Aberystwyth University

Multi-Criterion Mammographic Risk Analysis Supported with Multi-Label Fuzzy-Rough Feature Selection

Qu, Yanpeng; Yue, Guanli; Shang, Changjing; Yang, Longzhi; Zwiggelaar, Reyer; Shen, Qiang

Multi-criterion mammographic risk analysis supported with multi-label fuzzy-rough feature selection

Yanpeng Qu[a,b,*], Guanli Yue[a], Changjing Shang[b], Longzhi Yang[c], Reyer Zwiggelaar[b], Qiang Shen[b]

[a] *Information Technology College, Dalian Maritime University, Dalian 116026, China*
[b] *Department of Computer Science, Institute of Mathematics, Physics and Computer Science, Aberystwyth University, Aberystwyth, Ceredigion SY23 3DB, UK*
[c] *Department of Computer Science and Digital Technologies, Northumbria University, Newcastle upon Tyne NE1 8ST, UK*

ARTICLE INFO

ABSTRACT

Context and background: Breast cancer is one of the most common diseases threatening the human lives globally, requiring effective and early risk analysis for which learning classifiers supported with automated feature selection offer a potential robust solution.

Motivation: Computer aided risk analysis of breast cancer typically works with a set of extracted mammographic features which may contain significant redundancy and noise, thereby requiring technical developments to improve runtime performance in both computational efficiency and classification accuracy.

Hypothesis: Use of advanced feature selection methods based on multiple diagnosis criteria may lead to improved results for mammographic risk analysis.

Methods: An approach for multi-criterion based mammographic risk analysis is proposed, by adapting the recently developed multi-label fuzzy-rough feature selection mechanism.

Results: A system for multi-criterion mammographic risk analysis is implemented with the aid of multi-label fuzzy-rough feature selection and its performance is positively verified experimentally, in comparison with representative popular mechanisms.

Conclusions: The novel approach for mammographic risk analysis based on multiple criteria helps improve classification accuracy using selected informative features, without suffering from the redundancy caused by such complex criteria, with the implemented system demonstrating practical efficacy.

1. Introduction

Female breast cancer has an occurrence rate of 8% to 13% [1,2]. The cause of breast cancer is still unknown and currently, there is not any effective prevention measure. Breast cancer cannot be easily visualised with naked eyes or physically palpated with bare hands. Doctors had to make diagnosis based on their own experience before the emergence of computer-aided diagnosis (CAD) systems. Consequently, early development towards breast cancer had often been left unattended, leading to serious consequences to human lives. With the support of CAD, mammograms have been widely used in clinical practice. In particular, with the support of artificial intelligence (AI) technologies, automated CAD systems have been developed to detect the abnormality of breast masses, greatly improving the effectiveness of risk analysis and early diagnosis.

1.1. Background

Most of CAD systems are based on the strong correlation between breast cancer risk and breast tissue density or texture features, such as Boyd [3], BI-RADS [4], Tabár [5], Wolfe [6], but they vary in the single, underlying criterion used. For instance, Boyd evaluates the proportion of dense breast tissue in reference to the overall breast area, while BI-RADS directly exploits the density of the entire breast to determine the risk. Different AI techniques have been applied to building CAD systems, usually based on a single criterion, in an attempt to enable automated early diagnosis of breast cancer.

Typically, existing approaches firstly extract the features of mammograms using a labelled mammogram dataset, then, a learning classifier or risk assessment method [7–9] is trained with the dataset. For example, a technique based on multi-scale wavelet transformation for textural feature extraction was proposed in [10], which subsequently used the *k* nearest neighbour method (*k*NN) for classification. Also,

---

histogram equalisation and morphological operation were applied to enhance the mammograms, followed by an Otsu's thresholding mechanism for region of interest (ROI) segmentation in [11]. This work uses gray level co-occurrence matrix (GLCM) for feature extraction, and *k*NN, support vector machine (SVM) and artificial neural network (ANN) for classification. An evolutionary fuzzy extreme learning machine was employed for mammographic risk analysis [12], which combined evolutionary computation and extreme learning machine to efficiently build fuzzy classifiers. A kernel-based fuzzy-rough nearest-neighbour classification was used for mammographic risk analysis as well [13], which significantly improved the classification performance using kernels.

Such CAD systems work generally by predicting the occurrence probability of cancer or cancer risk, through exploiting a fixed set of features (usually representing the texture, brightness, shape, etc.) of captured mammograms. However, redundancy and noises are often present in the extracted features. This may adversely affect the final classification accuracy and runtime complexity. Feature selection can enhance CAD systems by choosing the most informative feature subset from the extracted feature set, as has been shown in the relevant literature. In particular, the work of [14] initially extracted 14 statistical features, but only 3 of them were selected using an ANN based on a simple try-and-employ strategy. Also, the approach presented in [15] extracted 61 original features from both spatial and spectral domain, while using just 25 and 11 selected features as the input of a general regression neural network classifier and a SVM classifier, respectively. On a larger scale, the t-test method was utilised to select 333 optimal features from those 46,080 extracted via curvelet transformation in [16], the selected features were fed into an SVM classifier for abnormal detection. Whilst being successful in their own right, all these systems are based on a single criterion or one fixed set of class labels (e.g., Boyd and BI-RADS, as mentioned previously).

1.2. Motivations

Although the existing CAD systems are normally based on one single criterion, there are many different criteria for evaluating the risk of breast cancer risk. Moreover, in real life, the eventual diagnosis for a certain case is often determined by multiple doctors from different perspectives. In order to investigate the potential underlying relations amongst the evaluation indicators for mammographic risk assessment, this paper proposes a computational intelligence system using multi-label fuzzy-rough feature selection (MLFRFS). It allows decision making based on multiple criteria, without introducing unnecessary redundancy to predict the breast cancer risk. This is developed on the basis of the seminal work as reported in [17].

Briefly, there are three categories of multi-label feature selection approaches: filter, embedded and wrapper. The wrapper and embedded approaches interact with the learning classification method to a full or large extent. Particularly, wrappers are directly integrated as part of the classification algorithm to evaluate the importance of features, while the embedded methods incorporate feature selection during the training process of a learning classifier, such as *k*NN. In contrast, filters are independent of the learning classifiers, especially with regard to feature evaluation during the selection process. Thus, the embedded approach appears to be one that compromises between wrappers and filters. MLFRFS is indeed such an approach, implemented by considering the inherent relationships between features and class labels while performing association rule learning.

1.3. Contributions

This proposed approach for multi-criterion mammographic risk analysis benefits from the novel MLFRFS algorithm which works in the following way. Given a dataset constrained with multiple criteria, the conventional method for label power set learning (LP) is adopted to

integrate the multiple criteria into one unified binary label representation [18]. From this, the association rules are learned on the basis of pre-defined support and confidence thresholds [19], which are then exploited to reduce the number of binary labels. This is enables a direct application of fuzzy-rough feature selection since the original multi-criterion classification problem has now been transformed to a single criterion one.

The resulting approach is fully implemented, supported with systematic experimental validation and evaluation. To reflect the potential in helping perform real-world breast cancer risk analysis, the dataset run is derived from images archived in the Mammographic Image Analysis Society (MIAS) database [20]. A comparative study is conducted in reference to popular and powerful feature selection methods such as FRFS [21], Consis [22], CFS [23] and FDMFS [21], using a range of learning classifiers, including: Naive Bayes (NB) [24], Logistic [25], Random Forests (RF) [26], and *k*NN [27]. The experimental results demonstrate the efficacy of the approach in offering effective risk analysis to aid in early diagnosis of breast cancer.

The remainder of the paper is structured as follows. The foundational theoretical aspects of this work are outlined in Section 2. The proposed approach and its implementation are detailed in Section 3. The experimental investigations are reported in Section 4 with results analysed. The paper concludes in Section 5 together with a brief discussion regarding relevant further research.

2. Technical underpinnings

The underpinning fundamentals of the research presently reported, including association rule learning, fuzzy-rough feature selection, and typical mammographic image analysis criteria are reviewed in this section.

2.1. Association rule learning

In general, association rule learning concerns with the computational mechanism that may be exploited to identify interesting association relations, termed association rules, between variables in a given dataset. The particular learning mechanism addressed herein was first proposed in [19], where the degrees of support and confidence are two important indicators which are defined below for completeness.

Definition 1. Let A be an item set, B another item set, $A \Rightarrow B$ an association rule and $\mathbb{U}$ a set of instances of a given dataset. $|A|$ represents the number of instances which contains A. The degree of support of A with respect to $\mathbb{U}$ is defined as the proportion of instances in the dataset which contains the item set A, similarly, the degree of support of $A \Rightarrow B$ is defined as the proportion of instances in the dataset which contains the item set A and B:

$$\text{supp}(A) = \frac{|A|}{|\mathbb{U}|}. \tag{1}$$

$$\text{supp}(A \Rightarrow B) = \frac{|A \cap B|}{|\mathbb{U}|}. \tag{2}$$

Definition 2. The degree of confidence of a rule, $A \Rightarrow B$, with respect to a set of instances $\mathbb{U}$, is the proportion of the instances which contains A and also contains B to the instances including A:

$$\text{conf}(A \Rightarrow B) = \frac{\text{supp}(A \cap B)}{\text{supp}(A)}. \tag{3}$$

The association rules meet certain pre-specified thresholds on the degrees of support and confidence are called strong rules, which reflect the most representative relations between the items in the dataset considered. Association rule learning in a dataset involves two procedures. Firstly, all frequent item sets are identified, which incurs most of the computation cost during the overall learning process. Denote a *k*-

item set as a collection containing k items. If the frequency of a k-item set is greater than the minimum support, the set is returned a frequent k-item set. Secondly, strong association rules are generated from frequent item sets, which are usually implemented by the Apriori algorithm as detailed in Algorithm 1 (taken from [28]).

Algorithm 1. Apriori

```
Apriori (𝕌, ε)
Input:
        𝕌, set of instances;
        ε, support threshold.
Output: F, Frequent item sets.
1       F₁ ⟵ {1-item sets} ; k ⟵ 2
2       while F_{k-1} ≠ ∅
3          F_k ⟵ ∅
4          C_k ⟵ {c = a ∪ {b} | a ∈ F_{k-1} ∧ b ∉ a} − {c | {s | s ⊆ c ∧ |s| = k − 1} ⊈ L_{k-1}}
5          for c in C_k
6          count[c] ⟵ 0
7          for n in 𝕌
8             if c ⊆ n
9          count[c] ⟵ count[c] + 1
10            end
11         end
12         if count[c] ≥ ε
13             F_k ⟵ F_k ∪ {c}
14         end
15            end
16         k ⟵ k + 1
17      end
18      F ⟵ ∪_k F_k
19      return F
```

This association rule learning algorithm works based on the apriori property that all non-empty subsets of frequent item sets are frequent. That is, if a collection of instances is an infrequent item set, any superset of this set must not be frequent. According to this property, Apriori can efficiently generate association rules from given instances by creating super-frequent item sets while pruning those with non-frequent subsets. As outlined in Algorithm 1, given a set of instances and the support threshold, the first action of this algorithm scans the dataset to identify all frequent 1-item sets F_1. The loop in Lines 2 to 17 generates all the frequent item sets. In particular, Line 4 creates the candidate set C_k based on F_{k-1}. The support count of the candidate set is calculated in Lines 5 to 11, which are then filtered with respect to the threshold in Lines 12 and 13 to generate F_k. From this, k is incremented in Line 16 to facilitate the creation of the frequent item set of size $k + 1$.

2.2. Fuzzy-rough feature selection

Fuzzy-rough feature selection (FRFS) is developed based on fuzzy-rough set theory, which can be defined by an axiomatic approach or a constructive approach. The most common definition of fuzzy-rough set uses the fuzzy T norm and implication operator I as follows [21]:

$$\mu_{\underline{R_P X}}(x) = \inf_{y \in \mathbb{U}} I(\mu_{R_P}(x, y), \mu_X(y)), \tag{4}$$

$$\mu_{\bar{R_P} X}(x) = \sup_{y \in \mathbb{U}} T(\mu_{R_P}(x, y), \mu_X(y)), \tag{5}$$

where $R_P(x, y)$ represents the fuzzy similarity relation induced by the subset of features P, and $\mu_{R_P}(x, y)$ is the fuzzy similarity of instance x and instance y:

$$\mu_{R_P}(x, y) = T_{a \in P}\{\mu_{R_a}(x, y)\}. \tag{6}$$

In traditional rough set approach, given a decision table $T = (\mathbb{U}, P, D)$, the P-positive domain of D is defined as:

$$\text{POS}_P(D) = \bigcup_{X \in \mathbb{U}/D} P(X), \tag{7}$$

which is the positive domain of the equivalence class $\mathbb{U}/D$ with respect to P. That is, the knowledge expressed by $\mathbb{U}/P$ can be assigned to the

instance set of $\mathbb{U}/D$.

Similar to the above case for crisp rough sets, the fuzzy positive region of the decision criteria D on an feature subset P can be defined by

$$\mu_{\text{POS}_{R_P}(D)}(x) = \sup_{X \in \mathbb{U}/D} \mu_{\underline{R_P X}}(x). \tag{8}$$

From this, the fuzzy-rough dependency is introduced as follows:

$$\gamma'_P(D) = \frac{\sum_{x \in \mathbb{U}} \mu_{\text{POS}_{R_P}(D)}(x)}{|\mathbb{U}|}. \tag{9}$$

As interpreted in the underlying mathematical theory for FRFS, this dependency measure signifies the importance of the feature subset upon which the relevant decision relies. The higher the dependency is measured, the more able the corresponding feature subset is to distinguish amongst the decision classes.

Note that FRFS is a heuristic feature selection method, so different search strategies can be used. For instance, the popular fuzzy rough fast reduction algorithm [21] is implemented by best first search, as summarised in Algorithm 2. It attempts to calculate the optimal feature subset incrementally without generating all possible subsets, starting with an empty set and iteratively adding the feature that returns the highest dependency measure amongst the remaining ones. As the number of features increases, the dependency of the feature subset will increase monotonically. When the dependency of the decision attribute upon the current feature subset is equal to that upon the entire feature set (Line 10), the algorithm terminates and outputs that feature subset as the selected subset. This selected feature subset can then be used to replace the original full set of features in any subsequent application.

Algorithm 2. Fuzzy-rough feature selection (FRFS)

```
FRFS (C, D)
Input:
                    C, all conditional features set;
                    D, set of decision attributes.
Output: R, Reduced feature subset.
1                   R = ∅
2                   do
3                      T ⟵ R
4                      foreach x ∈ (C − R)
5                   if γ'_{R∪{x}}(D) > γ'_T(D)
6                       T ⟵ R ∪ {x}
7                   end
8                      R ⟵ T
9                      end
10                  until γ'_R(D) = =γ'_C(D)
11                  return R
```

2.3. Mammographic image analysis

A number of evaluation indicators have been developed for breast cancer risk assessment, and the most common ones include Boyd, BI-RADS, Tabár, and Wolfe. In particular, Boyd introduces a quantitative classification of mammographic density, based on measuring the proportion of dense breast tissue relative to the overall breast area. The classification is known as Six-Class-Categories (SCC) where the density proportions are: Category 1: 0%, Category 2: (0–10%), Category 3: (10–25%), Category 4: (25–50%), Category 5: (50–75%) and Category 6: (75–100%). The increase in the level of breast tissue density is proportionally associated with an increase in the risk of developing breast cancer, specifically the relative numeric risk values for SCC 3–6 are 1.9, 2.2, 4.6 and 7.1, respectively.

The BI-RADS indicator categories a mammogram into one of four classes. BI-RADS I indicates the breast density is low; BI-RADS II represents certain fibroglandular tissue; BI-RADS III expresses high breast density; and BI-RADS IV reports extremely high density. Numerically, the risk values for BI-RADS II–IV are 1.6, 2.3 and 4.5, respectively.

Tabár describes breast composition using four building blocks: nodular density, linear density, homogeneous fibrous tissue and

radiolucent adipose tissue, which also define mammographic risk classification. Patterns I–III each represent a lower breast cancer risk and Patterns IV–V represent the higher risks as summarised below:

- Pattern I mammograms are composed of 25%, 16%, 35% and 24% of the four building blocks, respectively;
- Pattern II has approximate compositions of: 2%, 14%, 2% and 82%;
- Pattern III is quite similar in composition to Pattern II, except that the retroareolar prominent ducts are often associated with peri-ductal fibrosis;
- Pattern IV is dominated by proniment nodular and linear densities, with compositions of 49%, 19%, 15%, and 17%;
- Pattern V is dominated by extensive fibrosis and is composed as 2%, 2%, 89% and 7% of the building blocks.

Wolfe uses four categories (N1, P1, P2, DY) to represent mammogram risks, these four have occurrence rates of 0.1, 0.4, 1.7, 2.2 in developing breast cancer, respectively:

- N1 mainly concludes fatty tissue and a few fibrous tissue stands;
- P1 shows a prominent duct pattern, and a beaded appearance can be found either in the subareolar area or the upper axillary quadrant;
- P2 indicates server involvement of a prominent duct pattern which may occupy from one-half up to all of the volume of the par-enchyma, and often the connective tissue hyperplasia produces coalescence of ducts in certain areas;
- DY features a general increase in density of the parenchyma and there may, or may not, be a minor component of prominent duct.

Fig. 1 illustrates example results of using the aforementioned evaluation indicators.

3. Multi-criteria mammographic risk analysis

Mammographic images can be analysed in multiple ways from different perspectives, such as Boyd, BI-RADS, Tabár and Wolfe as introduced in the last section. In order to make more reliable decisions, multiple evaluation criteria should be considered collectively, whilst at the same time highly co-related ones should be pre-processed to avoid duplicated information processing. A new multi-criterion mammographic risk analysis approach is therefore proposed herein. Briefly, the proposed framework adapts the recently proposed multi-label fuzzy-rough feature selection approach [17] with the support of either of the following four popular classifiers: NB, Logistic, RF and *k*NN.

The proposed framework is illustrated in Fig. 2. Given a mammographic training dataset, a feature extraction method is applied first to represent images as feature values. Then, the multiple criteria problem is transformed into a concise single label problem using MLFRFS in three steps:

1 The multiple labels due to multiple evaluation criteria are transformed to a single-label problem using power set learning which captures every possible combination of label values as a new label in the transformed space.

2 Association rule learning is employed to remove redundant labels in the transformed single label space.

3 The fuzzy-rough feature selection process (as per Section 2.2) is applied to generate a concise set of features appearing in the learned associations to support the application of classifiers for breast cancer classification tasks.

From this, classifiers are trained recursively, with respect to the selected feature subset and each criterion. After this system is built, given any mammographic image, it will produce a risk criterion value as its risk analysis outcome. The key steps of system development is detailed in the following subsections.

3.1. Feature extraction

Image features can be divided into underlying features and semantic features. The underlying features include image intrinsic features such as color, texture and shape. Semantic features include advanced features such as behavior, emotion and spatial relationship. Most of the image extraction techniques for breast images are based on the underlying features. Of course, the underlying and semantic features can also be combined to achieve better image processing results by jointly using their advantages. A simple feature extraction process is illustrated in Fig. 3, where a square represents a pixel in the original image. In this example, if the gray value in the square is not 0, it is marked as 1. Subsequently each pixel is combined into one column vector.

Many feature extraction approaches for mammogram are available. For instance, visual features of the mammogram can be extracted as per the work of [29]. These may the relative position of the mass, the distance from the mass to the nipple, the size of the mass, the linear texture characteristics, and whether a glitch in included. Also, so-called bag of words features can be introduced using Latent Dirichlet Allocation for mammographic analysis [30], especially regarding image edges in divided image blocks in support of lesion diagnosis. The shape and margin of the breast mass are considered as important features to assist breast cancer diagnosis in [31], where 17 effective shape features are extracted. Advanced shape features may include eccentricity, equivalent diameter, entropy, standard deviation of mass edge, etc., and a tumor image can be divided into round, oval, lobulated shape, and four irregular shapes. In [32], to extract useful information for tumor diagnosis, K-means is utilised to recognise the hidden patterns of the benign and malignant tumors separately, with the computed membership of each tumor to these patterns treated as a feature.

The feature extraction method used in this paper is taken from the established work of [33]. Prior to extract features, all mammograms are preprocessed to determine the area of the breast while removing the background, label and pectoral muscle areas. See Fig. 4a and b for example results of such a segmentation process. Also, with the aim of avoiding adverse effects from microtexture that could appear in certain regions, the breast region is smoothed using a median filter of size 5×5. Gray-level information in combination with fuzzy C-means clustering is employed to group pixels into two separate categories:

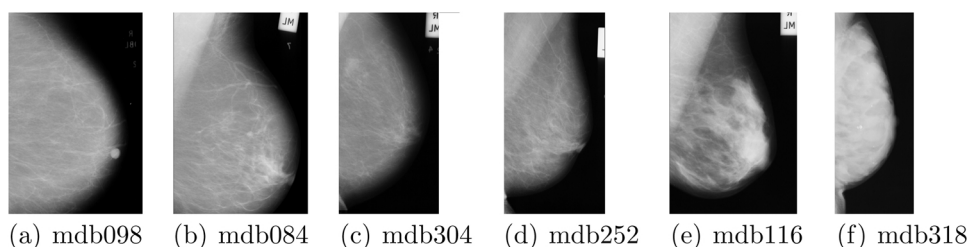(a) mdb098   (b) mdb084   (c) mdb304   (d) mdb252   (e) mdb116   (f) mdb318

Fig. 1. Example mammogram analysis based on Boyd, BI-RADS, Tabár, Wolfe, with results: (a) *SCC* 0%, I, Pattern II, N1; (b) *SCC* 0–10%, II, Pattern III, P1; (c) *SCC* 11–25%, II, Pattern III, P1; (d) *SCC* 26–50%, II, Pattern I, P1; (e) *SCC* 51–75%, III, Pattern IV, P2; (f) *SCC* > 75%, IV, Pattern V, DY; respectively.
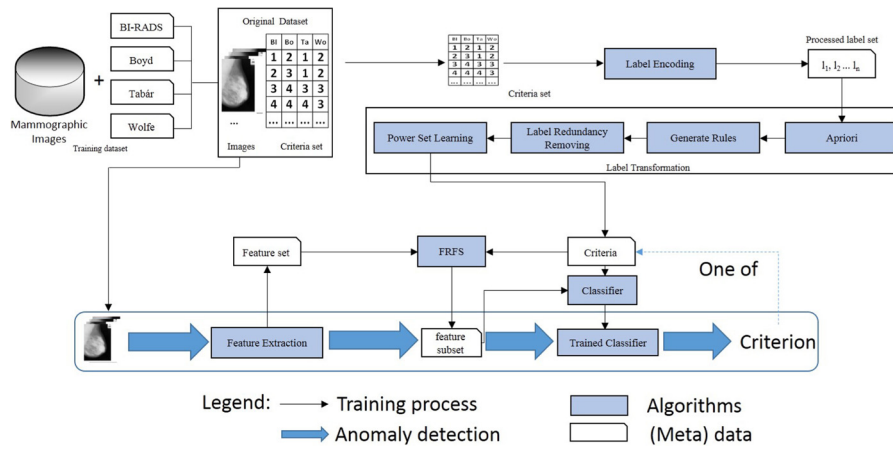
Fig. 2. Architecture of utilising MLFRFS for mammographic risk analysis.

fatty and dense tissues (see Fig. 4c). While running fuzzy C-means, the placement of the initial seed points is one of the central issues in the variation of segmentation results. Two classes are herein initialised with the gray-level values that represent 15% and 85% of the accumulative histogram of the breast pixels per mammogram (representing fatty and dense tissue, respectively). Morphollogical and texture features for both clusters are extracted. In particular, for morphological features, the relative area and the first four histogram moments are calculated, with the four histogram moments being related to the mean intensity, the standard deviation, the skewness, and the kurtosis per cluster. Another set of features derived from co-occurrence matrices are used as texture features. In so doing, each mammogram object is represented by 280 features, 10 derived from morphological characteristics, and the remaining 270 from the extracted texture information.

3.2. Multi-label fuzzy-rough feature selection

The MLFRFS algorithm is shown in Algorithm 3. Label encoding that uses power set learning transforms the multiple criteria into one unified set of label as per Line 2 in the algorithm Then, the multi-label dataset for each instance is converted into a single-label dataset through Line 3. This transformed single-label feature selection problem is then resolved by applying FRFS, with the final reduced feature subset as output. These key steps are detailed in the following.

Algorithm 3. Multi-label fuzzy-rough feature selection (MLFRFS)

MLFRFS $(\mathbb{U}, \varepsilon, \delta)$
Input:

$\mathbb{U}$, raining instances;
C, feature sets;
D, criteria sets;
ε: minimum threshold of support;
δ: minimum threshold of confidence.

Output: R, Reduced feature subset.
1   **Initialise** $R = \varnothing$;
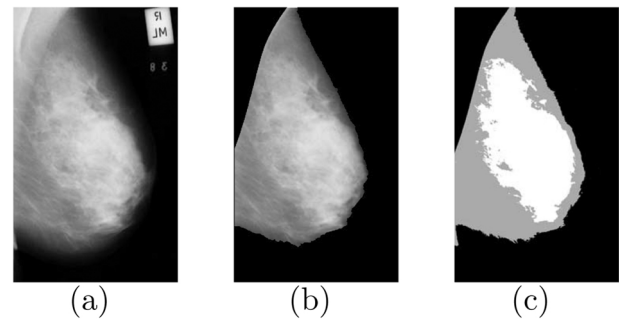
Fig. 4. Preprocessing steps for feature extraction: (a) Original mammogram; (b) Mammogram with redundant areas removed; (c) Mammogram with fuzzy c-mean applied.

2         $L \longleftarrow$ Label Encoding$(\mathbb{U}, \ D)$
3         $L_t \longleftarrow$ Label Transformation $(\{\mathbb{U}, C, L\}, \varepsilon, \delta)$ //Algorithm 4
4         $R \longleftarrow FRFS(\{C, L_t\})$ //Algorithm 2
5   **return** R

3.2.1. Label encoding by power set learning

Label powerset (LP) is a method which transforms a multi-label learning problem into a problem of a single class label. Suppose that a multi-criterion training set $T = (\mathbb{U}, C, D)$ is given, where $\mathbb{U}$ is the instances set, C the features set and D the decision criteria set; and d_j is a criterion within D with l_k being a certain specification of d_j. For instance, if two criteria are used, criterion d_1 with three possible values l_1, l_2 or l_3, and criterion d_2 with two possible values l_4 or l_5, then each image is labeled by one label out of $\{l_1, l_2, l_3\}$, and another out of $\{l_4, l_5\}$, as illustrated in Table 1. In this illustrative example, each possible label can be represented by a binary number as shown in Table 2.

Without losing generality, let L denote a multi-label set, consisting of all labels l_k representing each and every evaluation criterion concerned, and consider each label combination in the power space that
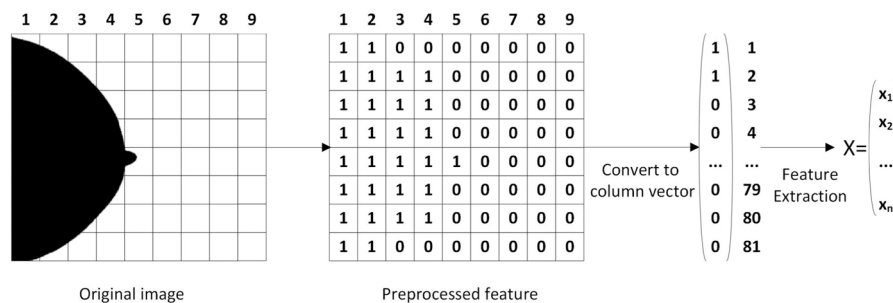
Fig. 3. Example of feature extraction.

Table 1
A simple original dataset.

Instance	Features				Criteria	
	f_1	f_2	f_3	f_4	d_1	d_2
1	0.4512	0.5488	0.3103	0.3732	l_1	l_4
2	0.5084	0.4916	0.3103	0.3567	l_2	l_4
3	0.4956	0.5044	0.3103	0.3275	l_3	l_4
4	0.1647	0.8353	0.4113	0.3897	l_3	l_5
5	0.2533	0.7467	0.3972	0.2625	l_1	l_4

Table 2
Simple dataset after preprocessing.

Instance	d_1			d_2		Label power space	Natural number space
	l_1	l_2	l_3	l_4	l_5		
1	1	0	0	1	0	l_1, l_4	1
2	0	1	0	1	0	l_2, l_4	2
3	0	0	1	1	0	l_3, l_4	3
4	0	0	1	0	1	l_3, l_5	4
5	1	0	0	1	0	l_1, l_4	1

appears in the training set as a new class. Let $\sigma(L): 2^{|L|} \to \mathbb{N}$ be the injective function from the power space of the label space L to the space of natural numbers. Continue the simple illustrative example above, then, the papping results of the given instances are those listed in the rightmost column of Table 2.

The LP method effectively converts the multi-label training set L into the following single-label training set:

$$T' = \{(\mathbf{x}_i, \sigma(L_i)) | 1 \leqslant i \leqslant u\}, \qquad (10)$$

where $L_i \subseteq L$ and u is the number of instances with the following new classes:

$$\wedge(T') = \{\sigma(L_i) | 1 \leqslant i \leqslant u\}. \qquad (11)$$

3.2.2. Label redundancy removal

Dependency between class labels may appear when a problem is transformed from multi-label space to a combined single-label space, which may cause unnecessary computation. In general, there are three cases regarding whether a potential correlation exists between any pair of labels in a multi-label dataset: (1) label l exists but l' does not; (2) labels l and l' appear together; and (3) labels l and l' exist but never occur together. Such relationships can be revealed using association rules. Based on discovered associations, the labels, which can be inferred by any other label, will be removed to reduce the size of the label set, thereby saving unnecessary computation. Following the running example in Section 3.2.1, suppose that two association rules have been generated: r_1 ($\{l_3\} \to \{l_4\}$) and r_2 ($\{l_3\} \to \{l_5\}$), with their effect shown in Table 3. Since the labels that are inferrable by the others need to be removed, label l_4 of data instance 3 and label l_5 of data instance 4 are both eliminated given these two discovered rules.

The label transformation method is summarised in Algorithm 4. Firstly, the internal variables of the algorithm are initialised in Line 1,

Table 3
Converted concise single label.

Instance	Label power space	Natural number space	Reduction by rules	Re-aligned natural number space
1	l_1, l_4	1	l_1, l_4	1
2	l_2, l_4	2	l_2, l_4	2
3	l_3, l_4	3	l_3	3
4	l_3, l_5	4	l_3	3
5	l_1, l_4	1	l_1, l_4	1

and frequent item sets are calculated in Line 2 by calling Apriori. In Lines 3 and 4, the association rules that satisfy minimum support and minimum confidence are generated by calculating the conditional probabilities according to the obtained frequent item sets and the given parameters. Then, the discovered association rules are sorted by their confidence levels. Lines 5 to 11 represent an iterative process of decrementing the label set size with regard to the association rules. In Line 12, the reduced label set for each instance is transformed into a single label by mapping them onto natural numbers. Finally, the algorithm returns the generated concise set of single labels as output.

Algorithm 4. Label transformation (LT)

|	
Label Transformation ($\mathbb{U}$, ε, δ)	
Input:	
	$\mathbb{U}$, instances set;
	L, set of encoding labels;
	ε, minimum threshold of support;
	δ, minimum threshold of confidence.
Output: L_t, Tranformed label set.	
1	$F = \varnothing$, $Rules = \varnothing$;
2	F = Apriori($\{\mathbb{U}, L\}$, ε) //Algorithm 1
3	**Generate** $Rules$ by F and δ
4	**Order** $Rules$ by $confidence$
5	**for** L_i in $\mathbb{U}$
6	**for** Ru (A, B) in $Rules$
7	**if** $A \subseteq L_i$
8	$L_i = L_i - B$
9	**end**
10	**end**
11	**end**
12	Transform labels L into decision class L_t in natural numbers space
13	**return** L_t

3.2.3. Feature selection

Having run the multiple label encoding and transformation processes, a concise set of single labels is generated. This completes the change of the originally multi-label problem into an equivalent single label problem. Therefrom, the FRFS method, as introduced in Section 2.2, is applied for feature selection. Note that, if preferred, any advanced feature selection approach may be applicable. However, FRFS is particularly chosen herein thanks to its popularity and availability, as well as due to its ability to support the handling of uncertain problems which breast cancer risk analysis always involves.

3.3. Classification

As the anomaly analysis based on multiple evaluation criteria has been converted into a concise single label classification problem, any single-label classifier can be applied to carry out the analysis of breast cancer risk. Various AI classifiers have been applied to evaluate breast cancer risks. In this work, four different learning classifiers, namely, NB [24], Logistic Regression [25], RF [26], and kNN [27], are utilised, which are trained using a benchmark mammographic dataset (see later). All these learning classifiers are commonly used in the relevant literature and their workings are generally well-understood. Hence, descriptions of the algorithm details are omitted here, but can be found in the respective references given (and indeed in many other sources).

3.4. Time complexity

The following presents an analysis of all key aspects affecting the time complexity for MLFRFS. Note that the effect of using a combination of different criterion specifications on time is mainly reflected in the Apriori algorithm. Obviously, the time required to run Apriori increases along with the number of criteria used.

- For each instance, suppose that d is the number of criteria to be involved, then generating these item sets requires $O(nd)$ time, where n is the total number of instances.

- In order to generate a k-item set, all $(k-1)$-item sets need to be considered for possible merging. Whether or not to merge depends on whether they have $(k-2)$ items that are in common. Each merging process requires at most $(k-2)$ comparisons. In the best case, each merge produces a viable candidate k-item set. In the worst case, the algorithm needs to merge all pairs of $(k-1)$-item sets [34]. Thus, the overall cost is

$$\sum_{k=2}^{d} (k-2)|C_k| < \text{Cost of merging} < \sum_{k=2}^{d} (k-2)|F_{k-1}|^2). \qquad (12)$$

A hash tree is constructed to store the candidate sets. Because the maximum depth of the tree is k, so the cost of generating the hash tree is $O(\sum_{k=2}^{d} k|C_k|)$. For candidate set pruning, the $(k-2)$ subsets of the k-item set need to be confirmed to be frequent, so the pruning step requires $O(\sum_{k=2}^{d} k(k-2)|C_k|)$.

- For d criterion specifications, Apriori may produce C_d^k item sets of size k, the cost for support counting is therefore, $O(N\sum_k C_d^k)$.

- Each frequent k-item set F_k, can lead up to the creation of $(2^k - 2)$ association rules, ignoring those that have an empty antecedent or consequent ($\varnothing \to F_k$ or $F_k \to \varnothing$). Hence, the total cost for producing association rules is $O(N\Sigma_k(2^k - 2))$.

- For a dimensionality of p, $(p^2 + p)/2$ evaluations of the dependency function may be performed in the worst case, indicating that the cost for computing FRFS is $O((p^2 + p)/2)$.

4. Experimentation

Four different evaluation criterion, including Boyd, BI-RADS, Tabár, and Wolfe as introduced in Section 2.3 are used in conducting this experimentation to validate and evaluate the proposed framework for breast cancer risk analysis.

4.1. Data sets and experimental set-up

All data used in this work is derived from images contained within the Mammographic Image Analysis Society (MIAS) database [20]. It includes a set of Medio-Lateral Oblique (MLO) left and right mammograms of 161 woman (322 samples). Each mammogram has been pre-processed and represented by 280 features. The spatial resolution of the image is $50\,\mu m \times 50\,\mu m$, which is quantised to 8 bits with a linear optical density in the range of [0–3.2].

In order to investigate the efficacy of the proposed approach, systematic comparative studies have been carried out, based on different sets of selected features involving an exhaustive set of combinations of criteria considered. In particular, the dataset using the labels with respect to the Boyd criterion is named as dataset Boyd; similarly, those using the labels regarding the BI-RADS, Tabár and Wolfe criteria are named as datasets BI-RADS, Tabár and Wolfe, respectively. Also, the dataset labeled on the basis of joint criteria Boyd and BI-RADS is denoted as Bo_BI; and that labeled with regards to the conjunctive use of the criteria Boyd, BI-RADS and Tabár is denoted as Bo_BI_Ta. In fact, many other combinations are examined and they are named in a similar manner as these examples, without being explicitly listed here to save space.

Note that in all experimental results reported here, stratified 10×10-fold cross-validation is used for testing. In each 10-FCV, a given dataset is randomly partitioned into 10 subsets, with one single subset retained as the testing data while the remaining 9 subsets for training. The 10×10-fold validation represents this process for 10 times. More specific details on the experimental set-up are given in the relevant subsections below.

4.2. Practical significance of mammographic datasets

The application of the four criteria (namely, Boyd, BI-RADS, Tabár,

Table 4
Labels per criterion.

Criterion	Labels
Boyd	$l_1, l_2, l_3, l_4, l_5, l_6$
BI-RADS	l_7, l_8, l_9, l_{10}
Tabár	$l_{11}, l_{12}, l_{13}, l_{14}, l_{15}$
Wolfe	$l_{16}, l_{17}, l_{18}, l_{19}$

Table 5
Top association rules ranked by confidence.

Order	Association rules	Confidence
1	$l_2, l_{12} \to l_7, l_{16}$	0.97
2	$l_{14}, l_{18} \to l_9, l_5$	0.90
3	$l_7, l_{12} \to l_2, l_{16}$	0.88
4	$l_{12}, l_{16} \to l_7, l_2$	0.84
5	$l_7, l_2 \to l_{12}, l_{16}$	0.83
6	$l_2, l_{16} \to l_7, l_{12}$	0.80

Wolfe) in deriving the associate rules has significant practical implications. To illustrate this and for clarity, Table 4 lists all possible decision labels in accordance to these criteria.

Running the implemented system using the above data and setting, 52 association rules are selected. The confidence levels of the resulting rules can be measured, with the several top-ranked rules listed in Table 5 as an example. Consider the first rule ($l_2, l_{12} \to l_7, l_{16}$). In this rule, l_2 represents the proportion of dense breast tissue in reference to the overall breast area is 10–25%; l_{12} shows the instance is composed by 2% in nodular density, 14% in linear density, 2% in homogeneous fibrous tissue and 82% in radiolucent adipose tissues, respectively; l_7 expresses the breast is almost entirely fatty; and l_{16} states that the instance mainly includes fatty tissue and a few fibrous tissue stands. The first two items indicate that the breast tissue density of this instance is relatively low, and the risk of breast cancer is low, while the latter two indicate that the breast is relatively fatty and the risk is also low. Other rules have their own meaningful practical interpretation too, showing the practical significance of these learned association rules that are well in line with the reality concerning beast cancer risks.

4.3. Effects of multiple criteria

The effects of using different combinations of 1 up to 4 evaluation criteria are looked at in this experiment, by the use of: learning classifier NB, best first search [35] (as the default search strategy of MLFRFS) and 10×10-fold cross validation. The results are shown in Fig. 5, where the line styles and colours encode different pathways to growing the combinations from one criterion to four. For instance, the green dot-dash line in Fig. 5(a) represents the classification accuracies achieved on the Boyd dataset by the use of: only the Boyd criterion, then both Boyd and Tabár, then three crieria of Boyd, Tabár and BI-RADS, and finally, all four criteria, in the process of selecting feature subsets.

The predicted results are directly compared with each underlying criterion for checking on accuracy. As an example, take the case while using the Boyd dataset. The classification accuracy is 53.41% when only the criterion Boyd is used. There are three ways to add one more criterion, BI-RADS, Tabár, or Wolfe, with the resultant classification accuracies being 55.59%, 53.41%, and 54.34%, respectively. Depending on which is the second criterion used, there are two separate cases when adding the next, or the third, criterion. Once the fourth criterion is added, the accuracy is raised to 55.59% for all four different routes. In general, it can be seen that for most cases, the use of more evaluation criteria leads to a higher classification accuracy, and that when all four criteria are combined, the classification accuracy is the highest for all datasets except Boyd.

(a) Boyd

(b) BI-RADS

(c) Tabár

(d) Wolfe
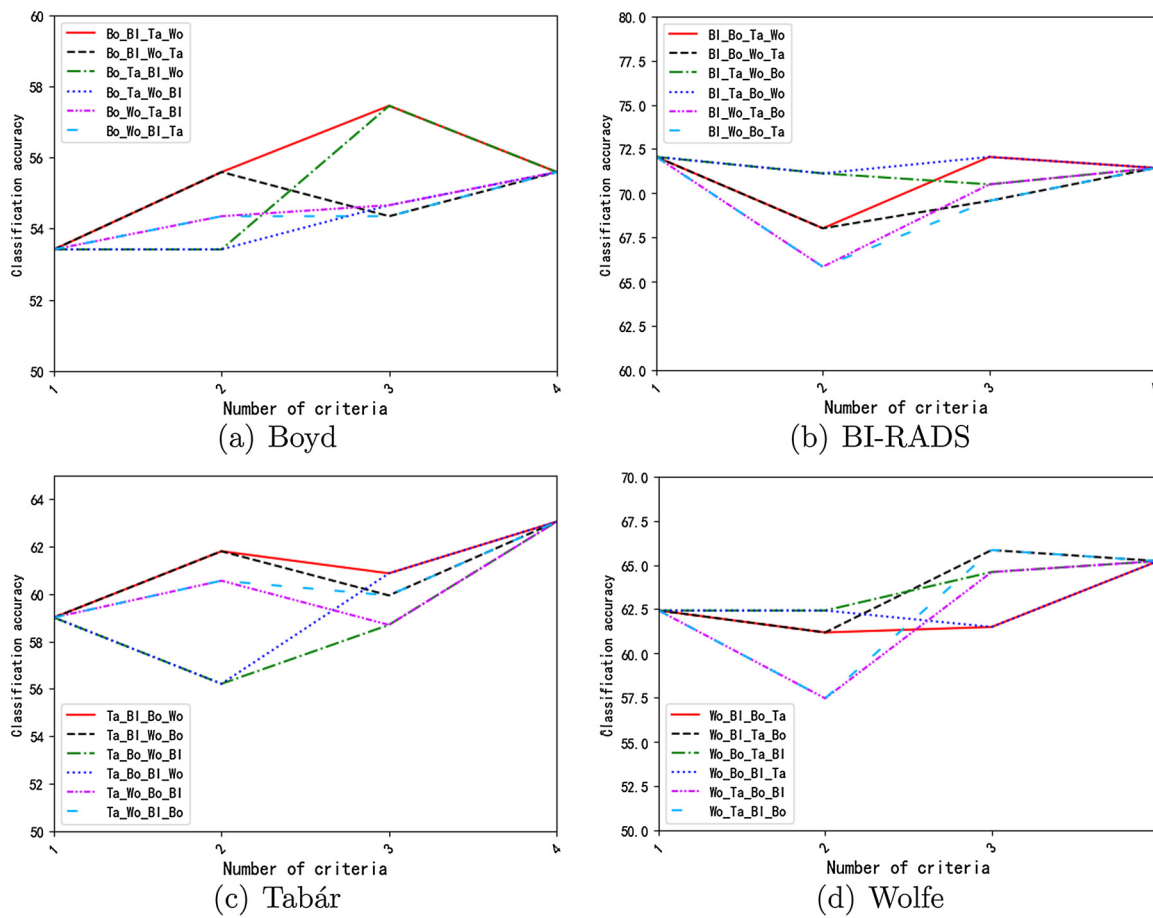
Fig. 5. Classification accuracy using different numbers of criteria.

Table 6
Feature subset size/time (second) for classifying each dataset.

Criterion	Search	MLFRFS	FRFS	Consis	CFS	FDMFS	ORI
Boyd	BF	7/28	6/26	14/1	35/1	3/19	280
	GS	7/18	6/16	14/1	32/1	3/10	280
	LF	7/6	7/7	29/1	14/1	3/4	280
	PSO	8/27	8/27	31/1	48/1	5/417	280
	ME	6/19	6/19	15/1	35/1	3/437	280
BI-RADS	BF	7/28	7/28	15/1	35/1	4/22	280
	GS	7/18	7/18	14/1	35/1	4/13	280
	LF	7/6	6/6	19/1	24/1	4/5	280
	PSO	8/27	7/28	26/1	59/1	9/422	280
	ME	6/19	6/20	15/1	39/1	4 /426	280
Tabár	BF	7/28	6/26	15/1	43/1	4/25	280
	GS	7/18	6/16	15/1	31/1	4/14	280
	LF	7/6	8/7	18/1	17/1	4/6	280
	PSO	8/27	8/28	18/2	48/1	6/414	280
	ME	6/19	6/19	14/1	43/1	4 /425	280
Wolfe	BF	7/28	6/25	14/1	30/1	4/25	280
	GS	7/18	6/16	14/1	30/1	4/13	280
	LF	7/6	7/6	22/1	20/1	4/6	280
	PSO	8/27	8/27	24/1	55/1	5/418	280
	ME	6/19	6/20	13/1	35/1	4 /415	280

4.4. Comparative study

4.4.1. Compared approaches

To further evaluate the performance of the proposed approach on the mammographic dataset, experimental comparisons are carried out against the use of the following popular single criterion feature selection methods, in addition to FRFS itself.

Table 7
Classification accuracy (%) and T-test on Boyd.

Classifier	Search	MLFRFS	FRFS	Consis	CFS	FDMFS	ORI
NB	BF	56.45	53.13	58.74	59.39	38.36 *	57.02
	GS	56.45	53.13 *	58.74 v	58.92 v	38.36 *	57.02
	LF	59.27	55.45 *	52.99 *	59.01	38.36 *	57.02 *
	PSO	49.60	51.40 v	57.03 v	58.40 v	56.01 v	57.02 v
	ME	53.11	55.12 v	59.48 v	60.44 v	38.36 *	57.02 v
Logistic	BF	59.12	58.39	59.44	52.54	41.53 *	53.34
	GS	59.12	58.39	59.44	56.21 *	41.53 *	53.34 *
	LF	58.05	59.56 v	57.08	56.14 *	41.53 *	53.34 *
	PSO	50.90	55.46 v	56.54 v	52.56	57.15 v	53.34 v
	ME	54.03	56.18 v	59.11 v	53.68	41.53 *	53.34
RF	BF	53.00	52.85	60.24	58.61	33.73 *	58.36
	GS	53.00	52.85	60.24 v	59.63 v	33.73 *	58.36 v
	LF	54.54	53.30	54.07	58.17 v	33.73 *	58.36 v
	PSO	46.89	49.51 v	58.37 v	59.63 v	53.11 v	58.36 v
	ME	51.40	53.94 v	59.66 v	59.34 v	33.73 *	58.36 v
kNN	BF	57.96	52.05	60.24	59.08	35.81 *	58.12
	GS	57.96	52.05 *	60.24 v	58.84	35.81 *	58.12
	LF	58.15	58.20	55.94 *	56.20 *	35.81 *	58.12
	PSO	49.29	51.55 v	59.98 v	59.92 v	56.57 v	58.12 v
	ME	52.66	54.01	59.29 v	59.42 v	35.81 *	58.12 v
Summary	BF	(v/ /*)	(0/4/0)	(0/4/0)	(0/4/0)	(0/0/4)	(0/4/0)
	GS	(v/ /*)	(0/2/2)	(3/1/0)	(2/1/1)	(0/0/4)	(1/2/1)
	LF	(v/ /*)	(1/2/1)	(0/2/2)	(1/1/2)	(0/0/4)	(1/1/2)
	PSO	(v/ /*)	(4/0/0)	(4/0/0)	(3/1/0)	(4/0/0)	(4/0/0)
	ME	(v/ /*)	(3/1/0)	(4/0/0)	(3/1/0)	(0/0/4)	(3/1/0)

Table 8
Classification accuracy (%) and T-test on BI-RADS.

Classifier	Search	MLFRFS	FRFS	Consis	CFS	FDMFS	ORI
NB	BF	71.12	72.07	69.17 *	71.87	57.91 *	70.13
	GS	71.12	72.07	67.99 *	71.87	57.91 *	70.13
	LF	69.98	68.57 *	63.70 *	72.82 v	57.91 *	70.13
	PSO	59.57	61.43 v	68.85 v	71.18 v	69.57 v	70.13 v
	ME	65.58	73.02 v	67.48 v	72.30 v	57.91 *	70.13 v
Logistic	BF	72.71	75.53 v	74.68 v	70.81 *	62.54 *	65.01 *
	GS	72.71	75.53 v	75.79 v	70.81 *	62.54 *	65.01 *
	LF	74.51	74.63	74.04	72.09 *	62.54 *	65.01 *
	PSO	62.34	65.12 v	72.23 v	70.61 v	74.07 v	65.01 *
	ME	67.44	75.66 v	74.32 v	71.05 v	62.54 *	65.01 *
RF	BF	71.25	71.89	72.71	74.35 v	62.19 *	73.02 v
	GS	71.25	71.89	72.89	74.35 v	62.19 *	73.02 v
	LF	71.39	70.96	70.69	75.13 v	62.19 *	73.02
	PSO	59.82	61.75 v	73.49 v	74.26 v	71.75 v	73.02 v
	ME	63.46	71.02 v	72.87 v	74.66 v	62.19	73.02 v
kNN	BF	72.08	70.65 *	71.24	74.07 v	62.59 *	71.43
	GS	72.08	70.65 *	72.73	74.07 v	62.59 *	71.43
	LF	70.09	72.82 *	68.88	74.48 v	62.59 *	71.43
	PSO	58.23	58.36	72.05 v	73.63 v	71.98 v	71.43 v
	ME	64.52	72.64 v	72.21 v	74.80 v	62.59 *	71.43 v
Summary	BF	(v/ /*)	(1/2/1)	(1/2/1)	(2/1/1)	(0/0/4)	(1/2/1)
	GS	(v/ /*)	(1/2/1)	(1/2/1)	(2/1/1)	(0/0/4)	(1/2/1)
	LF	(v/ /*)	(0/2/2)	(0/3/1)	(3/0/1)	(0/0/4)	(0/3/1)
	PSO	(v/ /*)	(3/1/0)	(4/0/0)	(4/0/0)	(4/0/0)	(4/0/0)
	ME	(v/ /*)	(4/0/0)	(4/0/0)	(4/0/0)	(0/1/3)	(3/0/1)

Table 10
Classification accuracy (%) and T-test on Wolfe.

Classifier	Search	MLFRFS	FRFS	Consis	CFS	FDMFS	ORI
NB	BF	64.11	62.54 *	66.36 v	68.18 v	51.89 *	66.21 v
	GS	64.11	62.54 *	66.36 v	68.18 v	51.89 *	66.21 v
	LF	65.29	52.46 *	61.99 *	63.48	51.89 *	66.21
	PSO	54.94	57.38 v	66.83 v	68.60 v	53.54	66.21 v
	ME	61.15	60.45	65.39 v	68.58 *	51.89 *	66.21 v
Logistic	BF	67.84	67.23	67.58	66.59	58.52 *	60.57 *
	GS	67.84	67.23	67.58	66.59	58.52 *	60.57 *
	LF	65.15	62.18 *	69.87 v	59.36 *	58.52 *	60.57 *
	PSO	55.84	62.65 v	68.16 v	64.09 v	63.29 v	60.57 v
	ME	65.66	61.55 *	69.16 v	65.89	58.52 *	60.57 *
RF	BF	63.48	60.01 *	65.57 v	67.03 v	54.47 *	65.11
	GS	63.48	60.01 *	65.57 v	67.03 v	54.47 *	65.11
	LF	63.95	53.79 *	64.00	60.03 *	54.47 *	65.11
	PSO	54.72	57.32 v	65.56 v	67.28 v	51.95 *	65.11 v
	ME	58.68	56.54	64.90 v	67.72 v	54.47 *	65.11 v
kNN	BF	65.52	60.28 *	64.44	67.36 v	53.10 *	65.54
	GS	65.52	60.28 *	64.44	67.36 v	53.10 *	65.54
	LF	63.17	53.71	60.28 *	61.88	53.10 *	65.54 v
	PSO	54.54	57.92 v	68.19 v	67.17 v	53.89	65.54 v
	ME	58.11	61.80 v	64.42 v	67.80 v	53.10 *	65.54 v
Summary	BF	(v/ /*)	(0/1/3)	(2/2/0)	(3/1/0)	(0/0/4)	(1/2/1)
	GS	(v/ /*)	(0/1/3)	(2/2/0)	(3/1/0)	(0/0/4)	(1/2/1)
	LF	(v/ /*)	(0/1/3)	(1/1/2)	(0/2/2)	(0/0/4)	(1/2/1)
	PSO	(v/ /*)	(4/0/0)	(4/0/0)	(4/0/0)	(1/2/1)	(4/0/0)
	ME	(v/ /*)	(1/2/1)	(4/0/0)	(3/1/0)	(0/0/4)	(3/0/1)

Table 9
Classification accuracy (%) and T-test on Tabár.

Classifier	Search	MLFRFS	FRFS	Consis	CFS	FDMFS	ORI
NB	BF	62.59	58.17 *	60.21 *	61.65	54.63 *	60.28 *
	GS	62.59	58.17 *	60.21 *	62.49	54.63 *	60.28 *
	LF	57.92	58.18	60.31 v	58.73	54.63 *	60.28 v
	PSO	54.27	54.27	63.28 v	62.06 v	61.10 v	60.28 v
	ME	57.14	59.46 v	62.95 v	62.27 v	54.63 *	60.28 v
Logistic	BF	62.71	59.58 *	64.33 v	58.36 *	56.84 *	55.90 *
	GS	62.71	59.58 *	64.33 v	61.33	56.84 *	55.90 *
	LF	61.12	63.63 v	65.54 v	65.35 v	56.84 *	55.90 *
	PSO	53.75	53.75	62.16 v	57.27 v	58.88 v	55.90 v
	ME	60.45	62.28 v	64.56 v	59.80	56.84 *	55.90 *
RF	BF	60.78	58.12 *	62.77 v	65.57 v	54.69 *	61.03
	GS	60.78	58.12 *	62.77 v	64.49 v	54.69 *	61.03
	LF	58.74	59.45	62.50 v	62.64 v	54.69 *	61.03 v
	PSO	54.57	54.57	62.93 v	64.42 v	60.64 v	61.03 v
	ME	54.35	60.10 v	61.14 v	64.08 v	54.69	61.03 v
kNN	BF	62.76	55.72 *	62.95	64.17	57.09 *	60.87 *
	GS	62.76	55.72 *	62.95	64.32 v	57.09 *	60.87 *
	LF	60.45	59.26 *	63.30 v	61.52	57.09 *	60.87
	PSO	53.58	53.58	63.61 v	63.31 v	60.14 v	60.87 v
	ME	54.11	60.85 v	63.05 v	62.99 v	57.09 v	60.87 v
Summary	BF	(v/ /*)	(0/0/4)	(2/1/1)	(1/2/1)	(0/0/4)	(0/1/3)
	GS	(v/ /*)	(0/0/4)	(2/1/1)	(2/2/0)	(0/0/4)	(0/1/3)
	LF	(v/ /*)	(1/2/1)	(4/0/0)	(2/2/0)	(0/0/4)	(2/1/1)
	PSO	(v/ /*)	(0/4/0)	(4/0/0)	(4/0/0)	(4/0/0)	(4/0/0)
	ME	(v/ /*)	(4/0/0)	(4/0/0)	(3/1/0)	(1/1/2)	(3/0/1)

features along with the degree of redundancy between them. Subsets of features that are highly correlated with the class while having low inter correlation are preferred.

- [37]: In the field of feature selection with fuzzy rough sets, apart from utilising dependency measures to select features, state-of-the-art techniques also include those exploring fuzzy discernibility matrices [21,38]. In FDMFS, a fuzzy identification matrix is constructed first. The initially empty feature subset is then incrementally enriched, by adding the best feature which is evaluated according to a certain heuristic measure over those features appearing in the discernibility function.

4.4.2. Classification accuracy

In order to evaluate the performance of the proposed approach, four different learning classifiers are utilised, namely, NB, Logistic, RF, kNN. This helps give a flavour of utilising MLFRFS to support diverse classification methods. At the same time, different search strategies such as Best First (BF) [35], Greedy Stepwise (GS) [39], Linear Forward (LF) [40], Particle Swarm Optimisation (PSO) [41] and Multi-objective evolutionary (ME) [42] are also used to run the experiments. The results are summarised in Tables 7–10.

Running MLFRFS requires the determination of the parameters of minimum support (*minSup*) and minimum confidence (*minConf*) within the interval [0,1]. Based on empirical investigations, these parameters are set such that *minSup* = 0.1 and *minConf* = 0.7. Recall previously reported results, MLFRFS works best using all four criteria. Thus, the experimental studies described herein are those obtained when MLFRFS uses the feature subset selected under all four criteria jointly. The number of features selected by different feature selection methods and the run time consumed are presented in Table 6.

Comparing the use of different search strategies in conjunction with that of different feature selection methods, the number of selected features by MLFRFS is much smaller than that returned by methods not based on fuzzy-rough sets (i.e., those except FRFS and FDMFS). This shows that MLFRFS is able to select a smaller number of features. In terms of runtime performance, MLFRFS and FRFS are similar while Consis and CFS both spend a shorter time. Nonetheless, a relatively rather large number of features selected by Consis and CFS will greatly

- [22]: This consistency-based approach evaluates the worth of a subset of features by the level of consistency with regard to the class values when the training instances are projected onto a subset of features. Consistency of any subset can never be lower than that of the original full set of features. Hence, the usual practice is to use this subset evaluator in conjunction with a random or exhaustive search to look for the smallest subset with consistency equal to that of the full set.

- [36]: Correlation feature selection is used to evaluate the worth of a subset of features, by considering the prediction ability of individual

Table 11
Confusion matrices and classification accuracies on Boyd.

MLFRFS
(Accuracy=60.87%)

	I	II	III	IV	V	VI
I	16.67	66.67	16.67	0	0	0
II	3.33	80	16.67	0	0	0
III	2.17	30.43	32.61	28.26	6.52	0
IV	0	5.33	6.67	57.33	26.67	4
V	0	0	1.1	24.18	65.93	8.79
VI	0	0	2.27	4.55	27.27	65.91

FRFS
(Accuracy=58.39%)

	I	II	III	IV	V	VI
I	0	83.33	0	0	0	16.67
II	3.33	78.33	13.33	3.33	1.67	0
III	2.17	32.61	21.74	30.43	13.04	0
IV	0	6.67	9.33	56	25.33	2.67
V	0	0	2.2	15.38	70.33	12.09
VI	2.27	2.27	0	6.82	31.82	56.82

Consis
(Accuracy=58.39%)

	I	II	III	IV	V	VI
I	0	83.33	0	0	0	16.67
II	11.67	73.33	15	0	0	0
III	4.35	21.74	39.13	23.91	10.87	0
IV	0	2.67	12	57.33	22.67	5.33
V	0	0	2.2	25.27	62.64	9.89
VI	0	0	2.27	6.82	31.82	59.09

CFS
(Accuracy=55.28%)

	I	II	III	IV	V	VI
I	16.67	50	16.67	0	0	16.67
II	10	70	16.67	3.33	0	0
III	8.7	26.09	34.78	21.74	8.7	0
IV	0	4	17.33	57.33	20	1.33
V	1.1	1.1	1.1	19.78	60.44	16.48
VI	0	0	4.55	4.55	43.18	47.73

FDMFS
(Accuracy=42.24%)

	I	II	III	IV	V	VI
I	0	66.67	0	0	16.67	16.67
II	0	60	5	15	18.33	1.67
III	0	34.78	8.7	21.74	34.78	0
IV	0	20	6.67	20	46.67	6.67
V	0	17.58	3.3	10.99	57.14	10.99
VI	0	6.82	0	2.27	25	65.91

Original
(Accuracy=54.34%)

	I	II	III	IV	V	VI
I	0	33.33	66.67	0	0	0
II	3.33	68.33	21.67	3.33	3.33	0
III	0	21.74	36.96	26.09	15.22	0
IV	0	2.67	16	52	26.67	2.67
V	1.1	2.2	3.3	18.68	59.34	15.38
VI	2.27	4.55	2.27	9.09	27.27	54.55

increase the time required to run the associated classifiers (e.g., Logistic).

Tables 7–10 show the classification performance of MLFRFS regarding different learning classifiers, with different feature selection methods and search strategies. In these tables, a result associated with the symbol "v" or "*" indicates that the corresponding approach performs better or worse in comparison to that utilises the feature subset selected by MLFRFS, given the same environment otherwise (in terms of what feature selection, search mechanism and learning classifier to use or whether the full original dataset is used). Where there is no symbol attached the result implies a statistical tie between the approaches compared.

While using the NB classifier for the dataset Boyd, MLFRFS surpasses all other feature selection methods and over the original dataset, run through the LF search strategy, with an accuracy of 59.27%. Unfortunately, this classifier does not show good classification results when used with other search strategies for the same dataset. Regarding NB for BI-RADS, employing MLFRFS ensures it surpassing the use of the other three feature selection methods when run with BF, GS and LF search strategies (having an accuracy of 71.12%, 71.12%, and 69.98%, respectively). Yet, if run with either PSO or ME search strategy, there is no favourable classification result attained. Using the NB classifier for Tabár, MLFRFS surpasses all other feature selection methods and also, over the original dataset with the BF or GS search strategy (both having an accuracy of 62.59%). However, it does not achieve good classification results using either of the other three search strategies. For NB and

Wolfe, in the LF search strategy, MLFRFS surpasses all other methods with an accuracy of 65.29%, but the use of selected feature subset does not beat that of the original dataset with full features. Although MLFRFS does not manage to return a good classification outcome when used with PSO, it surpasses two of the three alternative feature selection methods while used with either BF, GS or ME, MLFRFS.

For the classifier Logistic and dataset Boyd, adopting one of the BF, GS and LF search strategies, MLFRFS is able to surpass the other three feature selection methods and over the use of the full original dataset, with an accuracy of 59.12%. Running on ME, MLFRFS surpasses two other methods and over the original dataset, with an accuracy of 54.03%. For Logistic and BI-RADS, using the LF search strategy, MLFRFS surpasses the other three methods and over the original dataset, with an accuracy of 74.51%. With the use of BF or GS, MLFRFS exceeds two other methods and over the original dataset, with an accuracy of 72.71%. For Logistic and Tabár, running the BF or GS search strategy, MLFRFS surpasses the other three methods and the original dataset with an accuracy of 61.71%. For Logistic and Wolfe, using BF or GS, MLFRFS surpasses all other methods and over the original dataset with an accuracy of 67.84%. With LF and ME, MLFRFS also exceeds most of the rest.

Now, consider the RF classifier, run over different datasets and with different search strategies, MLFRFS still generally outperforms two of the other three feature selection methods. However, it does not offer a better solution when used with the PSO search strategy. This is likely because using this classifier for this particular dataset, a good number

Table 12
Confusion matrices and classification accuracies on BI-RADS.

MLFRFS (Accuracy=74.22%)

	I	II	III	IV
I	74.58	23.73	0	1.69
II	17.44	60.47	22.09	0
III	0	9.09	84.62	6.29
IV	0	0	35.29	64.71

FRFS (Accuracy=75.78%)

	I	II	III	IV
I	79.66	18.64	1.69	0
II	12.79	66.28	20.93	0
III	0	11.19	83.22	5.59
IV	0	0	38.24	61.76

Consis (Accuracy=75.15%)

	I	II	III	IV
I	77.97	20.34	0	1.69
II	11.63	62.79	25.58	0
III	0.7	9.79	83.22	6.29
IV	0	0	32.35	67.65

CFS (Accuracy=71.12%)

	I	II	III	IV
I	79.66	16.95	3.39	0
II	19.77	58.14	19.77	2.33
III	1.4	11.89	76.22	10.49
IV	0	8.82	23.53	67.65

FDMFS (Accuracy=62.42%)

	I	II	III	IV
I	61.02	20.34	15.25	3.39
II	20.93	39.53	39.53	0
III	0	18.88	76.22	4.9
IV	11.76	0	23.53	64.71

Original (Accuracy=64.91%)

	I	II	III	IV
I	61.02	32.2	6.78	0
II	20.93	46.51	32.56	0
III	1.4	14.69	76.92	6.99
IV	2.94	5.88	23.53	67.65

of complicated features may be required to perform classification, whilst the number of features selected by MLFRFS is rather small. Owing to the initial population of PSO being randomly set, running the particle swarm search strategy may be more capable of returning the required complex features though in bigger sizes.

For the kNN classifier and dataset Boyd, using the LF search strategy, MLFRFS exceeds three feature selection methods and over the use of the full original dataset, with an accuracy of 58.15%. With BF or GS search, MLFRFS exceeds two other methods, having an accuracy of 57.96%. For kNN and BI-RADS, adopting BF or GS, MLFRFS surpasses most of the other methods with an accuracy of 72.08%. However, when using other either of the other three search strategies, MLFRFS does not lead to significantly better classification results. For kNN and Tabár, MLFRFS outperforms the two other methods and the original dataset with an accuracy of 62.76% when using the BF or GS search strategy. For kNN and Wolfe, utilising LF, MLFRFS exceeds all other feature selection methods with an accuracy of 63.17%, though the use of selected features does not beat the use of the original full dataset. Using BF or GS, MLFRFS exceeds the three other methods with an accuracy of 65.52%.

Generally speaking, MLFRFS entails better classification results when the BF, GS or LF search strategy is adopted, while using its own returned feature subset. Such feature subsets are of a much smaller cardinality as compared to the full size of the original features. Using fewer features also saves run time for classification. However, in certain cases where the search strategy PSO or ME is adopted, MLFRFS does not show to offer excellent results. This is mainly because the number of features selected by MLFRFS may be too small, not sufficient to support such search which involves an initial randomly set population.

4.4.3. Confusion matrices

The above experimental results show that MLFRFS consistently offers superior results when used with the Logistic classifier, across a range of search strategies and datasets. In order to better understand how MLFRFS performs particularly with respect to different datasets, confusion matrices obtained from the results of running Logistic with 10×10-FCV are computed, as given in Tables 11–14.

Compared to the alternatives regarding the Boyd dataset, MLFRFS has reduced confusions amongst classes II, V and VI, as shown in Table 11. It successfully classifies 80%, 65.93% and 65.91% instances, for classes II and V and VI, respectively, beating other feature selection methods (with FRFS correctly classifying 78.33%, 70.33% and 56.82%; Consis 73.33%, 62.64% and 59.09%; CFS 70%, 60.44% and 47.73%; FDMFS 60%, 57.14% and 65.91%; and using the full original feature set only reaching 68.33%, 59.34% and 54.55%).

Regarding other datasets, MLFRFS also helps reduce confusions between classes I and III on BI-RADS; between II and IV on Tabár; and between I and III on Wolfe.

4.4.4. Area under ROC curve

Classification accuracy is not the only measure for performance evaluation and may sometimes fail to provide a comprehensive view of a learning classifier, especially for data involving imbalanced classes. A helpful alternative assessment is the receiver operating characteristic curve (ROC). It can be used to analyse and evaluate the predictive power of the learning algorithm through computing the Area-Curve (AUC) metric [43]. Being a statistically consistent measure, it can be particularly effective in assessing class discrimination outcomes [44,45]. Having recognised this, AUC is herein used to analyse the overall classification performance of utilising the feature subsets selected by different feature selection methods. To reduce repetition, this investigation is focussed on the use of the Logistic classifier, supported again with 10×10-fold FCV. The results of the AUC exercise are presented in Table 15, noting that a high value of AUC indicates a better performance. It can be seen from this table that in a great majority of cases, the use of MLFRFS leads to better results.

Table 13
Confusion matrices and classification accuracies on Tabárt.

MLFRFS
(Accuracy=64.60%)

	I	II	III	IV	V
I	76.47	3.36	4.2	15.13	0.84
II	11.32	79.25	9.43	0	0
III	37.5	40	15	7.5	0
IV	29.27	0	0	62.2	8.54
V	3.57	0	3.57	28.57	64.29

FRFS
(Accuracy=61.49%)

	I	II	III	IV	V
I	78.99	3.36	5.88	10.92	0.84
II	9.43	69.81	13.21	1.89	5.66
III	35	42.5	10	12.5	0
IV	34.15	1.22	0	56.1	8.54
V	3.57	0	3.57	32.14	60.71

Consis
(Accuracy=65.22%)

	I	II	III	IV	V
I	78.15	4.2	3.36	13.45	0.84
II	7.55	71.7	20.75	0	0
III	22.5	42.5	32.5	2.5	0
IV	28.05	0	0	59.76	12.2
IV	3.57	0	0	35.71	60.71

CFS
(Accuracy=59.94%)

	I	II	III	IV	V
I	62.18	5.88	11.76	17.65	2.52
II	15.09	60.38	24.53	0	0
III	22.5	35	40	2.5	0
IV	23.17	0	3.66	68.29	4.88
V	3.57	10.71	7.14	25	53.57

FDMFS
(Accuracy=59.01%)

	I	II	III	IV	V
I	80.67	2.52	0	15.97	0.84
II	26.42	64.15	0	7.55	1.89
III	35	47.5	7.5	10	0
IV	43.9	6.1	0	47.56	2.44
V	0	10.71	0	25	64.29

Original
(Accuracy=56.83%)

	I	II	III	IV	V
I	68.07	10.08	5.04	16.81	0
II	13.21	62.26	20.75	1.89	1.89
III	30	22.5	40	5	2.5
IV	30.49	4.88	7.32	52.44	4.88
V	3.57	10.71	3.57	46.43	35.71

4.4.5. T-test

As conventionally adopted in the literature [46], paired t-test is used throughout the present experimental studies to show any statistically significant differences between different approaches. This helps ensure that the results are not obtained by chance. As indicated previously, the baseline reference for the tests is the result obtainable by the feature subset selected by MLFRFS. In particular, for Tables 7–10 and 15, paired t-test results are summarised at the end of each table, counting the number of statistically better (v), equivalent (space) or worse (*) cases for each method in comparison to MLFRFS. In all experiments reported, the threshold of significance is set to 0.05 (as normally done in the literature).

T-test reveals significant differences in the classification accuracies between the use of MLFRFS and that of other feature selection methods (comparing with regard to the employment of the same learning classifier for the same dataset, of course). For example, in Table 10, (0/1/3) in the FRFS column indicates that the feature subset returned by this classical method performs better than MLFRFS in no case, equivalently well in one case, and worse than MLFRFS in three cases.

As shown in Tables 7–10, for search strategies BF, GS and LF, the number of occurrences of "*" in almost every table is much higher than that of "v". That is, MLFRFS leads to better performances in most cases, though the situation is less positive when the PSO or ME search strategy is utilised. Overall, the T-test shows that MLFRFS always entails better classification if there is statistical difference between compared approaches, unless no statistical significance (i.e., a tie) is detected. Note that the sizes of MLFRFS-returned feature subsets tend to be small. Together, the results demonstrate the efficacy of MLFRFS and thus, the potential of employing MLFRFS to perform mammographic risk analysis.

5. Conclusion

The risk analysis of mammographic images is of great practical significance to discover the potential danger of developing breast cancer, supporting doctors in performing generally rather difficult decision-making. The previous studies usually only used a single criterion, as a decision class, to forecast cancer risks. By taking the advantage of multiple criteria, this paper has introduced a novel approach to aiding in automated mammographic risk analysis, with the support of fuzzy rough sets theory. The approach combines feature selection, power set learning, and association rule learning to efficiently exploit the multiple criteria for better risk classification. In so doing, while reflecting a range of domain expertise, the implemented system has the ability of minimising redundancy otherwise caused by the use of such complex criteria. Comparative experimental results have positively demonstrated the efficacy of the proposed approach, in reference to popular existing techniques.

Topics for further research include a more comprehensive investigation into the hierarchical structure between the criteria to analyse their impact upon the performance of feature selection. A granulated representation of the decision indicators may further reveal any underlying relationships between the criteria. With granular representation of the class labels, effective search strategies that may optimise the selected feature subset, in terms of both subset size and quality of selected features for handling multi-label datasets are also very interesting to develop. In addition, to achieve improved risk analyses, how the proposed framework could be better integrated with state-of-the-art fuzzy-rough classifiers [13,47] forms a piece of active research.

Table 14
Confusion matrices and classification accuracies on Wolfe.

MLFRFS (Accuracy=68.94%)

	I	II	III	IV
I	79.03	20.97	0	0
II	13.04	63.04	23.91	0
III	0.93	17.59	69.44	12.04
IV	0	1.67	31.67	66.67

FRFS (Accuracy=67.39%)

	I	II	III	IV
I	77.42	20.97	1.61	0
II	14.13	60.87	23.91	1.09
III	0.93	15.74	70.37	12.96
IV	1.67	6.67	30	61.67

Consis (Accuracy=66.77%)

	I	II	III	IV
I	72.58	27.42	0	0
II	18.48	53.26	27.17	1.09
III	0	12.96	75.93	11.11
IV	3.33	5	26.67	65

CFS (Accuracy=68.32%)

	I	II	III	IV
I	74.19	24.19	1.61	0
II	10.87	66.3	19.57	3.26
III	0.93	17.59	66.67	14.81
IV	1.67	1.67	28.33	68.33

FDMFS (Accuracy=57.45%)

	I	II	III	IV
I	62.9	30.65	4.84	1.61
II	18.48	50	30.43	1.09
III	1.85	24.07	60.19	13.89
IV	6.67	0	35	58.33

Original (Accuracy=61.49%)

	I	II	III	IV
I	64.52	29.03	6.45	0
II	17.39	55.43	25	2.17
III	2.78	15.74	60.19	21.3
IV	1.67	3.33	25	70

Table 15
AUC and T-test.

Classifier	Search	MLFRFS	FRFS	Consis	CFS	FDMFS	ORI
Boyd	BF	0.94	0.64 *	0.83	0.47 *	0.36 *	0.51 *
	GS	0.94	0.64 *	0.83	0.56 *	0.36 *	0.51 *
	LF	0.93	0.74 *	0.79	0.92	0.36 *	0.51 *
	PSO	0.63	0.76 v	0.75	0.55	0.77 v	0.51
	ME	0.92	0.78	0.80	0.72	0.36 *	0.51 *
BI-RADS	BF	0.96	0.97	0.96	0.94 *	0.90 *	0.89 *
	GS	0.96	0.97	0.95	0.94 *	0.90 *	0.89 *
	LF	0.95	0.96	0.97	0.95	0.90 *	0.89 *
	PSO	0.89	0.92 v	0.95 v	0.93	0.95 v	0.89
	ME	0.95	0.96	0.94	0.94	0.90 *	0.89 *
Tabár	BF	0.85	0.83	0.86	0.80	0.80 *	0.79 *
	GS	0.85	0.83	0.86	0.86	0.80 *	0.79 *
	LF	0.84	0.86	0.87	0.87	0.80 *	0.79 *
	PSO	0.81	0.81	0.85	0.79	0.84 v	0.79
	ME	0.81	0.84 v	0.87 v	0.81	0.80	0.79
Wolfe	BF	0.97	0.97	0.96	0.95	0.92 *	0.87 *
	GS	0.97	0.97	0.96	0.95	0.92 *	0.87 *
	LF	0.96	0.94	0.98	0.83 *	0.92 *	0.87 *
	PSO	0.87	0.91 v	0.94 v	0.93 v	0.95 v	0.87
	ME	0.95	0.96	0.95	0.94	0.92 *	0.87 *
Summary	All	(v/ /*)	(4/13/3)	(3/17/0)	(1/14/5)	(4/1/15)	(0/5/15)

Conflicts of interest

The authors declared that they have no conflicts of interest to this work.

Acknowledgments

References

[1] Bray F, McCarron P, Parkin DM. The changing global patterns of female breast

cancer incidence and mortality. Breast Cancer Res 2004;6(6):229.

[2] Gabe R, Duffy SW, Mackay J, Anderson E, Duffy S, Ellis T, et al. The challenge of evaluating annual mammography screening for young women with a family history of breast cancer. J Med Screen 2006;13:177–82.

[3] Boyd N, Byng J, Jong R, Fishell E, Little L, Miller A, et al. Quantitative classification of mammographic densities and breast cancer risk: results from the Canadian national breast screening study. JNCI: J Natl Cancer Inst 1995;87(9):670–5.

[4] Liberman L, Menell JH. Breast imaging reporting and data system (bi-rads). Radiol Clin 2002;40(3):409–30.

[5] Tabár L, Tot T, Dean PB. Breast cancer: the art and science of early detection with mammography: perception, interpretation, histopathologic correlation. Thieme Stuttgart; 2005.

[6] Wolfe JN. Risk for breast cancer development determined by mammographic parenchymal pattern. Cancer 1976;37(5):2486–92.

[7] Chen S, Wang J-q, Zhang H-y. A hybrid pso-svm model based on clustering algorithm for short-term atmospheric pollutant concentration forecasting. Technol Forecast Soc Change 2019;146:41–54.

[8] Peng H-G, Shen K-W, He S-S, Zhang H-Y, Wang J-Q. Investment risk evaluation for new energy resources: an integrated decision support model based on regret theory and electre III. Energy Convers Manag 2019;183:332–48.

[9] Wang L, Peng J-J, Wang J-Q. A multi-criteria decision-making framework for risk ranking of energy performance contracting project under picture fuzzy environment. J Clean Prod 2018;191:105–18.

[10] Prathibha B, Sadasivam V. Breast tissue characterization using variants of nearest neighbour classifier in multi texture domain. J Inst Eng (India) Part CP: Comput Eng Div 2010;91:7–13.

[11] Mohamed H, Mabrouk MS, Sharawy A. Computer aided detection system for micro calcifications in digital mammograms. Comput Methods Programs Biomed 2014;116(3):226–35.

[12] Qu Y, Shang C, Wu W, Shen Q. Evolutionary fuzzy extreme learning machine for mammographic risk analysis. Int J Fuzzy Syst 2011;13(4):282–91.

[13] Qu Y, Shang C, Shen Q, Mac Parthaláin N, Wu W. Kernel-based fuzzy-rough nearest-neighbour classification for mammographic risk analysis. Int J Fuzzy Syst 2015;17(3):471–83.

[14] Verma B, Zakos J. A computer-aided diagnosis system for digital mammograms based on fuzzy-neural and feature extraction techniques. IEEE Trans Inf Technol Biomed 2001;5(1):46–54.

[15] Fu J, Lee S, Wong S, Yeh J, Wang A, Wu H. Image segmentation feature selection and pattern classification for mammographic microcalcifications. Comput Med Imaging Graph 2005;29(6):419–29.

[16] Meselhy EM, Faye I, Belhaouari SB. A statistical based feature extraction method for breast cancer diagnosis in digital mammogram using multiresolution representation. Comput Biol Med 2012;42(1):123–8.

[17] Qu Y, Rong Y, Deng A, Yang L. Associated multi-label fuzzy-rough feature selection. Fuzzy systems association and international conference on soft computing and intelligent systems 2017:1–6.

[18] Tsoumakas G, Vlahavas I. Random k-labelsets: an ensemble method for multilabel classification. European conference on machine learning 2007:406–17.

[19] Agrawal R, Imieliński T, Swami A. Mining association rules between sets of items in large databases. Proceedings of the 1993 ACM SIGMOD international conference on management of data 1993:207–16.

[20] Suckling J. The mammographic image analysis society digital mammogram database. Digit Mammo 1994:375–86.

[21] Jensen R, Shen Q. New approaches to fuzzy-rough feature selection. IEEE Trans Fuzzy Syst 2009;17(4):824–38.

[22] Liu H, Setiono R. Feature selection and classification – a probabilistic wrapper approach. International conference on industrial and engineering applications of

artificial intelligence and expert systems 1996:419–24.

[23] Hall MA. Correlation-based feature selection for machine learning. 1999.

[24] John GH, Langley P. Estimating continuous distributions in bayesian classifiers. Proceedings of the eleventh conference on uncertainty in artificial intelligence. 1995. p. 338–45.

[25] Le Cessie S, Van Houwelingen JC. Ridge estimators in logistic regression. J R Stat Soc Ser C (Appl Stat) 1992;41(1):191–201.

[26] Breiman L. Random forests. Mach Learn 2001;45(1):5–32.

[27] Aha DW, Kibler D, Albert MK. Instance-based learning algorithms. Mach Learn 1991;6(1):37–66.

[28] Agrawal R, Srikant R. Fast algorithms for mining association rules. Proc. VLDB conference 1994:487–99.

[29] Velikova M, Lucas PJF, Samulski M, Karssemeijer N. On the interplay of machine learning and background knowledge in image interpretation by bayesian networks. Artif Intell Med 2013;57(1):73–86.

[30] Wang Y, Li J, Gao X. Latent feature mining of spatial and marginal characteristics for mammographic mass classification. Neurocomputing 2014;144(1):107–18.

[31] Vadivel A, Surendiran B. A fuzzy rule-based approach for characterization of mammogram masses into bi-rads shape categories. Comput Biol Med 2013;43(4):259–67.

[32] Zheng B, Sang WY, Lam SS. Breast cancer diagnosis based on feature extraction using a hybrid of k-means and support vector machine algorithms. Expert Syst Appl 2014;41(4):1476–82.

[33] Oliver A, Freixenet J, Marti R, Pont J, Pérez E, Denton ER, et al. A novel breast tissue density classification methodology. IEEE Trans Inf Technol Biomed 2008;12(1):55–65.

[34] Tan P-N. Introduction to data mining. Pearson Education India; 2018.

[35] Dechter R, Pearl J. Generalized best-first search strategies and the optimality of a. J ACM (JACM) 1985;32(3):505–36.

[36] Lee C, Lee GG. Information gain and divergence-based feature selection for machine learning-based text categorization. Inf Process Manag 2006;42(1):155–65.

[37] Kononenko I. Analysis and extension of relief. The European conference on machine learning and principles and practice of knowledge discovery in databases. 1994. p. 171–82.

[38] Cornelis C, Martín GH, Jensen R, Ślęzak D. Feature selection with fuzzy decision reducts. International conference on rough sets and knowledge technology. 2008. p. 284–91.

[39] Caruana R, Freitag D. Greedy attribute selection. Machine learning proceedings 1994. 1994. p. 28–36.

[40] Gutlein M, Frank E, Hall M, Karwath A. Large-scale attribute selection using wrappers. 2009 IEEE symposium on computational intelligence and data mining. 2009. p. 332–9.

[41] Wang X, Yang J, Teng X, Xia W, Jensen R. Feature selection based on rough sets and particle swarm optimization. Pattern Recognit Lett 2007;28(4):459–71.

[42] Deb K. Multi-objective optimization using evolutionary algorithms vol. 16. John Wiley & Sons; 2001.

[43] Beck JR, Shultz EK. The use of relative operating characteristic (roc) curves in test performance evaluation. Arch Pathol Lab Med 1986;110(1):13–20.

[44] Huang J, Ling CX. Using auc and accuracy in evaluating learning algorithms. IEEE Trans Knowl Data Eng 2005;17(3):299–310.

[45] Mason SJ, Graham NE. Areas beneath the relative operating characteristics (roc) and relative operating levels (rol) curves: statistical significance and interpretation. Q J R Meteorol Soc 2010;128(584):2145–66.

[46] Nadeau C, Bengio Y. Inference for the generalization error. Mach Learn 2003;52(3):239–81.

[47] Qu Y, Shang C, Mac Parthaláin N, Wu W, Shen Q. Multi-functional nearest-neighbour classification. Soft Comput 2017;22(8):2717–30.