# *Aiding neural network based image classification with fuzzy-rough feature selection*

Shang, Changjing; Shen, Qiang

*Published in:*
IEEE International Conference on Fuzzy Systems, 2008. FUZZ-IEEE 2008

*DOI:*
10.1109/FUZZY.2008.4630488

*Publication date:*
2008

*Citation for published version (APA):*
Shang, C., & Shen, Q. (2008). Aiding neural network based image classification with fuzzy-rough feature selection. In IEEE International Conference on Fuzzy Systems, 2008. FUZZ-IEEE 2008 (pp. 976-982). IEEE Press. https://doi.org/10.1109/FUZZY.2008.4630488

# Aiding Neural Network Based Image Classification with Fuzzy-Rough Feature Selection

Changjing Shang and Qiang Shen

*Abstract*— **This paper presents a methodological approach for developing image classifiers that work by exploiting the technical potential of both fuzzy-rough feature selection and neural network-based classification. The use of fuzzy-rough feature selection allows the induction of low-dimensionality feature sets from sample descriptions of real-valued feature patterns of a (typically much) higher dimensionality. The employment of a neural network trained using the induced subset of features ensures the runtime classification performance. The reduction of feature sets reduces the sensitivity of such a neural network-based classifier to its structural complexity. It also minimises the impact of feature measurement noise to the classification accuracy. This work is evaluated by applying the approach to classifying real medical cell images, supported with comparative studies.**

## I. Introduction

Image classifiers implemented with a neural network have enjoyed much success in many application domains. However, complex application problems such as real-life medical image modelling and analysis have emphasised the issues of feature set dimensionality reduction and feature semantics preservation. In particular, to capture the essential characteristics of a real image, many features may have to be extracted without explicit knowledge of what properties might best represent the original image *a priori*. Yet, generating more features increases computational complexity and in the mean time, not all such features may be essential to perform classification. Due to measurement noise use of extra features may even cause the reduction of the overall representational power of the feature set and hence the classification accuracy. Thus, it is desirable to employ a method that can determine the most significant features, based on sample measurements, to simplify a neural network-based classifier.

The above observation reflects the need in solving many real-world classification problems. For example, comparing normal and abnormal blood vessel structures plays an important role in pathology and medicine [13]. Recent development of nuclear stains and Laser Scanning Confocal Microscopy (LSCM) has allowed the study of the structure of blood vessels at the cellular or sub-cellular level. Central to the classification of cell images is the capture and analysis of their underlying features. Many feature extraction methods are available to yield various kinds of characteristic description of a given image. However, little knowledge is available as to what features may be most useful to provide the discrimination power between normal and abnormal cells and between cells of a different type.

Computationally, it is impractical to generate many features and then to perform classification based on these features for rapid diagnosis. A common practice is therefore to generate a good number of features and select from them the most informative ones off-line, and then to use those selected only for classification on-line. For such medical applications, the features produced ought to have an embedded meaning and such meaning should not be altered during the selection process. This makes it difficult to utilise conventional dimensionality reduction techniques such as Principal Components Analysis (PCA) [3]. This is because PCA irreversibly destroys the underlying semantics of the original feature set.

This paper presents an alternative approach to aid building neural network-based classifiers by exploiting the potential of fuzzy-rough sets [6], [15] for semantics-preserving feature selection. The employment of a fuzzy-rough feature selection mechanism allows the induction of low-dimensionality feature sets from sample descriptions of feature patterns of a (typically much) higher dimensionality. Although crisp rough sets [11] might be adopted for the same purpose [14], they cannot work against real-valued image features unless further preprocessing mechanisms like data discretisation are used. This would require boolean partitions over the domain of the underlying features extracted from the original images. Unfortunately, for medical diagnoses, this requirement is generally very difficult to satisfy. Use of fuzzy-rough sets considerably reduces such difficulties.

The rest of this paper is organised as follows. Section II introduces the medical image classification problem considered herein. This, from the viewpoint of real-world application, justifies the need for the present research and sets up the background for the experimental investigations to be reported later. Section III describes the key techniques used in the work, including feature extraction and fuzzy-rough feature selection. For completeness, it also briefly outlines the structure and learning process of multi-layer feedforward neural network-based classifiers in the present context. Section IV shows the results of applying this work to the given medical application, supported by comparative studies. The paper is concluded in Section V with further work pointed out.

## II. Cell Images and their Classification

The samples of subcutaneous blood vessels used in this research were taken from patients suffering critical limb ischaemia immediately after leg amputation. The level of

Changjing Shang and Qiang Shen are with the Department of Computer Science, Abersytwyth University, SY23 3DB, Wales, UK (email: {cns, qqs}@aber.ac.uk).

978-1-4244-1819-0/08/$25.00©2008 IEEE

amputation was always taken to be in a non-ischaemic area. The vessel segments obtained from this area represent internal proximal (normal) arteries, whilst the distal portion of the limb shows ischaemic (abnormal) ones.

Images were collected using an inverted (Nikon Diaphot) microscope fitted with a Noran Odyssey LSCM of a x40 objective [13]. Serial optical slices were taken along the $z$ axis ($1\mu m$ apart), starting with the LSCM focussing on the top of a blood vessel in the $x$-$y$ plane, and moving down from the layer of adventitial cells, through the layer of smooth mussel cells, to the last layer of endothelial cells. Nine of these stacks were captured in different regions along the vessel length from different tissue samples. The resulting image database consists of 318 section images, each sized $512 \times 512$ with the grey levels ranging from 0 to 255. Among these images, 154 were obtained from 4 proximal, non-ischaemic vessels and the rest from 5 distal, ischaemic vessels.

Examples of the three types of cell image taken from non-ischaemic resistance arteries are shown in Fig. 1. Their counterparts taken from ischaemic resistance arteries are shown in Fig. 2. Note that many of such images for a given problem case may seem to be rather similar by eye. It is therefore a difficult task for visual inspection and classification. Building an image classifier to automatically classify such images forms the ultimate task of the present work.

## III. Techniques Employed

### A. Feature extraction with fractal models

To capture and represent many possibly essential characteristics of a given image, fractal models [1], [7] are used here for feature extraction. Of course, this does not affect the underlying approach taken in this paper, as any other feature extraction techniques may be equally applicable.

Fractal models are typically used to characterise the roughness of an image surface at various scales, which is generally greater than the topological (intuitive) dimension. Different definitions and their associated computational algorithms exist for determining the fractal dimensions (FDs). Within this work, FDs are computed via the estimation of the variograms of an image surface. A brief overview of this approach is given below.

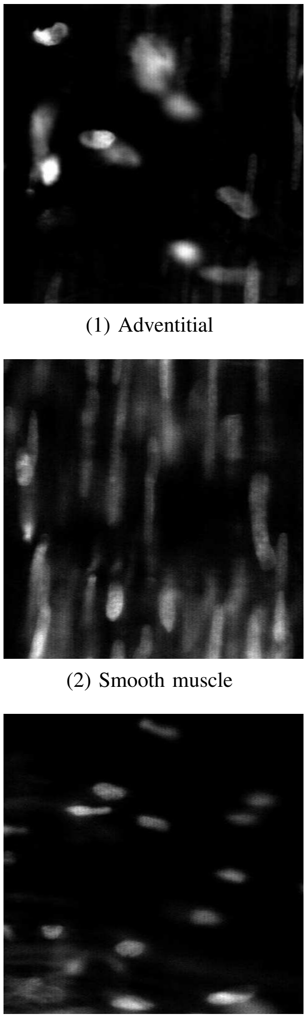
Without losing generality, an image $Y = \{y(s)\}$ is here assumed to be a Gaussian random field defined on an $M \times M$ lattice $\Omega$, where $y(s)$ denotes the grey level of a pixel at location $s = (i, j), i, j = 0, 1, \ldots, M-1$. Given an image $Y$, its fractal dimension $D$ approximately satisfies the following:

$$v(d) = c\ d^{(6-2D)} = c\ d^{a} \qquad (1)$$

where $a$ is termed the fractal index, $c$ is a constant and

$$v(d) = E\{y(s+d) - y(s)\} \qquad (2)$$

which is the variogram of the image, with $d$ denoting the distance between pairs of observations concerned.



(1) Adventitial

(2) Smooth muscle

(3) Endothelial

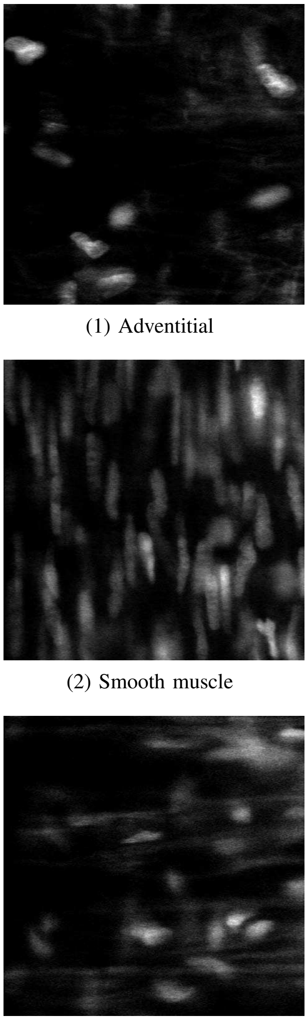
Fig. 1. Section cell images of proximal non-ischaemic subcutaneous blood vessels, taken from a human lower limb.

Applying the Least Squares fitting algorithm [7] to model (1), an estimate of the fractal index $\hat{a}$ $(0 \leq \hat{a} \leq 2)$ can be obtained. This leads to an estimation of the fractal dimension of $Y$ such that

$$\hat{D} = 3 - 0.5\hat{a} \qquad (3)$$

The estimated FD has a strong intuitive appeal: If the surface is very smooth, then the fractal dimension is two; if, however, the surface is extremely rough and irregular, then the fractal dimension approaches the limit of three.

Note that in the above, the variogram of an image and hence its FD are both estimated at a fixed image resolution level. This is done without specifying any spatial direction along which the set of pairs of observations is constructed. That is, the image is assumed to be isotropic. By varying the



(1) Adventitial

(2) Smooth muscle

(3) Endothelial

Fig. 2. Section cell images of distal ischaemic subcutaneous blood vessels, taken from a human lower limb.

resolution level [7] of the image, a set of isotropic fractal features can therefore be generated.

By imposing a constraint over the direction along which observations are obtained, a different variogram and fractal dimension can be estimated over any fixed resolution level. Such resulting fractal dimensions are termed directional fractals (DFs), as opposed to the conventional isotropic FDs that are measured over all possible directions. Obviously, specifying $N$ different directions leads to $N$ different DFs, assuming that the images under consideration are all aligned with respect to a common coordinate origin.

In addition to FDs, in order to capture other potentially significant information embedded in an image, conventional statistical measures such as the mean and standard deviation (STD) can also be utilised. In so doing, a given image is represented by a feature pattern consisting of a certain number of multi-resolution and directional fractals and of simple statistical measures. As to which of such features are indeed essential to perform classification is of course another matter. It is the determination of those most informative features that forms the start-point of this research.

### B. Fuzzy-rough sets and feature selection

Fuzzy-rough feature selection [6], [15] is concerned with the reduction of information or decision systems through the use of fuzzy-rough sets. Let $I = (\mathbb{U}, \mathbb{A})$ be an information system, where $\mathbb{U}$ is a non-empty set of finite objects (the universe of discourse) and $\mathbb{A}$ is a non-empty finite set of attributes such that $a : \mathbb{U} \to V_a$ for every $a \in \mathbb{A}$, with $V_a$ being the set of values that attribute $a$ may take. For decision systems, $\mathbb{A} = \{\mathbb{C} \cup \mathbb{D}\}$ where $\mathbb{C}$ is the set of conditional features and $\mathbb{D}$ is the set of decision values. Based on these notions, the basic concepts most relevant to the present work of fuzzy-rough feature selection are outlined below:

*1) Fuzzy equivalence classes:* Fuzzy equivalence classes [4], [10], [15] are central to the fuzzy-rough set approach in the same way that crisp equivalence classes are central to classical rough sets. For decision problems, this means that the decision values and the conditional values may all be fuzzy. The concept of crisp equivalence classes can be extended by the inclusion of a fuzzy similarity relation $S$ on the universe, which determines the extent to which two elements are similar in $S$. The following properties hold as usual:

- Reflexivity ($\mu_S(x,x)$ = 1)
- Symmetry ($\mu_S(x,y)$ = $\mu_S(y,x)$)
- Transitivity ($\mu_S(x,z) \geq \mu_S(x,y) \wedge \mu_S(y,z)$)

Using the fuzzy similarity relation, the fuzzy equivalence class $[x]_S$ for objects close to $x$ can be defined:

$$\mu_{[x]_S}(y) = \mu_S(x,y) \tag{4}$$

Obviously, this definition degenerates to the normal definition of equivalence classes when $S$ is crisp. Note that the family of normal fuzzy sets produced by a fuzzy partitioning of the universe of discourse can play the role of fuzzy equivalence classes [4].

*2) Fuzzy lower and upper approximations:* These are fuzzy extensions of their crisp counterparts. Informally, in crisp rough set theory, the lower approximation of a set contains those objects that belong to it with certainty. The upper approximation of a set contains the objects that possibly belong.

Formally, given a subset $P$ of features, the fuzzy $P$-lower and $P$-upper approximations are defined as:

$$\mu_{\underline{P}X}(x) = \sup_{F\in\mathbb{U}/P} min(\mu_F(x), \inf_{y\in\mathbb{U}} max\{1-\mu_F(y), \mu_X(y)\}) \tag{5}$$

$$\mu_{\overline{P}X}(x) = \sup_{F\in\mathbb{U}/P} min(\mu_F(x), \sup_{y\in\mathbb{U}} min\{\mu_F(y), \mu_X(y)\}) \tag{6}$$

where $\mathbb{U}/P$ stands for the partition of the universe of discourse, $\mathbb{U}$ with respect to $P$, and $F_i$ denotes a fuzzy equivalence class belonging to $\mathbb{U}/P$. Note that although the universe of discourse in feature reduction is finite, this is not the case in general, hence the use of $sup$ and $inf$ above. Incidentally, it is the tuple $< \underline{P}X, \overline{P}X >$ that is called a fuzzy-rough set.

*3) Partition of the Universe of Discourse:* For an individual feature, $a \in \mathbb{A}$, the partition of the universe by $\{a\}$ is defined by

$$\mathbb{U}/IND(\{a\}) = \{\{x|a(x) = \alpha,\ x \in \mathbb{U}\}|\alpha \in V_a\} \tag{7}$$

Clearly, this is the collection of fuzzy equivalence classes for that feature $a$ itself.

Of course, for feature selection purposes, it is necessary to find the dependency between various subsets of the original feature set. For instance, it may be necessary to be able to determine the degree of dependency of the decision feature(s) with respect to feature set $P = \{a, b\}, a, b \in \mathbb{A}$. In the crisp case, $\mathbb{U}/P$ contains sets of objects grouped together that are indiscernible according to both features $a$ and $b$. In the fuzzy case, objects may belong to many equivalence classes, so the cartesian product of $\mathbb{U}/IND(\{a\})$ and $\mathbb{U}/IND(\{b\})$ must be considered in determining $\mathbb{U}/P$. In general,

$$\mathbb{U}/P = \otimes\{a \in P : \mathbb{U}/IND(\{a\})\} \tag{8}$$

For example, if $P = \{a, b\}$, $\mathbb{U}/IND(\{a\}) = \{N_a, Z_a\}$ and $\mathbb{U}/IND(\{b\}) = \{N_b, Z_b\}$, then

$$\mathbb{U}/P = \{N_a \cap N_b, N_a \cap Z_b, Z_a \cap N_b, Z_a \cap Z_b\}$$

In so doing, each set in $\mathbb{U}/P$ denotes an equivalence class. The extent to which an object belongs to such an equivalence class is therefore calculated by using the conjunction of constituent fuzzy equivalence classes, say $F_i$, $i = 1, 2, ..., n$:

$$\mu_{F_1\cap...\cap F_n}(x) = min(\mu_{F_1}(x), \mu_{F_2}(x), ..., \mu_{F_n}(x)) \tag{9}$$

*4) Fuzzy-rough feature dependency:* The present research builds on the notion of fuzzy lower approximation to enable reduction of datasets containing real-valued features. Proposed as an extension of crisp rough feature selection, its working is expected to become identical to the crisp approach when dealing with discrete-valued features.

Thus, by the extension principle, the membership of an object $x \in \mathbb{U}$, belonging to the fuzzy positive region can be defined by (union of the lower approximations):

$$\mu_{POS_P(Q)}(x) = \sup_{X\in\mathbb{U}/Q} \mu_{\underline{P}X}(x) \tag{10}$$

Object $x$ will not belong to the positive region only if the equivalence class it belongs to is not a constituent of the positive region. This is equivalent to the crisp version where objects belong to the positive region only if their underlying equivalence class does so.

Using the definition of the fuzzy positive region, a useful dependency function between a set of features $Q$ and another set $P$ can be introduced as defined by:

$$\gamma'_P(Q) = \frac{|\mu_{POS_P(Q)}(x)|}{|\mathbb{U}|} = \frac{\sum_{x\in\mathbb{U}} \mu_{POS_P(Q)}(x)}{|\mathbb{U}|} \tag{11}$$

As with crisp rough sets, the dependency of $Q$ on $P$ is the proportion of objects that are discernible out of the entire dataset. In the present approach, this corresponds to determining the fuzzy cardinality of $\mu_{POS_P(Q)}(x)$ divided by the total number of objects in the universe.

*5) Fuzzy-rough* QuickReduct *algorithm:* The fuzzy-rough feature selection algorithm, named fuzzy-rough QuickReduct, is derived on the basis of the above fuzzy-rough dependency measure [15]. It borrows the ideas from the crisp version of QuickReduct originally proposed in [2], to direct the search for quality subset of features. The algorithm is given in Fig. 3. Fundamentally, it employs the fuzzy-rough dependency function $\gamma'$ to choose which features to add to the current subset of features. The algorithm terminates when the addition of any remaining feature does not increase the dependency.

```
FRQuickReduct($\mathbb{C}$,$\mathbb{D}$).
$\mathbb{C}$, the set of all conditional features;
$\mathbb{D}$, the set of decision features.

$R \leftarrow \{\}$, $\gamma'_{best} \leftarrow 0$, $\gamma'_{prev} \leftarrow 0$
do
    $T \leftarrow R$
    $\gamma'_{prev} \leftarrow \gamma'_{best}$
    $\forall x \in (\mathbb{C} - R)$
        if $\gamma'_{R\cup\{x\}}(\mathbb{D}) > \gamma'_T(\mathbb{D})$
            $T \leftarrow R \cup \{x\}$
            $\gamma'_{best} \leftarrow \gamma'_T(\mathbb{D})$
    $R \leftarrow T$
until $\gamma'_{best} == \gamma'_{prev}$
return $R$
```

Fig. 3. The fuzzy-rough QuickReduct algorithm

As with the original algorithm, for a dimensionality of $n$, the worst case dataset will result in $(n2 + n)/2$ evaluations of the dependency function. However, fuzzy-rough set-based feature selection is used off-line for dimensionality reduction prior to any involvement of an on-line system (e.g. a classifier) which will employ those features belonging to the resultant feature subset. Thus, this operation has no negative impact upon the run-time efficiency of the system.

### *C. Multilayer feedforward neural network for classification*

Each of the classifiers implemented herein consists of a feature extractor (see Section III-A) and a multilayer feedforward neural network (MFNN) based classifier, with these two sub-systems connected in series.

It is well-known that an MFNN accomplishes classification by mapping input feature patterns onto their underlying image classes. The design of each MFNN classifier is thus straightforward: The number of nodes in its input layer is set to that of the dimensionality of a given feature set produced by the feature extractor, and the number of nodes within its output layer is set to the number of underlying classes of interest. The internal structure of the network is designed to be flexible and may contain one or two hidden layers. (What actual number of internal layers and that of hidden nodes in each hidden layer would be better to use may be determined by experimental simulations given a fixed number of input features.)

The training of an MFNN-based classifier is essential to its runtime performance (done here by using the back-propagation algorithm [12]). For this, feature patterns that represent different images, coupled with their respective underlying image class (i.e. cell type) indices, are selected as the training data, with the input features being normalised into the range of 0 to 1.

In training an MFNN classifier, the feature extractor employed has the same functionality as its counterpart to be used in the resulting classifier. However, it generates more features at this stage (perhaps, many more), not knowing which features are more informative to use. The extracted features are passed through a subsystem that implements fuzzy-rough feature selection, removing redundant and less informative features. When applying such a trained classifier, only those features selected during the training phase are required to be extracted of course.

## IV. Experimental Results

### A. Experimental background

The image database used is the one summarised in Section II. Eighty-five images are used for training and the remaining 233 images are employed for testing.

During the training phase, for each image, five isotropic features are created, each having one of the following resolutions: 9 ($= log_2 512$), 8, 7, 6 and 5. That is, these isotropic features are created on the top five finest resolutions. To measure the directional fractals, the following four directions are used: horizontal (0°), first diagonal (45°), vertical (90°) and second diagonal (135°). In addition, in an attempt to capture basic statistical information, the mean and standard deviation (STD) that are readily available are also utilised. In so doing, a given image is represented by patterns of 11 features. For easy cross-referencing, Table I lists all the features and their reference numbers.

Different MFNN classifiers were built to accomplish classification by mapping feature patterns of a different dimensionality onto their underlying cell types, with explicit indications of whether they are normal or abnormal. There are a total of six output classes for the present problem case, representing adventitial, smooth mussel and endothelial cell types of normal tissues, and the same three types of abnormal ones. To limit the simulation cost, only networks with one hidden layer were considered. The number of hidden nodes were determined by systematically varying it during training. The structure of the best trained network, which has resulted in the least classification error over the training dataset with respect to a predefined number of iterations, was then chosen for use in testing.

| Ref. No. | Feature Meaning | Ref. No. | Feature Meaning |
|---|---|---|---|
| 1 | 0° direction | 7 | 3rd finest resolution |
| 2 | 45° direction | 8 | 4th finest resolution |
| 3 | 90° direction | 9 | 5th finest resolution |
| 4 | 135° direction | 10 | Mean |
| 5 | Finest resolution | 11 | STD |
| 6 | 2nd finest resolution | | |

TABLE I
Features and their reference numbers.

### B. Comparison with the use of unreduced features

It is important to show that, at least, the use of features selected does not significantly reduce the classification accuracy as compared to the use of the full set of original features. For this problem, the fuzzy-rough feature selection algorithm returns five features, namely, 0° DF, 95° DF, 5th finest resolution, mean and STD (i.e. features 1, 3, 9, 10 and 11), out of the original eleven. Table II lists the classification error rates produced by the best trained MFNNs.

| MFNN | Dim. | Features | Structure | Error |
|---|---|---|---|---|
| Reduced | 5 | 1,3,9,10,11 | $5\times10 + 10\times6$ | 7.55% |
| Original | 11 | 1,2,3,4,5,6,7,8,9,10,11 | $11\times24 + 24\times6$ | 9.44% |

TABLE II
Fuzzy-rough-selected vs. original full set of features.

It is very interesting to note that the error rate of using the five selected features is actually lower than that of using the full feature set. Further, this improvement of performance is obtained by a structurally much simpler network of 10 hidden nodes, as opposed to the classifier that requires 24 hidden nodes to achieve the optimal learning. This is indicative of the power of fuzzy-rough feature selection in helping reduce not only redundant feature measures but also the noise associated with such measurement, reflecting the usefulness of the present work.

### C. Comparison with the use of randomly selected features

The above comparison ensured that no information loss is incurred due to fuzzy-rough feature reduction. Actually, the selection process helps to remove measurement noise as a positive by-product. The question now is whether any other feature sets of a dimensionality 5 would perform similarly as those identified via fuzzy-rough selection. To avoid a biased answer to this, without resorting to exhaustive computation,

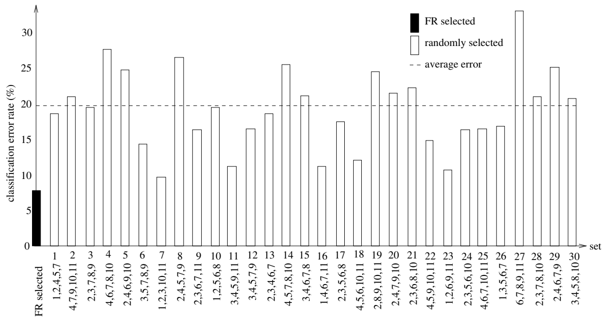


Fig. 4. Fuzzy-rough vs. randomly selected features.

30 sets of five features randomly chosen were used to see what classification results might be achieved.

Figure 4 shows the error rates of the corresponding 30 classifiers, along with the error rate of the classifier that uses fuzzy-rough (FR) selected features. The average error of the classifiers that each employ five randomly selected features is 19.1%, far higher than that attained by the classifier which utilises the FR-selected features of the same dimensionality. This implies that those randomly selected entail important information loss in the course of feature reduction; this is not the case for the fuzzy-rough selection-based approach.

### D. Comparison with the use of PCA-selected features

This study aimed at examining the performance of using different dimensionality reduction techniques. In particular, classifiers that are aided with fuzzy-rough feature selection are systematically compared to those supported by the use of PCA. The results are summarised in Table III. In this table, for the results of using PCA, feature number $i, i \in \{1, 2, ..., 11\}$, stands for the $i$th principal component, i.e. the transformed feature that is corresponding to the $i$th largest variance.

| MFNN | Dim. | Features | Structure | Error |
|---|---|---|---|---|
| FR | **5** | **1,3,9,10,11** | **$5\times10 + 10\times6$** | **7.7%** |
| PCA | 1 | 1 | $1\times12 + 12\times6$ | 57.1% |
| | 2 | 1,2 | $2\times12 + 12\times6$ | 32.2% |
| | 3 | 1,2,3 | $3\times12 + 12\times6$ | 31.3% |
| | 4 | 1,2,3,4 | $4\times24 + 24\times6$ | 28.8% |
| | **5** | **1,2,3,4,5** | **$5\times20 + 20\times6$** | **18.9%** |
| | 6 | 1,2,3,4,5,6 | $6\times18 + 18\times6$ | 15.4% |
| | 7 | 1,2,3,4,5,6,7 | $7\times24 + 24\times6$ | 11.6% |
| | 8 | 1,2,3,4,5,6,7,8 | $8\times24 + 24\times6$ | 13.7% |
| | 9 | 1,2,3,4,5,6,7,8,9 | $9\times12 + 12\times6$ | 9.9% |
| | 10 | 1,2,3,4,5,6,7,8,9,10 | $10\times20 + 20\times6$ | 7.3% |
| | 11 | 1,2,3,4,5,6,7,8,9,10,11 | $11\times8 + 8\times6$ | 7.3% |

TABLE III
FUZZY-ROUGH VS. PCA-RETURNED FEATURES.

These results show that, of the same dimensionality (i.e., 5), the classifier using the features selected by the fuzzy-rough mechanism has a substantially higher classification accuracy and moreover, this is achieved via a considerably simpler network. Further, it is worth recalling that PCA alters the underlying semantics of the features during its transformation process. That is, those features marked with 1, 2, ..., 11 in Table III are not the original 11 features, but their linear combinations.

If more principal features are employed, the error rate may generally be reduced. However, as compared to the classifier that uses FR-selected features, an MFNN using PCA-selected features still generally underperforms, until almost the full set of principal features is used. Yet, the overall structural complexity of all such classifiers are more complex than that of the fuzzy-rough based classifier. The best of them involves $11 \times 8 + 8 \times 6 = 136$ weights as compared to $5 \times 10 + 10 \times 6 = 110$. Additionally, the use of those classifiers based on PCA-returned features would require many more feature measurements to achieve comparable classification results.

### E. Comparison with the use of crisp rough-selected features

It is interesting to note that the results of applying fuzzy-rough feature selection to aid the MFNN-based classification appear to be very similar to those of using crisp rough set-based selection [2]. In fact, there happened to be only five features being chosen when crisp rough set-based method was used at its best [14]. In particular, four of the five features were the same as those chosen by fuzzy-rough selection, namely features 1, 9, 10, 11, with the only other different one being feature number 4 (instead of the present feature number 3).

However, the crisp approach requires an additional, and rather subjectively defined, quantity discretisation mechanism to convert real-valued image features into discrete nominal values prior to feature selection. Different discretisation schemes may lead to a rather different choice of feature subsets, often one with a higher dimensionality (rather than 5). As opposed to this, fuzzy-rough feature selection is directly applied to the real-valued features, with fuzzy equivalence classes being automatically computed from the feature values. In addition, the result that the same number of features was obtained using the crisp rough set-based approach might have also been affected by the characteristics of the cell-type classification problem itself because the dimensionality of the original feature patterns is not very large.

For a more scaled-up application, with the increase of the dimensionality of the original feature patterns and the use of different feature extraction mechanisms, subjective discretisation may become much harder to optimise. This will then lead to the loss of important information, thereby affecting the selection of the smallest subset of quality features and hence the subsequent complexity of the MFNN structure and their classification accuracy. A more meaningful comparison between these two approaches however, remains as active research.

## V. CONCLUSIONS

This paper has presented an approach which supports the potentially powerful neural network classification systems with a fuzzy-rough set-based feature reduction method. Unlike transformation-based dimensionality reduction techniques, this approach retains the underlying semantics of the selected feature subset. This is very important to help ensure that the classification results are understandable by the user. Following this approach, the conventional multi-layer feedforward networks, which are sensitive to the dimensionality of feature patterns, can be expected to become effective on classification of images whose pattern representation may otherwise involve a large number of features.

The work has been applied to the real problem of normal and abnormal blood vessel classification involving different cell types. Although the application problems encountered are complex, the resulting selected features are manageable and the classifier built upon such features generally outperforms those using more features or an equal number of features obtained by conventional approaches represented by PCA. Experimental results have clearly demonstrated this.

Note that comparisons between the use of fuzzy-rough selected features against that of those obtained by PCA form a focus of this paper. This is mainly due to the observation that PCA is a representative approach commonly taken to perform dimensionality reduction. However, there exist many alternative methods for dimensionality reduction (e.g., [16], [17]) which may also outperform PCA for the present application. Further comparisons to such alternative techniques will therefore help reveal more details about the strengths and limitations of the present approach. Work is ongoing in this direction.

Finally, it is worth indicate that, although the present research on fuzzy-rough feature selection is incorporated with neural network-based classifiers, it can be extended to work with other types of intelligent classification system such as classical decision trees [9] and fuzzy classifiers [5], [8]. This forms a piece of very interesting further research.

## ACKNOWLEDGEMENT

The authors are grateful to Dr Richard Jensen for his contribution, while taking full responsibility for the views expressed in this paper.

## REFERENCES

[1] S. Chen, J. Keller and J. Crownover. On the calculation of fractal features from images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15 (1993) 1087–1090.

[2] A. Chouchoulas and Q. Shen. Rough set-aided keyword reduction for text categorisation. *Applied Artificial Intelligence*, 15 (9) (2001) 843–873.

[3] P. Devijver and J. Kittler. *Pattern Recognition: a Statistical Approach*. Prentice Hall, 1982.

[4] D. Dubois and H. Prade. Putting rough sets and fuzzy sets together. In R. Slowinski (Ed.). *Intelligent Decision Support*. Kluwer Academic Publishers, (1992) 203–232.

[5] C. Janikow. Fuzzy decision trees: Issues and methods. *IEEE Transactions on Systems, Man, and Cybernetics – Part B: Cybernetics*, 28 (1) (1998) 1–14.

[6] R. Jensen and Q. Shen. New approaches to fuzzy-rough feature selection. To appear in: IEEE Transactions on Fuzzy Systems.

[7] L. Kaplan. Extended fractal analysis for texture classification and segmentation. *IEEE Transactions on Image Processing*, 8 (1999) 1572–1585.

[8] J. Marin-Blazquez and Q. Shen. From approximative to descriptive fuzzy classifiers. *IEEE Transactions on Fuzzy Systems*, 10 (4) (2002) 484–497.

[9] T. Mitchell. *Machine Learning*. McGraw-Hill, 1997.

[10] S.K. Pal and A. Skowron (Eds.). *Rough-Fuzzy Hybridization: A New Trend in Decision Making*. Singapore: Springer Verlag. 1999.

[11] Z. Pawlak. *Rough Sets: Theoretical Aspects of Reasoning About Data*. Kluwer Academic Publishers, Dordrecht, 1991.

[12] D. Rumelhart, E. Hinton and R. Williams. Learning internal representations by error propagating. In: D. Rumelhart and J. McClelland (Eds.), *Parallel Distributed Processing*. MIT Press, 1986.

[13] C. Shang, J. McGrath, C. Daly and J. Barker. Modelling and classification of vascular smooth muscle cell images. *IEE Electronics Letters*, 36(18) (2000) 1532–1533.

[14] C. Shang and Q. Shen. Rough feature selection for neural network based image classification. *International Journal of Image and Graphics*, 2 (4) (2002) 541–555.

[15] Q. Shen and R. Jensen. Selecting Informative Features with Fuzzy-Rough Sets and its Application for Complex Systems Monitoring. *Pattern Recognition*, 37 (7) (2004) 1351–1363.

[16] J. Tenenbaum, V. de Silva and J. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290 (5500) (2000) 2319–2323.

[17] F. Young and R. Hamer. *Theory and Applications of Multidimensional Scaling*. Eribaum Associates, Hillsdale, 1994.