Transformation-Based Fuzzy Rule Interpolation with Mahalanobis Distance Measures Supported by Choquet Integral

Mou Zhou, Changjing Shang, Guobin Li, Liang Shen, Nitin Naik, Shangzhu Jin, Jun Peng and Qiang Shen

***Abstract*—Fuzzy rule interpolation (FRI) strongly supports approximate inference when a new observation matches no rules, through selecting and subsequently interpolating appropriate rules close to the observation from the given (sparse) rule base. Traditional ways of implementing the critical rule selection process are typically based on the exploitation of Euclidean distances between the observation and rules. It is conceptually straightforward for implementation but applying this distance metric may systematically lead to inferior results because it fails to reflect the variations of the relevance or significance levels amongst different domain features. To address this important issue, a novel transformation-based FRI approach is presented, on the basis of utilising the Mahalanobis distance metric. The new FRI method works by transforming a given sparse rule base into a coordinates system where the distance between instances of the same category becomes closer while that between different categories becomes further apart. In so doing, when an observation is present that matches no rules, the most relevant neighbouring rules to implement the required interpolation are more likely to be selected. Following this, the scale and move factors within the classical transformation-based FRI procedure are also modified by Choquet integral. Systematic experimental investigation over a range of classification problems demonstrates that the proposed approach remarkably outperforms the existing state-of-the-art FRI methods in both accuracy and efficiency.**

***Index Terms*—Fuzzy rule interpolation, transformation-based FRI, approximate inference, Mahalanobis distance, Choquet integral.**

I. Introduction

THANKS to the capability of performing approximate inference with a sparse rule base, fuzzy rule interpolation (FRI) [1] greatly expands the scope of applications of the classical compositional rule of inference (CRI) [2], which would otherwise collapse when an observation does not match any rule antecedent from the rule base. FRI has gained considerable developments for the past two decades. Its core working principle is to implement linear interpolative reasoning by manipulating selected rules that flank the unmatched observation, or to perform extrapolative reasoning if the antecedent variables of certain neighbouring fuzzy rules do not flank the observation [3].

In general, many rules in a given (sparse) rule base may be used to implement interpolative or extrapolative inferences. However, (at least) because of computational complexity, it is not advisable to make use of all the rules from the rule base for interpolation (or extrapolation, but for presentational simplicity only interpolation is hereafter referred to given the mathematical dual form of both types of reasoning). Therefore, just close neighbouring rules to the observation are employed to participate in interpolative reasoning for most FRI methods. The rationale for using neighbouring rules is their similarity to the observation. Whilst attempt exists to automatically select rules for interpolation without manually setting the number of closest rules [4], a great majority of FRI approaches [5]–[15] typically utilise Euclidean distances between an unmatched observation and the given rule base to select such closest rules. The procedure of rule selection plays very critical roles in the subsequent inferential process, in the vicinity of the observation. Although the use of Euclidean distance metric is classical and conceptually straightforward to implement, its employment can lead to utilising an inferior subset of rules when multiple antecedent variables have different levels of, or weights on, contribution to the reasoning outcomes [1].

Feature selection tools [16] can help a reasoning system learn different weighting scores of antecedent variables automatically. In particular, the potential of feature evaluation has been exploited through integrating such a tool within FRI [17], [18]. These applications have revealed that by considering degrees of relative feature importance in calculating distances from an observation to the rules is useful for the system to find the most relevant rules to perform FRI. The resulting weighted FRI methods can attain better performance in addressing classification problems than their original unweighted ones. Whilst effective, this type of weighted approach may become void for problems where there is no clear distinction of importance between the features. In addition, the use of weighting scores assigned to every domain feature for distance calculation inevitably increases computational complexity. Such techniques do not eliminate any coupling between features, which means that after feature evaluation, variables may remain interrelated in the feature space, thereby (adversely) retaining redundant information [19].

Unlike Euclidean distance, Mahalanobis distance [20] is a distance measure that incorporates the dealing of correlations

This work was supported in part by the Strategic Partner Acceleration Award (80761-AU201), funded under the Sêr Cymru II programme, UK.

M. Zhou, C. Shang (*Corresponding Author*), G. Li and Q. Shen are with Department of Computer Science, Aberystwyth University, Aberystwyth SY23 3DB, U.K. (e-mail: {moz3, cns, gul12, qqs}@aber.ac.uk).

L. Shen is with School of Information Engineering, Fujian Business University, Fuzhou 350506, China. (e-mail: liang.shen.18@fjbu.edu.cn).

N. Naik is with School of Engineering and Applied Science, Aston University, Birmingham B4 7ET, UK. (e-mail: n.naik1@aston.ac.uk).

M. Zhou, S. Jin and J. Peng are with School of Intelligent Technology and Engineering, Chongqing University of Science and Technology, Chongqing 401331, China (e-mail: {zhoumou, szjin, jpeng}@cqust.edu.cn).

between features. It was originally proposed to introduce a form of covariance measure that considers the distribution of data in a given feature space. However, in employing it to cope with classification problems, it has been recognised that a distance metric that examines just the internal relationship of features is not sufficient. A desirable metric should also reflect the relationships between domain attributes and data labels. This has led to the development of the modern Mahalanobis distance metric that possesses an inherent learning capacity, via so-called Mahalanobis metric learning [21]. The learning process aims to acquire a Mahalanobis matrix $\mathcal{M}$ to transform data samples to a new coordinates system where domain features from the original feature space are reconstructed. In the resulting feature space, data distribution becomes significantly more distinct, facilitating classification, because instances of the same category are gathered together while instances with different labels are separated far apart. Inspired by this observation, a novel FRI method is introduced herein by the use of Mahalanobis distance with metric learning that helps select more suitable rules involving weighted attributes to perform FRI.

As a popular approach, scale and move transformation-based FRI (T-FRI) [7] has exceptional merits of yielding unique, normal and convex interpolative consequences. It enables the utilisation of sophisticated fuzzy representations, such as complex polygon, Gaussian or other bell-shaped fuzzy membership functions as well as simple triangular or trapizoidal ones. Therefore, T-FRI is employed in this work to serve as the underlying FRI platform. In addition, Choquet integral, a particular non-additive aggregation function, is adopted in an effort to support an effective integration of weights into the T-FRI inference process. Furthermore, a systematic experimental evaluation of five distinct metric learning algorithms is conducted to ensure that the proposed approach does not rely on a certain particular metric learning technique. The results of running over a wide range of classification problems (in comparison with state-of-the-art T-FRI methods) demonstrate that the approach presented herein substantially facilitates the improvement of the underlying T-FRI.

The paper is organised as follows. Section II outlines the relevant background of T-FRI and the basic ideas of Mahalanobis distance. Section III describes the proposed framework of Mahalanobis distance-supported T-FRI with metric learning and provides a theoretical analysis of the proposed approach. Section IV discusses the results of comparative experimental studies. Section V concludes the paper and suggests future enhancements.

II. Background

This section presents the relevant background work, including an outline of FRI based on scale and move transformations and a brief description of Mahalanobis distance as well as Mahalanobis metric learning.

A. Transformation-Based FRI (T-FRI)

Without losing generality, suppose that a fuzzy rule base with multiple multi-antecedent rules is expressed as follows:

$$\begin{aligned} &\text{Rule } R_i: \\ &\text{If } x_1 \text{ is } A_{i1} \text{ and } x_2 \text{ is } A_{i2} \text{ and } \cdots \text{ and } x_m \text{ is } A_{im}, \\ &\qquad\qquad \text{then } y \text{ is } B_i \end{aligned} \tag{1}$$

where $i = 1, 2, \ldots, N$ with N being the number of rules for this rule base; R_i is the ith rule; m is the number of antecedent attributes; A_{ij} and B_i represent the value of the jth ($j \in [1, m]$) antecedent variable and that of the consequent in R_i, respectively, each defined by a fuzzy set. An observation (or input) for this fuzzy reasoning system is given by

$$\text{Observation } O^*: A_1^*, A_2^*, \cdots, A_j^*, \cdots, A_m^* \tag{2}$$

where A_j^* denotes the fuzzy set of the jth antecedent variable.

As an important notion in T-FRI, the representative value (Rep) of a fuzzy set is widely used to guide fuzzy interpolative reasoning. The Rep value of a fuzzy set reflects the essential information embedded within both the overall location of its domain range and the geometric shape of its membership function. For instance, the general form of Rep for an arbitrary polygonal fuzzy set $A = (a_1, a_2, \ldots, a_n)$ is defined by

$$\text{Rep}(A) = \sum_{t=1}^{n} w_t a_t \tag{3}$$

where $a_t, t = 1, 2, \ldots, n$ are the abscissas of vertices depicting the polygonal with their ordinates defining the membership values, and w_t denotes the weight assigned to a_t.

The triangular membership function is very popular in encoding fuzzy sets within fuzzy systems owing to its computational simplicity. The abscissas of the three vertices for a fuzzy triangular membership function A are a_1, a_2, a_3, and the representative value of such a fuzzy set can be defined as

$$\text{Rep}(A) = \frac{a_1 + a_2 + a_3}{3} \tag{4}$$

with the $w_t, t = 1, 2, 3$ all being set to $\frac{1}{3}$.

Given the above preliminaries, when an observation does not match any of the rules from the given rule base, T-FRI performs interpolative inference through four core procedures as graphically illustrated in Fig. 1 and outlined below. Note that, for simple depiction, the fuzzy rule base is portrayed in this figure by a small number of points (with different shapes representing different consequent classes) and projected onto a two-dimensional space.

1) Neighbouring Rule Selection: Considering computation efficacy, not all the rules in the rule base are necessarily taken for taking part in interpolation. As the first step of T-FRI, n closest or nearest neighbouring rules to an unmatched observation O^* are selected from the rule base, which have the n smallest distances to O^*. Specifically, while computing the distance from O^* to $R_i, i = 1, 2, \ldots, N$, the distance between the value pair of each relevant antecedent variable (per observation) is defined by

$$d(A_j^*, A_{ij}) = \frac{|\text{Rep}(A_j^*) - \text{Rep}(A_{ij})|}{range_j} \tag{5}$$

This article has been accepted for publication in IEEE Transactions on Fuzzy Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TFUZZ.2022.3194368

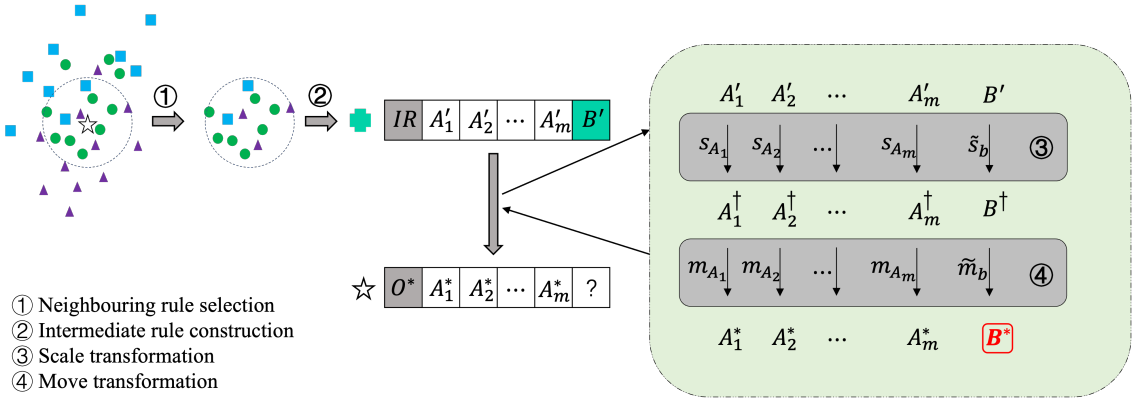

Fig. 1. Framework of T-FRI.

where $range_j = \max_{A_j} - \min_{A_j}, j = 1, 2, \ldots, m$ represents the domain range of the jth antecedent feature. Then, the distance between O^* and R_i is formulated by

$$d(O^*, R_i) = \sqrt{\sum_{j=1}^{m} d(A_j^*, A_{ij})^2}. \tag{6}$$

Based on this, the n closest rules having the least distance measurements with regard to O^* are selected to be employed in the next step.

2) Intermediate Rule Construction: This step is concerned with the process of constructing a required intermediate rule, comprised of both antecedent and consequent parts, mimicking the general representation format of the rules in the given rule base. This is implemented through the use of the principle of analogical reasoning [7], which basically states that if there exists a certain degree of similarity between the values of antecedent variables A_j' and A_j^*, then the consequent parts B' and B^* should share the same similarity degree. This principle forms the intuitive justification not just for this step, but throughout the entire subsequent FRI procedures.

Let $w_{ij}, i = 1, 2, \ldots, n, j = 1, 2, \ldots, m$, represent the weight degree to which the jth antecedent variable of the ith rule contributes to building the jth antecedent fuzzy set A_j' of the intermediate rule. It is negatively related to the distance between A_j^* and A_{ij}

$$w_{ij} = \frac{1}{1 + d(A_j^*, A_{ij})} \tag{7}$$

where $d(A_j^*, A_{ij})$ is defined by Eqn. (5). To guarantee that the sum over attribute j equals to 1, this term needs to be normalised such that

$$\tilde{w}_{ij} = \frac{w_{ij}}{\sum_{t=1,2,\ldots,n} w_{tj}}. \tag{8}$$

Thus, the intermediate antecedent A_j'' is obtained by

$$A_j'' = \sum_{i=1,2,\ldots,n} \tilde{w}_{ij} A_{ij}. \tag{9}$$

Note that the fuzzy terms A_j'' calculated via Eqn. (9) are not the required values of the intermediate antecedent features since they do not have the same representative values as A_j^* from the observation. In order to ensure the consistency of these vital Rep values before and after transformation, new intermediate fuzzy antecedent values $A_j', j = 1, 2, \ldots, m$, are obtained by

$$A_j' = A_j'' + \delta_j range_j \tag{10}$$

where $range_j = \max_{A_j} - \min_{A_j}$ as with Eqn. (5) and δ_j is defined by

$$\delta_j = \frac{\text{Rep}(A_j^*) - \text{Rep}(A_j'')}{range_j}. \tag{11}$$

Similar to the antecedent part, the intermediate consequent part B' is then calculated by

$$B' = \sum_{i=1,2,\ldots,n} \tilde{w}_{ib} B_i + \tilde{\delta}_b range_B \tag{12}$$

where B_i is the consequent value of the ith rule; $range_B = \max_B - \min_B$; and the two significant factors $\tilde{w}_{ib}$ and $\tilde{\delta}_b$ are computed as follows:

$$\tilde{w}_{ib} = \frac{1}{m} \sum_{j=1}^{m} \tilde{w}_{ij}, \tag{13}$$

$$\tilde{\delta}_b = \frac{1}{m} \sum_{j=1}^{m} \delta_j \tag{14}$$

where $\tilde{w}_{ij}$ and δ_j are calculated from Eqn. (8) and Eqn. (11), respectively.

After the above two steps, the most similar n fuzzy rules to the observation are aggregated into a single intermediate rule (IR):

Intermediate Rule:

$$\text{If } x_1 \text{ is } A_1' \text{ and } x_2 \text{ is } A_2' \text{ and } \cdots \text{ and } x_m \text{ is } A_m', \quad \text{then } y \text{ is } B'. \tag{15}$$

The rest of the interpolative inference is again, based on the exploitation of the analogical reasoning. In order to achieve this, the subsequent inference procedures firstly transform the intermediate rule as per Eqn. (15) into a scaled intermediate rule comprised of $A_1^\dagger, A_2^\dagger, \ldots, A_m^\dagger$ and $B^\dagger$, where fuzzy terms $A_j^\dagger, j = 1, 2, \ldots, m$, and $B^\dagger$ denote the scaled intermediate fuzzy sets for the antecedent and the consequent part respectively. Secondly, the scaled intermediate rule is

This article has been accepted for publication in IEEE Transactions on Fuzzy Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TFUZZ.2022.3194368

further transformed into one that governs the relationships between the given observation in the form of Eqn. (2) and an interpolative consequent value B^*. These two steps are called scale and move transformations, controlled by two critical factors (namely, scale and move factors) that ensure the reasoning system attaining the similarity degree between A'_j and A^*_j.

3) Scale Transformation: Throughout this paper, following the mainstream T-FRI implementations, triangular membership functions are utilised to represent fuzzy sets (purely for computational simplicity). For an arbitrary fuzzy term $A'_j(a'_{j1}, a'_{j2}, a'_{j3})$ employed by the intermediate rule, the scale rate is computed by

$$s_{A_j} = \frac{a^*_{j3} - a^*_{j1}}{a'_{j3} - a'_{j1}}. \tag{16}$$

Applying s_{A_j} to an antecedent feature value, the corresponding scaled intermediate antecedent fuzzy set $A^\dagger_j(a^\dagger_{j1}, a^\dagger_{j2}, a^\dagger_{j3})$ is given by

$$\begin{bmatrix} a^\dagger_{j1} \\ a^\dagger_{j2} \\ a^\dagger_{j3} \end{bmatrix} = \frac{1}{3}\begin{bmatrix} 1+2s_{A_j} & 1-s_{A_j} & 1-s_{A_j} \\ 1-s_{A_j} & 1+2s_{A_j} & 1-s_{A_j} \\ 1-s_{A_j} & 1-s_{A_j} & 1+2s_{A_j} \end{bmatrix}\begin{bmatrix} a'_{j1} \\ a'_{j2} \\ a'_{j3} \end{bmatrix}. \tag{17}$$

In doing so, the representative values for every feature remain consistent throughout the transformation. Note that the definitions of s_{A_j} and the computation of $A^\dagger_j$ for other types of complex membership functions can be found in [7].

According to the aforementioned analogical reasoning principle, it is intuitive to acquire the consequent scaled intermediate fuzzy set $B^\dagger(b^\dagger_1, b^\dagger_2, b^\dagger_3)$ such that

$$\begin{bmatrix} b^\dagger_1 \\ b^\dagger_2 \\ b^\dagger_3 \end{bmatrix} = \frac{1}{3}\begin{bmatrix} 1+2\tilde{s}_b & 1-\tilde{s}_b & 1-\tilde{s}_b \\ 1-\tilde{s}_b & 1+2\tilde{s}_b & 1-\tilde{s}_b \\ 1-\tilde{s}_b & 1-\tilde{s}_b & 1+2\tilde{s}_b \end{bmatrix}\begin{bmatrix} b'_1 \\ b'_2 \\ b'_3 \end{bmatrix} \tag{18}$$

where $\tilde{s}_b$ is the average of s_{A_j}:

$$\tilde{s}_b = \frac{1}{m}\sum_{j=1}^{m} s_{A_j} \tag{19}$$

4) Move Transformation: Following scale transformation, this step strives to move $A^\dagger_j$ to the position that coincides with the position of the original observation A^*_j, and similarly shifts $B^\dagger$ to yield the desirable analogical reasoning outcome B^*. This procedure is accomplished by applying the following move ratio to $A^\dagger_j$:

$$m_{A_j} = \begin{cases} \frac{3(a^*_{j1} - a^\dagger_{j1})}{a^\dagger_{j2} - a^\dagger_{j1}}, & \text{if } a^*_{j1} \geq a^\dagger_{j1} \\ \frac{3(a^*_{j1} - a^\dagger_{j1})}{a^\dagger_{j3} - a^\dagger_{j2}}, & \text{otherwise;} \end{cases} \tag{20}$$

with a similar application to $B^\dagger$. Akin to scale transformation, the move rate $\tilde{m}_b$ for the consequent attribute is calculated by averaging those of the antecedent variables, such that

$$\tilde{m}_b = \frac{1}{m}\sum_{j=1}^{m} m_{A_j}. \tag{21}$$

Finally, the required interpolative reasoning consequence $B^*(b^*_1, b^*_2, b^*_3)$ is computed by

$$\begin{bmatrix} b^*_1 \\ b^*_2 \\ b^*_3 \end{bmatrix} = \begin{cases} \frac{1}{3}\begin{bmatrix} 3-\tilde{m}_b & \tilde{m}_b & 0 \\ 2\tilde{m}_b & 3-2\tilde{m}_b & 0 \\ -\tilde{m}_b & \tilde{m}_b & 3 \end{bmatrix}\begin{bmatrix} b^\dagger_1 \\ b^\dagger_2 \\ b^\dagger_3 \end{bmatrix}, & \text{if } \tilde{m}_b \geq 0 \\ \frac{1}{3}\begin{bmatrix} 3 & -\tilde{m}_b & \tilde{m}_b \\ 0 & 3+2\tilde{m}_b & -2\tilde{m}_b \\ 0 & -\tilde{m}_b & 3+\tilde{m}_b \end{bmatrix}\begin{bmatrix} b^\dagger_1 \\ b^\dagger_2 \\ b^\dagger_3 \end{bmatrix}, & \text{otherwise.} \end{cases} \tag{22}$$

B. Mahalanobis Distance and Metric Learning

This subsection presents the basic ideas of the Mahalanobis distance metric and introduces five typical metric learning methods, any of which may be utilised to support adapting Mahalanobis distance measures.

1) Mahalanobis Distance: Dissimilar to Euclidean distance, Mahalanobis distance [20] measures relationships between data instances of a given problem domain, by considering the correlation between features. Suppose that $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are two samples from a dataset. The calculation of Euclidean distance between them can be expressed by

$$d(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sqrt{(\boldsymbol{x}_1 - \boldsymbol{x}_2)^{\boldsymbol{T}}(\boldsymbol{x}_1 - \boldsymbol{x}_2)}. \tag{23}$$

If however, both are projected onto a new linear space through linear transformations ($\boldsymbol{x}_1 \mapsto \boldsymbol{\mathcal{A}}\boldsymbol{x}_1$ and $\boldsymbol{x}_2 \mapsto \boldsymbol{\mathcal{A}}\boldsymbol{x}_2$, where $\boldsymbol{\mathcal{A}}$ is a transformation matrix), then the distance in the new space becomes:

$$\begin{aligned} d(\boldsymbol{x_1}, \boldsymbol{x_2}) &= \sqrt{(\boldsymbol{\mathcal{A}}\boldsymbol{x}_1 - \boldsymbol{\mathcal{A}}\boldsymbol{x}_2)^{\boldsymbol{T}}(\boldsymbol{\mathcal{A}}\boldsymbol{x}_1 - \boldsymbol{\mathcal{A}}\boldsymbol{x}_2)} \\ &= \sqrt{(\boldsymbol{x}_1 - \boldsymbol{x}_2)^{\boldsymbol{T}}\boldsymbol{\mathcal{A}}^{\boldsymbol{T}}\boldsymbol{\mathcal{A}}(\boldsymbol{x}_1 - \boldsymbol{x}_2)} \\ &= \sqrt{(\boldsymbol{x}_1 - \boldsymbol{x}_2)^{\boldsymbol{T}}\boldsymbol{\mathcal{M}}(\boldsymbol{x}_1 - \boldsymbol{x}_2)} \end{aligned} \tag{24}$$

where $\boldsymbol{\mathcal{M}}$ is termed a Mahalanobis matrix.

To induce a distance metric, $\boldsymbol{\mathcal{M}}$ should be a positive semi definite (PSD) matrix. By imposing singular value decomposition on $\boldsymbol{\mathcal{M}}$, it can be decomposed into $\boldsymbol{\mathcal{M}} = \boldsymbol{\mathcal{P}}^{\boldsymbol{T}}\boldsymbol{\Sigma}\boldsymbol{\mathcal{P}}$, where $\boldsymbol{\mathcal{P}}$ is a unitary matrix, satisfying $\boldsymbol{\mathcal{P}}^{\boldsymbol{T}}\boldsymbol{\mathcal{P}} = \boldsymbol{\mathcal{I}}$, with $\boldsymbol{\mathcal{I}}$ denoting the identity matrix; and $\boldsymbol{\Sigma}$ is a diagonal matrix with the diagonal elements being singular values [22]. From this, the Mahalanobis distance is defined as:

$$\begin{aligned} d_{\boldsymbol{\mathcal{M}}}(\boldsymbol{x_1}, \boldsymbol{x_2}) &= \sqrt{(\boldsymbol{x_1} - \boldsymbol{x_2})^T\boldsymbol{\mathcal{P}}^{\boldsymbol{T}}\boldsymbol{\Sigma}\boldsymbol{\mathcal{P}}(\boldsymbol{x_1} - \boldsymbol{x_2})} \\ &= \sqrt{(\boldsymbol{\mathcal{P}}\boldsymbol{x_1} - \boldsymbol{\mathcal{P}}\boldsymbol{x_2})^T\boldsymbol{\Sigma}(\boldsymbol{\mathcal{P}}\boldsymbol{x_1} - \boldsymbol{\mathcal{P}}\boldsymbol{x_2})}. \end{aligned} \tag{25}$$

As such, in the context of T-FRI, if replacing Euclidean distance with Mahalanobis distance, two main functionalities of the Mahalanobis distance metric can be exploited. One is to discover an optimal orthogonal matrix $\boldsymbol{\mathcal{P}}$ that removes the couplings amongst antecedent features, mapping the original samples onto a new coordinates system; and the other is to assign weights from the associated diagonal matrix $\boldsymbol{\Sigma}$ to the transformed features, reflecting the relationship between them and the consequent within the resulting coordinates system.

The traditional Mahalanobis matrix $\boldsymbol{\mathcal{M}} = \boldsymbol{\mathcal{S}}^{-1}$, where $\boldsymbol{\mathcal{S}}$ is the covariance matrix of the dataset, takes into account

This article has been accepted for publication in IEEE Transactions on Fuzzy Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TFUZZ.2022.3194368

the distribution of data on the original feature space. This can be expanded and strengthened through the application of modern metric learning techniques. In particular, as introduced in the seminal work of [21], metric learning can be formulated as a convex optimisation problem, such that the relationship between instances of the same class becomes closer to one another, whilst instances of different classes are farther away from each other. The following briefly outlines five metric learning methods for deriving the Mahalanobis matrix that each can help achieve such Mahalanobis distance measures.

2) Large Margin Nearest Neighbours (LMNN): LMNN works by enabling $\boldsymbol{\mathcal{M}}$ to pull the k nearest examples belonging to the same class together while pushing off the examples from different classes [23]. In order to achieve this, two types of constraint are imposed: (i) a set of must-link constraints, $\mathbb{S}_{lmnn}$, such that one of the k nearest points $\boldsymbol{x_j}$ to $\boldsymbol{x_i}$ must belong to the same category of $\boldsymbol{x_i}$ (e.g., $y_j = y_i$, where y_i and y_j are the labels of $\boldsymbol{x_i}$ and $\boldsymbol{x_j}$ respectively); and (ii) a set of ternary constraints, $\mathbb{R}_{lmnn}$, such that in the three-member set $(\boldsymbol{x_i}, \boldsymbol{x_j}, \boldsymbol{x_k})$, $(\boldsymbol{x_i}, \boldsymbol{x_j})$ fall into the same category but $\boldsymbol{x_k}$ does not belong to it (e.g., $y_j = y_i, y_k \neq y_j$, where y_k is the label of $\boldsymbol{x_k}$). From these, Mahalanobis matrix is learned by a program for convex optimisation:

$$\begin{aligned} &\min_{\boldsymbol{\mathcal{M}} \succeq 0, \xi \geq 0} (1-\mu) \sum_{(\boldsymbol{x_i}, \boldsymbol{x_j}) \in \mathbb{S}_{lmnn}} d^2_{\boldsymbol{\mathcal{M}}}(\boldsymbol{x_i}, \boldsymbol{x_j}) + \mu \sum_{i,j,k} \xi_{ijk} \\ &s.t. \\ &d^2_{\boldsymbol{\mathcal{M}}}(\boldsymbol{x_i}, \boldsymbol{x_k}) - d^2_{\boldsymbol{\mathcal{M}}}(\boldsymbol{x_i}, \boldsymbol{x_j}) \geq 1 - \xi_{ijk} \ \forall (\boldsymbol{x_i}, \boldsymbol{x_j}, \boldsymbol{x_k}) \in \mathbb{R}_{lmnn} \end{aligned} \tag{26}$$

where ξ_{ijk} is a nonnegative slack variable, aimed at measuring the amount by which a differently labelled input $\boldsymbol{x_k}$ invades the area around $\boldsymbol{x_i}$, the boundary of which is defined by the same class for the input $\boldsymbol{x_j}$. Therefore, similar to linear support vector machine algorithm (SVM), ξ_{ijk} works as a penalty parameter to adjust the objective function, with $\mu \in [0, 1]$ balancing the above two sets of constraints of the objective function.

3) Information-Theoretic Metric Learning (ITML): ITML works based on exploiting Information Theory [24]. Suppose that the distance between two multivariate Gaussian distributions is generally defined by

$$KL\big(p(\boldsymbol{x}; \boldsymbol{\mathcal{M}_0}) \parallel p(\boldsymbol{x}; \boldsymbol{\mathcal{M}})\big) = \int p(\boldsymbol{x}; \boldsymbol{\mathcal{M}_0}) \log \frac{p(\boldsymbol{x}; \boldsymbol{\mathcal{M}_0})}{p(\boldsymbol{x}; \boldsymbol{\mathcal{M}})} dx \tag{27}$$

where $\boldsymbol{\mathcal{M}}$ is a metric matrix to be learned and $\boldsymbol{\mathcal{M}_0}$ is a priori metric matrix (usually $\boldsymbol{\mathcal{M}_0} = \boldsymbol{\mathcal{S}^{-1}}$); $p(\boldsymbol{x}; \boldsymbol{\mathcal{M}_0}) = \frac{1}{z} exp\big(-\frac{1}{2} d_{\boldsymbol{\mathcal{M}_0}}(\boldsymbol{x}, \boldsymbol{\mu})\big)$ and $p(\boldsymbol{x}; \boldsymbol{\mathcal{M}}) = \frac{1}{z} exp\big(-\frac{1}{2} d_{\boldsymbol{\mathcal{M}}}(\boldsymbol{x}, \boldsymbol{\mu})\big)$ are two Gaussian distributions, where $\boldsymbol{\mu}$ is the mean of Gaussians and z is a normalising constant; and $KL(\cdot)$ measures relative entropy. Then, the metric learning problem can be formulated as one of Bregman optimisation [25] by computing the following:

$$\begin{aligned} &\min_{\boldsymbol{\mathcal{M}}} \ KL\big(p(\boldsymbol{x}; \boldsymbol{\mathcal{M}_0}) \parallel p(\boldsymbol{x}; \boldsymbol{\mathcal{M}})\big) \\ &s.t. \ d_{\boldsymbol{\mathcal{M}}}(\boldsymbol{x_i}, \boldsymbol{x_j}) \leq \mu, \quad (\boldsymbol{x_i}, \boldsymbol{x_j}) \in \mathbb{S} \\ &\quad\ d_{\boldsymbol{\mathcal{M}}}(\boldsymbol{x_i}, \boldsymbol{x_j}) \geq l, \quad (\boldsymbol{x_i}, \boldsymbol{x_j}) \in \mathbb{D} \end{aligned} \tag{28}$$

where μ and l are parameters; $\mathbb{S}$ is a set of pairwise similarity constraints; $\mathbb{D}$ is a set of pairwise dissimilarity constraints; and $KL(\cdot)$ is computed by

$$\begin{aligned} &KL\big(p(\boldsymbol{x}; \boldsymbol{\mathcal{M}_0}) \parallel p(\boldsymbol{x}; \boldsymbol{\mathcal{M}})\big) = \frac{1}{2} D_{ld}(\boldsymbol{\mathcal{M}}, \boldsymbol{\mathcal{M}_0}) \\ &\frac{1}{2} D_{ld}(\boldsymbol{\mathcal{M}}, \boldsymbol{\mathcal{M}_0}) = tr(\boldsymbol{\mathcal{M}} \boldsymbol{\mathcal{M}_0^{-1}}) - \log det(\boldsymbol{\mathcal{M}} \boldsymbol{\mathcal{M}_0^{-1}}) - d \end{aligned} \tag{29}$$

where $D_{ld}(\cdot)$ is termed Bregman divergence; $tr(\cdot)$ is a matrix trace; and $\log det(\cdot)$ is the logarithm of the determinant of a matrix. The main purpose of Eqn. (28) is to regularise the matrix $\boldsymbol{\mathcal{M}}$ to remain possibly close to $\boldsymbol{\mathcal{M}_0}$, under soft constraints on keeping distances between points belonging to $\mathbb{S}$ smaller than μ and those between dissimilar points larger than l.

4) Sparse Determinant Metric Learning (SDML): SDML aims to deal with the problems where the dimensionality of the feature space is much greater than the sample size [26]. Compared with ITML, it utilises a double regularisation (namely, $\log det(\cdot)$- and l_1-norm) on the off-diagonal elements of $\boldsymbol{\mathcal{M}}$. The optimisation problem concerned can be described as:

$$\begin{aligned} \min_{\boldsymbol{\mathcal{M}} \succeq 0} \quad & tr(\boldsymbol{\mathcal{M}_0^{-1}} \boldsymbol{\mathcal{M}}) - \log det(\boldsymbol{\mathcal{M}}) + \\ & \lambda \left\| \boldsymbol{\mathcal{M}} \right\|_{1,off} + \eta L(\mathbb{S}, \mathbb{D}) \end{aligned} \tag{30}$$

where $\left\| \boldsymbol{\mathcal{M}} \right\|_{1,off}$ is the off-diagonal l_1-norm of $\boldsymbol{\mathcal{M}}$ ($\left\| \boldsymbol{\mathcal{M}} \right\|_{1,off} = \sum_{i \neq j} |\boldsymbol{\mathcal{M}}_{ij}|$); λ is a balance parameter; $L(\mathbb{S}, \mathbb{D})$ is a loss function defined on the constraints of $\mathbb{S}$ and $\mathbb{D}$, as defined in Eqn. (28); and η is a positive balance parameter trading off between the loss function and the regulariser. As $\boldsymbol{\mathcal{M}_0}$ is a constant matrix, the relative entropy defined by Eqn. (29) is simplified to the first two terms of Eqn. (30), making $\boldsymbol{\mathcal{M}}$ as close as possible to the given $\boldsymbol{\mathcal{M}_0}$.

5) Least Squares Metric Learning (LSML): LSML learns a Mahalanobis matrix from training data by comparing their relative distances [27]. Suppose that a set of data samples can be arranged such that

$$\mathbb{C} = \{(\boldsymbol{x_i}, \boldsymbol{x_j}, \boldsymbol{x_k}, \boldsymbol{x_l}) : d(\boldsymbol{x_i}, \boldsymbol{x_j}) < d(\boldsymbol{x_k}, \boldsymbol{x_l})\}.$$

The optimisation problem is given in the form of

$$\begin{aligned} \min_{\boldsymbol{\mathcal{M}} \succeq 0} \quad & D_{ld}(\boldsymbol{\mathcal{M}}, \boldsymbol{\mathcal{M}_0}) + \\ & \sum_{(\boldsymbol{x_i}, \boldsymbol{x_j}, \boldsymbol{x_k}, \boldsymbol{x_l}) \in \mathbb{C}} \omega_{i,j,k,l} H(d_{\boldsymbol{\mathcal{M}}}(\boldsymbol{x_i}, \boldsymbol{x_j}) - d_{\boldsymbol{\mathcal{M}}}(\boldsymbol{x_k}, \boldsymbol{x_l})) \end{aligned} \tag{31}$$

where $D_{ld}(\boldsymbol{\mathcal{M}}, \boldsymbol{\mathcal{M}_0})$ is defined in Eqn. (29) and $H(\cdot)$ is the squared hinge function defined as follows:

$$H(x) = \begin{cases} x^2, & \text{if } x > 0 \\ 0, & \text{if } x \leq 0. \end{cases} \tag{32}$$

6) Relevant Component Analysis (RCA): RCA was originally designed for image retrieval [28]. As with principal component analysis (PCA), RCA compresses data along the axes with the greatest irrelevant variability. Particularly, the Mahalanobis matrix $\boldsymbol{\mathcal{M}}$ learned by RCA is based on a weighted sum of in-chunklets covariance matrices, assigning

weights depending on the perceived "relevance", which is estimated using "chunklets" (namely, groups of points of the same class). It is of high computational efficiency for implementation by calculating the following matrix:

$$\widehat{\mathcal{C}} = \frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ji} - \widehat{m}_j)(x_{ji} - \widehat{m}_j)^T \tag{33}$$

where n is the total number of points in the k chunklets; chunklet j is composed of $\{x_{ji}\}_{i=1}^{n_j}$ with $\widehat{m}_j$ being its mean; the number of chunklets and each chunklet size are both randomly set initially, with the former normally set to be much larger than the number of classes of the underlying problem concerned; and $\widehat{\mathcal{C}}^{-1} = \mathcal{M}$. Importantly, the within-chunklet variability can be significantly reduced by identifying features irrelevant to the task, leading to the optimal solution to the optimisation problem which minimises the distances between data of the same class (as proven in [28]).

The above descriptions provide five different approaches; any one of them may be utilised for implementation. This shows the flexibility of the present work. However, there is no established rule for making a choice of which method to use under what conditions. Empirically, LMNN is excellent at coping with various supervised metric learning tasks because it makes no assumptions about the data distribution. ITML is capable to handle a wide variety of constraints under weakly supervised conditions due to its independence of eigenvalue calculation and semi-definite programming. SDML can address sparse metric learning within a high-dimensional feature space. LSML is particularly useful when the pairwise constraints are not obtained naturally. RCA enjoys good performance in performing specific tasks such as face recognition. Thus, when presented with a certain application, a trial-and-error approach may be taken to determine which of the methods would be best suitable for addressing the problem at hand.

III. FRI BASED ON MAHALANOBIS DISTANCE

Given the above preliminaries, this section presents a novel framework of T-FRI based on redefining the distance metric employed, through Mahalanobis metric learning, as illustrated in Fig. 2. Again, for illustrative simplicity, the fuzzy rule base is depicted by a few points (herein with just three types of shape for class labels, a number of axes for various dimensions, and a sphere for a hypersphere in describing a sub-problem space). Of course, this does not mean that the relevant theoretical development and practical implementation are subject to such a simplified version. The following specifies the transformation process of a fuzzy rule base and that of an observation, and the utilisation of Choquet Integral [29] for the effective implementation of the aggregation operations in T-FRI. Note that any of the metric learning methods introduced above can be employed to implement the learning of the required Mahalanobis matrix $\mathcal{M}$.

A. Transformation of Fuzzy Rule Base and Observation

There are four steps to implement the transformation of a fuzzy rule base and an observation, three for rule bases and one for observations.

Firstly, for a given fuzzy rule base generally represented in the form of Eqn. (1), representative values of the fuzzy sets involved within the rules are utilised to facilitate the learning of the Mahalanobis matrix. From this, the fuzzy rule base is translated into:

$$\begin{bmatrix} Rep(A_{11}) & Rep(A_{12}) & \cdots & Rep(A_{1m}) & Rep(B_1) \\ Rep(A_{21}) & Rep(A_{22}) & \cdots & Rep(A_{2m}) & Rep(B_2) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ Rep(A_{N1}) & Rep(A_{N2}) & \cdots & Rep(A_{Nm}) & Rep(B_N) \end{bmatrix}, \tag{34}$$

which is artificially considered as a set of training data $(\boldsymbol{x_i}, y_i), i \in [1, N]$, in preparation for use in Mahalanobis metric learning.

Secondly, in order to obtain the required Mahalanobis matrix $\mathcal{M}$ any one of the previously reviewed metric learning methods is applied to the artificial training data resulting from the first step. As a direct application of such learning methods, no attempt is herein made to optimise the hyper-parameters in these methods (e.g., the number of neighbours in LMNN, the maximum number of iterations in ITML, and the number of chunks to generate in RCA). Instead, their default settings as proposed in the respective literature are adopted.

Thirdly, the original (sparse) fuzzy rule base is transformed into a reformulated representation within a new feature space, through the application of the learned Mahalanobis matrix. Recall that triangular membership functions are employed to define fuzzy sets in this work. Hence, suppose that an arbitrary fuzzy set A_{ij} given in a rule expressed by Eqn. (1) is represented as a ternary vector $(a_{ij_1}, a_{ij_2}, a_{ij_3})^T$, where the elements are the abscissas of the three vertices of the fuzzy membership function (s.t., $a_{ij_1} \leq a_{ij_2} \leq a_{ij_3}$). Then, by left multiplying such vectors collectively, with the learned transformation matrix $\mathcal{M}$, the antecedent values (respectively defined by A_{ij}) of the original fuzzy rules are mapped onto a new linear space. For example, the antecedent feature values of the ith rule R_i which are defined by fuzzy sets $(A_{i1}, A_{i2}, \ldots, A_{im})$ are transformed into

$$\begin{aligned} &\mathcal{M}\begin{bmatrix} a_{i1_1} & a_{i2_1} & \ldots & a_{im_1} \\ a_{i1_2} & a_{i2_2} & \ldots & a_{im_2} \\ a_{i1_3} & a_{i2_3} & \ldots & a_{im_3} \end{bmatrix}^T \\ &= \begin{bmatrix} \widehat{a_{i1_1}} & \widehat{a_{i2_1}} & \ldots & \widehat{a_{im_1}} \\ \widehat{a_{i1_2}} & \widehat{a_{i2_2}} & \ldots & \widehat{a_{im_2}} \\ \widehat{a_{i1_3}} & \widehat{a_{i2_3}} & \ldots & \widehat{a_{im_3}} \end{bmatrix}^T \end{aligned} \tag{35}$$

where $(\widehat{a_{ij_1}}, \widehat{a_{ij_2}}, \widehat{a_{ij_3}})^T$ denotes the transformed fuzzy set with respect to A_{ij}. The consequent part of R_i can be acquired in the same way, but for classification problems and therefore, the crisp output, this procedure is omitted to save computational effort. As the outcome of this third procedure,

This article has been accepted for publication in IEEE Transactions on Fuzzy Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TFUZZ.2022.3194368

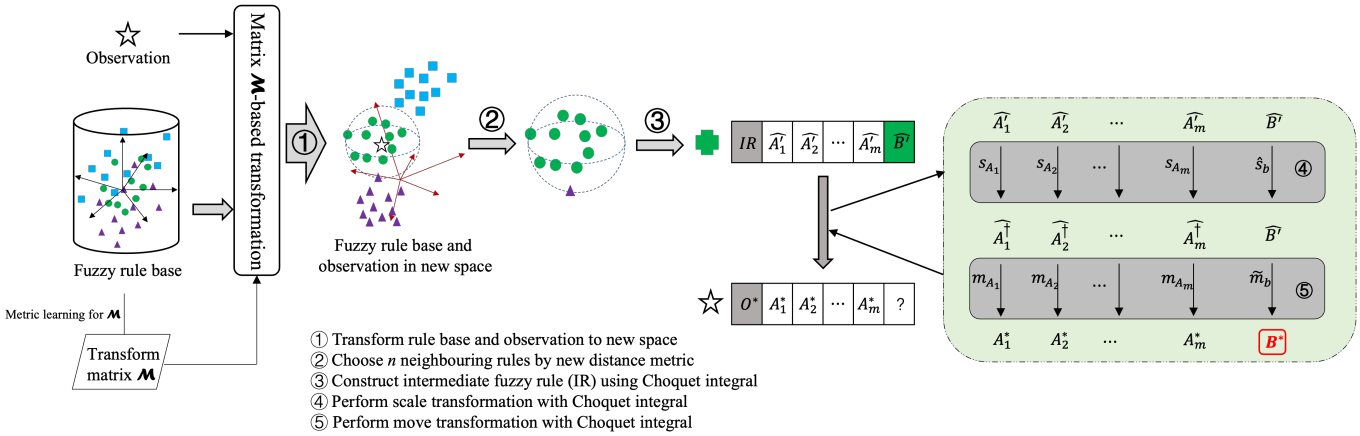

Fig. 2. Framework of proposed method.

the transformed rule base is developed in the new space as follows:

$$\text{Rule } \widehat{R}_i: \quad \text{If } x_1 \text{ is } \widehat{A_{i1}} \text{ and } x_2 \text{ is } \widehat{A_{i2}} \text{ and } \cdots \text{ and } x_m \text{ is } \widehat{A_{im}}, \text{ then } y \text{ is } B_i \tag{36}$$

where $\widehat{A_{ij}} = (\widehat{a_{ij_1}}, \widehat{a_{ij_2}}, \widehat{a_{ij_3}})^T$.

Lastly, given an observation if it does not match any rule within the original sparse rule base (and hence, no conventional fuzzy reasoning can be carried out using CRI [2]), it is then projected onto the new coordinates system with the transformation matrix by the same procedure that transforms rule antecedents as given in the last step. That is, the transformed observation with respect to an unmatched observation O^* is

$$\text{Observation } \widehat{O}^*: \widehat{A_1^*}, \widehat{A_2^*}, \cdots, \widehat{A_j^*}, \cdots, \widehat{A_m^*} \tag{37}$$

where $\widehat{A_j^*} = (\widehat{a_{j1}^*}, \widehat{a_{j2}^*}, \widehat{a_{j3}^*})^T$ denotes the fuzzy set value of the jth transformed antecedent feature.

To exemplify the above transformation processes, consider a trivial sparse fuzzy rule base concerning the Iris dataset (simplified from the KEEL dataset repository [30]). The rule base before and that after the transformation are illustrated in Fig. 3 and Fig. 4, respectively. In particular, Fig. 3 depicts the pairwise relationships between each original feature, and Fig. 4 shows their corresponding relationships after being transformed into the new space. Note that all antecedent fuzzy variables are expressed using their representative values, with different colours representing different consequent labels. Importantly, it can be seen from Fig. 4 that each pair of transformed features, as well as every individual transformed feature, have almost equal ability for use to distinguish various classes, forming a sharp contrast with their originals in Fig. 3.

The introduction of Mahalanobis distance metric is (mainly) to help the underlying FRI process to improve the selection of the n nearest rules in an effort to derive an intermediate rule. Having accomplished the above four steps, fuzzy rule interpolation-based inference could be performed as done with the traditional T-FRI, using the fuzzy rules $\widehat{R}_i, i \in [1, N]$ and any observation $\widehat{O}^*$. Such a process would then start with

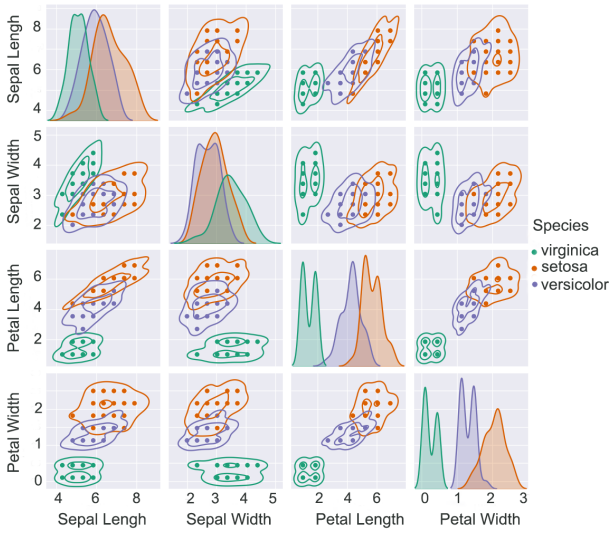

Fig. 3. Pairwise relationships of original rule base (Iris dataset).

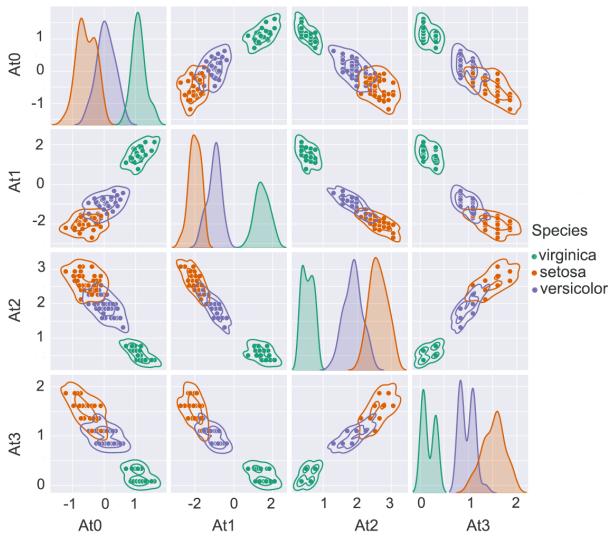

Fig. 4. Pairwise relationships of transformed rule base (Iris dataset).

This article has been accepted for publication in IEEE Transactions on Fuzzy Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TFUZZ.2022.3194368

the selection of n nearest transformed rules to construct the imperative intermediate fuzzy rule, determined by Euclidean distance as below:

$$d(\widehat{O}^*, \widehat{R}_i) = \sqrt{\sum_{j=1}^{m} d(\widehat{A_j^*}, \widehat{A_{ij}})^2}. \tag{38}$$

For classification problems, the inference process would be straightforward, since while transforming the original fuzzy rule antecedent attributes, no change is made to the consequent part of any rule. The inferred class for an unmatched observation using rules from the new space is the final output of the entire reasoning system. For regression problems, as the transformation processes are linear, the inferred results of the overall system could be calculated by left multiplying $\boldsymbol{\mathcal{M}}^{-1}$ to the output produced. However, to further exploit the potential of utilising transformed rules and observation gained in the resulting new feature space, the remaining steps of T-FRI that involve the use of four key reasoning parameters are modified, as follows.

As outlined in Section II-A, the interpolative consequence of T-FRI is inferred via manipulation of four key parameters: $\tilde{w}_{ib}$, $\tilde{\delta}_b$, $\tilde{s}_b$ and $\tilde{m}_b$ (referring to Eqns. (13), (14), (19) and (21) respectively). For existing T-FRI approaches, the overall inference process and hence, the ultimately interpolated results are highly dependent on the use of what aggregation function is adopted to accomplish the required operations. For instance, arithmetic mean is applied in the seminal T-FRI [7] and weighted arithmetic mean in the most recently developed weighted T-FRI (denoted WT-FRI hereafter) [17].

Generally speaking, the operation of combining or merging different values into a single compound one is termed aggregation, and the function performing this operation is referred to as an aggregation function [31]. Mathematically, a large family of operators can be employed to serve as a powerful aggregator to integrate diverse domain attribute values. Arithmetic mean and weighted arithmetic mean are just two ones popularly used in T-FRI. This work adopts the aggregation function named Choquet integral for use with the transformed new feature space. It seeks to enhance the efficacy of constructing the intermediate rules in the implementation of required scale and move transformations.

Choquet integral is capable of assessing and hence, exploiting contributions of elements being compounded, taking into consideration not only the significance of individual attributes but also their underlying groups (be they clusters or classes) [29]. Thus, it is of particular relevance to performing classification tasks. It is because of this recognition that Choquet integral-based aggregation is herein implemented to modify the aforementioned four key factors as follows:

$$\widehat{\tilde{w}_{ib}} = \sum_{j=1}^{m} (\widehat{\tilde{w}_{i\sigma(j)}} - \widehat{\tilde{w}_{i\sigma(j-1)}}) \mu(U_{\widehat{\tilde{w}_{i\sigma(j)}}}) \tag{39}$$

$$\widehat{\tilde{\delta}_b} = \sum_{j=1}^{m} (\widehat{\delta_{\sigma(j)}} - \widehat{\delta_{\sigma(j-1)}}) \mu(U_{\widehat{\delta_{\sigma(j)}}}) \tag{40}$$

$$\widehat{\tilde{s}_b} = \sum_{j=1}^{m} (\widehat{s_{A_{\sigma(j)}}}) - s_{\widehat{A_{\sigma(j-1)}}}) \mu(U_{\widehat{s_{A_{\sigma(j)}}}}) \tag{41}$$

$$\widehat{\tilde{m}_b} = \sum_{j=1}^{m} (\widehat{m_{A_{\sigma(j)}}}) - \widehat{m_{A_{\sigma(j-1)}}}) \mu(U_{\widehat{m_{A_{\sigma(j)}}}}) \tag{42}$$

where $\widehat{\cdot}$ denotes any value or parameter considered in the newly transformed coordinates system; $(\sigma(1), \sigma(2), \ldots, \sigma(m))$ is a non-decreasing permutation, e.g., $\{\widehat{\tilde{w}_{i\sigma(1)}}, \widehat{\tilde{w}_{i\sigma(2)}}, \ldots, \widehat{\tilde{w}_{i\sigma(m)}}\}$ is a non-decreasing value permutation of antecedent features $\{\widehat{\tilde{w}_{i1}}, \widehat{\tilde{w}_{i2}}, \ldots, \widehat{\tilde{w}_{im}}\}$ with $\cdot_{\sigma(0)} = 0$ by convention and m being the number of the features; $U_{\widehat{\cdot_{\sigma(j)}}} = \{\widehat{\cdot_{\sigma(j)}}, \widehat{\cdot_{\sigma(j+1)}}, \ldots, \widehat{\cdot_{\sigma(m)}}\}$ is the subset of indices of the $m - j + 1$ largest components of $\widehat{\cdot_{\sigma(\cdot)}}$; and μ is a fuzzy measure function. Note that these modifications on the four parameters within T-FRI are linear and so, the computational complexity of the original T-FRI is not adversely affected.

The use of an aggregation operator (Choquet integral or else) can nonetheless be cumbersome with the 2^m elements to address, especially when there are numerous variables (or m is large). To reduce computational complexity, the power measure [32] is exploited to define the required fuzzy measure function, which is formulated by

$$\mu(U) = (\frac{|U|}{m})^q, \quad \text{with } q > 0 \tag{43}$$

where $|U|$ stands for the cardinality of the set U (i.e., the number of elements in U). Note that when $q = 1$, the power measure degenerates to an additive fuzzy measure. In general, if q is a fixed real value, then the number of required elements for calculating Choquet integral is $m-1$ that is much smaller than 2^m. In this work, for computational simplicity, q is set to 2 (unless otherwise stated).

Having obtained the four key T-FRI parameters with representations in the new feature space, the following execution of the FRI process remains the same as the traditional T-FRI (or WT-FRI).

B. Theoretical analysis

As the above proposed approach is based on the existing work of T-FRI, it is interesting to theoretically compare it with the seminal T-FRI algorithm and the state-of-the-art WT-FRI method. According to Eqn. (25), the Mahalanobis distance between a given rule R_p and an observation as defined per Eqns. (1) and (2) can be rewritten by

$$\begin{aligned} & d^2_{\boldsymbol{\mathcal{M}}}(O^*, R_p) \\ & \quad = (\boldsymbol{x^*} - \boldsymbol{x_p})^T \boldsymbol{\mathcal{P}}^{\boldsymbol{T}} \boldsymbol{\Sigma} \boldsymbol{\mathcal{P}} (\boldsymbol{x^*} - \boldsymbol{x_p}) \\ & \quad = (\boldsymbol{\mathcal{P}}(\boldsymbol{x^*} - \boldsymbol{x_p}))^T \boldsymbol{\Sigma} (\boldsymbol{\mathcal{P}}(\boldsymbol{x^*} - \boldsymbol{x_p})) \\ & \quad = (\boldsymbol{\mathcal{P}}(\boldsymbol{x^*} - \boldsymbol{x_p}))^T \begin{bmatrix} \Sigma_1 & 0 & \cdots & 0 \\ 0 & \Sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \Sigma_m \end{bmatrix} (\boldsymbol{\mathcal{P}}(\boldsymbol{x^*} - \boldsymbol{x_p})) \end{aligned} \tag{44}$$

where $\boldsymbol{\Sigma}$ is a $m \times m$ diagonal matrix; $\boldsymbol{x^*} = (\text{Rep}(A_1^*), \text{Rep}(A_2^*), \ldots, \text{Rep}(A_m^*))^T = (x_1^*, x_2^*, \ldots, x_m^*)^T$ and $\boldsymbol{x_p} = (\text{Rep}(A_{p1}), \text{Rep}(A_{p2}), \ldots, \text{Rep}(A_{pm}))^T = (x_{p1}, x_{p2}, \ldots, x_{pm})^T$.

This article has been accepted for publication in IEEE Transactions on Fuzzy Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TFUZZ.2022.3194368

According to WT-FRI (which degenerates to the standard T-FRI if all feature weights are equal) [17], the distance between O^* and R_p is calculated by aggregating weights of all features:

$$\widetilde{d}^2_W(O^*, R_p) = \frac{\sum_{j=1}^{m}((1-W_j)d(O^*_j, A_{pj}))^2}{\sum_{j=1}^{m}(1-W_j)^2} \tag{45}$$

where W_j represents the weight associated with the jth feature and the distance between two fuzzy sets O^*_j and A_{pj} is given as Eqn. (5). Since the normalisation term $\sum_{j=1}^{m}(1-W_j)^2$ in Eqn. (45) is a constant, it can be omitted in the process of calculating the distances as only information on the relevant distance measures is of use for selecting the nearest neighbouring rules. Therefore, Eqn. (45) defined in [17] can be simplified by

$$\begin{aligned} &d^2_W(O^*, R_p) \\ &\quad= \sum_{j=1}^{m}\left((1-W_j)\frac{|Rep(A^*_j) - Rep(A_{pj})|}{range_j}\right)^2 \\ &\quad= \sum_{j=1}^{m}\left((1-W_j)\frac{x^*_j - x_{pj}}{range_j}\right)^2. \end{aligned} \tag{46}$$

Note that W_j and $range_j$ are constants for a given rule base. Note also that with the above proposed new approach (and indeed, for any T-FRI method), a more general version of the inference mechanism directly using fuzzy sets themselves instead of fuzzy representative values could be introduced if preferred. Nonetheless, this would incur a significant decrease in FRI efficiency. Hence, given the underlying inference is of an approximate nature in the first place, such development is left out of the scope of the present work.

C. T-FRI with Aggregation Function

Let $C_j = \frac{1-W_j}{range_j}$, $j = 1, 2, \ldots, m$. Then, Eqn. and (46) can be rewritten in matrix form as follows:

$$\begin{aligned} &d^2_W(O^*, R_p) \\ &= (C_1(x^*_1 - x_{p1}))^2 + \cdots + (C_m(x^*_m - x_{pm}))^2 \\ &= \begin{bmatrix} C_1(x^*_1 - x_{p1}), \cdots, C_m(x^*_m - x_{pm}) \end{bmatrix} \begin{bmatrix} C_1(x^*_1 - x_{p1}) \\ \vdots \\ C_m(x^*_m - x_{pm}) \end{bmatrix} \\ &= \left(\begin{bmatrix} (x^*_1 - x_{p1}), \cdots, (x^*_m - x_{pm}) \end{bmatrix} \begin{bmatrix} C_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & C_m \end{bmatrix}^T \right) \\ &\qquad \left(\begin{bmatrix} C_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & C_m \end{bmatrix} \begin{bmatrix} (x^*_1 - x_{p1}) \\ \vdots \\ (x^*_m - x_{pm}) \end{bmatrix} \right) \\ &= (\boldsymbol{x^*} - \boldsymbol{x_p})^T \begin{bmatrix} C_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & C_m^2 \end{bmatrix} (\boldsymbol{x^*} - \boldsymbol{x_p}) \end{aligned} \tag{47}$$

where $\boldsymbol{x^*} - \boldsymbol{x_p} = \begin{bmatrix} (x^*_1 - x_{p1}), \cdots, (x^*_m - x_{pm}) \end{bmatrix}^T$.

As WT-FRI is an extension to T-FRI, the above distance degenerates to the simple Euclidean distance measure that is used in T-FRI and is simply formulated by

$$d^2_T(O^*, R_p) = (\boldsymbol{x^*} - \boldsymbol{x_p})^T(\boldsymbol{x^*} - \boldsymbol{x_p}). \tag{48}$$

Based on the above discussions, especially through comparing Eqns. (44), (47) and (48), the following points can be observed:

1) : Distances between an observation and a rule in T-FRI and WT-FRI are both computed in the original feature space. The difference is that, unlike T-FRI, the distance used in WT-FRI reinforces the role of important features through assigning to the features different weights W_j (see Eqn. (45) or C_j in Eqn. (47)) learned by a feature ranking mechanism. In so doing, WT-FRI gains a great advantage of choosing the closest rules to perform interpolation, yielding impressive improvement over the original T-FRI. However, for certain problems, one or several features from the original feature space may not have any prominent contributions to the interpolative inference process, or there may not be a clear distinction of importance between different features. Thus, WT-FRI may not achieve an intended improvement. Moreover, statistical correlations among some (original) features are almost universal in real-world applications. Retaining all of them often leads to information redundancy, thereby restricting and reducing the efficiency of interpolative inference, especially when the number of features is large. if there exists noise in the redundant features, then the effectiveness of the reasoning process is also adversely affected.

2) : Distance between an observation and a rule in the present approach is calculated in the newly transformed feature space. With all data transformed through a unitary matrix $\boldsymbol{\mathcal{P}}$, where the original features no longer play a direct role, the relationships between the transformed features are reestablished. Particularly, the new features are linearly independent of each other and may reflect various degrees of significance. As such, in calculating the distances, different weights Σ_j (as per Eqn. 44) can be readily assigned to the features. In addition, $Rank(\boldsymbol{\mathcal{P}}) \leq m$, where m is the number of original features. This means that there may be zero value(s) in $\Sigma_1, \Sigma_2, \ldots, \Sigma_m$, which will enable the computation of the Mahalanobis distances on a r-dimensional space ($r \leq m$). Obviously, this is very helpful when the number of features becomes large.

3) : The complexity of selecting the nearest rules by calculating distances between an unmatched observation and each individual rule incurred by the proposed approach is rather different from that associated with WT-FRI. Here, it is computed by measuring the simple Euclidean distances within the transformed coordinates system. However, the corresponding computing process for WT-FRI is not simply using Euclidean distance but through aggregating weights of all features involved. Remarkably, the computational cost of the overall fuzzy rule interpolative inference is largely dependent upon the selection of the neighbouring rules nearest to the observation, whilst the subsequent computational procedures are of the same order of complexity amongst the family of T-FRI-based methods. Thus, it becomes evident that the present

TABLE I
DATASETS EMPLOYED

Datasets	#Instances	#Attributes	#Classes
Wine	178	13	3
Iris	150	4	3
NewThyroid	215	5	3
Balance	625	4	3
Phoneme	5404	5	2
Pima	768	8	2
Appendicitis	106	7	2
WDBC	569	30	2
Ionosphere	351	33	2
Sonar	208	60	2
Glass	214	9	7

approach is of a lower complexity than WT-FRI, while being of an equivalent complexity to that of the original T-FRI.

IV. EXPERIMENTAL RESULTS

This section conducts systematic experimental evaluations to assess the classification performance of an FRI system implementing the proposed approach. This is carried out via comparing with the aforementioned two FRI techniques, on a wide range of benchmark problems [30]. The 11 datasets concerned involve multi-class and multivariate classifications, and cover tasks in rather different domains (including medical, chemical, and morphological). Additionally, extra experiments are implemented to investigate the effect of dimensionality reduction in conjunction with the novel approach.

A. Experimental Set-up

All experiments are tested in Pycharm Professional 2020 implemented on a MacBook Pro with M1 Chip and MacOS Big Sur. The details of these datasets are summarised in Table I. As indicated previously, the fuzzy values of all domain variables are represented herein by triangular membership functions (but more sophisticated ones such as trapezoidal, complex polygon or other bell-shaped fuzzy membership functions could be employed if preferred [1]). A practical approach [33] devised to tackle classification problems is utilised herein to generate a dense fuzzy rule base from data (which has the potential of deleting redundant rules by considering the significance degrees of emerging rules). Of course, alternative rule induction methods (e.g., [34], [35]) may be employed if desired, which may be particularly useful for dealing with regression problems.

To minimise any bias in performance comparison, all feature value domains are normalised into the common range of 0 to 1, and each is uniformly partitioned into five fuzzy sets. Then, 80% rules from the resulting dense fuzzy rule base are randomly selected to constitute a sparse rule base to validate the performance of FRI. Note that in resolving real-world problems, should there be a dense rule base available then no FRI would be required, but the application of conventional CRI. The present setup (of deliberately removing 20% of learned rules) is purely for the purpose of evaluating FRI methods when facing a sparse rule base. Reasoning results are compared with the underlying ground truth to assess the average accuracy through 10×10-fold cross-validation.

Each metric learning method reviewed previously is employed in the comparative studies, and the results are compared with those attainable by the state-of-the-art T-FRI and WT-FRI. The weights of features are derived from Information Gain (IG) as with the common approach typically exploited in the literature [17]. In order to optimise the selection of the nearest neighbouring rules, the weighted degree of each rule produced in the process of fuzzy rule induction is used to modify each distance metric such that

$$\widehat{d}^2(O^*, R_p) = d^2(O^*, R_p) \times \frac{1}{1 + RW_p} \tag{49}$$

where $d^2(\cdot)$ denotes any of the three types of distance metric (namely, d_T^2, d_W^2 and $d_{\boldsymbol{\mathcal{M}}}^2$); and RW_p is the weight degree of the rule R_p. Empirically, for many existing FRI techniques, an increase in the number of the nearest rules used to construct the intermediate rule does not necessarily lead to a noticeable improvement in accuracy, whilst the computational efficiency may sharply deteriorate [18]. Thus, only the least number, i.e., two of the nearest rules are exploited for interpolation, following the common practice in T-FRI.

B. Results and Discussion

Interpolative reasoning outcomes are reported and analysed here, covering studies on both effectiveness and efficiency.

1) Accuracy: Experimental results consisting of the average classification accuracies and standard deviations (SDs) on the 11 benchmark datasets run are summarised in Table II. Note that the results reported are those obtained by an integrated application of both FRI and CRI. This does not affect fair comparison amongst different FRI techniques as they each runs CRI over the same original (sparse) rule base. The underperformance of running CRI alone without the support of FRI is not presented here since it is obvious that such an approach would not be able to yield a reasonable inference outcome given the high sparsity of the original rule base (which has also been generally proven in the relevant literature [1]).

According to Table II, the proposed approach offers the highest classification accuracy with a low SD for all cases, beating both the original and weighted T-FRI method (albeit not each of the five implementations performs equally excellently). It remarkably surpasses the original T-FRI on the *Wine*, *Iris*, *WDBC* and *Ionosphere* datasets (with 10%-20% improvements over the performance of T-FRI). Although WT-FRI is the state-of-the-art strengthened version of T-FRI, providing an excellent method for handling fuzzy classification problems (see the results on *Iris*, *NewThyroid* and *Appendicitis*), it only has a marginal improvement on the *Wine*, *WDBC*, *Ionosphere* and *Sonar* datasets. The superior performance of the present experimental investigation conforms to the outcome of the theoretical analysis reported earlier. This implies that it is more significant to distribute different weights to antecedent features when they are independent, and that the aggregation operation

This article has been accepted for publication in IEEE Transactions on Fuzzy Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TFUZZ.2022.3194368

TABLE II
AVERAGE CLASSIFICATION ACCURACIES WITH STANDARD DEVIATION OVER 10×10-FOLD CROSS VALIDATION

Datasets	T-FRI	WT-FRI	LMNN-T-FRI	ITML-T-FRI	SDML-T-FRI	LSML-T-FRI	RCA-T-FRI
Wine	0.7657 ± 0.0247	0.7681 ± 0.0128	0.9340 ± 0.0073	**0.9499 ± 0.0115**	0.9214 ± 0.0196	0.8480 ± 0.0165	0.9256 ± 0.0164
Iris	0.7854 ± 0.0371	0.8904 ± 0.0142	**0.9174 ± 0.0091**	0.8094 ± 0.0131	0.9000 ± 0.0137	0.8280 ± 0.0173	0.8812 ± 0.0159
NewThyroid	0.8081 ± 0.0362	0.8471 ± 0.0203	**0.8929 ± 0.0120**	0.8829 ± 0.0186	0.8671 ± 0.0148	0.8069 ± 0.0222	0.8572 ± 0.0235
Balance	0.6861 ± 0.0110	0.6176 ± 0.0145	0.7514 ± 0.0107	0.7811 ± 0.0071	0.7459 ± 0.0050	0.7559 ± 0.0095	**0.7890 ± 0.0089**
Phoneme	0.7652 ± 0.0062	0.7628 ± 0.0047	**0.7879 ± 0.0021**	0.7734 ± 0.0075	0.7844 ± 0.0032	0.7669 ± 0.0024	0.7595 ± 0.0028
Pima	0.6685 ± 0.0108	0.6784 ± 0.0120	0.6871 ± 0.0140	**0.7073 ± 0.0078**	0.6808 ± 0.0074	0.6943 ± 0.0165	0.6995 ± 0.0133
Appendicitis	0.7620 ± 0.0097	0.7960 ± 0.0263	**0.8335 ± 0.0212**	0.8202 ± 0.0154	0.8224 ± 0.0072	0.8198 ± 0.0175	0.8311 ± 0.0132
WDBC	0.7435 ± 0.0118	0.7628 ± 0.0108	0.9391 ± 0.0060	**0.9498 ± 0.0056**	0.9310 ± 0.0019	0.7756 ± 0.0098	0.8250 ± 0.0099
Ionosphere	0.6777 ± 0.0137	0.6829 ± 0.0129	**0.8735 ± 0.0081**	0.8672 ± 0.0118	0.8338 ± 0.0067	0.8477 ± 0.0134	0.8154 ± 0.0050
Sonar	0.6848 ± 0.0088	0.6968 ± 0.0090	**0.8353 ± 0.0115**	0.7974 ± 0.0108	0.7663 ± 0.0210	0.7004 ± 0.0148	0.7785 ± 0.0110
Glass	0.4085 ± 0.0334	0.4039 ± 0.0225	0.4744 ± 0.0176	**0.5100 ± 0.0123**	0.4850 ± 0.0187	0.4565 ± 0.0282	0.4688 ± 0.0429

within WT-FRI over the selected nearest rules may become less effective if there is no marked difference in the relative importance levels amongst features.

Particularly, LMNN-T-FRI and ITML-T-FRI exhibit great power in doing their jobs, for they nearly occupy the top position in terms of accuracy rank across all bar one dataset (with the corresponding mean and SD figures highlighted in bold in Table II). However, the test record of LSML-T-FRI is mediocre, and its performance may not even be so good as WT-FRI for two datasets (*Iris* and *NewThyroid*). Nonetheless, when dealing with harder, multi-feature tasks (e.g., *WDBC*, *Ionosphere* and *Sonar*), the proposed approach facilitates the underlying T-FRI system to outperform its competitors. Note that given the sparse rule base, all types of T-FRI system examined herein are encountered with a challenge on the *Glass* dataset, which a fairly significant number (i.e., 7) of classes. Notwithstanding this general observation, it is still a thrill to find that the proposed approach achieves better results than the other two, no matter which metric learning method is employed.

2) Friedman and Nemenyi Tests: The above results only show the simple averaged inference accuracies for each individual dataset. To have a more in-depth comparative examination of the performance concerning the proposed approach, statistical tests are done, in terms of both Friedman test and Nemenyi test (in recognition of their suitability for comparing two or more classifiers on multiple datasets [36]).

Friedman test compares multiple algorithms starting with the null hypothesis that all algorithms concerned have the same performance. All candidate methods are sorted and ranked with regard to their performances on different datasets, as shown in Table III, where the best is ranked number 1 and those algorithms that have the same performance share the average of the otherwise individual ranking values. The test is based on the evaluation of the following parameter [37]:

$$\tau_F = \frac{(Num-1)\tau_{\chi^2}}{Num(Alm-1)-\tau_{\chi^2}} \tag{50}$$

where Alm and Num are the number of algorithms and that of datasets, respectively, and τ_{χ^2} is computed by

$$\begin{aligned}\tau_{\chi^2} &= \frac{Alm-1}{Alm} \times \frac{12Num}{Alm^2-1}\sum_{i=1}^{Alm}(r_i - \frac{Alm+1}{2})^2 \\ &= \frac{12Num}{Alm(Alm+1)}(\sum_{i=1}^{Alm} r_i^2 - \frac{Alm(Alm+1)^2}{4})\end{aligned} \tag{51}$$

with r_i representing the average ranking number of algorithm i. If the value of τ_F is greater than the critical threshold (obtained from the *scipy.stats.f.ppf* function [37] in response to a given confidence level), the null hypothesis is rejected and therefore, the performances of these algorithms are judged to be not the same.

For the present application, given that $Alm = 7$ and $Num = 11$, $\tau_F = 17.153$. Following the common practice in the literature, suppose that the p-value for hypothesis test is 5% (or the level of confidence is 95%). Then, by referring to *scipy.stats.f.ppf* it is found that the critical value is 2.2541 (which is less than τ_F). Thus, the null hypothesis that all the seven algorithms compared have the same performance is rejected. In other words, it can be claimed with a significant confidence that different methods investigated herein do perform differently.

The conclusion drawn by the Friedman test can only suggest that there are significant differences between these algorithms, but it cannot indicate which of them are different. Nemenyi test [38] is then applied to show the differences between the individual methods. For this, the value of the so-called critical difference (CD) is first calculated by:

$$CD = q_\alpha \sqrt{\frac{Alm(Alm+1)}{6Num}} \tag{52}$$

where the critical value q_α is obtained from a given look-up table [36], with the α reflecting the confidence level $p = 1-\alpha$. For instance, given a confidence level of 95%, $\alpha = 0.05$, $q_\alpha = 2.949$ and therefore, $CD = 2.716$. From this, differences in the average ranks between the methods compared are calculated as summarised in Table IV. These figures are then compared against CD: If the difference is greater than CD,

TABLE III
ORDER OF PERFORMANCE OF EACH ALGORITHM

Datasets	T-FRI	WT-FRI	LMNN-T-FRI	ITML-T-FRI	SDML-T-FRI	LSML-T-FRI	RCA-T-FRI
Wine	6.5	6.5	2	1	3.5	5	3.5
Iris	7	2.5	1	6	2.5	5	4
NewThyroid	6.5	5	1	2	3	6.5	4
Balance	6	7	3.5	1.5	5	3.5	1.5
Phoneme	5	5	1.5	3	1.5	5	7
Pima	7	6	4.5	1	4.5	2.5	2.5
Appendicitis	7	6	1.5	3.5	3.5	5	1.5
WDBC	7	6	2	1	3	5	4
Ionosphere	7	6	1	2	4	3	5
Sonar	7	5.5	1	2	4	5.5	3
Glass	6.5	6.5	3	1	2	5	4
Average	6.591	5.636	2	2.182	3.318	4.636	3.636

TABLE IV
ABSOLUTE DIFFERENCES OF AVERAGE RANKS BETWEEN METHODS

Datasets	T-FRI	WT-FRI	LMNN-T-FRI	ITML-T-FRI	SDML-T-FRI	LSML-T-FRI	RCA-T-FRI
T-FRI	0	0.955	**4.591**	**4.409**	**3.273**	1.955	**2.955**
WT-FRI	0.955	0	**3.636**	**3.454**	2.318	1	2
LMNN-T-FRI	**4.591**	**3.636**	0	0.182	1.318	2.636	1.636
ITML-T-FRI	**4.409**	**3.454**	0.182	0	1.136	2.454	1.454
SDML-T-FRI	**3.273**	2.318	1.318	1.136	0	1.318	0.318
LSML-T-FRI	1.955	1	2.636	2.454	1.318	0	1
RCA-T-FRI	**2.955**	2	1.636	1.454	0.318	1	0

* Figures bigger than CD are highlighted in bold type.

the corresponding two methods are considered being of a significant difference.

From Table IV, it can be seen that there are six pairs of algorithms of a difference larger than CD. This means that the test should reject the null hypothesis given a significant level of p=5%, while asserting with confidence that fair distinction exists in performance between each of these six pairs of algorithms. In particular, for the five algorithms implementing the proposed approach, all but LSML-T-FRI are strikingly different from the original T-FRI. Also, LMNN-T-FRI and ITML-T-FRI are dissimilar to WT-FRI. Amongst the five new methods themselves, whilst different metric learning methods are employed, their performances do not show any significant difference.

The above results are reinforced by Fig. 5, where the central dot and short line regarding each method show the corresponding average rank and CD range, respectively. If the lines of the two algorithms overlap, then they are not strongly distinct, and vice versa. This forms a further testimony to both the theoretical and empirical results attained previously.

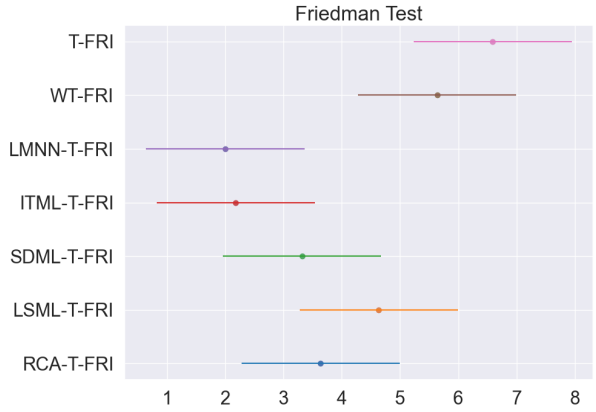

Fig. 5. Friedman test.

3) Efficiency: It is theoretically presumed that the proposed approach is as efficient as T-FRI, being more efficient than WT-FRI. This requires experimental confirmation. For this purpose, the time complexity of each of the algorithms concerned is assessed here. To reduce computational overloads, without loss of generality, five datasets from the previous 11 are randomly selected to carry out this verification. Fig. 6 presents the bar chart showing the results of running each method on these datasets with the time consumption averaged through 10×10-fold cross validation. It demonstrates that WT-FRI has a lower efficiency than the other six methods across all the five datasets. Forming sharp contrast with the computational cost of WT-FRI, that of each algorithm implementing the proposed approach is consistently comparable to that of

This article has been accepted for publication in IEEE Transactions on Fuzzy Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TFUZZ.2022.3194368

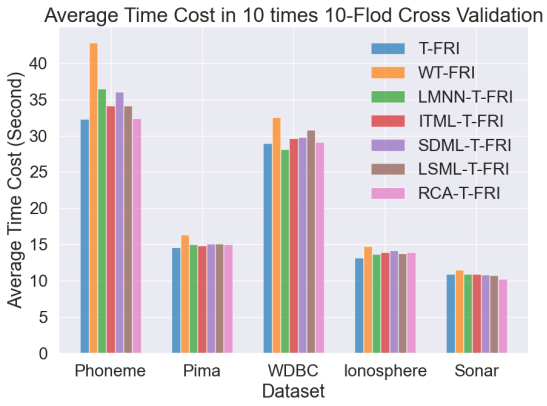

Fig. 6. Average time cost.

the original T-FRI.

4) Robust Performance against Dimensionality Reduction: From Fig. 6 it can also be seen that whilst the number of instances of the *WDBC* dataset is smaller than that of the *Pima* dataset, the average time cost for all tested T-FRI methods is much greater than that required for the *Pima* dataset. This is due to the higher dimensionality possessed by the *WDBC* dataset. Fortunately, as discussed previously (in Section III-B), when the number of features is large, the Mahalanobis distance can be computed on a lower-dimensional space by introducing a low-rank matrix $\mathcal{P}$ to transform the original rules, enabled by a dimensionality reduction scheme.

Owing to its popularity and availability, Local Fisher Discriminant Analysis (LFDA) is herein employed to address the issue of dimensionality reduction (mathematical details of this algorithm are beyond the scope of this paper but can be found in [39]). For this experimental investigation, only three datasets of a relatively large number of features are utilised (namely, *WDBC*, *Ionosphere* and *Sonar*). Fig. 7 depicts the average classification accuracies and time costs on these three datasets. This figure shows the achieved performance indices against a different (reduced) number of antecedent features via the use of LFDA. Obviously, with the use of LFDA, the corresponding time cost over these datasets drops sharply with the decrease of the number of features. However, the accuracy rate does not have a marked fall until the number of antecedent variables has reduced to a low figure, but it is still higher than that attainable by T-FRI or WT-FRI.

V. Conclusion

This paper has presented a novel transformation-based fuzzy rule interpolation (T-FRI) technique that considerably enhances the performance of fuzzy interpolative reasoning, by the use of Mahalanobis distance metric. The metric is introduced in the crucial step of choosing the closest rules neighbouring an unmatched observation to implement rule interpolation. A number of metric learning methods are addressed, each of which can be exploited to learn the required Mahalanobis matrix, indicating the flexibility of this approach. Additionally, Choquet integral is applied as the aggregation function to strengthen the performance of the underlying T-FRI method. This paper has provided both theoretical and experimental comparisons with the state-of-the-art T-FRI mechanisms, demonstrating the significant potential of the novel approach in terms of both accuracy and efficiency.

The clear benefits of utilising Mahalanobis distance gives rise to a question of whether other alternatives for distance metric (e.g., Cosine similarity [40], Bilinear similarity [41] and Histogram distance [42]) may be employed to bring similar positive improvements over the existing T-FRI methods. Also, more aggregation functions may be utilised to empower the interpolation process, such as Sugeno integral [43], Penalty functions [44], and modified Choquet Integral [45]. Therefore, much can be done to explore the possibility of further consolidating the efficacies of T-FRI. Whilst the present experimental investigations have covered statistical analyses over a wide range of datasets, practical investigations concerning complicated real-world problems such as network security [15] and medical diagnosis [46] remain as active research. Furthermore, the present approach is verified only against Mamdani style fuzzy models [47], studies of how this approach may be extended to addressing neuro-fuzzy models such as ANFIS [48] and TSK-type fuzzy models [49] form a piece of interesting future work. This would have the potential to strengthen the most recent development in performing approximate reasoning with such models [50].

This article has been accepted for publication in IEEE Transactions on Fuzzy Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TFUZZ.2022.3194368

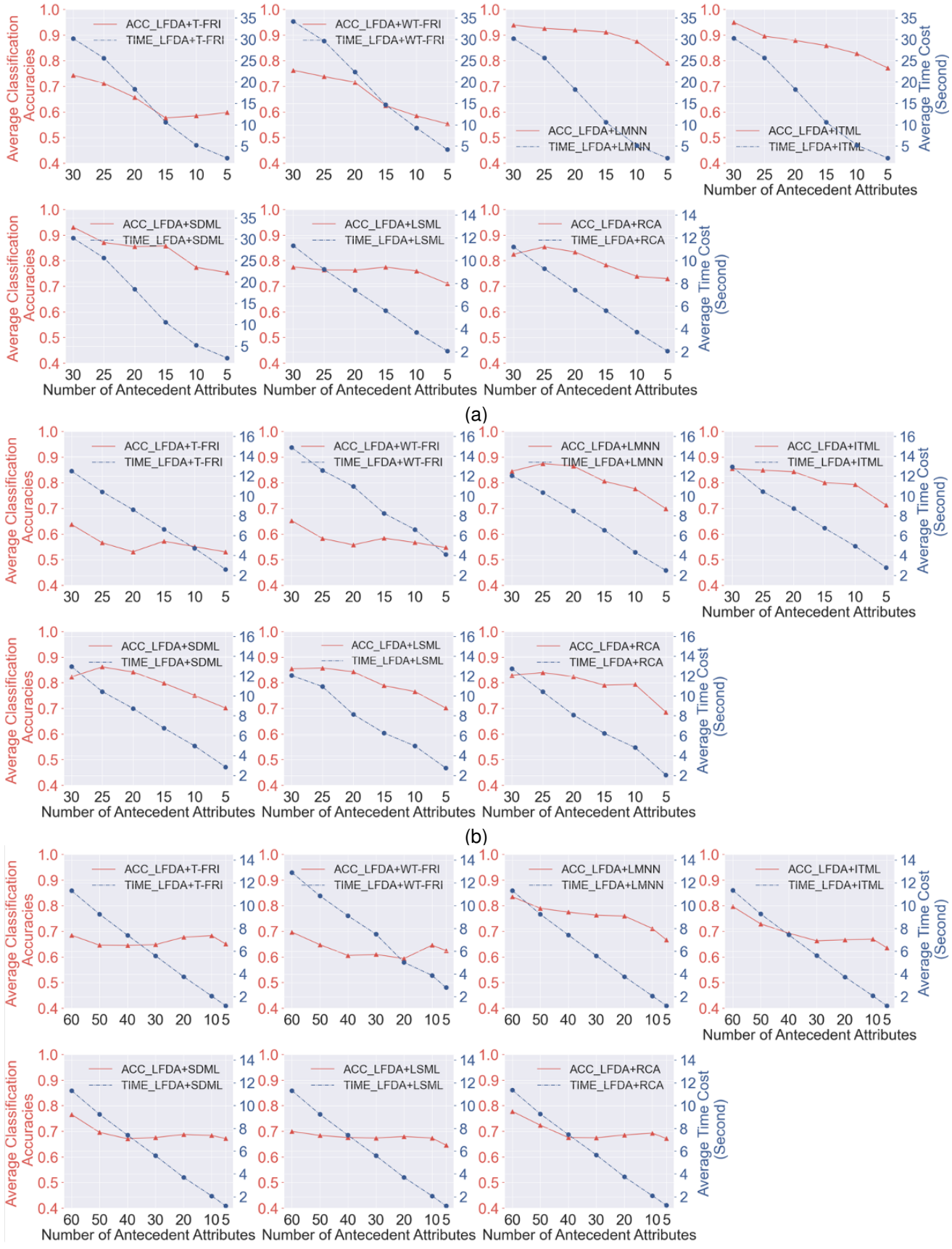

Fig. 7. Average classification accuracies and time costs via 10×10-fold cross validation: (a) WDBC. (b) Ionosphere. (c) Sonar.

REFERENCES

[1] F. Li, C. Shang, Y. Li, J. Yang, and Q. Shen, "Approximate reasoning with fuzzy rule interpolation: background and recent advances," *Artificial Intelligence Review*, vol. 54, pp. 4543–4590, 2021.

[2] L. A. Zadeh, "Outline of a new approach to the analysis of complex systems and decision processes," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-3, no. 1, pp. 28–44, 1973.

[3] L. T. Kóczy and K. Hirota, "Approximate reasoning by linear rule interpolation and general approximation," *International Journal of Approximate Reasoning*, vol. 9, no. 3, pp. 197–225, 1993.

[4] T. Chen, C. Shang, J. Yang, F. Li, and Q. Shen, "A new approach for transformation-based fuzzy rule interpolation," *IEEE Transactions on Fuzzy Systems*, vol. 28, no. 12, pp. 3330–3344, 2020.

[5] W. H. Hsiao, S. M. Chen, and C. H. Lee, "A new interpolative reasoning method in sparse rule-based systems," *Fuzzy Sets and Systems*, vol. 93, no. 1, pp. 17–22, 1998.

[6] P. Baranyi, L. T. Kóczy, and T. D. Gedeon, "A generalized concept for fuzzy rule interpolation," *IEEE Transactions on Fuzzy Systems*, vol. 12, no. 6, pp. 820–837, 2004.

[7] Z. Huang and Q. Shen, "Fuzzy interpolative reasoning via scale and move transformations," *IEEE Transactions on Fuzzy Systems*, vol. 14, no. 2, pp. 340–359, 2006.

[8] Y. Yam, M. L. Wong, and P. Baranyi, "Interpolation with function space representation of membership functions," *IEEE Transactions on Fuzzy Systems*, vol. 14, no. 3, pp. 398–411, 2006.

[9] Y. C. Chang, S. M. Chen, and C. J. Liau, "Fuzzy interpolative reasoning for sparse fuzzy-rule-based systems based on the areas of fuzzy sets," *IEEE Transactions on Fuzzy Systems*, vol. 16, no. 5, pp. 1285–1301, 2008.

[10] L. Yang, F. Chao, and Q. Shen, "Generalized adaptive fuzzy rule interpolation," *IEEE Transactions on Fuzzy Systems*, vol. 25, no. 4, pp. 839–853, 2017.

[11] S. M. Chen and Y. C. Chang, "Weighted fuzzy rule interpolation based on ga-based weight-learning techniques," *IEEE Transactions on Fuzzy Systems*, vol. 19, no. 4, pp. 729–744, 2011.

[12] S. M. Chen, Y. C. Chang, and J. S. Pan, "Fuzzy rules interpolation for sparse fuzzy rule-based systems based on interval type-2 gaussian fuzzy sets and genetic algorithms," *IEEE Transactions on Fuzzy Systems*, vol. 21, no. 3, pp. 412–425, 2013.

[13] S. M. Chen, S. H. Cheng, and Z. J. Chen, "Fuzzy interpolative reasoning based on the ratio of fuzziness of rough-fuzzy sets," *Information Sciences*, vol. 299, pp. 394–411, 2015.

[14] S. Jin, R. Diao, C. Quek, and Q. Shen, "Backward fuzzy rule interpolation," *IEEE Transactions on Fuzzy Systems*, vol. 22, no. 6, pp. 1682–1698, 2014.

[15] N. Naik, R. Diao, and Q. Shen, "Dynamic fuzzy rule interpolation and its application to intrusion detection," *IEEE Transactions on Fuzzy Systems*, vol. 26, no. 4, pp. 1878–1892, 2018.

[16] M. Dash and H. Liu, "Feature selection for classification," *Intelligent Data Analysis*, vol. 1, no. 1, pp. 131–156, 1997.

[17] F. Li, C. Shang, Y. Li, J. Yang, and Q. Shen, "Fuzzy rule based interpolative reasoning supported by attribute ranking," *IEEE Transactions on Fuzzy Systems*, vol. 26, no. 5, pp. 2758–2773, 2018.

[18] F. Li, C. Shang, Y. Li, J. Yang, and Q. Shen, "Interpolation with just two nearest neighboring weighted fuzzy rules," *IEEE Transactions on Fuzzy Systems*, vol. 28, no. 9, pp. 2255–2262, 2020.

[19] A. Bellet, A. Habrard, and M. Sebban, *Metric Learning*. Morgan & Claypool Publishers, 2015.

[20] R. De Maesschalck, D. Jouan-Rimbaud, and D. Massart, "The mahalanobis distance," *Chemometrics and Intelligent Laboratory Systems*, vol. 50, no. 1, pp. 1–18, 2000.

[21] E. Xing, A. Ng, M. Jordan, and S. J. Russell, "Distance metric learning with application to clustering with side-information," *Advances in Neural Information Processing Systems*, vol. 15, pp. 505–512, 2002.

[22] J. Mei, M. Liu, H. R. Karimi, and H. Gao, "Logdet divergence-based metric learning with triplet constraints and its applications," *IEEE Transactions on Image Processing*, vol. 23, no. 11, pp. 4920–4931, 2014.

[23] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *Journal of Machine Learning Research*, vol. 10, no. 2, pp. 207–244, 2009.

[24] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, "Information-theoretic metric learning," in *Proceedings of 24th International Conference on Machine learning*, 2007, pp. 209–216.

[25] L. Bregman, "The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming," *USSR Computational Mathematics and Mathematical Physics*, vol. 7, no. 3, pp. 200–217, 1967.

[26] G. J. Qi, J. Tang, Z. J. Zha, T. S. Chua, and H. J. Zhang, "An efficient sparse metric learning in high-dimensional space via l 1-penalized log-determinant regularization," in *Proceedings of 26th International Conference on Machine Learning*, 2009, pp. 841–848.

[27] E. Y. Liu, Z. Guo, X. Zhang, V. Jojic, and W. Wang, "Metric learning from relative comparisons by minimizing squared residual," in *Proceedings of 12th International Conference on Data Mining*, 2012, pp. 978–983.

[28] N. Shental, T. Hertz, D. Weinshall, and M. Pavel, "Adjustment learning and relevant component analysis," in *European Conference on Computer Vision*. Springer, 2002, pp. 776–790.

[29] G. Beliakov, A. Pradera, T. Calvo *et al.*, *Aggregation functions: A guide for practitioners*. Springer, 2007, vol. 221.

[30] J. Alcalá-Fdez, A. Fernández, J. Luengo, J. Derrac, S. García, L. Sánchez, and F. Herrera, "Keel data-mining software tool: data set repository, integration of algorithms and experimental analysis framework." *Multiple-Valued Logic & Soft Computing*, vol. 17, 2011.

[31] M. Grabisch, J. L. Marichal, R. Mesiar, and E. Pap, *Aggregation functions*. Cambridge University Press, 2009, vol. 127.

[32] G. Lucca, J. A. Sanz, G. P. Dimuro, B. Bedregal, R. Mesiar, A. Kolesárová, and H. Bustince, "Preaggregation functions: Construction and an application," *IEEE Transactions on Fuzzy Systems*, vol. 24, no. 2, pp. 260–272, 2016.

[33] Z. Chi, H. Yan, and T. Pham, *Fuzzy algorithms: with applications to image processing and pattern recognition*, 1996, vol. 10.

[34] L. X. Wang and J. Mendel, "Generating fuzzy rules by learning from examples," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 22, no. 6, pp. 1414–1427, 1992.

[35] T. Chen, C. Shang, P. Su, and Q. Shen, "Induction of accurate and interpretable fuzzy rules from preliminary crisp representation," *Knowledge-Based Systems*, vol. 146, pp. 152–166, 2018.

[36] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *The Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.

[37] M. Friedman, "The use of ranks to avoid the assumption of normality implicit in the analysis of variance," *Journal of the American Statistical Association*, vol. 32, no. 200, pp. 675–701, 1937.

[38] P. B. Nemenyi, *Distribution-free multiple comparisons*. Princeton University, 1963.

[39] M. Sugiyama, "Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis." *Journal of Machine Learning Research*, vol. 8, no. 5, pp. 1027–1061, 2007.

[40] R. Baeza-Yates, B. Ribeiro-Neto *et al.*, *Modern information retrieval*. ACM Press New York, 1999, vol. 463.

[41] G. Chechik, V. Sharma, U. Shalit, and S. Bengio, "Large scale online learning of image similarity through ranking." *Journal of Machine Learning Research*, vol. 11, no. 3, 2010.

[42] J. Hafner, H. Sawhney, W. Equitz, M. Flickner, and W. Niblack, "Efficient color histogram indexing for quadratic form distance functions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, no. 7, pp. 729–736, 1995.

[43] M. Sugeno, "Fuzzy measure and fuzzy integral," *Transactions on Instrument and Control Engineers*, vol. 8, no. 2, pp. 218–226, 1972.

[44] H. Bustince, G. Beliakov, G. Pereira Dimuro, B. Bedregal, and R. Mesiar, "On the definition of penalty functions in data aggregation," *Fuzzy Sets and Systems*, vol. 323, pp. 1–18, 2017.

[45] G. P. Dimuro, G. Lucca, B. Bedregal, R. Mesiar, J. A. Sanz, C.-T. Lin, and H. Bustince, "Generalized cf1f2-integrals: From choquet-like aggregation to ordered directionally monotone functions," *Fuzzy Sets and Systems*, vol. 378, pp. 44–67, 2020.

[46] F. Li, C. Shang, Y. Li, and Q. Shen, "Interpretable mammographic mass classification with fuzzy interpolative reasoning," *Knowledge-Based Systems*, vol. 191, p. 105279, 2020.

[47] E. Mamdani and S. Assilian, "An experiment in linguistic synthesis with a fuzzy logic controller," *International Journal of Man-Machine Studies*, vol. 7, no. 1, pp. 1–13, 1975.

[48] J.-S. R. Jang, "Anfis: adaptive-network-based fuzzy inference system," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 23, no. 3, pp. 665–685, 1993.

[49] T. Takagi and M. Sugeno, "Fuzzy identification of systems and its applications to modeling and control," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-15, no. 1, pp. 116–132, 1985.

[50] J. Yang, C. Shang, Y. Li, F. Li, L. Shen, and Q. Shen, "Constructing anfis with sparse data through group-based rule interpolation: An evolutionary approach," *IEEE Transactions on Fuzzy Systems*, vol. 30, no. 4, pp. 893–907, 2022.