## *Interpretable mammographic mass classification with fuzzy interpolative reasoning*

Li, Fangyi; Shang, Changjing; Li, Ying; Shen, Qiang

*Published in:*
Knowledge-Based Systems

*DOI:*
10.1016/j.knosys.2019.105279

*Publication date:*
2020

*Citation for published version (APA):*
Li, F., Shang, C., Li, Y., & Shen, Q. (2020). Interpretable mammographic mass classification with fuzzy interpolative reasoning. *Knowledge-Based Systems*, *191*, Article 105279. https://doi.org/10.1016/j.knosys.2019.105279

# Interpretable Mammographic Mass Classification with Fuzzy Interpolative Reasoning

Fangyi Li[a,b], Changjing Shang[b], Ying Li[a], Qiang Shen[b,*]

[a]*School of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an, P.R. China, 710072*
[b]*Department of Computer Science, Faculty of Business and Physical Sciences, Aberystwyth University, Aberystwyth, UK, SY23 3DB*

**Abstract**

Breast mass cancer remains a great challenge for developing advanced computer-aided diagnosis (CADx) systems, to assist medical professionals for the determination of benignancy or malignancy of masses. This paper presents a novel approach to building fuzzy rule-based CADx systems for mass classification of mammographic images, via the use of weighted fuzzy rule interpolation. It describes an integrated implementation of such a classification system that ensures interpretable classification of masses through firing the rules that match given observations, while having the capability of classifying unmatched observations through fuzzy rule interpolation (FRI). In particular, a feature weight-guided FRI scheme is exploited to enable such inference. The work is implemented through integrating feature weights with a popular scale and move transformation-based FRI, with the individual feature weights derived from feature selection as a preprocessing process. The efficacy of the proposed CADx system is systematically evaluated using two real-world mammographic image datasets, demonstrating its explicit interpretability and potential classification performance.

*Keywords:* Mammographic mass classification, Fuzzy rule-based system, Weighted interpolative reasoning, Inference interpretability

*Corresponding author
*Email addresses:* fal2@aber.ac.uk (Fangyi Li), cns@aber.ac.uk (Changjing Shang), lybyp@nwpu.edu.cn (Ying Li), qqs@aber.ac.uk (Qiang Shen)

## 1. Introduction

Breast cancer is one of the severest threats for women around the world. Early detection of breast lesions has been shown to provide an essential means to reduce the possibility of deterioration of patients' health conditions or even death. Amongst various tools available, mammography screening offers a particularly popular technique for identifying the presence of abnormalities in breasts. As a result, mammographic images are produced, in the form of films or more advanced recently, in that of full field digital mammograms, which are helpful to effectively detect and diagnose breast cancer by medical professionals.

Mass and microcalcification are two important early signs of abnormalities for detecting developing breast cancer, which are normally present in mammographic images. Masses are often indistinguishable from the surrounding parenchymal, resulting in more significant challenges for mass detection and classification. In general, an abnormal mass can be categorised into either benign or malignant. For instance, the standardised Breast Imaging Reporting and Data System (BI-RADS) [1] characterises masses for determination of benign or malignant in terms of their shapes, margins and densities. This reflects how radiologists visualise the mammographic images for diagnosis. Benign masses are frequently found to be in round or oval shapes, having well-defined margins and low densities, whilst malignant masses are more likely in irregular shapes and have spicule margins with relatively high densities.

Reading mammograms is a very demanding task for radiologists, and the determination of whether an image shows a benign mass or malignant may be affected by the experience and subjective criteria of a certain radiologist who handles a given case. The development of Computer-Aided Diagnosis (CADx) techniques plays an effective supporting role in assisting medical professionals in the interpretation of medical images. Especially, a combination of using a CADx system and exploiting human expertise directly would greatly improve diagnostic accuracy and efficiency. A number of CADx systems have been studied and

applied to support mammographic abnormality diagnosis (e.g., [2, 3, 4, 5, 6, 7]). Most developed techniques can be referred to in the recent survey of such research in [8, 9, 10].

Existing computational techniques may provide a second opinion for mammographic mass diagnosis, by dealing with the mammograms using pathological related knowledge. In general, most CADx systems for mammogram mass classification build their structures by following a number of key phases, including: image preprocessing, region of interest (ROI) extraction, mass segmentation, feature extraction and selection, and class determination. Various image features have been found in the literature for characterising mass properties, such as traditional features in terms of intensity, morphology, texture, etc. and features generated from advanced computational mechanisms like deep neutral networks [11]. Morphological (aka. geometric) features are one of the most common types used to discriminate mammographic masses [12], typically extracted to represent the shape and boundary characteristics of masses. They are commonly adopted to support precise mass segmentation carried out by radiologists or CADx systems. This is because such features depict what radiologists visualise a mass lesion, which are essential to enable subsequent interpretation of the classification or diagnostic outcome.

From medical viewpoint, interpreting mammogram masses visually is a very demanding task for radiologists. It would therefore be a great assistance to be able to produce interpretable diagnoses from any CADx system in use. Recently, efforts have been made for improving accuracy of CADx systems for mammogram classification, such as those achieved by deep convolutional neural networks (DCNNs, e.g., [13, 11, 14]), which have been seen to make great progress in meeting the visual recognition challenges. In such work, informative features are extracted to generate potential explanations for mammogram classification, by visually showing the edge of mass in saliency maps for example. However, to ensure interpretable feature representations requires the annotations of radiologists (or other alternative means) to correlate the DCNNs features with radiological features that reflect clinically relevant phenomena.

This makes the interpretation progress and hence, the entire diagnostic system more complicated. It remains a difficult problem to discover clinically explainable interpretations for machine learning-based CADx systems.

The question now is what intelligent classification methods can be better developed to facilitate the use of semantics-rich geometric mass features, in an effort to enhance CADx systems' explainability explicitly. Fuzzy rule-based systems are known to be able to simulate human reasoning in decision support. Inference made by firing fuzzy if-then rules can be readily interpreted by human users. Such systems provide an effective tool to deal with the impreciseness and vagueness commonly incurred in real-world problems, including the description of mammographic mass characteristics. Fuzzy rule-based techniques therefore, have a natural appeal in establishing a CADx system for mammographic mass diagnosis. For example, Adaptive Neuro-Fuzzy Inference System (ANFIS) has been applied to classifying normal/abnormal mammograms, as well as to determining abnormal severity [15]. Also, the classical Compositional Rule of Inference (CRI) [16] has been employed to perform mammogram diagnostic reasoning (e.g., classifying mammogram mass lesions into the well-known BI-RADS shape categories) [17, 7].

Little work exists to explicitly interpret radiological phenomena of mass lesions in mammograms with the use of fuzzy rules, however. In addition, there may not be sufficient mammographic image data to enable the full exploitation of traditional fuzzy systems to perform required diagnostic tasks. As such, a fuzzy rule base inducted from the data may not cover the entire problem domain, resulting in the situations where certain observations can not match any of the rules in the rule base, thereby deriving no or wrong conclusions [17]. Fuzzy interpolative reasoning through fuzzy rule interpolation (FRI) can help to deal with exactly such sparse knowledge-based problems [18, 19, 20, 21, 22, 23, 24, 25]. The efficacy of classical FRI techniques have been significantly strengthened with the recent advances in the literature, e.g., by the so-called weighted FRI approach [26, 27, 28], which no longer imposes the constraint that the rule antecedent features are of equal significance in decision-making. Instead, fea-

tures are ranked with their relative weights exploited in the procedures of a conventional FRI method (e.g., the scale and move transformation-based FRI, T-FRI [29]). The resultant techniques have been successfully applied in tackling classification and prediction problems, inspiring the development reported herein.

This paper presents two key contributions to the relevant literature: 1) an implemented fuzzy rule-based inference system for mass classification in mammograms, where fuzzy interpolative reasoning is embedded for the first time in a CADx system (for coping with sparse rule bases), supported by feature weight-guided FRI; and 2) an explicit explanation output from the CADx system, in the form of clinically interpretable rules using features of doamin semantics, thereby providing a "second opinion" for assisting radiologists to read mammograms. The remainder of this paper is organised as follows. Section 2 describes the mammographic image data addressed in this work. Section 3 presents the fuzzy rule-based interpolative reasoning system for mammographic mass classification. Section 4 provides experimental evaluation of the implemented system with systematically statistical comparisons. Finally, Section 5 concludes the paper and points out issues for further research.

## 2. Databases

The benchmark mammographic image datasets used in this work are adopted from the Breast Cancer Digital Repository (BCDR) [30]. It is a wide-ranging and comprehensively annotated public database for mammographic disease study, especially for the development of breast cancer CADx techniques and for training medical physicians involved in the diagnostic, treatment or research of breast cancer and associated technologies. This repository is continuously being enriched and currently, contains cases of 1734 patients with mammography and ultrasound images, clinical history, lesion segmentation and selected pre-computed image-based descriptors.

BCDR consists of two different types of sub-repository: 1) a digitalised film

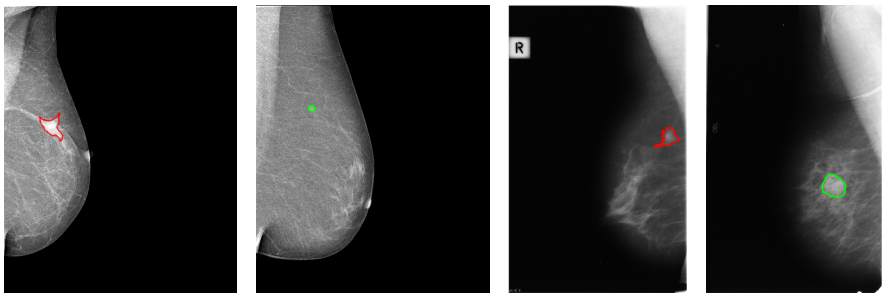
mammography (FM)-based repository, and 2) a full field digital mammography (DM)-based repository. Both FM and DM repositories are divided into several sub-datasets including different number of cases, which form a common ground for fair comparison between various CADx systems for mammographic disease analysis. As with other established mammographic databases, digitalised film mammogram images have rather lower resolution whilst full field digital mammogram images are much more common nowadays (because of their higher spatial resolution and permitting more image manipulation to enable better visualisation). Without biases, the present work takes samples from both FM and DM sub-datasets, containing the following types of mass:

- BCDR-D01: comprised of 79 biopsy-proven lesions of 64 women, rendering 141 segmentations. All of them present suspicious mass, of which 85 are benign and 56 are malignant. Each image is a grey level mammogram in 14 bits with a resolution of 3328×4084 pixels.
- BCDR-F01: comprised of 200 biopsy-proven lesions of 190 women, rendering 362 segmentations, with mass lesions occurring in 231 segmented images where the number of benign and malignant masses are 112 and 119, respectively. Each image is a grey level digitalised mammogram in 8 bits with a resolution of 720×1168 pixels.

Note that each mammogram image considered has a precise segmentation of identified lesion. In particular, the contour of mass is manually annotated by medical specialists. Fig. 1 shows examples of benign and malignant mass lesions with respective mass segmentations, taken from each of the two datasets.

## 3. Fuzzy Rule-based Interpolative Classifier

This section details the design and implementation of a rule-based system that works through the assistance of fuzzy interpolative reasoning, for classifying mammographic mass in mammogram images.




(a) Malignant mass in BCDR-D01 (b) Benign mass in BCDR-D01 (c) Malignant mass in BCDR-F01 (d) Benign mass in BCDR-F01

Figure 1: Samples of mass lesions with mass contours annotated.

### *3.1. System Framework*

The workflow of the entire diagnostic system is specified as illustrated in Fig. 2. The general working process is as follows. Having identified a general region of interest (ROI) and segmented mass lesion from a given original mammogram image, a set of potentially descriptive features are extracted for characterising the properties of the image (particularly regarding the geometric shape, margin, density of mass lesion). The resulting mass features are evaluated by a feature ranking method, of two-fold objectives: 1) selection of more informative top features, and 2) assignment of weights to those selected ones in terms of their relative ranking scores. A fuzzy semantic rule base is generated from the given image database through the use of selected mass features as rule conditionals, by employing a certain standard fuzzy rule induction method (which is beyond the scope of this paper).

Following the aforementioned preparation, the primary work for mass classification is highlighted in the dashed box in Fig. 2. In particular, when a novel observed mass is present (represented with selected features) it is regarded as a new observation to be checked against the rules within the rule base. If it is matched by any existing rule, the rule is fired by the use of conventional compositional rule of inference (CRI). If there is no rule matching the observation, weighted fuzzy rule interpolation (where T-FRI is used for implementation in

this work, though others may be used as an alternative [28]) to perform interpolative reasoning, estimating the benignancy or malignancy of the given mass. Technical details are provided in the following.

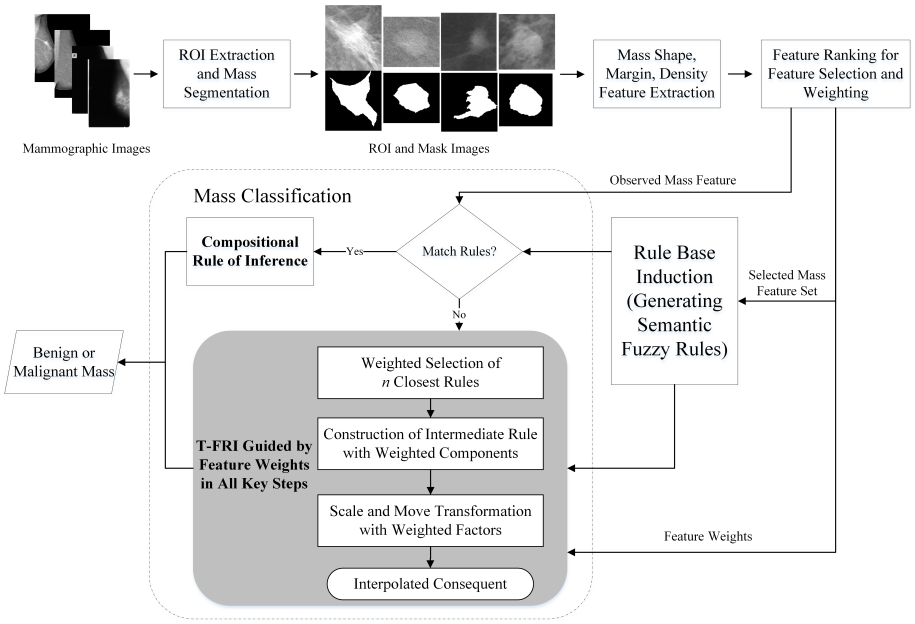
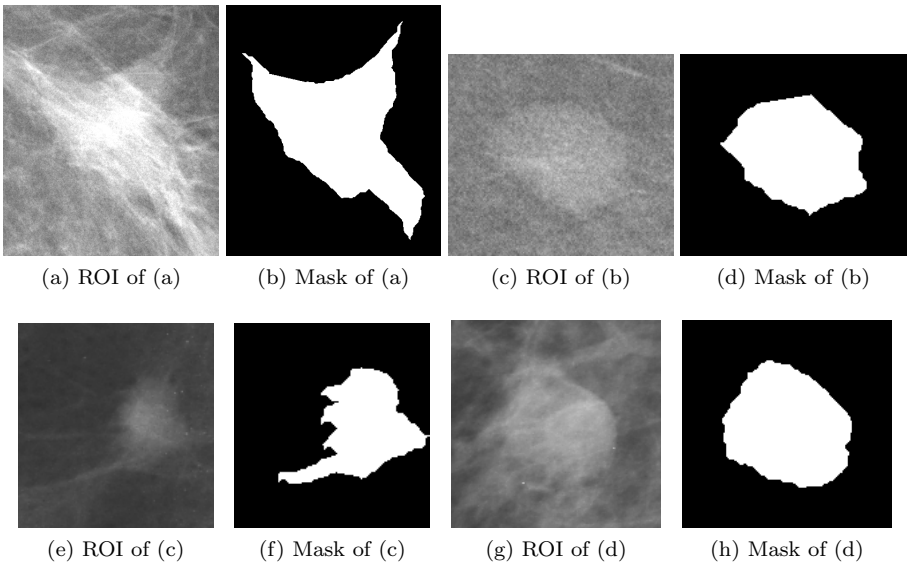



Figure 2: Fuzzy interpolative reasoning for mammographic mass classification.

### *3.2. ROI Extraction and Mass Segmentation*

In BCDR, each mammogram is associated with a precise segmentation of the underlying mass lesion. Since the focus of this work is on mass classification, the available contours of masses are adopted here for generating the ROI image and subsequently, the mass-segmented mask image of each given mammogram. These two images are chopped from the original mammogram, such that the observed mass locates in the centre. The resultant images consolidate the basis upon which to extract features in terms of mass shapes, margins and densities. Fig 3 shows examples of the ROI and mass-segmented mask images as per those mammogram samples displayed in Fig 1.



(a) ROI of (a) (b) Mask of (a) (c) ROI of (b) (d) Mask of (b)

(e) ROI of (c) (f) Mask of (c) (g) ROI of (d) (h) Mask of (d)

Figure 3: ROI and mass-segmented mask images of mass samples given in Fig. 1.

### *3.3. Mass Feature Extraction and Ranking*

Given the ROI image and mass-segmented image of a mammogram, a set of features are extracted for characterising mass lesion in terms of the image properties such as mass shape, margin and density. Generally, the benign masses are frequently found to be in round or oval shapes, having well-defined margins and low densities, while the malignant masses are more likely in irregular shapes and have spicule margins with relatively high densities. Inspired by this observation, in this work, a total of 18 features are taken as the possible ones to distinguish benign and malignant masses, as listed in Table 1. This intuitive approach is based on the understanding of medical professionals practice, in that these two types of mass are often differentiated from their geometrical shape and boundary as well as density.

Benign and malignant masses may be found in rather different shapes. To reflect this viewpoint, six geometry features are extracted from the mask images of mass, including: mass area (F1), mass perimeter (F2), circularity measure (F3), convexity measure (F4), mass eccentricity (F5) and dispersion (F6). In

particular, area and perimeter are basic shape descriptors for measuring the size of a mass. The features F3-F6 are metrics which define the morphological characteristics of masses in different shapes, potentially helpful to differentiate masses of regular shape from those of irregular, and to quantify the circularity and ellipticity of regular masses.

The margin of a mass offers another view for depicting the geometric properties of masses. Margin features can be grouped in two sub-categories. One is used to determine the degree of boundary roughness. Herein, five normalised radial length (NRL)-based statistical features (F7-F11) and compactness measure (F12) are employed to cover this aspect. The other group is to quantify the sharpness of margin intensity, with three margin gradient features (F13-F15) adopted to measure the pixel intensity variations over the boundaries of masses.

Mass shape and margin features characterise the morphological properties of mass regions, while the density features of mass reveal the intensity of mass region compared against its surrounding tissue. The last three features are therefore adopted to exploit the pixel intensity within a mass involved in the ROI images. In particular, the features F16 and F17 are computed with respect to the statistics relevant to the moments which measure the intensity of suspicious mass region. The contrast measure (F18) is the difference between the average grey level of the ROI and that of the surrounding region, evaluating the intensity variation within masses in contrast to that outside.

Note that there may exist redundant features among the extracted combinatorial properties of mass shape, margin and density. Obviously, such redundancy should be removed, not only to improve the performance of classifier (via the use of less features gaining efficiency and the reduction of measurement noise gaining effectiveness), but also to enhance interpretability of the diagnostic system (with less complex rules). In this work, a feature ranking mechanism taken from the core of the popular Relief-F algorithm [34] is employed to evaluate individual mass features. This results in a set of scores that indicate the relative importance of each feature in the determination of benign and malignant mass. Intuitively, those features which have relatively lower scores may have poorer

Table 1: Mass Features in Different Category for Characterising Mass Lesion

| Mass Features | | Physical Meaning |
|---|---|---|
| Shape | Area (F1) [8, 31]<br>Perimeter (F2) [8, 2] | Size<br>Small values indicate small mass lesion |
| | Circularity (F3) [8, 32] | Degree of roundness/circularity<br>F3=1 for a circular mass and less than 1 for mass that departs from circularity |
| | Convexity (F4) [8] | Relative amount that an object differs from a convex object<br>F4=1 for convex mass (as with many benign masses) and less than 1 for nonconvex mass (as with many spiculated or malignant masses) |
| | Eccentricity (F5) [31, 17] | Degree of ellipticity<br>Small values for circle-like ellipse and large for line segment-like ellipse |
| | Dispersion (F6) [31, 17] | Degree of irregularity (Density of region)<br>Small values indicate regular masses while large values for irregular masses |
| Margin | Statistical normalised radial length (NRL) features (F7-F11) [8, 32]:<br>mean, SD, entropy, area ratio, zero-crossing count | Degree of boundary roughness |
| | Compactness (F12) [31, 33] | Small values indicate smooth contour (as with benign masses) |
| | Margin statistical gradient features (F13-F15) [2]:<br>mean, SD, entropy | Intensity variations across the boundaries of mass<br>Small values indicate flat edges while large values for sharp boundary |
| Density | Mass Intensity Mean (F16) [2, 8] | Average intensity value inside mass<br>Small values indicate low density mass |
| | Mass Intensity Standard Deviation (F17) [8] | Intensity variation inside mass<br>Small values indicate little intensity variation within mass |
| | Contrast measure of ROIs (F18) [8, 32] | Intensity variation between inside and outside of mass<br>Small values indicate low density contrast |

capability in the discrimination of different classes, and thus a subset of features are selected whose score values are higher than the average. The average score is herein utilised in order to ensure the process is automated; otherwise, if desirable, a pre-defined threshold may be used for the removal of low-ranking features.

Without losing generality, suppose that there are $m$ ($m \leq 18$) features being selected, each of which has a ranking score $RS_i, i = 1, 2, \ldots, m$. These different score values can then be normalised as weights associated with each of the selected individual features, as follows:

$$W_i = \frac{RS_i}{\sum_{t=1,\ldots,m} RS_t} \tag{1}$$

Given their underlying definition, the resulting normalised ranking scores have a natural appeal to be interpreted as the relative significance degrees of the contribution that a remaining feature may make to the decision, regarding the benignancy or malignancy of the mass. Such weights will also be utilised to guide the fuzzy rule-based interpolative inference system for mass classification as to be discussed later.

### *3.4. Generation of Fuzzy Classification Rules*

Having represented mass lesions in mammograms with selected mass shape, margin and density features, fuzzy rules for mass classification can be generated from given images whose decision classes are known. More specifically, the fuzzy rule base for mass classification consists of fuzzy *if-then* rules whose antecedent attributes are the ranked mass features selected in Section 3.3 and consequent attribute is the corresponding mass lesion type (i.e., Benign or Malignant).

The fuzzy values for each antecedent feature are fuzzified linguistic terms, which are defined in terms of the physical meaning of the underlying mass features (that are given in Table 1). Different values of the numerical metrics defining the features indicate different properties of a certain mass (including: shape, margin and density). Generally, the linguistic terms describing the fea-

tures can be given in order, such as "*..., Small ,..., Medium ,..., Large ,...*". Table 2 lists the linguistic values used in this work, mimicking the terms used by the medical professionals in the field concerned. From this definition, a fuzzy rule base is inducted from the extracted feature data, by promoting any hype-grid delimited by the fuzzy feature values that is hit by at least one given data. Note that any standard fuzzy rule induction method may be employed to create the rules, which is not the focus of this work. Unless stated otherwise, rules are herein learned from the selected mass features based on the use of the classical method of [35].

A possible rule, for example, from the learned rule base may be represented such that

If *Area* is *Small* and ... *Circularity* is *Large* and ... *NRL zero-crossing* is *Small* and *Margin gradient mean* is *Large* and ... *Density contrast* is *Small*, then *Mass* is *Benign*.

From the underlying semantics of the morphological and density features, this rule can be directly mapped onto the following, using the linguistic terms given in Table 2:

If *Mass* is *Small* and ... *Mass shape* is *Very Like Circular* and ... *Mass margin* is *Smooth* and *Margin* is *Circumscribed* and ... *Mass density contrast (between inside and outside mass)* is *Low*, then *Mass* is *Benign*.

Using fuzzy rules like the above helps facilitate the understanding of any conclusion drawn regarding whether a new mammogram stands for a benign or malignant mass, through the use of the fuzzy interpolative reasoning system as described next.

### *3.5. Feature Weight-Guided Interpolative Reasoning*

When a new observation is present, in terms of a set of measured feature values (representing an unknown mass lesion), all rules in the rule base are used to match against it in order to derive a diagnostic conclusion. However, the rule base learned from previously given data may be sparse, especially when only limited source data (or classified medical mammographic images) are available.

Table 2: Gradual Linguistic Terms Defined for Mass Features as Numerical Values Vary from *Small* to *Large*

| Mass Features | | Linguistic Terms |
|---|---|---|
| Shape | Area (F1)<br>Perimeter (F2) | Small,..., Medium,..., Large (mass) |
| | Circularity (F3) | Unlike,..., Less Like,..., Very Like (circular mass) |
| | Convexity (F4) | Unlike,..., Less Like,..., Very Like (convex regular mass) |
| | Eccentricity (F5) | Very Like,..., Less Like,..., Unlike (oval mass) |
| | Dispersion (F6) | Regular,..., Less Regular,..., Irregular (mass) |
| Margin | Statistical normalised radial length (NRL) features (F7-F11)<br>Compactness (F12) | Smooth,..., Slight undulated,..., Irregular (contours) |
| | Margin statistical gradient features (F13-F15) | Obscured,..., Blurred,..., Circumscribed (Well-/ Sharply-Defined) (margins) |
| Density | Mass Intensity Mean (F16) | Low,..., Isodense,..., High (mass density) |
| | Mass Intensity Standard Deviation (F17) | Low,..., Medium,..., High (mass density variation) |
| | Contrast measure of ROIs (F18) | Low,..., Isodense,..., High (density contrast) |

Thus, checking against all the available rules cannot fully cover the entire problem domain. That is, there exist situations where no rules can be found that match the new observation, leading to no conclusion to be drawn. To enable approximate inference on the unmatched observation, FRI is utilised. In this work, the recently developed feature weighted FRI is adapted to implement the required interpolative reasoning for mass classification. This adaptation is the first practical application of weighted FRI techniques.

For a formal illustration of the feature weighted FRI approach, without losing generality, suppose that a (sparse) fuzzy rule base $R = \{r^1, r^2, \ldots, r^N\}$ has been learned, each involving multiple antecedent attributes. Also, suppose that through feature ranking each attribute is now associated with a weight. Thus, a certain rule $r^i$ may be represented as follows:

$r^i$ : If $a_1(W_1)$ is $A_1^i$ and $a_2(W_2)$ is $A_2^i$ and $\cdots$ and $a_m(W_m)$ is $A_m^i$, then $z$ is $B^i$

where $a_j, j = 1, 2, \ldots, m$ are the rule antecedent attributes or features; $W_j$ are the weights of these features; $z$ is the consequent attribute; $A_j^i$ denotes the fuzzy set value taken by $a_j$ in the rule $r^i$; and $B^i$ represents the fuzzy set value of the consequent attribute $z$ in $r^i$. In addition, let an observation $o^*$ be represented by

$o^*$ : $a_1$ is $A_1^*$ and $a_2$ is $A_2^*$ and $\cdots$ and $a_m$ is $A_m^*$ where $A_j^*$ denotes the observed value for the feature $a_j$. In practice, each of such observed feature value may be a real number, but for generality, in this work it is assumed to be a fuzzy value that is fuzzified from the underlying real number, using a conventional fuzzification method.

For simplicity, triangular membership functions are employed for implementing the CADx system in this work, with each fuzzy set represented by three characteristic points (CPs). Namely, the fuzzy values $A_j^i$, $A_j^*$, $B^i$ appearing in the rule antecedents and observations, as well as the consequent $B^*$ to be computed $(i = 1, 2, \ldots, N, j = 1, 2, \ldots, m)$ are expressed as $(a_{j1}^i, a_{j2}^i, a_{j3}^i)$, $(a_{j1}^*, a_{j2}^*, a_{j3}^*)$, $(b_1^i, b_2^i, b_3^i)$, and $(b_1^*, b_2^*, b_3^*)$, respectively. Within each triangular membership function, the first and third CP stand for the two extreme points of the support

with a membership value of 0 and the middle one stands for the normal point of the fuzzy set with a membership of 1.

Using the above notations, the algorithm for feature weight-guided FRI can be summarised as follows.

*3.5.1. Step 1: Weighted Selection of Closest Rules for Interpolation*

Any FRI process starts as an observation $o^*$ being newly presented to the fuzzy system does not activate any rule in the sparse rule base, due to no matching (or in certain FRI-based systems, due to too low level a partial matching). Then, for interpolation, $n$ ($n \geq 2$) rules closest to the observation are sought in order to implement the interpolation. The similarity measure defined for such rule selection is based on weighted aggregation of distances between individual antecedent attributes of a given rule and their corresponding values in the observation. This takes into consideration of the individual antecedent weights evaluated from the feature ranking scores, leading to the selection of rules which have one of the first $n$ minimal distances to the observation.

Euclidean distance metric is typically used, as defined below:

$$d(o^*, r^i, W) = \frac{1}{\sqrt{\sum_{t=1}^{m}(1-W_t)^2}} \sqrt{\sum_{j=1}^{m} \left((1-W_j) d(A_j^*, A_j^i)\right)^2} \tag{2}$$

with $d(A_j^*, A_j^i)$ being computed via the representative value [20] such that

$$d(A_j^*, A_j^i) = \frac{\left| Rep(A_j^*) - Rep(A_j^i) \right|}{max_{A_j} - min_{A_j}} \tag{3}$$

where $d(A_j^*, A_j^i)$ represents the normalised result of the otherwise absolute distance; $max_{A_j}$ and $min_{A_j}$ denote the maximal and minimal value of the attribute $a_j$, respectively; and $m$ is the number of all antecedent attributes involved in all the given rules. As indicated previously, triangular membership functions are used throughout and therefore, the representative value of a fuzzy set $(a_1, a_2, a_3)$ may be simply calculated by averaging the CPs of the triangular fuzzy set, such

that

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

In this work, the number of closest rules, $n$ is set to 2 for conducting rule interpolation. This is supported by the recent studies [28] in that the adoption of the least number of closest rules (i.e., $n = 2$) is able to achieve a superior performance for feature weighted T-FRI. Such a set up normally has a high classification accuracy while saving computational costs.

*3.5.2. Step 2: Construction of Intermediate Fuzzy Rule*

From the preceding procedure, two closest rules to a given observation are chosen (which are of the shortest distances amongst all the rules to the observation). Using these two rules, an intermediate fuzzy rule $r'$ is constructed, forming the starting point of the transformation process of T-FRI. In implementation, the construction procedure computes the antecedent fuzzy sets $A'_j, j = 1, \ldots, m$ and the corresponding consequent fuzzy set $Z'$ of the intermediate rule:

$r'$ : If $a_1$ is $A'_1$ and $a_2$ is $A'_2$ and $\cdots$ and $a_m$ is $A'_m$, then $z$ is $Z'$

which is a weighted aggregation of the two selected closest rules.

Let $w_j^i, i \in \{1, 2\}$, denote the weight to which the $j$th antecedent of the $i$th fuzzy rule contributes to the construction of the $j$th antecedent $A'_j$ of the intermediate fuzzy rule:

$$w_j^i = \frac{1}{1 + d(A_j^i, A_j^*)} \tag{5}$$

where $d(A_j^i, A_j^*)$ is calculated as per Eqn. (3). Then,

$$A'_j = \sum_{i=1,2} \hat{w_j^i} A_j^i \tag{6}$$

where $\hat{w_j^i}$ is the normalised weight defined by

$$\hat{w_j^i} = \frac{w_j^i}{\sum_{t=1,2} w_j^t} \tag{7}$$

The consequent value of the intermediate rule is constructed in the same manner as above:

$$Z' = \sum_{i=1,2} \hat{w_z^i} Z^i \tag{8}$$

which is the weighted aggregation of the consequent values of the two closest rules $Z^i, i = 1, 2$, where $\hat{w_z^i}$ is the weighted average of those weights associated with the antecedents $\hat{w_j^i}$ in all rules, by the use of the feature weights ($W_j, j = 1, \ldots, m$):

$$\hat{w_z^i} = \sum_{j=1}^{m} W_j \hat{w_j^i} \tag{9}$$

*3.5.3. Step 3: Computation of Scale and Move Factors*

The initial goal of a transformation process $T$ in T-FRI is to scale and move intermediate antecedent fuzzy sest $A'_j$. This is in order that the transformed shapes and representative values coincide with those of the respective observed values $A^*_j$. Over the process of this transformation, move and scale factors are recorded so that the consequent of the intermediate rule can be modified accordingly, to produce the required interpolated result by following the sound intuition that similar antecedents should lead to similar consequent. This process is implemented in two stages: (i) scale operation from $A'_j$ to $\hat{A}'_j$ (denoting the scaled intermediate fuzzy set), and (ii) move operation from $\hat{A}'_j$ to $A^*_j$. From these operations, the required scale rate $s_{A_j}$ and move ratio $m_{A_j}$ are determined.

In particular, given a triangular fuzzy set $A'_j = (a'_{j1}, a'_{j2}, a'_{j3})$, the scale rate $s_{A_j}$ is:

$$s_{A_j} = \frac{a^*_{j3} - a^*_{j1}}{a'_{j3} - a'_{j1}} \tag{10}$$

which essentially expands or contracts the support length of $A'_j : a'_{j3} - a'_{j1}$ so that it becomes the same as that of $A^*_j$. The scaled intermediate fuzzy set $\hat{A'_j}$, which has the same representative value as $A'_j$, is then obtained such that

$$\begin{aligned}
\hat{a'_{j1}} &= \frac{(1+2s_{A_j})a'_{j1} + (1-s_{A_j})a'_{j2} + (1-s_{A_j})a'_{j3}}{3} \\
\hat{a'_{j2}} &= \frac{(1-s_{A_j})a'_{j1} + (1+2s_{A_j})a'_{j2} + (1-s_{A_j})a'_{j3}}{3} \\
\hat{a'_{j3}} &= \frac{(1-s_{A_j})a'_{j1} + (1-s_{A_j})a'_{j2} + (1+2s_{A_j})a'_{j3}}{3}
\end{aligned} \tag{11}$$

Similarly, the move operation shifts the position of $\hat{A'_j}$ to becoming the same as that of $A^*_j$, while maintaining its representative value $Rep(\hat{A'_j})$. This is achieved using the move ratio $m_{A_j}$:

$$m_{A_j} = \begin{cases} \frac{3(a^*_{j1} - \hat{a'_{j1}})}{\hat{a'_{j2}} - \hat{a'_{j1}}}, \text{ if } a^*_{j1} \geq \hat{a'_{j1}} \\ \frac{3(a^*_{j1} - \hat{a'_{j1}})}{\hat{a'_{j3}} - \hat{a'_{j2}}}, \text{ otherwise} \end{cases} \tag{12}$$

*3.5.4. Step 4: Scale and Move Transformation for Interpolated Result*

Having calculated the necessary scale and move factors (i.e., $s_{A_j}$ and $m_{A_j}, j = 1, \ldots, m$), this procedure completes the T-FRI process, deriving the required consequent of $Z^*$. This follows the intuition of similar observations leading to similar consequents, a heuristic fundamental to analogical approximate reasoning. For this, the transformation factors on the antecedent attributes are aggregated. This is implemented by taking into consideration of feature weights assigned for each of the individual antecedent attributes, as shown below:

$$s_z = \sum_{j=1}^{m} W_j s_{A_j} \qquad m_z = \sum_{j=1}^{m} W_j m_{A_j} \tag{13}$$

This entails the computation of scaled $\hat{Z'} = (\hat{z'_1}, \hat{z'_2}, \hat{z'_3})$:

$$\begin{aligned}
\hat{z_1'} &= \frac{(1+2s_z)z_1' + (1-s_z)z_2' + (1-s_z)z_3'}{3} \\
\hat{z_2'} &= \frac{(1-s_z)z_1' + (1+2s_z)z_2' + (1-s_z)z_3'}{3} \\
\hat{z_3'} &= \frac{(1-s_z)z_1' + (1-s_z)z_2' + (1+2s_z)z_3'}{3}
\end{aligned} \tag{14}$$

where $Z' = (z_1', z_2', z_3')$ is the fuzzy value of the intermediate consequent previously computed. From this, again, by analogy to the transformation required for the antecedent to match the observation, move transformation is applied, resulting in the final, required interpolated consequent $Z^* = (z_1^*, z_2^*, z_3^*)$:

$$\begin{aligned}
z_1^* &= \hat{z_1'} + m_z\gamma \\
z_2^* &= \hat{z_2'} - 2m_z\gamma \qquad \gamma = \begin{cases} \frac{\hat{z_2'} - \hat{z_1'}}{3}, \text{ if } m_z \geq 0 \\ \frac{\hat{z_3'} - \hat{z_2'}}{3}, \text{ otherwise} \end{cases} \\
z_3^* &= \hat{z_3'} + m_z\gamma
\end{aligned} \tag{15}$$

Based on the above, when a sparse rule base is learned from source data and a novel observation finds no rules to match, the required consequent can be derived. The entire interpolative process is guided by the feature weights. Note that for those matched observations, the classification results are directly obtained by firing the matched rules without going through interpolation. As with many fuzzy rule-based systems, the resultant consequent fuzzy sets are required to be defuzzified for providing a class label, returning the conclusion on classification. Obviously, in the present CADx system, the conclusion drawn over the given mass is whether its type is benignancy or malignancy.

### *3.6. Pseudocode of Mass Classification System*

Finally, to reinforce the understanding and to help implement the proposed mammographic diagnostic system, Algorithms 1 and 2 present the pseudocode

for the training and application (or testing) of the classification system, respectively. They jointly reflect the overall system framework as illustrated in Fig. 2. Note that the subroutine implementing the core shaded part of Fig. 2, i.e., the procedure for feature weight-guided interpolation, is simply outlined in Line 8 of Algorithm 2 without comprehensively detailing it. This is because the work presented herein is aimed to offer a practical application of weighted FRI, detailed pseudocode of which is beyond the scope of this paper but can be found in [36].

**Algorithm 1** Pseudodode of Mammographic Mass Classifier under Training

**Input:** • Training dataset with mammographic images labelled with mass type

**Output:** • Selected conditional attributes and their relative weights

• Rule base

1: Identify mass ROI images for each of input mammographic images;
2: Segment mass aided with available contours provided in dataset, resulting in mass-segmented mask images;
3: Extract $K$ mass shape (F1-F6), margin (F7-F15) and density (F16-F18) features ($K = 18$, as specified in Table 1) for each mammogram using pairs of ROI and mass-segmented mask images;
4: Rank extracted mass features (F1-F18) of training dataset to obtain ranking score $RS_i, i = 1, 2, \ldots, K$;
5: Select top $m$ features $F = \{RS_i, i = 1, \ldots, m\}$ such that $RS_i > \frac{1}{K}\sum_{t=1}^{K} RS_t$;
6: Calculate feature weights $W$ in terms of Eqn. (1);
7: Generate fuzzy rule base $R$ using selected mass features and mass types;
8: **Return** $F$, $W$ and $R$

## 4. Experimental Evaluation

This section presents a systematic experimental evaluation of the proposed fuzzy rule-based interpolative system for mammographic mass classification. The results are reported on the classification accuracy, sensitivity, specificity, the area under the Receiver Operating Characteristic (ROC) curve, and the ratios of false positives and false negatives over the size of the testing data. These are supported by running nonparametric Wilcoxon signed-rank tests for validating the statistical significance of the classification performance.

**Algorithm 2** Pseudocode of Mammographic Mass Classifier in Action

**Input:**
- Rule base ($R$) generated from training
- Selected features ($F$) and their relative weights ($W$) produced from training
- Unknown mammogram to be classified

**Output:**
- Mass category (i.e., benignancy or malignancy)

```
Identify mass ROI images of given mammogram;
Segment mass, resulting in mass-segmented mask images;
Extract |F| features (as specified in F, where |F| stands for F's cardinality), serving as observation o* to be classified;
Match o* against each rule in rule base R;
if matched with at least one rule then
    Fire matched rule(s) using CRI to obtain required consequent Z* for o*;
else
    Execute weighted FRI to compute Z* = WeightedTFRI(R, o*, W);
end if
Defuzzify Z* as a class label;
Return Benign or Malignant mass
```

### *4.1. Experimental Setup*

To have a common ground for fair comparison, all of the given mammographic images which contain mass lesions provided in BCDR-D01 and BCDR-F01 datasets are employed to conduct the evaluation, for full field digital mammograms and digitalised film mammograms, respectively. As indicated previously, the mass contours annotated by medical specialists are used for the generation of mass-segmented mask images, where the distance between the margin of the chopped box and the provided mass boundary is empirically set as 30 pixels. The corresponding ROI images are of the same size as that of the mask images, while each sharing the same location as their respective original. Again, examples of those can be found in Fig. 3 of Section 3.2.

The classification performance is herein evaluated by 10-fold cross validation randomly repeated for 10 times for both datasets. The partition of each antecedent attribute domain (which is normalised) into triangular membership fuzzy values is achieved by approximating what is learned by the use of Fuzzy C-Means (FCM) [37]. The number of triangular membership functions (i.e., clusters) for each attribute tuned by FCM is determined by the standard method

of [38].

Comparative experimental studies are carried out for classifying mammographic masses, amongst the following three situations: 1) matching the rules in the learned rule base using CRI only for classification (as per classical fuzzy inference systems without FRI), 2) performing CRI for those matched testing observations and conventional unweighted T-FRI for those unmatched ones, and 3) running CRI for matched rule firing and the implemented feature weighted T-FRI for interpolative rule-based classification.

To comprehensively evaluate the classification performance, the following four commonly used metrics are first adopted: classification accuracy, sensitivity, specificity and AUC (i.e., area under ROC curve). These performance indices are computed as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{16}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{17}$$

$$Specificity = \frac{TN}{TN + FP} \tag{18}$$

where TP, FP, TN and FN stand for the number of: true positives, false positives, true negatives and false negatives, respectively. The ROC curve is created by plotting the true positive rate against the false positive rate at various threshold settings and then, the area encompassed by the plotted curve is computed. All of the four evaluation metrics take values between 0 and 1, and a good diagnostic test is obtained when these are close to 1. In addition, two ratio-based performance criteria are also checked, namely FP ratio and FN ratio, which are defined as the ratio between the number of FP over the data size, and that between the number of FN over the data size, respectively. Here, data size stands

for the number of images tested. These two ratios are computed as follows:

$$\text{FP ratio} = \frac{FP}{\text{Number of testing images}} \tag{19}$$

$$\text{FN ratio} = \frac{FN}{\text{Number of testing images}} \tag{20}$$

Smaller values of these ratios indicate better classification, of course.

### *4.2. Results and Discussion*

Comparative experimental results are reported and discussed in this section, including aspects regarding classification interpretability as well as performance measurements.

#### *4.2.1. Interpretability of Fuzzy Rules for Diagnosis*

The mechanism for mammographic mass classification in this work is achieved by the use of semantic fuzzy rules, through rule firing for novel observations that match a certain given rule or rule interpolation for those that match no rules. As indicated before, such fuzzy rules are human interpretable because of the employment of selected semantics-rich, morphological and density features as rule antecedent attributes. In the following, two examples are provided to show the interpretable diagnostic procedure of rule matching (i.e., CRI) and that of rule interpolation for mass classification, respectively.

**(a) Running CRI over matching rule(s)**

For BCDR-D01 dataset, nine top-ranked features are selected to generate the fuzzy rule base. These are: Perimeter (F2), NRL entropy (F9), Mass intensity standard deviation (F17), Margin gradient entropy (F15), Compactness (F12), Mass intensity mean (F16), Margin gradient SD (F14), Convexity (F4), Margin gradient mean (F13). All types of mass feature are involved. In particular, F2, F4 and F17, F16 are mass shape and density features, respectively, while the remaining are mass margin descriptors. These features utilise 3, 3, 4, 4, 3, 3, 3, 4, and 4 fuzzy membership sets representing the underlying linguistic

terms, returned by the application of FCM. In particular, the terms used for three-membership features are "*Small, Medium, Large*", and those for four-membership features are "*Small, Medium-small, Medium, Large*" or "*Small, Medium, Medium-large, Large*", depending on which end over the normalised interval [0,1] the partitions are closer to.

Consider as an illustrative example, the observation consisting of the following feature values:

[F2, F9, F17, F15, F12, F16, F14, F4, F13] =

[0.0760, 0.3178, 0.0368, 0.4178, 0.2690, 0.1181, 0.0709, 0.9179, 0.0865]

with the original mammogram, mass-segmented mask and ROI images shown in Fig. 4. There are four fuzzy rules in total which match this observation, of which the one shown below has the largest matching degree:

If *Perimeter (F2)* is *Small* and *NRL entropy (F9)* is *Medium* and *Mass intensity SD (F17)* is *Small* and *Margin gradient entropy (F15)* is *Medium* and *Compactness (F12)* is *Small* and *Mass intensity mean (F16)* is *Small* and *Margin gradient SD (F14)* is *Small* and *Convexity (F4)* is *Large* and *Margin gradient mean (F13)* is *Small*, then *Mass* is *Benign*.

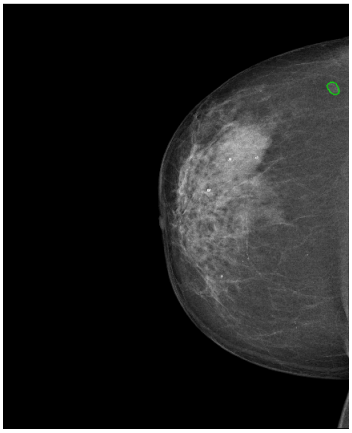
Considering the semantic meaning of each feature given in Table 2, the above rule can be directly translated into:

If *Mass size* is *Small* and *Mass contour* is *Smooth and Blurred* and *Mass density (and its variation)* is *Low* and *Mass* is *Very Like a convex regular region*, then *Mass* is *Benign*.

Firing this rule successfully classifies the mass as Benign, as shown in Fig. 4. It visually recognises the mass lesion in terms of its geometrical shape, contour and density properties, which can be readily understood by medical specialists or explained to the patient.

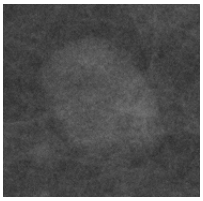
**(b) Running weighted rule interpolation due to no matching rules**

As illustrated in Fig. 2 of Section 3.1, feature weight-guided FRI sub-procedure is triggered by any observation that matches no rules in the sparse rule base, deriving an interpolative classification of the mass category. In this case, selecting two closest neighbouring rules forms the starting point and sets the foundation



(a) Original mammogram

(b) Mass-segmented mask



(c) Chopped ROI

Figure 4: Benign mass classified by matched fuzzy rules.

for rule interpolation.

As with the case for BCDR-D01, in BCDR-F01, all extracted features are ranked first, resulting in the top six being selected. These are: Compactness (F12), Convexity (F4), Circularity (F3), NRL entropy (F9), NRL zero-crossing count (F11) and Mass intensity mean (F16). In particular, F4 and F3 are mass shape features, F12, F9 and F11 are mass margin features, and F16 is selected again as the density descriptor in this dataset.

The number of fuzzy membership functions learned for these selected features are 4, 4, 5, 4, 2, 4, respectively. The fuzzy terms taken by the four-membership features attain the same rule as set in BCDR-D01, while the (only) two-membership feature has two alternatives (i.e., "*Small, Large*") and the remaining one has five fuzzy values, taking from "*Small, Medium-small, Medium, Medium-large, Large*".

Consider the case where the following observation is given which has no rules matched:

[F12, F4, F3, F9, F11, F16] =

[0.9184, 0.2868, 0.2456, 0.8442, 0.4595, 0.4882]

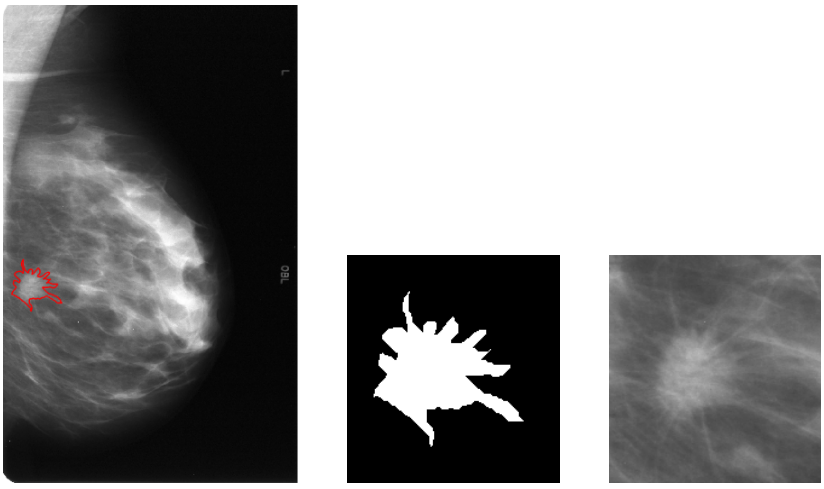
The original mammogram, mass-segmented mask and ROI images for this case

are shown in Fig. 5. From this, two fuzzy rules are found to be the closest to the given observation according to Eqn. (2) in Section 3.5, which are:

*Rule 1:* If *Compactness (F12)* is *Large* and *Convexity (F4)* is *Medium-small* and *Circularity (F3)* is *Medium-small* and *NRL entropy (F9)* is *Medium-large* and *NRL zero-crossing count (F11)* is *Large* and *Mass intensity mean (F16)* is *Medium*, then *Mass* is *Malignant*.

*Rule 2:* If *Compactness (F12)* is *Medium-large* and *Convexity (F4)* is *Medium* and *Circularity (F3)* is *Medium* and *NRL entropy (F9)* is *Medium-large* and *NRL zero-crossing count (F11)* is *Large* and *Mass intensity mean (F16)* is *Medium*, then *Mass* is *Malignant*.

Both rules give malignancy as the conclusion. Having taken into account the semantic linguistic values used for each mass feature in Table 2, these two selected rules jointly lead to the following interpolated rule, with detailed computational process omitted to save space:



(a) Original mammogram (b) Mass-segmented mask (c) Chopped ROI

Figure 5: Malignant mass classified by feature weight-guided FRI.

If *Mass* is *Less Like a circular regular region* and *Mass contour* is *Irregular* and *Mass density* is *Slightly high*, then *Mass* is *Malignant*.

The final interpolated consequent also indicates malignancy for the observed

mass. As can be seen, classifying mammographic mass through interpolating two semantic fuzzy rules offers clear interpretability.

Collectively, the interpretability of the proposed fuzzy rule-based diagnostic system is shown by the process of inferring the category of mammogram mass, running either CRI over matched rule(s) against a given observation or weighted rule interpolation when there is no matching rule. Such interpretability is empowered by the employment of selected semantics-rich, morphological and density features as rule antecedent attributes, in conjunction with the underlying logic relationships between these attributes and the classification outcome. Only clinically explainable fuzzy rules are used for classification. This forms a significant contrast with existing techniques for addressing the problem of mass classification. For instance, in attempting to building an interpretable CADx system using a deep convolution neural network (DCNN)-based framework, such as DeepMiner [13], great effort has been devolved to discovering interpretable representations in deep neural networks so as to provide explanations for medical predictions. Unfortunately, generation of explanations for DCNN-based mammogram classification requires sophisticated expert annotation regarding any interpretable network units. Another attempt is to reveal visually interpretable images extracted from a DCNN, being only concerned with the edge of masses in saliency maps [11]. Yet, no human-like linguistic explanation is produced automatically, unlike what is facilitated in the present rule-based approach.

#### *4.2.2. Performance Based on Fairly Dense Rule Base*

In this part of investigation, all fuzzy rules in the learned rule base are used for mammographic mass classification. Table 3 shows the results with respect to the six performance criteria, namely classification accuracies, sensitivities, specificities, AUC, FP ratio and FN ratio, which are obtained by averaging the outcomes of $10\times10$-fold cross validation. In particular, results on the row named CRI are those achieved by firing matched rules only, those on T-FRI by aiding CRI with classical T-FRI, those on W-T-FRI by combining CRI and feature weighted T-FRI. The classification outcome is obviously unknown for

cases where CRI is used alone to deal with any unmatched observation, in which case an error is recorded while calculating the accuracy, sensitivity, specificity, FP ratio and FN ratio, but this does not apply to the computation of AUC.

The performance of CRI provides the baseline for comparison. As can be seen in Table 3, most of the testing samples are matched with the learned rule(s), resulting in reasonable classified results for both datasets. This is not surprising as the datasets used for training have been fairly comprehensive. Nevertheless, the rule base is not complete, there are uncovered problem spaces for which T-FRI and W-T-FRI can help improve the performance. Indeed, the use of either FRI method significantly strengthens the effectiveness of CRI on BCDR-D01, in terms of the improvement on classification accuracy, sensitivity and specificity, and in the reduction of both false positive ratio and false negative ratio. This shows the potential of fuzzy interpolative inference for coping with challenging situations where the given rule base fails to include rules matching a novel observation.

Applying the feature weighted FRI method has shown a slight further enhancement over the use of the popular T-FRI. The statistical significance is herein verified by Wilcoxon signed-rank test (with the parameter $p = 0.0312$). This demonstrates that the best AUC performance is attained by the use of W-T-FRI for both datasets, with 0.9614 and 0.9023 for the two datasets, respectively. This performance is comparable to that of the state-of-the-art CADx systems for mammographic mass classification, where the recorded best AUC measures are 0.9650 and 0.8940, respectively for BCDR-D01 and BCDR-F01 (see [39, 40]). Yet, the classification process, and hence, the results of running the existing methods are not so easy to interpret as their counterparts of the proposed approach. More importantly, the performance improvement becomes much more significant when considering situations where only a sparse rule base is available, as to be shown next.

Table 3: Performance on BCDR-D01 and BCDR-F01 by 10×10 Cross Validation

| BCDR-D01 | | | | | | |
|---|---|---|---|---|---|---|
| Schemes | Accuracy (%) | Sensitivity (%) | Specificity (%) | AUC | FP ratio | FN ratio |
| CRI | 83.44 | 78.57 | 86.59 | - | 0.0814 | 0.0857 |
| T-FRI | 91.22 | 87.85 | 93.41 | 0.9607 | 0.0400 | 0.0485 |
| W-T-FRI | 91.65 | 88.93 | 93.41 | 0.9614 | 0.0400 | 0.0442 |
| BCDR-F01 | | | | | | |
| Schemes | Accuracy (%) | Sensitivity (%) | Specificity (%) | AUC | FP ratio | FN ratio |
| CRI | 83.73 | 81.30 | 86.27 | - | 0.0668 | 0.0967 |
| T-FRI | 84.28 | 82.14 | 86.49 | 0.9019 | 0.0657 | 0.0924 |
| W-T-FRI | 84.28 | 82.14 | 86.49 | 0.9023 | 0.0657 | 0.0924 |

*4.2.3. Performance Based on Very Sparse Rule Base*

The classification results presented in the previous part of experimental evaluation are achieved by the use of the entire rule base learned from the data available. This is the situation that a real-world application would encounter. Even for the examined problem where a good amount of training data is exploited to generate a fairly dense rule base, as with the investigated case, sparseness may exist. This itself already shows the need for the employment of FRI techniques. However, there are practical situations where not sufficient training data is obtainable, especially when dealing with certain novel medical cases. It is therefore very interesting to investigate how the T-FRI in general and the W-T-FRI method in particular may bring forward any benefits in such situations. For this purpose, without complicating the experimental studies by introducing different datasets, here, two rule bases which are much sparser than the one used previously are artificially generated, by randomly removing a number of learned rules from the original used. Note that this artificially imposed removal is for academic investigation only; in real application, unless there is inconsistency or redundancy, learned rules are not to be removed.

Table 4 shows the averaged results of this investigation, in relation to the percentage of rules removed. Particularly, the two sparser rule bases run are created by randomly deleting 30% and 70% of the learned rules, respectively. As expected, and reflected by this table, the performance of applying CRI alone declines dramatically as the proportion of rules remaining available decreases. The accuracies drop 30.01% (=83.44%-53.43%) and 60% (=83.44%-23.44%) for BCDR-D01 and 45.31% (=83.73%-38.42%) and 67.4% (=83.73%-16.33%) for BCDR-F01, respectively. The resultant performance deteriorates so much that such an approach is no longer acceptable in practice.

On the contrast, both FRI methods have shown to be able to alleviate such performance decline. With the employment of a FRI mechanism present CADx system maintains a strong capability in distinguishing suspicious mass lesions when CRI performs poorly, given only a considerably sparse rule base. Even

when just a small proportion of rules remains available (for the cases where 70% of the rules are removed), the classification performance (regarding accuracies, sensitivities and specificities) is still at an approximate rate of 80% on the BCDR-D01 dataset and at high 60% on BCDR-F01. Regarding the FP and FN ratios, a significant reduction in these for both datasets has been shown by the use of either T-FRI or W-T-FRI as compared to the use of just CRI. Together, these results strongly demonstrate the significant effectiveness of fuzzy interpolative reasoning for resolving the problems involving a sparse rule base.

Examining more closely by comparing the performance of T-FRI and that of W-T-FRI, as the rule base is reduced to be much sparser, the improvement of W-T-FRI over T-FRI becomes more notable. In particular, the classification accuracy is enhanced by 2.36% and 4.08% with regards to 30% and 70% reduction of the rules on BCDR-D01, and by 1.62% and 3.73% on BCDR-F01. Furthermore, Table 5 summarises the average number of testing samples that require rule interpolation in each of the three inference situations (namely, running CRI alone, and CRI with T-FRI or W-T-FRI). Note that RB in this table and the next, stands for Rule Base. It is evident that the more unmatched samples, the more opportunities there are for the FRI methods to perform.

Comparative performance is also measured through ROC analysis. The ROC curves resulting from running the standard T-FRI and feature weighted T-FRI over the use of different rule bases are given in Fig. 6, on both BCDR-D01 and BCDR-F01. Whilst it is not surprising that the best performance is achieved using the fairly dense rule base for both methods, W-T-FRI is shown to be less sensitive to the deterioration of sparsity of the rule base.

Last but not least, as indicated previously, to further determine the statistical significance in performance improvement of T-FRI over CRI, and that of W-T-FRI over T-FRI, the nonparametric Wilcoxon signed-rank tests are conducted. This is carried out for the classification accuracies obtained from the use of three different inference mechanisms implemented, with three different sparsities of the rule base on both datasets. Table 6 lists the $p$-value of each pairwise test. As can be seen in this table, all but one expectable test show

Table 4: Performance Based on Sparse Rule Base by 10×10 Cross Validation

| | Sparser Rule Base 1 (30% removed) | | | | | | |
|---|---|---|---|---|---|---|---|
| BCDR-D01 | Schemes | Accuracy (%) | Sensitivity (%) | Specificity (%) | AUC | FP ratio | FN ratio |
| | CRI | 53.43 | 44.04 | 59.60 | - | 0.2452 | 0.2238 |
| | T-FRI | 83.71 | 78.57 | 87.06 | 0.8918 | 0.0786 | 0.0857 |
| | W-T-FRI | 86.07 | 81.55 | 89.02 | 0.9010 | 0.0666 | 0.0738 |
| | Sparser Rule Base 2 (70% removed) | | | | | | |
| | Schemes | Accuracy (%) | Sensitivity (%) | Specificity (%) | AUC | FP ratio | FN ratio |
| | CRI | 23.44 | 12.95 | 30.29 | - | 0.4232 | 0.3482 |
| | T-FRI | 77.67 | 74.55 | 79.70 | 0.8589 | 0.1232 | 0.1018 |
| | W-T-FRI | 81.75 | 79.02 | 83.53 | 0.8784 | 0.1000 | 0.0839 |
| | Sparser Rule Base 1 (30% removed) | | | | | | |
| BCDR-F01 | Schemes | Accuracy (%) | Sensitivity (%) | Specificity (%) | AUC | FP ratio | FN ratio |
| | CRI | 38.42 | 29.41 | 47.99 | - | 0.2532 | 0.3652 |
| | T-FRI | 73.47 | 71.22 | 75.89 | 0.7948 | 0.1174 | 0.1489 |
| | W-T-FRI | 75.09 | 73.95 | 76.34 | 0.8238 | 0.1152 | 0.1347 |
| | Sparser Rule Base 2 (70% removed) | | | | | | |
| | Schemes | Accuracy (%) | Sensitivity (%) | Specificity (%) | AUC | FP ratio | FN ratio |
| | CRI | 16.33 | 12.60 | 20.24 | - | 0.3884 | 0.4521 |
| | T-FRI | 62.60 | 63.30 | 61.91 | 0.6833 | 0.1855 | 0.1898 |
| | W-T-FRI | 66.33 | 68.06 | 64.58 | 0.7040 | 0.1724 | 0.1652 |

Table 5: Average Number of Testing Samples

| Dataset | Number of Samples Requiring Interpolation/Total (per Fold) | | |
|---|---|---|---|
| | Fairly Dense RB | Sparser RB 1 | Sparser RB 2 |
| BCDR-D01 | 1.24/14 | 5.85/14 | 10.50/14 |
| BCDR-F01 | 0.21/23 | 12.77/23 | 18.70/23 |

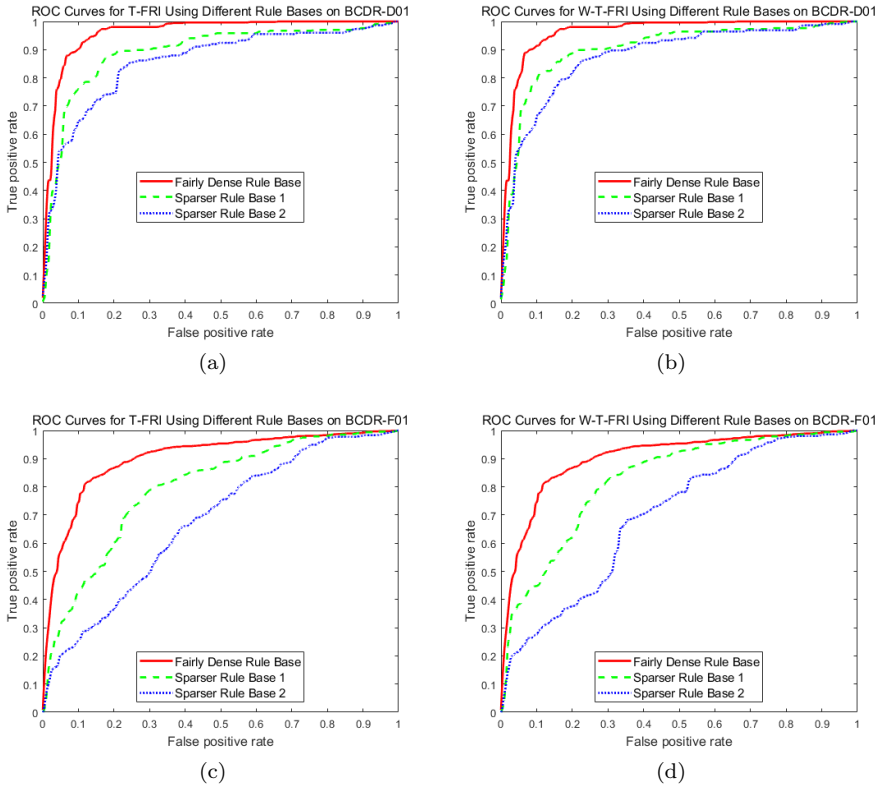



Figure 6: ROC for T-FRI and W-T-FRI using rule bases of different sparsity.

relative small $p$-values (e.g., $p < 0.05$), which reflects the statistical significance of outperformance in each comparison. The only exception (with $p = 1$) is for the case comparing W-T-FRI against T-FRI on BCDR-F01 when the originally learned, fairly dense rule base is employed.

Table 6: $P$-value in Statistical Wilcoxon Signed Rank Test

| BCDR-D01 | | | |
|---|---|---|---|
| | Original RB | 30% Reduced RB | 70% Reduced RB |
| CRI vs. T-FRI | $5.65 \times 10^{-13}$ | $1.34 \times 10^{-11}$ | $3.31 \times 10^{-8}$ |
| T-FRI vs. W-T-FRI | 0.0312 | $8.03 \times 10^{-5}$ | $1.71 \times 10^{-4}$ |
| BCDR-F01 | | | |
| | Original RB | 30% Reduced RB | 70% Reduced RB |
| CRI vs. T-FRI | 0.0019 | $3.36 \times 10^{-8}$ | $1.56 \times 10^{-6}$ |
| T-FRI vs. W-T-FRI | 1 | 0.0169 | $4.88 \times 10^{-4}$ |

## 5. Conclusion

In this paper, a novel fuzzy rule-based diagnostic system for mammographic mass classification has been presented. The system is able to derive a conclusion for unknown observed masses that have no rules to match. The diagnostic outcomes are interpretable as the rules are learned over selected features in terms of mass geometric and density properties, with feature values represented in linguistic terms. The effectiveness of adapting feature weighted fuzzy rule interpolation as the core of the implemented system has been systematically evaluated and demonstrated, capable of dealing with rather sparse rule bases. This has been accomplished through comparison with the state-of-the-art work on mammogram datasets.

The present work utilises transformation-based FRI algorithm, especially its weighted version for implementation. However, other established approaches for FRI may be modified for use as alternatives [18, 25]. In the implemented CADx system, individual feature weights are derived by ranking scores using a common feature ranking mechanism. It would be interesting to investigate how different feature selection techniques (e.g., those reported in [41, 42]) may be exploited to evaluate extracted features. If successful, this would help effectively reduce the dimensionality without destroying the underlying semantics of feature description, and of course rank the features to produce the required weights. Finally, to reinforce the motivations for this development, applying the proposed approach to more complicated breast cancer diagnostic problems

remains an active research.

## References


[1] T.H. Samuels. "Illustrated Breast Imaging Reporting and Data System BIRADS." American College of Radiology Publications, (1998).

[2] W. Xie, Y. Li, and Y. Ma. "Breast mass classification in digital mammography based on extreme learning machine." *Neurocomputing* **173** (2016): 930-941.

[3] X. Liu, and J. Tang. "Mass classification in mammograms using selected geometry and texture features, and a new SVM-based feature selection method." *IEEE Systems Journal* **8**.3 (2014): 910-920.

[4] N.P. Pérez, M.A.G. López, A. Silva, and I. Ramos. "Improving the Mann–Whitney statistical test for feature selection: An approach in breast cancer diagnosis on mammography." *Artificial intelligence in medicine* **63**.1 (2015): 19-31.

[5] G. Magna, P. Casti, S.V. Jayaraman, M. Salmeri, A. Mencattini, E. Martinelli, and C.D. Natale. "Identification of mammography anomalies for breast cancer detection by an ensemble of classification models based on artificial immune system." *Knowledge-Based Systems* **101** (2016): 60-70.

[6] A. Oliver, A. Torrent, X. Lladó, M. Tortajada, L. Tortajada, M. Sentís, J. Freixenet, and R. Zwiggelaar. "Automatic microcalcification and cluster detection for digital and digitised mammograms." *Knowledge-Based Systems* **28** (2012): 68-75.

[7] G.H.B. Miranda, and J.C. Felipe. "Computer-aided diagnosis system based on fuzzy logic for breast cancer categorization." *Computers in biology and medicine* **64**, (2015): 334-346.

[8] H.D. Cheng, X.J. Shi, R. Min, L.M. Hu, X.P. Cai, and H.N. Du. "Approaches for automated detection and classification of masses in mammograms." *Pattern recognition* **39**.4 (2006): 646-668.

[9] N.I. Yassin, S. Omran, E.M.El Houby, and H. Allam. "Machine learning techniques for breast cancer computer aided diagnosis using different image modalities: A systematic review." *Computer methods and programs in biomedicine* **156** (2018): 25-45.

[10] A. Oliver, J. Freixenet, J. Martí, E. Pérez, J. Pont, E. R.E. Denton, and R. Zwiggelaar. "A review of automatic mass detection and segmentation in mammographic images." *Medical Image Analysis* **14** (2010): 87–110.

[11] D. Lévy, and A. Jain. "Breast mass classification from mammograms using deep convolutional neural networks." *arXiv preprint* arXiv:1612.00542 (2016).

[12] R.W.D. Pedro, A. Machado-Lima, and F.LS Nunes. "Is mass classification in mammograms a solved problem?-a critical review over the last 20 years." *Expert Systems with Applications* **119** (2019): 90-103.

[13] J. Wu, B. Zhou, D. Peck, S. Hsieh, V. Dialani, L. Mackey, and G. Patterson. "DeepMiner: Discovering Interpretable Representations for Mammogram Classification and Explanation." *arXiv preprint* arXiv:1805.12323 (2018).

[14] D. Yi, R.L. Sawyer, D.C. III, J. Dunnmon, C. Lam, X. Xiao, and D. Rubin. "Optimizing and visualizing deep learning for benign/malignant classification in breast tumors." *arXiv preprint* arXiv:1705.06362 (2017).

[15] R. Mousa, Q. Munib, and A. Moussa. "Breast cancer diagnosis system based on wavelet analysis and fuzzy-neural." *Expert systems with Applications* **28**.4 (2005): 713-723.

[16] L.A. Zadeh "Outline of a new approach to the analysis of complex systems and decision processes." *IEEE Transactions on systems, Man, and Cybernetics* **1**, (1973): 28-44.

[17] A. Vadivel, and B. Surendiran. "A fuzzy rule-based approach for characterization of mammogram masses into BI-RADS shape categories." *Computers in biology and medicine* **43**.4 (2013): 259-267.

[18] L.T. Kóczy, K. Hirota. "Approximate reasoning by linear rule interpolation and general approximation." *International Journal of Approximate Reasoning* **9**.3, 197-225 (1993).

[19] P. Baranyi, L.T. Kóczy, T.D. Gedeon. "A generalized concept for fuzzy rule interpolation." *IEEE Transactions on Fuzzy Systems* **12**.6, 820-837 (2004).

[20] Z. Huang, and Q. Shen. "Fuzzy interpolative and extrapolative reasoning: a practical approach." *IEEE Transactions on Fuzzy Systems* **14**.2, (2006): 340-359.

[21] L. Yang, F. Chao, and Q. Shen. "Generalised adaptive fuzzy rule interpolation." *IEEE Transactions on Fuzzy Systems* **25**.4 (2017): 839-853.

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

[23] S. Jin, R. Diao, C. Quek and Q. Shen. "Backward fuzzy rule interpolation." *IEEE Transactions on Fuzzy Systems* **22**.6 (2014): 1682-1698.

[24] N. Naik, R. Diao and Q. Shen. "Dynamic fuzzy rule interpolation and its application to intrusion detection." *IEEE Transactions on Fuzzy Systems* **26**.4 (2018): 1878-1892.

[25] Y.C. Chang, S.M. Chen, C.J. Liau. "Fuzzy interpolative reasoning for sparse fuzzy-rule-based systems based on the areas of fuzzy sets." *IEEE Transactions on Fuzzy Systems* **16**.5 (2008): 1285-1301.

[26] 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* **26**.5 (2018): 2758-2773.

[27] S.M. Chen, S.I. Adam. "Weighted fuzzy interpolated reasoning based on ranking values of polygonal fuzzy sets and new scale and move transformation techniques." *Information Sciences* **435**, 184-202 (2018).

[28] F. Li, C. Shang, Y. Li, and Q. Shen. "Interpolation with just two nearest neighbouring weighted fuzzy rules." *IEEE Transactions on Fuzzy Systems* (2019).

[29] Z. Huang, and Q. Shen. "Fuzzy interpolative reasoning via scale and move transformations." *IEEE Transactions on Fuzzy Systems* **16**.1, (2008): 13-28.

[30] M.G. Lopez, N.G. Posada, D.C. Moura, R.R. Pollán, J.M.F. Valiente, C.S. Ortega, M. Solar et al. "BCDR: a breast cancer digital repository." *In 15th International Conference on Experimental Mechanics*. 2012. Available on `https://bcdr.ceta-ciemat.es/`.

[31] A.R. Dominguez, and A.K. Nandi. "Detection of masses in mammograms via statistically based enhancement, multilevel-thresholding segmentation, and region selection." *Computerized Medical Imaging and Graphics* **32**.4 (2008): 304-315.

[32] N. Petrick, H.P. Chan, B. Sahiner, and M.A. Helvie. "Combined adaptive enhancement and region-growing segmentation of breast masses on digitized mammograms." *Medical physics* **26**.8 (1999): 1642-1654.

[33] T. Mu, A.K. Nandi, and R.M. Rangayyan. "Classification of breast masses using selected shape, edge-sharpness, and texture features with linear and kernel-based classifiers." *Journal of Digital Imaging* **21**.2 (2008): 153-169.

[34] I. Kononenko. "Estimating attributes: analysis and extensions of RELIEF." *In European conference on machine learning*, pp. 171-182. Springer, Berlin, Heidelberg, 1994.

[35] L. Wang, and J.M. Mendel. "Generating fuzzy rules by learning from examples." *IEEE Transactions on systems, man, and cybernetics* **22**6 (1992): 1414-1427.

[36] F. Li, Y. Li, C. Shang, and Q. Shen. "Fuzzy knowledge-based prediction through weighted rule interpolation." *IEEE Transactions on Cybernetics* (2019).

[37] J.C. Bezdek, R. Ehrlich, and W. Full. "FCM: The fuzzy c-means clustering algorithm." *Computers & Geosciences* **10**.2-3 (1984): 191-203.

[38] M. Chen, and S. Wang. "Fuzzy clustering analysis for optimizing fuzzy membership functions." *Fuzzy sets and systems* **103**.2 (1999): 239-254.

[39] D.C. Moura, and M.A. Guevara López. "An evaluation of image descriptors combined with clinical data for breast cancer diagnosis." *International journal of computer assisted radiology and surgery* **8**.4 (2013): 561-574.

[40] D.C. Moura, M.A. Guevara López, P. Cunha, N.G Posada, R.R. Pollan, I. Ramos, J.P. Loureiro, I.C. Moreira, B.M.F.Araújo, and T.C. Fernandes. "Benchmarking datasets for breast cancer computer-aided diagnosis (CADx)." *In Iberoamerican Congress on Pattern Recognition*, pp. 326-333. Springer, Berlin, Heidelberg, 2013.

[41] R. Diao, F. Chao, T. Peng, N. Snook, and Q. Shen. "Feature selection inspired classifier ensemble reduction." *IEEE Transactions on Cybernetics* **44**.8, (2014): 1259-1268.

[42] R. Jensen, and Q. Shen. "New approaches to fuzzy-rough feature selection." *IEEE Transactions on Fuzzy Systems* **17**.4 (2009): 824-838.