theta,i,l}$, which indicates that $\omega_{i,l} = 1$.

In each classification experiment of this paper, the associated data set is obtained from a single data source. Hence, in each classification experiment, the reliability $\left(r_{\theta,i,l}\right)$ and weight $\left(w_{\theta,i,l}\right)$ for any evidential element are essentially the same. To make it simple, $r_{\theta,i,l} = w_{\theta,i,l}$, which is used in the process of evidence combination.

Based upon the above definitions and discussions, the conjunctive MAKER rule is employed to combine multiple pieces of evidence to generate the combined probabilities for an evidence combination or antecedent of a belief rule. Eqs. (17) and (18) display the conjunctive MAKER rule to obtain $p_{\theta,(2)}$, which represents the combined probability for the proposition θ being jointly supported by a pair of evidence $e_{i,l}$ and $e_{j,m}$.

$$p_{\theta,e(2)} = \begin{cases} 0, \ \theta = \varnothing \\ \dfrac{m_{\theta,e(2)}}{\sum_{D\subseteq\Theta} m_{D,e(2)}}, \ \theta \subseteq \Theta, \ \theta \neq \varnothing \end{cases} \tag{17}$$

$$m_{\theta,e(2)} = \left[\left(1-r_{j,m}\right)m_{\theta,i,l} + \left(1-r_{i,l}\right)m_{\theta,j,m}\right] + \sum_{A\cap B=\theta} \gamma_{A,B,il,jm}\alpha_{A,B,il,jm} m_{A,i,l} m_{B,j,m} \tag{18}$$

In Eqs. (17) and (18), $m_{\theta,e(2)}$ denotes the combined probability mass for both $e_{i,l}$ and $e_{j,m}$ jointly supporting proposition θ. $\gamma_{A,B,i,j}$ is referred to as a nonnegative parameter reflecting the degree of joint support that both $e_{i,l}$ and $e_{j,m}$ provide to θ,

which is relative to the individual support from $e_{i,l}$ and $e_{j,m}$ that point to propositions A and B, respectively [17]. It is assumed that $\gamma_{A,B,i,j}$ is 1 in the classification experiments of this paper. If $w_{\theta,i,l} = w_{i,l}$ for any A, B, and $\theta \subseteq \Theta$, we can apply the above conjunctive MAKER rule to combination of independent evidence, which reduces to the evidential reasoning rule [25]. When $w_{i,l} = r_{i,l} = 1$, the MAKER rule can be further reduced to Dempster's rule [24]. It can be reduced even further to Bayes's rule if there is no ambiguity in data [17]. It should be noted that Eq. (18) is recursively applied for evidence before Eq. (17) is used.

Using the conjunctive MAKER rule, the combined probabilities of classes of output variable can be generated for all the possible combinations of multiple pieces of evidence in a data set. Each possible evidence combination and its associated combined probabilities are used to generate a belief rule. All the possible belief rules constitutes a BRB. In the *Iris* data set, there are four input variables. We can define three pieces of evidence at the associated referential values of each input variable, which are displayed in Table 1. Thus, there are $3^4 = 81$ possible combinations of four pieces of evidence. Each evidence combination and its associated combined probabilities are used to form a belief rule which is exhibited in Table S2 of the supplementary materials. For example, four pieces of evidence defined at 7.7000 of *sepal length*, 4.4000 of *sepal width*, 6.7000 of *petal length*, and 2.5000 of *petal width* are combined to generate probabilities: 0.0066, 0.0015, and 0.9919 matching classes: "1", "2", and "3", respectively. Hence, the associated belief rule is that if sepal length is 7.7000, and sepal width is 4.4000, and petal length is 6.7000, and petal width is 2.5000, then the probability for *Iris setosa* is 0.0066, and that for *Iris versicolor* is 0.0015, and that for *Iris virginica* is 0.9919.

3.4. Rule combination for classification

Based on a BRB, the conjunctive MAKER rule is further used to generate the predicted probabilities of classes of output variable for an instance of a data set. Each observation of an input variable can activate two adjacent referential values it is in between. An instance featuring two input variables can activate $2^2 = 4$ combinations of referential values from two input variables. In other words, it can activate 4 belief

rules featuring two input variables. Similarly, in the example of "training set", an instance: $\{5.0000,\ 2.3000,\ 3.3000,\ 1.0000\}$ can activate $2^4=16$ belief rules out of the BRB, which are displayed in Table 8.

Table 8

The belief rules activated by an instance: $\{5.0000,\ 2.3000,\ 3.3000,\ 1.0000\}$ out of a belief rule base

Rule No.	**If** the values of features of a flower are				**Then** the probabilities of the flower being *Iris setosa*, *Iris versicolor*, and *Iris virginica* are		
	Sepal length	*Sepal width*	*Petal length*	*Petal width*	*Iris setosa*	*Iris versicolor*	*Iris virginica*
1	4.9991	2.0000	1.0000	0.1000	0.0736	0.9099	0.0165
2	4.9991	2.0000	1.0000	1.3389	0.1235	0.8385	0.0380
3	4.9991	2.0000	4.4044	0.1000	0.0658	0.8800	0.0542
4	4.9991	2.0000	4.4044	1.3389	0.0000	0.8470	0.1530
5	4.9991	2.8969	1.0000	0.1000	0.9561	0.0413	0.0026
6	4.9991	2.8969	1.0000	1.3389	0.6668	0.3259	0.0073
7	4.9991	2.8969	4.4044	0.1000	0.5540	0.4297	0.0163
8	4.9991	2.8969	4.4044	1.3389	0.0000	0.9595	0.0405
9	7.7000	2.0000	1.0000	0.1000	0.0552	0.9346	0.0102
10	7.7000	2.0000	1.0000	1.3389	0.0950	0.8808	0.0242
11	7.7000	2.0000	4.4044	0.1000	0.0530	0.9109	0.0361
12	7.7000	2.0000	4.4044	1.3389	0.0006	0.8876	0.1118
13	7.7000	2.8969	1.0000	0.1000	0.9538	0.0456	0.0006
14	7.7000	2.8969	1.0000	1.3389	0.7236	0.2740	0.0024
15	7.7000	2.8969	4.4044	0.1000	0.3998	0.5969	0.0033
16	7.7000	2.8969	4.4044	1.3389	0.0039	0.9832	0.0129

Eq. (19) is used to generate $\alpha_{n,il,jm}$, which denotes a joint matching degree for an instance characterized by two input variables ($\{x_{n,l}, x_{n,m}\}$) matching referential values combinations of a belief rule ($\{A_{i,l}, A_{j,m}\}$). It can be further extended to the instances featuring more input variables. A joint matching degree indicates the degree to which we should use the activated belief rules to predict the probability for each class of output variable for an instance.

$$\alpha_{n,il,jm} = \alpha_{n,i,l}\alpha_{n,j,m}$$

where

$$\alpha_{n,i,l} = \frac{A_{i+1,l} - x_{n,l}}{A_{i+1,l} - A_{i,l}},\ \textit{and}\ \alpha_{n,i+1,l} = 1 - \alpha_{n,i,l},\ \textit{if}\ A_{i,l} \le x_{n,l} \le A_{i+1,l};$$
$$\alpha_{n,i',l} = 0,\ \textit{if}\ i' = 1,\ldots,T_l,\ \textit{and}\ i' \ne i, i+1. \tag{19}$$

In the instance: $\{5.0000,\ 2.3000,\ 3.3000,\ 1.0000\}$, 5.0000, 2.3000, 3.3000, and 1.0000 activates the sets of two adjacent referential values: $\{4.9991,\ 7.7000\}$, $\{2.0000,\ 2.8969\}$, $\{1.0000,\ 4.4044\}$, and $\{0.1000,\ 1.3389\}$, respectively. The matching degree of 5.0000 to 4.9991, that of 2.3000 to 2.8969, that of 3.3000 to 1.0000, and that of 1.0000 to 1.3389 are generated by $\frac{7.7000-5.0000}{7.7000-4.9991} \approx 0.9997$, $1-\frac{2.8969-2.3000}{2.8969-2.0000} \approx 0.3345$, $\frac{4.4044-3.3000}{4.4044-1.0000} \approx 0.3244$, and $1-\frac{1.3389-1.0000}{1.3389-0.1000} \approx 0.7265$, respectively. Thus, the matching degree that the instance: $\{5.0000,\ 2.3000,\ 3.3000,\ 1.0000\}$ matches the referential values combination: $\{4.9991,\ 2.8969\ ,1.0000\ ,1.3389\}$ upon which a belief rule is based, is generated by $0.9997 \times 0.3345 \times 0.3244 \times 0.7265 \approx 0.0788$. This indicates that the instance matches the referential values combination on which a belief rule is based to a low degree, and that the belief rule plays a small role in the rules combination for classification of an instance.

Having generated all the associated joint matching degrees to which an instance matches referential values combinations of belief rules, we can combine the activated belief rules to predict the probabilities of classes of output variable for an instance. To combine these belief rules, their reliabilities and weights need to be considered. Let $e_{(L)}$ represent L pieces of evidence. The combination of referential values that $e_{(L)}$ are defined at constitutes the antecedent of a belief rule. Let $r_{e(L)}$ and $w_{e(L)}$ be the reliability and weight, respectively, of a belief rule. $r_{e(L)}$ and $w_{e(L)}$ ($r_{e(L)} = w_{e(L)}$ in this paper) can be initially determined based on expert knowledge or trained using data sets.

Based on the joint matching degree for an instance matching an activated belief rule and the activated one's reliability (weight), we are able to generate the updated reliability (weight) of the activated belief rule for an instance (represented by '$r'_{e(L)}$'), which is shown in Eq. (20).

$$r'_{e(L)} = \alpha_{n,e(L)} r_{e(L)} \tag{20}$$

In Eq. (20), $\alpha_{n,e(L)}$ indicates the matching degree for an instance matching an activated belief rule, which is generated using the method extended from Eq. (20). The updated reliability (weight) helps us consider how reliable and important an activated belief rule is in the combination of activated belief rules. With the associated updated reliability (weight) of a belief rule, we can use the conjunctive MAKER rule to combine the activated belief rules to predict the probabilities for classes of output variable assigned to an instance. For example, based on the MAKER rule, the belief rules activated by the instance: $\{5.0000, 2.3000, 3.3000, 1.0000\}$ can be combined to generate probabilities: 0.1079, 0.8288, and 0.0632 for class: "1", "2", and "3", respectively. In other words, if a flower has a *sepal length*: 5.0000, *sepal width*: 2.3000, *petal length*: 3.3000, and *petal width*: 1.0000, the predicted probabilities of a flower being *Iris setosa*, *Iris versicolor*, and *Iris virginica* are 0.1079, 0.8288, and 0.0632, respectively.

3.5. Training of model parameters

In the above process, $A_{i,l}$, $r_{\theta,i,l}$, $w_{\theta,i,l}$, etc., are the adjustable parameters of models assigned for inference and prediction. These parameters can be trained based on the data sets for classification. An optimal learning model is proposed for parameters training based on the principle of maximizing the likelihood of true class of output variable, which is displayed in Eq. (21).

$$\begin{aligned} &\min\ \delta \\ &s.t.\ A_{i,l},\ r_{\theta,i,l},\ w_{\theta,i,l}\ \gamma_{A,B,il,jm} \in \Omega \end{aligned} \tag{21}$$

In Eq. (22), $\delta=\frac{1}{KN}\sum_{n=1}^{N}\sum_{\theta\subseteq\Theta}\left(p_n(\theta)-\hat{p}_n(\theta)\right)^2$, $p_n(\theta)$ and $\hat{p}_n(\theta)$ indicate the predicted and observed probability for the proposition θ being true, respectively, which is provided in the n^{th} instance of a classification data set. K represents the number of hypotheses in a frame of discernment or the number of classes in an output variable. The target of optimal learning model is to minimize the mean squared error (MSE) to make $p_n(\theta)$ as close to $\hat{p}_n(\theta)$ as possible. Ω is referred to as a feasible space of parameters including the constraints e.g., $0\le r_{\theta,i,l}\le 1$. Based on the optimal learning model, an adapted genetic algorithm [38] is employed to train the parameters of a model built by the proposed probabilistic modeling approach, which is based on the MAKER framework (hereinafter referred to as the MAKER-based model). In the algorithm, each individual of a population contains both the referential values where some pieces of evidence are located and the weights for evidential elements, which is suitable for parallel computing. The optimal training on the data sets in this study is highly complex, and the conventional mathematical methods are not efficient. There are a few advantages in applying the genetic algorithm to optimization problems, which are shown as follows [40]. (1) The genetic algorithm does not have many mathematical requirements about optimization problems, and it can handle various types of objective function and constraints (i.e., linear or nonlinear) defined on discrete, continuous, or mixed search spaces. (2) The ergodicity of evolution operators makes genetic algorithms very effective at performing a global search. (3) Genetic algorithms provide us with great flexibility to hybridize with domain-dependent heuristics to achieve efficient implementation for a specific optimization problem. In the further studies, other optimization algorithms will be used for the optimization of the MAKER-based model and compared with the adapted genetic algorithm. The MAKER-based model is capable of capturing complex nonlinear causal relationship between inputs and output of a numerical system, which has been validated by a number of functions approximation experiments [38].

4. Analysis and discussion

As detailed in Section 3.1, the decomposition of the input space is implemented based on the referential-value-based discretization [29,38]. Continuous data from an input space are discretized using referential values, whereby evidence is generated for inference. For example, as presented in Fig. 2, a three-dimensional input space: $x_{(1)} \times x_{(2)} \times x_{(3)}$ can be decomposed into $2 \times 2 \times 2 = 8$ cubic local regions using three referential values (including those at the minima and maxima of input variables, e.g., $A_{1,3}$ and $A_{3,3}$) for each input variable (signified by an axis of the plot, e.g., the axis of $x_{(2)}$). Each observation of an instance (x_n) of a data set, which is denoted by a data point in Fig. 2, lies in between two adjacent referential values of an input variable. A piece of evidence is directly acquired from a referential value using statistical analysis: sample casting and likelihoods normalization as shown in Section 3.1, which requires no assumptions about specific statistical input distributions and input-output relationships.

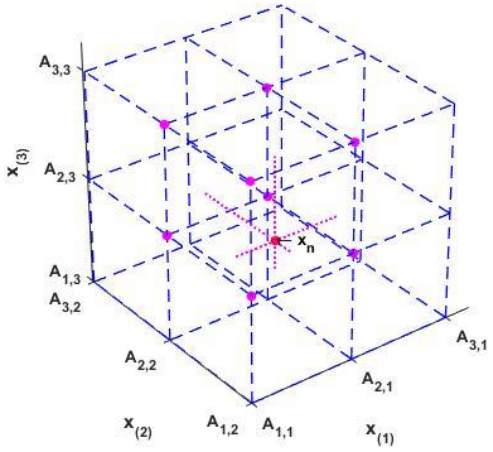

Fig. 2. Decomposition of three-dimensional input space

Data point (x_n) in Fig. 2, which represents an instance of a data set, can be located within a local cubic region determined by the magenta points (at the intersections of dotted lines) that denote referential values combinations. Thus, in order to produce the predicted probabilities for an instance, it is necessary to generate the probabilities of classes of output variable for these referential values combinations. Namely, we need to

generate the probabilities for the consequences of belief rules located at the magenta points. Using the conjunctive MAKER rule (Section 3.3), we can combine multiple pieces of evidence relating to a magenta point to generate associated belief rules, while considering interrelationship between a pair of evidence to be combined. This allows us to determine the probabilities of belief rules at the magenta points. All such belief rules located at magenta points in an input space constitute a BRB or a MAKER-based model for inference and prediction. The BRB is used to further generate predicted probabilities of classes of output variable for an instance by combining the activated belief rules at the magenta points using the MAKER rule. In the process to generate predicted probabilities, the matching degree of an instance matching belief rules (shown in Eq. 19) is used to measure the proximity of a data point in Fig. 2 to the magenta points (combinations of referential values). Thus, based on referential values and matching degrees, a complete description is provided for the relative location of a data point in an input space, which represents an instance of a data set. Following this, MAKER-based model parameters can be trained via a machine learning algorithm to minimize the MSE, thus reducing the difference between predicted and observed probability of a proposition being true.

Under the above-described structure, a MAKER-based model is essentially an approximator combining decomposed submodels denoted by local regions to describe the general pattern of a numerical system [41]. For each submodel (i.e., a local region), we can formulate an explicit input-output relationship [41]. As such, the MAKER-based model established by the proposed probabilistic modeling approach features a unique strong interpretability, which is specified in the following aspects.

(1) Evidence acquisition is interpretable. Evidence is directly acquired from referential values of continuous data by statistical analysis including sample casting and likelihoods normalization. Under the MAKER framework, we can combine multiple pieces of evidence, and capture interdependence between a pair of evidence. The capture of interdependence is achieved by using interdependence index based on marginal and joint likelihood functions, rather than assuming interdependence between a pair of evidence.

(2) Inference mechanism is interpretable. An instance of a continuous classification data set can activate multiple pieces of evidence from different input variables. It is

highly necessary to combine activated evidence to generate belief rules reflecting actual information that an instance contains. From a BRB based on belief rules, each given instance is able to activate belief rules, which can be further combined to generate a predicted output distribution. As we can record how changes in input variables influence output variable, the BRB inference process is essentially transparent. The BRB inference guarantees that the MAKER-based models are totally transparent and interpretable.

(3) Parameters determination is interpretable. The parameters of MAKER-based models consist of referential values of input variables and reliabilities (weights) of evidential elements. Both of them can be trained based on a machine learning algorithm to make difference between predicted and observed probability for a proposition being true as small as possible.

The above specific definitions provide a clear path for understanding and evaluating the interpretability under the context of machine learning. They may be further developed to guide the way to improve model interpretability.

5. Experimental study

Classification experiments for the MAKER-based model and other models are carried out on the classification data sets including *Banana, Haberman's survival,* and *Iris* data set (these data sets are downloaded from the website of KEEL: https://sci2s.ugr.es/keel/category.php?cat=clas). These benchmark data sets are useful to validate the MAKER-based model constructed by the proposed modeling approach, while a larger number of data sets will be included in the classification experiments of further research. The associated performance comparative analysis is conducted to compare the performance of these models in these experiments. Each of the classification data sets has already been divided into five subsets using distribution optimally balanced stratified cross-validation [42]. This ensures all the subsets have a similar class distribution, resembling that of the entire data set. One of the five subsets can be retained as a test set, and the remaining subsets are used as a training set. Such a process can be repeated for five times (folds), with each of the five subsets used exactly

once as the test set. Thus, each classification data set is partitioned into five folds of training and test sets for cross-validation.

In the comparative analysis, on one side of the comparison is the MAKER-based model constructed using the proposed probabilistic approach; on the other side are the conventional alternative models (displayed in Table 9), which consist of several groups of candidate submodels: decision tree, discriminant analysis, logistic regression, support vector machine (SVM), k-nearest neighbor (KNN), ensembles, and naïve Bayes. The candidate submodels of MAKER-based and conventional alternative models are trained based on each training set. Within each group of candidate submodels, the submodel with the highest average training accuracy is chosen as the group representative model. The training of candidate submodels of conventional alternative models is implemented in the application of "Classification Learner" in Matlab. The parameters used for the training of these models are the default parameters of the application. For example, regarding the submodel of "Complex Tree", the maximum number of splits is 100, and the split criterion is Gini's diversity index. In terms of "Quadratic Discriminant", the associated regularization is diagonal covariance. With regards to "Fine Gaussian SVM", the box constraint level is 1, and the manual kernel scale is 0.5. In the respect of "Weighted KNN", the associated number of neighbors is 10, and the distance metric is Euclidean, and the distance weight is squared inverse. The testing of candidate submodels of conventional alternative models and the visualization of the testing results are implemented by the built-in functions of Matlab. The training and testing of MAKER-based submodels are implemented by self-developed codes in Matlab. Of note is that following the stopping criteria proposed by Yao [25], the training for the MAKER-based submodels continues until each input variable of the training sets contains five trained referential values. This allows for a balance between the complexity and accuracy of the model. The representative MAKER-based submodel for the Banana data set has five trained referential values for each input variable, while those of the *Haberman's survival* and *Iris* data set have one trained referential value for each input variable.

Table 9

675 The candidate submodels of conventional alternative models for the classification of the
676 data sets

Alternative models	Candidate submodels	Group representative models
Decision tree	Simple tree, medium tree, and complex tree	Complex tree
Discriminant analysis	Linear discriminant, and quadratic discriminant	Quadratic discriminant
Logistic regression	Logistic regression	Logistic regression
Support vector machine (SVM)	Linear SVM, quadratic SVM, cubic SVM, fine Gaussian SVM, medium Gaussian SVM, and coarse Gaussian SVM	Fine Gaussian SVM
K-nearest neighbor (KNN)	Fine KNN, medium KNN, coarse KNN, cosine KNN, cubic KNN, and weighted KNN	Fine KNN, and weighted KNN
Ensembles	Booted trees, bagged trees subspace discriminant, subspace KNN, and RUSBoosted trees	Bagged trees, and subspace KNN
Naïve Bayes	Naïve Bayes	Naïve Bayes

677

678 All the group representative models are then tested based on each test set, and the
679 predicted outcomes are generated in the form of probabilities or scores. The predicted
680 outcomes are then used to generate the area under the receiver operating curve
681 (AUROC), which is subsequently employed for the comparison of the classification
682 models. AUROC is one of the most commonly used global index for classifiers
683 evaluations [43]. Although accuracy is employed widely to compare the predictive

capability of different classifiers, it completely ignores probability estimations of classification that most classifiers generate [44]. AUROC is argued to be an improved measure, whereby higher values indicate greater classification capabilities (1.0 is optimum) [44]. A general rule of thumb for using AUROC to judge the classification capability of a classifier [45,46] is that an AUROC between 0.7 and 0.8 is considered acceptable, between 0.8 and 0.9 indicates excellent discrimination, and larger than 0.9 implies outstanding discrimination.

Tables 10-12 report the AUROCs associated with each representative model for the *Banana*, *Haberman's survival*, and *Iris* data set, respectively, as well as the associated average AUROCs of the models across the five test sets. A comparison of the receiver operating characteristics (ROC) curves of the MAKER-based models with those of the optimum conventional models (in terms of AUROC) are presented in the supplementary material.

Table 10

The AUROCs of alternative models for the *Banana* data set

Models/Measures	AUROC					
	Test1	Test2	Test3	Test4	Test5	Avg.
MAKER	0.95158	0.96072	0.96254	0.96060	0.96296	**0.95968**
Complex tree	0.93494	0.93956	0.93735	0.94741	0.94917	0.94169
Quadratic discriminant	0.64853	0.64710	0.65017	0.65160	0.65386	0.65025
Logistic regression	0.54892	0.54909	0.54775	0.54950	0.55027	0.54911
Fine Gaussian SVM	0.94302	0.95581	0.96107	0.95938	0.96611	0.95708
Fine KNN	0.87788	0.89556	0.89293	0.86254	0.86118	0.87802
Weighted KNN	0.95471	0.96592	0.96757	0.95900	0.96363	0.96217
Ensemble: bagged trees	0.94867	0.96139	0.96088	0.95910	0.96396	0.95880
Ensemble: subspace KNN	0.63064	0.60058	0.62289	0.60923	0.62424	0.61752
Naive Bayes	0.66185	0.66148	0.66561	0.66770	0.67017	0.66536

The average AUROC of the MAKER-based model across the five test sets is 0.95968, which is the second largest one among all the AUROCs of the alternative models for the *Banana* data set (Table 5). The weighted KNN model achieves the optimum AUROC for this data set. In addition, both the logistic regression and naïve Bayes model are capable of being interpreted, their average AUROCs are much lower than that of the MAKER-based models. Furthermore, the complex tree model has a slightly lower average AUROC than the MAKER-based model. This indicates that for the complex *Banana* data set, simple interpretable models (e.g., logistic regression and naïve Bayes model) are unable to perform as well as their complex counterparts (e.g., complex tree and MAKER-based model).

Table 11

The AUROCs of alternative models for the *Haberman's survival* data set

Classifiers/Measures	AUROC					
	Test1	Test2	Test3	Test4	Test5	Avg.
MAKER	0.61046	0.77778	0.74028	0.65139	0.67292	**0.69057**
Complex tree	0.53464	0.62361	0.59931	0.50069	0.56597	0.56484
Quadratic discriminant	0.71634	0.73333	0.62222	0.68750	0.80069	0.71202
Logistic regression	0.67843	0.71806	0.64583	0.63750	0.73542	0.68305
Fine Gaussian SVM	0.71503	0.65694	0.56528	0.71528	0.71736	0.67398
Fine KNN	0.59869	0.57639	0.56528	0.62986	0.56528	0.58710
Weighted KNN	0.69673	0.64236	0.62778	0.69306	0.77569	0.68712
Ensemble: bagged trees	0.73464	0.67014	0.62361	0.66806	0.68819	0.67693
Ensemble: subspace KNN	0.55948	0.66806	0.63611	0.62500	0.58333	0.61440
Naive Bayes	0.69542	0.70139	0.62222	0.58333	0.69097	0.65867

The average AUROC of the MAKER-based model for the *Haberman's survival* data set is 0.69057, again reaching the second place amongst all models in terms of AUROCs (Table 6). Based on these AUROCs, the classification performance of the MAKER-based model is considered acceptable. Note that the *Haberman's survival* data

set is imbalanced, where the ratio of the number of positive to negative samples is approximately 1:3. This can have an impact on the classification results. Moreover, the AUROC of MAKER-based model surpasses that of the complex tree and logistic regression model. Results demonstrate the acceptable classification performance of the MAKER-based model for the *Haberman's survival* data set.

Table 12

The AUROCs of alternative models for the *Iris* data set

Classifiers/Measures	AUROC					
	Test1	Test2	Test3	Test4	Test5	Avg.
MAKER	0.99000	1.00000	0.99250	1.00000	0.99500	**0.99550**
Complex tree	0.90000	0.97500	0.97250	0.97500	0.91750	0.94800
Quadratic discriminant	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000
Fine Gaussian SVM	0.99500	0.99000	0.99500	1.00000	0.95000	0.98600
Fine KNN	0.95000	0.97500	0.95000	0.95000	0.87500	0.94000
Weighted KNN	1.00000	0.99500	1.00000	1.00000	0.98500	0.99600
Ensemble: bagged trees	0.99500	1.00000	0.99500	1.00000	0.96500	0.99100
Ensemble: subspace KNN	1.00000	1.00000	0.99250	0.98750	0.97250	0.99050
Naive Bayes	0.99500	0.99500	0.99500	0.99000	0.99000	0.99300

In order to determine the ROC curves and AUROCs of each model for the classification of the *Iris* data set, *Iris Versicolor* was taken as the positive class, while *Iris Setosa* and *Iris Virginica* are combined as the negative class. Table 7 indicates an average AUROC of 0.9955 for the MAKER-based model, which is the third largest one among all the AUROCs for the *Iris* data set. This indicates the outstanding classification performance of the MAKER-based model for the *Iris* data set.

The AUROCs in Tables 5-7 indicate that the MAKER-based model is an outstanding classifier for the *Banana* and *Haberman's survival* data set, and a generally acceptable one for the *Iris* data set. In addition, it generally performs better than other interpretable models such as complex tree, logistic regression, and naïve Bayes.

However, higher computational complexity is involved in the interpretable MAKER-based models constructed by the proposed approach, as there is a high multiplicative complexity on the number of referential values of input variables in a BRB [47]. It is necessary to conduct further research to improve the training efficiency of the MAKER-based models.

6. Conclusions

This paper presents a new probabilistic modeling approach to conduct a MAKER-based classifier for interpretable inference and classification. A comparative analysis is conducted between the MAKER-based model built by the proposed modeling approach and conventional alternative ones to evaluate their classification performance on the *Banana*, *Haberman's survival*, and *Iris* data set. Experimental results demonstrate the general robustness of the MAKER-based model in classifying the data sets. For example, AUROCs of 0.95968, 0.69507, and 0.99550 were determined for the *Banana*, *Haberman's survival*, and *Iris* data set. The lower value associated with the *Haberman's survival* data set may be attributed to the lack of balance between negative and positive samples of the data set.

Furthermore, the MAKER-based model is characterized by a unique strong interpretability, which is specified in three aspects: (1) interpretable evidence acquisition, (2) interpretable inference mechanism, and (3) interpretable parameters determination. This provides a clear definition of "interpretability" under the context of machine learning. The proposed probabilistic modeling approach has a great potential in solving different types of modeling and prediction problems in complex systems. However, further research is necessary for handling high multiplicative complexity of referential values numbers of input variables in a BRB [47], and dealing with the relatively poor sensitivity for classification of imbalanced data sets (e.g., *Haberman's survival* data set), and establishing MAKER-based models based on the data sets with "unknown" class.

References

[1] W.J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, B. Yu, Interpretable machine learning: definitions, methods, and applications, PNAS. (2019). https://doi.org/10.1073/pnas.1900654116.

[2] M. Du, N. Liu, X. Hu, Techniques for Interpretable Machine Learning, Commun. ACM. 63 (2018) 68–77. https://doi.org/10.1145/3359786.

[3] F. Doshi-Velez, B. Kim, Considerations for Evaluation and Generalization in Interpretable Machine Learning, in: Explain. Interpret. Model. Comput. Vis. Mach. Learn., Springer, Cham, 2018: pp. 3–17. https://doi.org/10.1007/978-3-319-98131-4_1.

[4] N. Papernot, P. McDaniel, Deep k-Nearest Neighbors: Towards Confident, Interpretable and Robust Deep Learning, (2018). http://arxiv.org/abs/1803.04765 (accessed January 21, 2020).

[5] C. Molnar, Interpretable machine learning, Lulu.com, 2019.

[6] M.A. Ahmad, A. Teredesai, C. Eckert, Interpretable machine learning in healthcare, in: Proc. - 2018 IEEE Int. Conf. Healthc. Informatics, ICHI 2018, 2018. https://doi.org/10.1109/ICHI.2018.00095.

[7] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, N. Elhadad, Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission, in: Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., 2015. https://doi.org/10.1145/2783258.2788613.

[8] S. Kaufman, S. Rosset, C. Perlich, O. Stitelman, Leakage in data mining: Formulation, detection, and avoidance, in: ACM Trans. Knowl. Discov. Data, 2012. https://doi.org/10.1145/2382577.2382579.

[9] W. Caicedo-Torres, J. Gutierrez, ISeeU: Visually interpretable deep learning for mortality prediction inside the ICU, J. Biomed. Inform. (2019). https://doi.org/10.1016/j.jbi.2019.103269.

[10] F. Doshi-Velez, B. Kim, Towards A Rigorous Science of Interpretable Machine Learning, (2017). http://arxiv.org/abs/1702.08608 (accessed January 21, 2020).

[11] C. Fan, F. Xiao, C. Yan, C. Liu, Z. Li, J. Wang, A novel methodology to explain and evaluate data-driven building energy performance models based on interpretable machine learning, Appl. Energy. (2019). https://doi.org/10.1016/j.apenergy.2018.11.081.

[12] H. Lakkaraju, S.H. Bach, J. Leskovec, Interpretable decision sets: A joint framework for description and prediction, in: Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., 2016. https://doi.org/10.1145/2939672.2939874.

[13] W. Samek, T. Wiegand, K.-R. Müller, Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models, ITU J. ICT Discov. - Spec. Issue 1 - Impact Artif. Intell. Commun. Networks Serv. 1 (2017) 1–10.

[14] S. Pereira, R. Meier, R. McKinley, R. Wiest, V. Alves, C.A. Silva, M. Reyes, Enhancing interpretability of automatically extracted machine learning features: application to a RBM-Random Forest system on brain lesion segmentation, Med. Image Anal. (2018). https://doi.org/10.1016/j.media.2017.12.009.

[15] M.D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), 2014. https://doi.org/10.1007/978-3-319-10590-1_53.

[16] D. Nauck, R. Kruse, Obtaining interpretable fuzzy classification rules from medical data, Artif. Intell. Med. (1999). https://doi.org/10.1016/S0933-3657(98)00070-0.

[17] J.B. Yang, D.L. Xu, Inferential modelling and decision making with data, in: ICAC 2017 - 2017 23rd IEEE Int. Conf. Autom. Comput. Addressing Glob. Challenges through Autom. Comput., 2017. https://doi.org/10.23919/IConAC.2017.8082048.

[18] J.B. Yang, J. Liu, J. Wang, H.S. Sii, H.W. Wang, Belief rule-base inference methodology using the evidential reasoning approach - RIMER, IEEE Trans. Syst. Man, Cybern. Part ASystems Humans. (2006). https://doi.org/10.1109/TSMCA.2005.851270.

[19] G. Kong, D.L. Xu, R. Body, J.B. Yang, K. MacKway-Jones, S. Carley, A belief rule-based decision support system for clinical risk assessment of cardiac chest pain, Eur. J. Oper. Res. (2012). https://doi.org/10.1016/j.ejor.2011.10.044.

[20] G. Kong, D.L. Xu, J.B. Yang, X. Yin, T. Wang, B. Jiang, Y. Hu, Belief rule-based inference for predicting trauma outcome, Knowledge-Based Syst. (2016). https://doi.org/10.1016/j.knosys.2015.12.002.

[21] L.H. Yang, F.F. Ye, Y.M. Wang, Ensemble belief rule base modeling with

diverse attribute selection and cautious conjunctive rule for classification problems, Expert Syst. Appl. (2020). https://doi.org/10.1016/j.eswa.2019.113161.

[22] A.P. Dempster, Upper and Lower Probabilities Induced by a Multivalued Mapping, Ann. Math. Stat. (1967). https://doi.org/10.1214/aoms/1177698950.

[23] A.P. Dempster, A GENERALIZATION OF BAYESIAN INFERENCE, J. R. Stat. Soc. Ser. B. 30(2) (1968) 205–232.

[24] G. Shafer, A Mathematical Theory of Evidence, illustrate, 1976.

[25] J.B. Yang, D.L. Xu, Evidential reasoning rule for evidence combination, Artif. Intell. (2013). https://doi.org/10.1016/j.artint.2013.09.003.

[26] J.B. Yang, D.L. Xu, Evidential reasoning rule for evidence combination, Artif. Intell. 205 (2013) 1–29. https://doi.org/10.1016/j.artint.2013.09.003.

[27] L. Chang, Z. Zhou, Y. You, L. Yang, Z. Zhou, Belief rule based expert system for classification problems with new rule activation and weight calculation procedures, Inf. Sci. (Ny). (2016). https://doi.org/10.1016/j.ins.2015.12.009.

[28] L. Jiao, Q. Pan, T. Denœux, Y. Liang, X. Feng, Belief rule-based classification system: Extension of FRBCS in belief functions framework, Inf. Sci. (Ny). (2015). https://doi.org/10.1016/j.ins.2015.03.005.

[29] X. Xu, J. Zheng, J. bo Yang, D. ling Xu, Y. wang Chen, Data classification using evidence reasoning rule, Knowledge-Based Syst. (2017). https://doi.org/10.1016/j.knosys.2016.11.001.

[30] Z.G. Zhou, F. Liu, L.C. Jiao, Z.J. Zhou, J.B. Yang, M.G. Gong, X.P. Zhang, A bi-level belief rule based decision support system for diagnosis of lymph node metastasis in gastric cancer, Knowledge-Based Syst. (2013). https://doi.org/10.1016/j.knosys.2013.09.001.

[31] Z.G. Zhou, F. Liu, L.L. Li, L.C. Jiao, Z.J. Zhou, J.B. Yang, Z.L. Wang, A cooperative belief rule based decision support system for lymph node metastasis diagnosis in gastric cancer, Knowledge-Based Syst. (2015). https://doi.org/10.1016/j.knosys.2015.04.019.

[32] M. Du, N. Liu, Q. Song, X. Hu, Towards explanation of DNN-based prediction with guided feature inversion, in: Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., 2018. https://doi.org/10.1145/3219819.3220099.

[33] O. Biran, K. McKeown, Human-centric justification of machine learning

predictions, in: IJCAI Int. Jt. Conf. Artif. Intell., 2017. https://doi.org/10.24963/ijcai.2017/202.

[34] L.A. Hendricks, Z. Akata, M. Rohrbach, J. Donahue, B. Schiele, T. Darrell, Generating visual explanations, in: Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), 2016. https://doi.org/10.1007/978-3-319-46493-0_1.

[35] R. Tkachenko, I. Izonin, P. Vitynskyi, N. Lotoshynska, O. Pavlyuk, Development of the non-iterative supervised learning predictor based on the ito decomposition and sgtm neural-like structure for managing medical insurance costs, Data. (2018). https://doi.org/10.3390/data3040046.

[36] I. Izonin, R. Tkachenko, N. Kryvinska, P. Tkachenko, M. Greguš ml, Multiple Linear Regression Based on Coefficients Identification Using Non-iterative SGTM Neural-like Structure, in: Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), 2019. https://doi.org/10.1007/978-3-030-20521-8_39.

[37] J.B. Yang, Rule and utility based evidential reasoning approach for multiattribute decision analysis under uncertainties, Eur. J. Oper. Res. 131 (2001) 31–61. https://doi.org/10.1016/S0377-2217(99)00441-5.

[38] S. Yao, Investigation into Rule-based Inferential Modelling and Prediction with Application in Healthcare, University of Manchester, 2019. https://www.research.manchester.ac.uk/portal/en/theses/investigation-into-rulebased-inferential-modelling-and-prediction-with-application-in-healthcare(e73ae49a-887e-4305-8973-728c1bbe251e).html (accessed January 25, 2020).

[39] J.B. Yang, D.L. Xu, A study on generalising bayesian inference to evidential reasoning, Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics). (2014).

[40] M. Gen, R. Cheng, Genetic Algorithms and Engineering Design, John Wiley & Sons, Inc., Hoboken, NJ, USA, 1997. https://doi.org/10.1002/9780470172254.

[41] Y.W. Chen, J.B. Yang, D.L. Xu, S.L. Yang, On the inference and approximation properties of belief rule based systems, Inf. Sci. (Ny). (2013). https://doi.org/10.1016/j.ins.2013.01.022.

[42] J. Alcalá-Fdez, A. Fernández, J. Luengo, J. Derrac, S. García, L. Sánchez, F.

898 Herrera, KEEL data-mining software tool: Data set repository, integration of
899 algorithms and experimental analysis framework, J. Mult. Log. Soft Comput.
900 (2011).
901 [43] D. Faraggi, B. Reiser, Estimation of the area under the ROC curve, Stat. Med.
902 (2002). https://doi.org/10.1002/sim.1228.
903 [44] C.X. Ling, J. Huang, H. Zhang, AUC: A better measure than accuracy in
904 comparing learning algorithms, in: Lect. Notes Comput. Sci. (Including Subser.
905 Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), 2003.
906 https://doi.org/10.1007/3-540-44886-1_25.
907 [45] M. Smithson, E.C. Merkle, Generalized Linear Models for Categorical and
908 Continuous Limited Dependent Variables, Chapman and Hall/CRC, 2013.
909 https://doi.org/10.1201/b15694.
910 [46] D.W. Hosmer, S. Lemeshow, Applied logistic regression, 2nd ed, Wiley, New
911 York, 2000.
912 [47] Y. Chen, Y.W. Chen, X. Bin Xu, C.C. Pan, J.B. Yang, G.K. Yang, A data-driven
913 approximate causal inference model using the evidential reasoning rule,
914 Knowledge-Based Syst. (2015). https://doi.org/10.1016/j.knosys.2015.07.026.
915