A kernelised fuzzy-Support Vector Machine CAD system for the diagnosis of lung cancer from tissue images

Walker H. Land, Jr.*

Department of Bioengineering, Binghamton University,
Binghamton, NY, 13903-6000, USA
E-mail: wland@binghamton.edu
*Corresponding author

Daniel W. McKee

Department of Mathematics
and Computer and Information Sciences,
Mansfield University, Mansfield, PA, 16933, USA
E-mail: dmckee@mansfield.edu

Tatyana Zhukov, Dansheng Song
and Wei Qian

Department of Interdisciplinary Oncology,
Colleges of Medicine and H. Lee Moffitt Cancer Center
and Research Institute,
University of South Florida, Tampa, FL, 33612, USA
E-mail: Zhukov@moffitt.usf.edu
E-mail: dsong2@mail.usf.edu
E-mail: wqian@health.usf.edu

Abstract: This research describes a non-interactive process that applies several forms of computational intelligence to classifying biopsy lung tissue samples. Three types of lung cancer evaluated (squamous cell carcinoma, adenocarcinoma, and bronchioalveolar carcinoma) together account for 65–70% of diagnoses. Accuracy achieved supports hypothesis that an accurate predictive model is generated from training images, and performance achieved is an accurate baseline for the process's potential scaling to larger datasets. Feature vector performance is good or better than Thiran and Macq's in every case. Except bronchioalveolar carcinomas, each individual cancer classification task experienced improvement, with two groupings showing nearly 20% classification accuracy.

Keywords: Computer-Aided Diagnosis; CAD; lung cancer; segmentation; feature selection; classification; microscopy images; kernel methods; Support Vector Machine; SVM.

Reference to this paper should be made as follows: Land Jr., W.H., McKee, D.W., Zhukov, T., Song, D. and Qian, W. (2008) 'A kernelised fuzzy-Support Vector Machine CAD system for the diagnosis of lung cancer from tissue images', *Int. J. Functional Informatics and Personalised Medicine*, Vol. 1, No. 1, pp.26–52.

Biographical notes: Walker H. Land, Jr. is currently a Research Professor in the Department of Bioengineering as well a Principal Investigator and Director of a Computational Intelligence group there. He has over 30 years of industrial research experience and over 25 years of academic research and teaching experience. He is the author/co-author of over 200 peer reviewed research as well as several other publications.

Daniel W. McKee holds a PhD and Master of Science in Computer Science, both from Binghamton University. He is an Assistant Professor at Mansfield University, and has been performing cancer-related computational intelligence research since 2000.

Tatyana Zhukov is an Assistant Professor, Cancer Prevention and Control Division at the H. Lee Moffitt Cancer Center and Research Institute. He was trained in clinical biochemistry and clinical cytology and has worked for the last ten years to apply these skills in the development of molecular markers for human cancer. Her professional insight into cell image cytometry, proteomics, and biomarker development has led to substantial advances in methods used and proposed for lung and breast cancer screening largely through her participation in the Early Detection Research Network (EDRN).

Dansheng Song is a Research Associate at the Department of Interdisciplinary Oncology, College of Medicine, University of South Florida. He has extensive experience over the last ten years in the applications of advanced Computer-Aided Diagnosis (CAD) methods on medical imaging, biomedical imaging and molecular imaging to clinical trials. He has published more than 20 paper in these field.

Wei Qian is an Associate Professor, Director of the Biomedical Imaging Program at the Department of Interdisciplinary Oncology and Radiology, College of Medicine, University of South Florida. He has extensive experience over the last ten years in the applications of advanced Computer-Aided Diagnosis (CAD) methods on medical imaging, biomedical imaging and molecular imaging to clinical trials. He has published more than 100 papers in these field.

1 Introduction

The American Cancer Society (2007a) estimates that in the USA in 2007, over 213,000 people will be newly diagnosed with lung cancer and over 160,000 people will die from lung cancer. This makes lung cancer the second most prevalent type of cancer to be diagnosed in both men and women, and the leading cancer-related cause of death in both sexes. Lung cancers will account for approximately 15% of new cancer diagnoses in 2007 (American Cancer Society, 2007b). Lung cancers can be divided into two main groups: small cell lung cancer, which accounts for approximately 10–15% of lung cancers, and non-small cell lung cancer, which accounts for the other 85–90%.

The lung cancers evaluated in this study primarily fall into the second category. Squamous cell carcinomas represent 25–30% of new lung cancer diagnoses, while adenocarcinomas (including bronchioalveolar carcinoma) represent about 40% (American Cancer Society, 2007c).

This research develops a non-interactive process that applies several forms of computational intelligence to the task of classifying biopsy lung tissue samples based on visual data in the form of raw digital photographs of those samples. The three types of lung cancer evaluated (squamous cell carcinoma, adenocarcinoma, and bronchioalveolar carcinoma) together account for 65–70% of lung cancer diagnoses.

Computing technology has often proven useful in performing tedious or complex tasks quickly, accurately, and consistently. In extensive previous work by this author and others, computational intelligence was used to identify cancerous breast lesions based on radiologist's impressions of various features visible on a mammogram. This previous work includes, but is not limited to: Fogel et al. (1995, 1997, 1998a, 1998b), Land et al. (2000, 2001a, 2001b, 2001c, 2001d, 2002a, 2002b, 2003a, 2003b, 2003c, 2004a, 2004b, 2004c, 2004d), Lo et al. (1997, 1999) and McKee (2001). In other areas of medicine, computational intelligence has been used to separate (or segment) an image (such as an MRI display) into its constituent parts, allowing automated estimations of tissue volumes (Pham, 2003).

While related research forms a basis for this work in many of the areas being applied, providing a self-tuning end-to-end process is unique to our work. In addition, the segmentation procedure developed for this research is based on the well known and widely used Fuzzy C-Means (FCM) algorithm, but we introduce a new kernel-based extension to this algorithm to improve its accuracy and enable the segmentation to self-adapt to the image at hand.

The end-to-end process developed can be broken down into three consecutive steps, each with its own unique challenges and dependent on the quality of the processing in the previous step(s). The first step is segmentation, or identifying physical features such as nuclei, cytoplasm, and background within the source image. The second step is the extraction and measurement of features, based on the segmented image. Finally, the third step is to combine the measurements to classify the cells in the image as cancerous or non-cancerous.

The applications of computational intelligence used in this study are all directed toward classification problems. The goal of a classification problem is to accurately identify to which set (or class) an unknown item belongs. In terms of the image segmentation problem, the classes are known – nucleus, cytoplasm, and background, and we need to determine the correct classification for each pixel in the image. Deciding whether an image corresponds to normal or cancerous tissue is also a classification problem – the classes are 'normal' and 'cancer' (or perhaps a specific type of cancer), and based on the metrics collected from the segmented image, we must decide to which of these classes the image belongs.

The classification methods used in this study can be divided into two main categories: rule-based systems and those based in Statistical Learning Theory (SLT). Rule-based systems use a series of rules to classify an unknown datum. These rules are usually developed by an 'expert' in the classification domain.

An opposing approach for classification stems from SLT, which uses mathematical methods to attempt to reduce data with known classifications to a model. This model can then be used to determine the statistical probability that an unknown datum belongs

to a particular given class, and those probabilities can in turn be used to provide a classification and often a measure of confidence in that classification.

2 Problem foundation

This section contains a summary of several key knowledge domain areas that serve as a foundation for this work.

2.1 Medical background

This summary discussion is provided because a basic understanding of the structure and form of each of these cancers is a prerequisite to developing intelligent software to perform accurate classifications, and is sometimes absent from papers describing research of this type.

2.1.1 Cancer

Cancer in general is a type of *neoplasm*, or 'new growth'. The word *tumour* is often used to refer to neoplasms of all sorts, benign and malignant, while *cancer* is a general term for malignant neoplasms. All neoplasms (including the lung cancers represented in this study) have certain characteristics in common. They consist of an independent, abnormal tissue growth, uncoordinated with the surrounding tissues. The *parenchyma* contains the growing outer edge of the neoplasm, and the *stroma* contains connective tissues and blood vessels needed to support the development of the neoplasm. Cancers may be broadly classified into two groups: those which develop in epithelial tissue are carcinomas (e.g., adenocarcinoma), and those which develop in non-epithelial tissue are sarcomas (e.g., fibrosarcoma). Sarcomas generally have a minimal stroma, while carcinomas can have a significant network of blood vessels and other supportive tissue (Cotran, 1999, pp.260–262).

Cotran (1999, pp.264–265) describes four categories that are useful in classifying a tumour as benign or malignant: differentiation, rate of growth, local invasion, and metastasis. Differentiation describes how well the tumour cells replicate the normal cells from their tissue of origin. Well-differentiated tumours bear a strong resemblance to normal tissues, both in the appearance and the functioning of the cells. Poorly differentiated tumours have 'primitive-appearing, unspecialised cells'. Benign tumours are usually well-differentiated, while malignant tumours span the entire range from well-differentiated to undifferentiated.

The remaining categories are useful from a clinical perspective, but are beyond the scope of cytological examination. The rate of growth describes the change in the neoplasm with time. Although a number of factors may influence the rate of growth, benign tumours generally grow slowly over several years, whereas cancers may grow rapidly or erratically. Local invasion describes whether a neoplasm grows as a separate cohesive mass or whether it infiltrates surrounding tissues. Benign tumours may displace surrounding tissues, but generally respect the tissue boundaries and will not 'grow into' neighbouring structures. Cancers, however, will infiltrate, invade, and eventually destroy surrounding tissue. Finally, metastasis (a tumour implant not connected to the original

tumour) is always an indicator of cancer, since benign neoplasms do not metastasise, while almost all cancers can metastasise (Cotran, 1999, pp.267, 268).

2.1.2 Lung cancer

Most lung cancers develop in the bronchial epithelium, and are therefore carcinomas (Koss, 1979, p.608). The cells of the parenchyma allow us to classify the tumour (Cotran, 1999, p.261) and we will focus the remainder of this discussion on that layer.

Tumours are named for the type of cells in the parenchyma. Squamous cell carcinoma consists of cells that mimic the morphology and function of stratified squamous epithelia. Adenocarcinomas mimic glands or ducts in the epithelial lining. Bronchogenic carcinomas (a subset of adenocarcinomas) mimic respiratory passages, and more specifically bronchioalveolar carcinoma imitates the terminal bronchioles and alveoli (Cotran, 1999, p.263; Koss, 1979, p.608; Takahashi, 1971, p.179). Several other types of lung cancer also exist, but are outside the scope of this study. For a more thorough examination of additional types of lung cancer and their properties, see Koss (1979, p.608), Graham (1972, pp.259–284) and Cotran (1999, pp.697–753).

2.1.3 Cytology of lung tumours

This section discusses some of the specific features visible at the cellular level, which may be useful for determining whether a tumour is cancerous. An overview of general cytological features is presented, followed by a specific discussion of each type of cancer represented in this study.

2.1.3.1 Cytological features common to multiple types of cancer

Poor differentiation is also called *anaplasia*, and is accompanied by a number of visible changes in the cells. The cells and the nuclei often have a wide variation in size and shape (pleomorphism). Anaplastic nuclei are extremely dark staining (hyperchromatic) due to large quantities of DNA. The nuclei are larger than in normal cells, with a nucleocytoplasmic ratio approaching 1 : 1 (normal cells have a ratio around 1 : 4 to 1 : 6). There can be wide variations in nucleus shape, with coarse chromatin granules distributed along the nuclear membrane, and large nucleoli.

2.1.3.2 Cytological presentation of squamous cell carcinoma

Squamous cell carcinoma is characterised by sheets of cells attempting to form squamous epithelium. These sheets often contain keratin ‘pearls’ – small round nests of keratin-producing cells. Necrosis is common in the centre of this type of tumour (Koss, 1979, p.610).

Squamous carcinoma cells can vary widely in size and shape, with giant cells adjacent to small cells. Spindly cancer cells are common as well. Abnormal nuclear shape is common with this type of cancer, and the nuclei are often hyperchromatic, staining evenly to a deep colour resembling droplets of India ink. The increased receptivity to staining is caused by pyknosis, a degenerative change to the nucleus. Some nuclei may also appear relatively pale, particularly within highly keratinised cells. While the nucleus is generally large for the size of the cell, there are also very small pyknotic nuclei,

making the nucleocytoplasmic ratio of limited value in diagnosing this type of cancer (Koss, 1979, pp.610–613).

When cells are forcibly removed from the tumour (which is the case in all of the samples used in this study), the cancer cells appear in sheets or clusters. The nuclei are granular, often with only slight to moderate hyperchromasia, often with large nucleoli (Koss, 1979, p.616). Irregular distribution of chromatin within the nucleus is one of the most significant factors in identifying individual malignant cells, and undifferentiated cells often have no cytoplasm and lack cell borders (Graham, 1972, p.267).

2.1.3.3 Cytological presentation of adenocarcinoma and bronchioalveolar carcinoma

Adenocarcinomas of central bronchial origin are tumours whose precise tissue of origin has not been clearly determined. They may be derived from the epithelium of the secondary bronchi, bronchioles, or related mucus glands. These tumours typically contain large, irregular nuclei which appear in groups (Graham, 1972, p.269), and have a variable configuration, which may include papillary structures, lined with cells that are polygonal or columnar (Koss, 1979, p.629). Otherwise, the nucleus appears similar to any other malignant cell.

Bronchoalveolar carcinoma is a particular type of adenocarcinoma (Takahashi, 1971, p.179) that originates in the epithelium of the terminal bronchioles and alveoli. These tumours are usually peripherally located; but the cytological presentation is similar to adenocarcinoma (Koss, 1979, pp.625–629).

2.1.3.4 Cytological impact of tissue specimen collection techniques

In order to view and evaluate the cells from a tumour, a tissue biopsy is needed. The biopsy process involves collecting a tissue sample, which must be transferred quickly to one or more microscope slides and fixed and stained. The prepared slide is then viewed under a microscope by an expert who uses a number of factors to decide whether the cells are from normal or cancerous tissue, and if cancerous, to determine the type of cancer.

There are several methods used to collect a tissue sample for biopsy, and the method used impacts the decision rules employed by the expert to develop a classification. For example, a sputum sample may be collected from the lungs through a deep cough, possibly triggered by inhaling an aerosol. Similar ‘loose’ material may also be collected by inserting a catheter into a bronchus in the area of a suspicious tumour and the bronchial wall scraped with a fine nylon brush to collect cells (Takahashi, 1971, pp.166–171). Both of these mechanisms normally contain a number of ‘dead’ cells which have been shed naturally by the lining of the lungs and are in varying states of necrosis (Koss, 1979, p.609). Squamous cells from the lining of the mouth and pharynx may also appear in sputum samples and must be differentiated from squamous cancer cells (Koss, 1979, p.551).

As radiographic techniques have improved our ability to precisely locate tumours, fine needle aspiration has become more widely used. This collection method uses a fine needle with jagged edges toward the tip to collect tissue directly from the tumour in question. Though somewhat more invasive than sputum collection or bronchial scrapings, this technique has several advantages – including the ability to collect cells from

peripherally located tumours which may not produce sputum. Unlike the sputum samples mentioned above, the presence of necrotic cells in a sample obtained in this manner would be highly unusual and indicate a (possibly cancerous) pathology (Koss, 1979, pp.607–609).

Finally, tissue samples may be collected through an open incision, where the tumour or a portion of it is surgically removed. For the purposes of this study, it is sufficient to say that the cytological features of such tissue samples are similar to those obtained through fine needle aspiration.

There are a number of non-cancerous external factors which can cause substantial changes in lung tissue as well, such as smoking or exposure to asbestos or other contaminants. An accurate diagnosis can only be made by considering the cytological evidence in combination with the patient's history and other factors (Koss, 1979, pp.551–560).

2.2 Fuzzy C-Means (FCM)

The well-known FCM clustering algorithm was chosen as a starting point for our segmentation process. This algorithm uses an iterative process to divide the pixels in the source image into an arbitrary number of classes (clusters) by solving a constrained optimisation problem. We want to minimise

$$J(\mathbf{U},\mathbf{v}) = \sum_{k=1}^{n}\sum_{i=1}^{c}(u_{ki})^2(d_{ki})^2 \tag{1}$$

subject to

$$u_{ki} \in [0,1] \quad \forall k,i, \tag{2}$$

$$\sum_{i=1}^{c} u_{ki} = 1 \quad \forall k, \tag{3}$$

and

$$0 < \sum_{k=1}^{n} u_{ki} < n \quad \forall i. . \tag{4}$$

In equation (1), u_{ki} is the degree of membership of pixel k in cluster i, d_{ki} is the 'distance' (typically the Euclidean norm) between pixel k and the centroid for cluster i, n is the number of pixels in the image, c is the number of clusters (classes), $\mathbf{U}$ is an $n \times c$ matrix of the membership values for each pixel and cluster, and $\mathbf{v}$ is the collection of centroids, one for each cluster. The constraints listed ensure that each pixel's membership in each cluster is bound by [0, 1] (equation (2)), and that the membership values across all clusters sum to one for each pixel (equation (3)). Finally, equation (4) ensures that each class has some pixels with a non-zero degree of membership, and that no single cluster can contain all pixels. This constraint forces the solution to have the desired number of distinct clusters.

The optimisation problem described above can be solved by Bezdek's (1981) FCM clustering algorithm. The initial membership matrix in Step 1 can be assigned at random; however, the algorithm will converge much more rapidly if we make an intelligent guess at the classification for each pixel. Once the initial membership matrix has been

established, it can be used in Step 2 to provide a new estimation of the position of the centre of each cluster. Step 3 uses these new centres to refine the membership matrix. These two steps (using the membership matrix to refine the cluster centres, and using the centres to refine the membership matrix) are alternated until the changes to the membership matrix become insignificant.

A key element affecting the accuracy of FCM's classification is the distance measurement, d_{ki}. Bezdek defines this distance as the Euclidean norm between a sample $\mathbf{x}_k$, and a cluster center, $\mathbf{v}_i$, so that

$$d_{ki} = \| \mathbf{x}_k - \mathbf{v}_i \| . \tag{5}$$

However, the Euclidean norm may not be the most accurate representation of the 'distance' between these points. By substituting a kernel function for the distance calculation in equation (1) we obtain

$$J(\mathbf{U}, \mathbf{v}) = \sum_{k=1}^{n} \sum_{i=1}^{c} (u_{ki})^2 (\mathrm{K}(k,i))^2 . \tag{6}$$

Now the clustering can be based on the kernel's calculation of the similarity between the sample and the cluster centre. It is important to note that the distance we are referring to here is the difference in the Red, Green, Blue (RGB) colour space between the sample pixel and the cluster centre – it has nothing to do with the physical location of the sample pixel within the image. Likewise, the cluster centre is the prototypical set of RGB values for a pixel belonging to that cluster – physical location within the image is not considered. In fact, FCM treats the image as a set of pixel values, without respect for the coordinate system used to assemble those pixels into a two-dimensional image.

2.3 Kernel based learning methods

Kernel-based learning methods comprise a subset of SLT. This learning theory is based on the premise that we can learn about an unknown datum by comparison with known data. Kernel-based methods are those which use a kernel as a non-linear similarity measure to perform this comparison.

2.3.1 Support Vector Machines (SVMs)

The Support Vector Machine (SVM) is one of the most well-known kernel-based classifiers. While several texts (Cristianini and Shawe-Taylor, 2000; Haykin, 1999; Gunn, 1998; Vapnik, 1995; Burges, 1998) provide extensive development of the mathematical foundation of SVM s, this section will present an overview of the nature and operation of these machines. In the context of cancer identification, the main purpose of an SVM is to construct an 'optimal hyperplane' as the decision surface such that the margin of separation between positive (cancer) and negative (normal tissue) cases is maximised. SVMs are based on four fundamental ideas:

- Structural and Empirical Risk Minimisation (SRM/ERM)
- the Vapnik-Chervonenkis (VC) dimension

- the constrained optimisation problem
- the SVM decision rule.

In the explanation of these concepts, vectors and matrices will be labelled with bold typeface, while vector and matrix *elements* will be labelled with normal typeface. The interested reader can find a discussion of Structural and Empirical Risk Minimisation and the VC dimension in the above references on SVMs. These concepts relate to the theoretical foundation of SVMs, an understanding of which is not essential background for the remainder of this study. As such, that discussion will not be duplicated here. However, since much of the analysis and results in this study assume a basic familiarity with the constrained optimisation problem and the SVM decision rule, these concepts will be reviewed briefly.

2.3.1.1 The constrained optimisation problem

Obtaining a solution in the most general case where the environment is non-linear and non-separable requires the use of inner-product kernel functions. The inner-product kernel function provides a mapping from the input space to a higher-dimensional feature space. This kernel mapping is used to construct a decision surface that is non-linear in the input space, but has a linear image in the feature space. The inner product kernel function must be symmetric and must satisfy Mercer's Theorem (additional discussion of Mercer's Theorem can be found in Cristianini and Shawe-Taylor (2000). The solution to the Lagrangian dual problem for this most general case is given by:

$$\max \mathbf{W}(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \, \mathrm{K}(\mathbf{x}_i, \mathbf{x}_j) \tag{7}$$

subject to the constraints:

$$\sum_{i=1}^{N} \alpha_i y_i = 0 \tag{8}$$

and

$$0 \le \alpha_i \le C. \tag{9}$$

Equation (7) is also referred to as the objective function, equation (8) as the linear constraint. Equation (9) provides an upper bound, C, on the Lagrange multipliers. This bound is called the regularisation parameter and limits the effect of outliers in the training data.

2.3.1.2 The SVM decision rule and soft limiting

Together, equations (7)–(9) describe the Quadratic Programming (QP) problem. QP problems are well-founded in SLT. Simply stated, SLT proves that bounds on the expected (or true) error for 'future' points (such as those not found in the training set) may be obtained. It may be shown that these bounds are a function of the classification error on the training data, expressed in terms that measure the complexity (or capacity) of the classification function (Vapnik, 1995). For example, maximising the separating margin for linear functions reduces the function complexity or capacity.

Consequently, by this explicit margin maximisation, one accomplishes the minimisation of bounds on the generalisation error, which means the learning machine can expect better generalisation with high probability.

Secondly, these QP problems may be solved by several methods such as gradient ascent methods, conjugate direction methods, interior point methods, or Platt's (1998) Sequential Minimal Optimisation (SMO) method. This study used SMO to solve the QP problem, because of its ability to find the solution much more quickly than the other methods.

Finally, when the optimal solution to the QP problem has been found, a new point is classified by using the SVM decision rule and the hyperbolic tangent soft limiter:

$$\mathrm{F}(\mathbf{x}) = \tanh\left(\sum_{i=1}^{N} y_i \alpha_i \, \mathrm{K}(\mathbf{x}_i, \mathbf{x}) + b\right). \tag{10}$$

2.3.2 Kernel functions

We now turn our attention to kernel functions, represented as K(**x**, **y**) in equations (7) and (10). The purpose of a kernel function in both training (equation (7)) and classification (equation (10)) using an SVM is to represent the similarity (in a higher-dimensioned feature space) of the two input vectors; however, the idea of kernel functions is not limited to SVMs, and the general principles discussed here apply to other kernel-based classifiers as well, including our kernel-based extension to FCM, discussed in the previous section.

A kernel function should yield a higher output from input vectors which are very similar than from input vectors which are less similar. An ideal kernel would provide an exact mapping from the input space to a feature space which was a precise, separable model of the two input classes; however, such a model is usually unobtainable, particularly for complex, real-world problems, and those problems in which the input vector provided contains only a subset of the information content needed to make the classes completely separable. As such, a number of statistically-based kernel functions have been developed, each providing a mapping into a generic feature space that provides a reasonable approximation to the true feature space for a wide variety of problem domains. The kernel function that best represents the true similarity between the input vectors will yield the best results, and kernel functions that poorly discriminate between similar and dissimilar input vectors will yield poor results. As such, intelligent kernel selection requires at least a basic understanding of the source data and the ways different kernels will interpret that data. Some of the more popular kernel functions are the (linear) dot product, the Gaussian Radial Basis Function (GRBF), the Exponential Radial Basis Function (ERBF), and the polynomial kernel. The dot product and polynomial kernels use a similar mechanism to establish similarity, while the two radial basis functions share a different common measure.

2.4 Related research

The cell image segmentation problem has been around for a long time, and has been studied by several other researchers. However, most existing research on cell image segmentation has focused on identifying cells in a noisy background (Garbay et al., 1986; Jiang and Yang, 2002; Nedzved et al., 2000) or solely on nucleus isolation and evaluation

(Sammouda et al., 2002; Thiran and Macq, 1996; Schnorrenberg et al., 1997). Sammouda et al. (2002) work with cell images similar to ours to detect cancer; however, that work focused only on nucleus segmentation and analysis, whereas we extract information from the entire cell image.

A good overview of current segmentation approaches can be found in Bengtsson et al. (2004), where techniques including thresholding, using edge and shape information, and seeded watershed transforms are discussed. Wählby et al. (2004) present a seeded watershed technique which takes into account a number of morphological features to improve the quality of the segmentation, achieving approximately 90% accuracy by comparison with manual counts from the same image fields. While the 2D images processed in that study were based on florescence microscopy, a number of the principles used could be applied to optical microscopy images as well.

Other possible approaches for improving the accuracy of the segmentation by using edge information or other neighbouring pixel information are discussed in Morrison and Attikouzel (1992) and Pham (2003). These are presented in the context of MRI image segmentation, however, and would require significant adaptation in order to work with cell images.

A very thorough survey of the current research into automated cancer diagnosis based on histopathological images can be found in Demir and Yener (2005). This survey includes a discussion of segmentation, feature extraction and selection, and the various classifier types that have been applied to this class of problems, as well as a summary of the evaluation methods used in the various studies that have been performed.

Schnorrenberg et al. (1997) outline a system similar to ours for the analysis of biopsy specimens, automating an assessment currently performed by a histopathologist; however, the specimens in that study were stained to identify specific markers for cancerous cells. The processing done by their system is limited to identifying the percentage of positive nuclei and the stain intensity, and using those to assign a diagnostic index. While they still had to address variations in the collection process, the cancer identification is much more narrowly defined than in our system, which uses biopsy samples stained to bring out general cell features rather than specific cancer markers and uses entire cell information rather than evaluating the nuclei only.

Sanei and Lee (2003) present an interesting technique for multi-class identification of blood cells by applying and extending face recognition work to the task of cell recognition. With this approach each cell is classified as a particular type (basophile, monocyte, lymphocyte, etc.). While this is a significant departure from our methods, if such an approach could be extended to identify particular types and stages of normal cells and cancer cells, it could in theory provide a similar functionality to that developed in this study.

One other study (Zhou et al., 2002) closely parallels our work, but has some significant differences. This study deals with identifying lung cancer from biopsy images very similar to ours (digitised 400X light microscope images of hematoxylin-eosin stained needle biopsy tissue samples); however, their classification is based on classifying individual cell images extracted from those large field images – our process evaluates an entire field of view containing hundreds of cells. The above study also performs multi-class classification on the images, whereas we have limited our classification analysis at this point to several binary classification scenarios using normal tissue and one or more types of cancer. The Zhou study achieves classification error rates

varying from 46% for single artificial neural networks, and 14–21% for various artificial neural network ensembles.

3 Lung cancer data

The source data used in this study consists of 162 surgical biopsy samples provided by the Moffitt Cancer Centre and Research Institute. The tissue in each image was stained using r-H2AX, PX-DAB, and hematoxylin counterstain. A light microscope field of view at 400X magnification was then imaged, at a resolution of 1520×1080 pixels. The images are separated into two independent groups, a set of 83 images used for training and process development, and 79 images used for validating the performance of the system.

In the training set, 37 of the images are of normal tissues, 21 are of adenocarcinomas, ten are of bronchioalveolar carcinomas, and 15 are of squamous cell carcinomas.

In the validation set, 44 of the images are of normal tissues, 13 are of adenocarcinomas, eight are of bronchioalveolar carcinomas, eight are of squamous cell carcinomas, and six are of undifferentiated (large cell) carcinomas.

The field of view in each image contains between 200 and 1,000 cells. The cancer images may have connecting tissue or other tissues visible as well as the cancer cells. As some tissues absorb the stains more readily, there are colour intensity differences from one image to the next and even within the same type of tissue in the same image. This darker cytoplasm presents a challenge for the segmentation process, as it is often much closer in intensity to the nuclei than to the remaining cytoplasm. Several images also have excess cytoplasm or other cell matter not associated with a complete, intact cell. This is an artifact of the collection process, and should be ignored (i.e., classified as background, even though it is identical in appearance to the remaining cytoplasm in the image) when evaluating the cells in the image. Figure 1 shows a portion of one of the images illustrating these challenges.

Figure 1 Cytoplasm not associated with any cell (a) and differences in cytoplasm stain intensity (b1, b2) (see online version for colours)

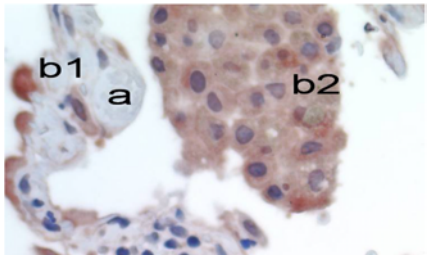
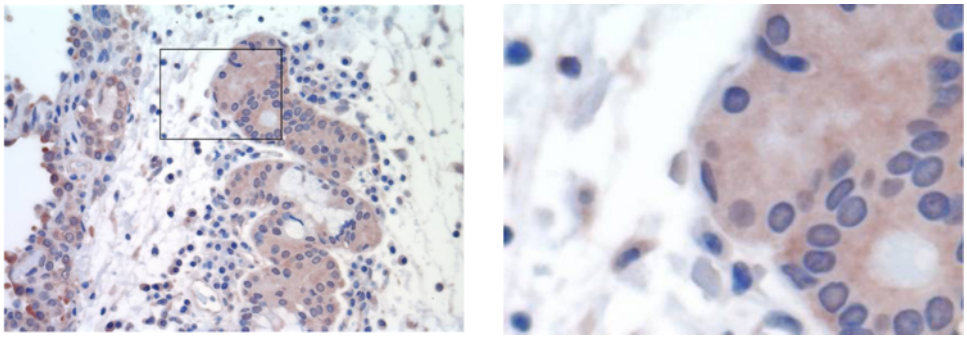


Figure 2(a) shows the field of view for a non-cancerous sample, and Figure 2(b) shows the full resolution detail of the marked area in that image. Figure 4 shows a similar pair for an adenocarcinoma image. Comparing Figures 2(b) and 3(b), it is clear that the appearance of a typical cell from each class (nucleus, cytoplasm, and background) varies substantially between images. In images such as Figure 3(b), there is very little contrast between the nuclei and the cytoplasm, presenting yet another challenge for the segmentation process.

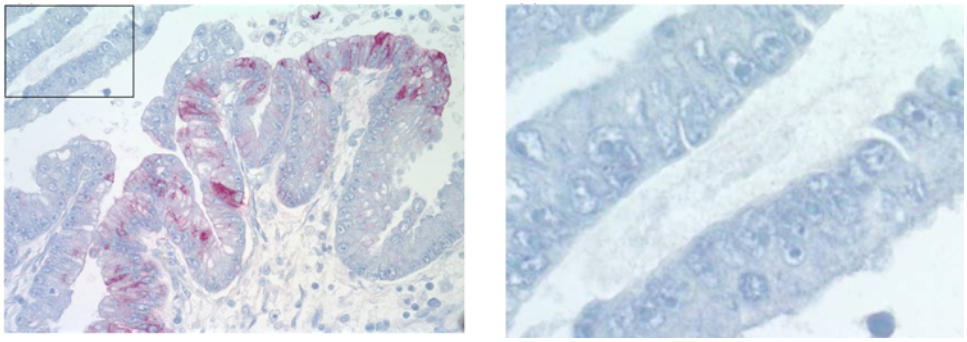
The images in the validation set are similar in appearance and composition to the training images, and present the same challenges.

Figure 2 (a) Microscope field of view for a normal tissue sample and (b) enlarged detail of marked area (see online version for colours)



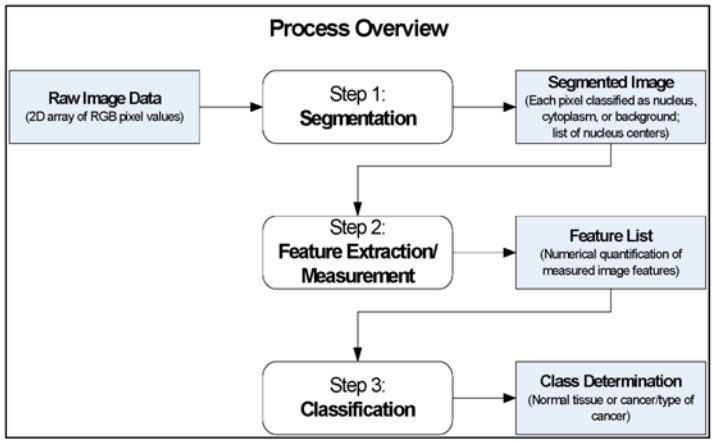
(a) (b)

Figure 3 (a) Microscope field of view for an adenocarcinoma tissue sample and (b) enlarged detail of marked area (see online version for colours)



(a) (b)

Figure 4 End-to-end process overview



4 Method

This section summarises the three sub-processes constituting the CAD of the lung cancer images.

4.1 Classification process overview

The end-to-end process developed in this study focuses on approximating the decisions made by an expert in cytology. A few simplifications have been made to reduce the variability of the inputs into this decision process. First, all samples being used were collected through surgical biopsy, so a more precise set of assumptions about normal tissue conditions can be applied. Second, all of the images used were taken at the same magnification level (400X) – this provides some level of consistency in the meaning of the cell sizes and the amount of tissue being analysed. This still leaves a number of significant variations to overcome. In addition to normal tissue, there are three different types of cancer represented in the training data, each with its own characteristics. While a common process was used as a basis for the stains applied to the cells, there are substantial variations in colour and contrast from one image to the next. There are also similar variations caused by lighting and exposure differences during the image capture process. In order to approximate an expert's decision based on an image, we must break down the reasoning process used by the expert and develop a parallel process. The expert uses visual clues such as the size of the cells, variability in the size of the cells, relative size of the nucleus within the cell, the appearance of the cell components (how darkly or evenly stained different parts of the cell are), and the organisation of the cells in the image. Clearly we must be able to make objective measurements to approximate the expert's subjective impressions of these qualities; however, in order to take measurements, we must first be able to identify the constituent parts of the image. Each image contains several hundred cells, possibly some clear background, and possibly extraneous cytoplasm not associated with any cell (this cytoplasm is an artefact of the collection and slide preparation processes). We need to accurately separate the nuclei, the cytoplasm, and the background portions of the image. This step of the process is segmentation.

Once we have a segmented image, we can begin to analyse the image to extract properties of the cells pictured. This is the second step in the process. By counting the number of pixels in each nucleus, we can estimate the area of that nucleus. Colour variations within the nucleus can give us clues as to its internal structure, and variations between nuclei can also be significant. Once the visual data in the image has been reduced to a series of numerical measurements, those measurements can be used in conjunction with a computational intelligence classifier (the third and final step) to characterise the image as representing normal or cancerous cells (see Figure 4).

4.2 Segmentation

As noted in Section 3, the properties of each image are unique. In order to obtain an accurate segmentation, the process must adapt to the properties of the image at hand. This section will summarise the development of a robust segmentation process for these images.

Demir and Yener (2005) break the segmentation algorithms currently used in histopathological images into two broad categories: region-based approaches and boundary-based approaches. Region-based approaches attempt to determine whether a given pixel belongs to a cell or nucleus, while boundary-based approaches attempt to find the boundary points which circumscribe a cell or nucleus. The automated boundary-based approaches typically involve minimising an energy function over a deformable spline. The energy function can be defined to penalise undesirable properties in the spline (Lee and Street, 1999). Other boundary-based approaches (Nielsen et al., 1999; Einstein et al., 1997) require significant user interaction to define the boundary points, and as such are unfeasible for large-scale images of tissue.

Region-based approaches include thresholding (where pixels above and below a certain threshold value are separated into separate classes) and learning algorithms. Gunduz et al. (2004) use c-means clustering with the pixel's colour information in order to identify the regions of interest, but the clusters still had to be assigned to 'cell' or 'non-cell' classes by a human expert.

Although interactive boundary-based approaches are generally more precise, we opted to use a region-based approach – FCM – as a starting point for the segmentation. Pham (2003) successfully used an edge-based adaptive extension to FCM for MRI image segmentation, and in other work had proposed other various extensions to FCM for dealing with various artifacts of the MRI imaging process. FCM also matched the parameters of our problem well – each image has nuclei, cytoplasm, and background pixels, but we do not have any specific information about the way each of those pixel types will appear in a given image. Since FCM only needs to know how many clusters the image data should be divided into, and computes and refines the cluster centre based on the image data, it was a good fit for this problem.

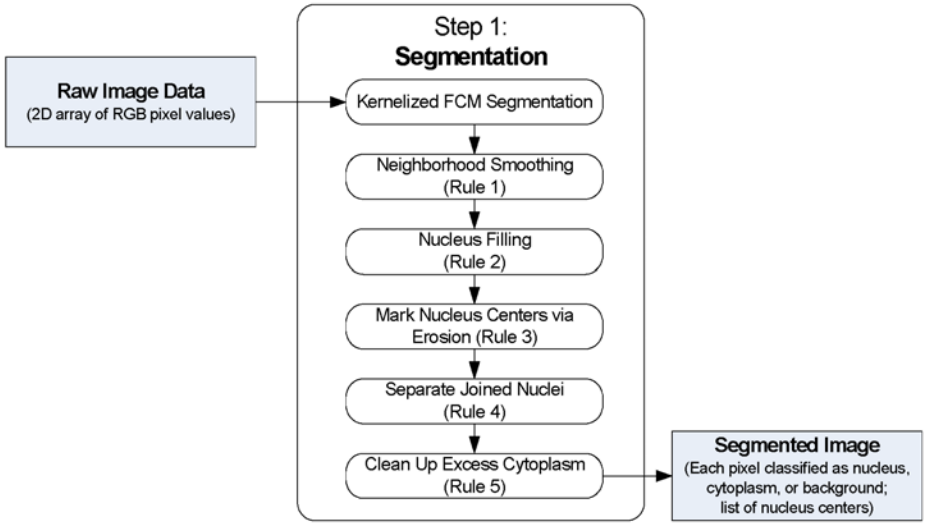
Our initial FCM implementation assigned pixels to clusters randomly, and allowed the FCM algorithm to separate the classes. While this approach worked reasonably well as far as the segmentation was concerned, it took several iterations for the algorithm to converge, and the clusters were not always in the same order (e.g., for a given run classes 0, 1, and 2 could represent the background, cytoplasm, and nucleus, in that order; for the following run on the same image, the classes could represent nucleus, background, and cytoplasm, in that order). Gunduz et al. (2004) solved this problem by using human intervention to determine the class to which each cluster belonged. In order to keep the process non-interactive, we leverage the defining information we do have for each class. Since we expect the background to be the lightest, the nuclei to be the darkest, and the cytoplasm to fall somewhere between those two, we can give the algorithm a head start by using rough estimates for each cluster centre and assigning each pixel initially to the closest cluster centre estimate (instead of assigning at random). This dramatically reduces the time the algorithm takes to converge, since it no longer has to self-organise the pixels into clusters – the basic structure is already there and simply needs to be refined. This also has the benefit of providing consistent class assignment – since we provided the basic structure in terms of initial cluster centre estimates, the algorithm will refine the clusters such that the classes appear in the order of those estimates.

Consequently, the images were first segmented based on the information in the individual pixels, and then further refined by evaluating the pixels in context within the image, which improved the accuracy of the FCM segmented image. The following Expert System Rules were used to refine the segmentation accuracy:

- pixels in geographic proximity are likely to have the same classification
- nuclei are geometrically closed solids
- nuclei are bound by a convex curve and are at least several pixels in diameter
- shapes with simple geometric centres represent multiple nuclei
- cytoplasm associated with a cell is near a nucleus.

Therefore, the kernelised FCM and rule-based segmentation subsystem is depicted in Figure 5. The raw image data is first processed by the Kernelised FCM to provide an initial, coarse image segmentation. This segmentation is then refined by the application of the five rules discussed above, yielding a final segmented image and a list of nucleus centres. This segmented image and list of centres is then used as the input to the next step in the overall process.

Figure 5 Details of the segmentation sub-process



4.3 Feature extraction and measurement

The features described here correlate in some way with the cell characteristics discussed in the Medical Background section. The 15 features selected are listed in Table 1. This is by no means an exhaustive list of possible features, but represents a variety of characteristics that are relatively straightforward to quantify. These features can be broadly categorised into four areas: those related to the size of the cells and nuclei, those related to the nucleocytoplasmic ratio, those related to the texture of the stained nuclei, and those relating to the shape of the nuclei. Metrics 1–5 and 12–15 all fall into the broad category of morphological features (those dealing with size and shape). Metrics 7 and 9–11 describe textural features (those which quantify variations in the intensity of a surface – smoothness, coarseness, regularity). Metrics 6 and 8 (and to a lesser degree 7 and 9) are intensity-based features (providing information on the colour intensity histogram of the pixels within a nucleus).

Table 1 List of features measured in segmented images

Metric	*Description*
Cell size/anisonucleosis	
1	Average nucleus area
2	Standard deviation of nucleus area
3	Average cytoplasm area
4	Average cell area
Nucleocytoplasmic ratio	
5	Nucleocytoplasmic ratio
Nuclear texture/hyperchromasia	
6	Average nucleus pixel intensity (measured across entire image)
7	Standard deviation corresponding to Metric 6
8	Average of nucleus average intensity (each nucleus averaged separately)
9	Average of Standard Deviation of nucleus pixel intensity (SD of each nucleus measured separately)
10	Standard Deviation corresponding to Metric 8
11	Standard Deviation corresponding to Metric 9
Nuclear shape/deformity	
12	Average nucleus radius (each nucleus measured separately)
13	Average of Standard Deviation of nucleus radius (SD of nucleus radius measurements measured separately for each nucleus)
14	Standard Deviation corresponding to Metric 12
15	Standard Deviation corresponding to Metric 13

The first category addresses both the size of the cells and nuclei (Metrics 1, 3, and 4). Metric 1 is measured simply by counting the number of pixels classified as nucleus and dividing by the number of marked centres to find the average number of pixels per nucleus. Metric 3 can be measured in the same way, counting cytoplasm pixels instead of nucleus pixels, and Metric 4 is the sum of the Metrics 1 and 3. Metric 2 measures variations in size between nuclei, since a lack of uniformity between nuclei can be an indicator of cancer. This measure requires a list of the area for each individual nucleus. By using the marked centres to iterate through the nuclei in the image, and counting the number of pixels in each nucleus, we can produce this list, from which we can calculate the standard deviation.

The nucleocytoplasmic ratio can also be calculated from Metrics 1 and 3 by dividing the nucleus area (Metric 1) by the cytoplasm area (Metric 3). Unfortunately, since the boundaries between the cytoplasm from one cell to the next are not visible, we can not evaluate any of the cytoplasm related metrics on a cell by cell basis, but only across the entire image.

The third category deals with the texture of the nuclei in the image. Variations in pixel intensity within a given nucleus can indicate a granular texture (Metrics 7 and 9), and the overall intensity within a nucleus (Metrics 6 and 8) can also be meaningful, since certain cancerous nuclei stain very darkly (hyperchromasia). Metrics 10 and 11 are intended to measure the degree of uniformity across all of the nuclei in the image – Metric 10 measuring the consistency of the average nucleus' intensity, while Metric 11 measures whether all nuclei have a similar level of granularity, or whether some have a very smooth texture while others are coarse and granular.

Finally, the fourth category evaluates the shape or roundness of the nuclei in the image, and the consistency of that shape throughout the image. For Metric 12 and 13, the radius of each nucleus was measured in eight directions from the marked centre (N, S, E, W, NE, NW, SE, and SW). These eight measurements were averaged to provide the radius for that nucleus (which was the basis for Metric 12), and the standard deviation of those eight measurements was recorded for each nucleus (which provided the basis for Metric 13). Metric 13 therefore tells us how round the typical nucleus is (a round nucleus having a lower standard deviation of radius measurements than that of an oblong or irregular nucleus). Metrics 14 and 15 tell us how much the nuclei in the image vary from the typical values given by Metrics 12 and 13, and thus address the degree of consistency in shape of the nuclei in the image.

4.4 Classification

The primary classifier used in this study was the SVM. While the SVM has been successfully used for cancer identification in several studies, it has not previously been applied to histopathological cell images. The SVM was chosen for its ability to perform well in the presence of noise in the source data, as well as its guarantee of solving to a global minimum for the parameters provided. Since a number of the measurements taken from the image are related (e.g., the nucleus area, cytoplasm area, and average nuclear diameter are all distinct measures, but related to the cell size), we want a classifier that will be able to extract the most information from the measurement data provided, but not allow the lack of orthogonality in the data to obscure the correct classification.

5 Results

This section summarised the results obtained from the kernelised FCM-SVM approach described here as well as the approach using Thiran and Macq's feature vector for comparison.

5.1 Results from the Kernelised FCM-SVM approach

Several machine learning algorithms have been applied to cancer diagnosis from histopathological images, including neural networks, *k*-nearest neighbourhood, logistic regression, fuzzy systems, linear discriminant analysis, and decision trees (Demir and Yener, 2002). Statistical tests have also been used in some studies in order to provide classification.

The primary classifier used in this study was the SVM. While the SVM has been successfully used for cancer identification in several studies, it has not previously been applied to histopathological cell images. The SVM was chosen for its ability to perform well in the presence of noise in the source data, as well as its guarantee of solving to a global minimum for the parameters provided. Since a number of the measurements taken from the image are related (e.g., the nucleus area, cytoplasm area, and average nuclear diameter are all distinct measures, but related to the cell size), we want a classifier that will be able to extract the most information from the measurement data provided, but not allow the lack of orthogonally in the data to obscure the correct classification.

5.1.1 Normal tissue vs. all cancers

The SVM configurations tested yielded a classification accuracy of up to 63% when tested against all of the training images. This corresponds to incorrectly classifying 31 of the 83 images. The best ROC A_Z achieved was 0.61. The kernel configuration used to achieve each of these is shown in Table 2.

Table 2 Best performing kernels for normal vs. all cancers (using training data only)

Kernel used	*Regularisation parameter* (*C*)	*Classification accuracy*	*ROC* A_Z
RBF, $\sigma = 2$	5	52/83 (63%)	0.609
Hyperbolic tangent, multiplier = 200	1100	52/83 (63%)	0.610

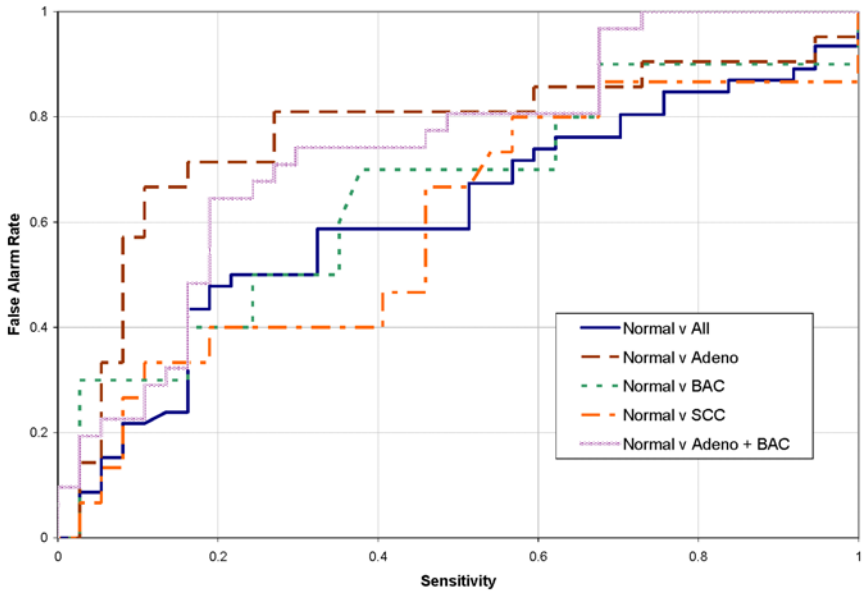
5.1.2 Normal tissue vs. combined adenocarcinomas and bronchioalveolar carcinomas

When classifying the normal images against the combined adenocarcinomas and bronchioalveolar carcinomas from the training set, the maximum classification accuracy of 72% (missing 19 out of 68 images) was slightly lower than for either type of cancer individually. The maximum A_Z achieved, 0.738, fell between the maximum values for these two cancer types when evaluated individually. Table 3 summarises the kernel parameters used to achieve these results. Figure 6 shows the actual ROC curves for the best performer (based on ROC A_Z) in each combination.

Table 3 Best performing kernels for normal vs. combined adenocarcinoma and bronchioalveolar carcinoma images (using training data only)

Kernel used	*Regularisation parameter* (*C*)	*Classification accuracy*	*ROC* A_Z
RBF, $\sigma = 1$	2	49/68 (72%)	0.738
RBF, $\sigma = 2$	5	47/68 (69%)	0.727

Figure 6 ROC curves for best performing classifiers using only the training data



5.1.3 Classification results using test dataset

Once the baseline performance had been established using one-hold-out cross-validation on the training data, a second series of tests was done, training the SVM with the entire training set.

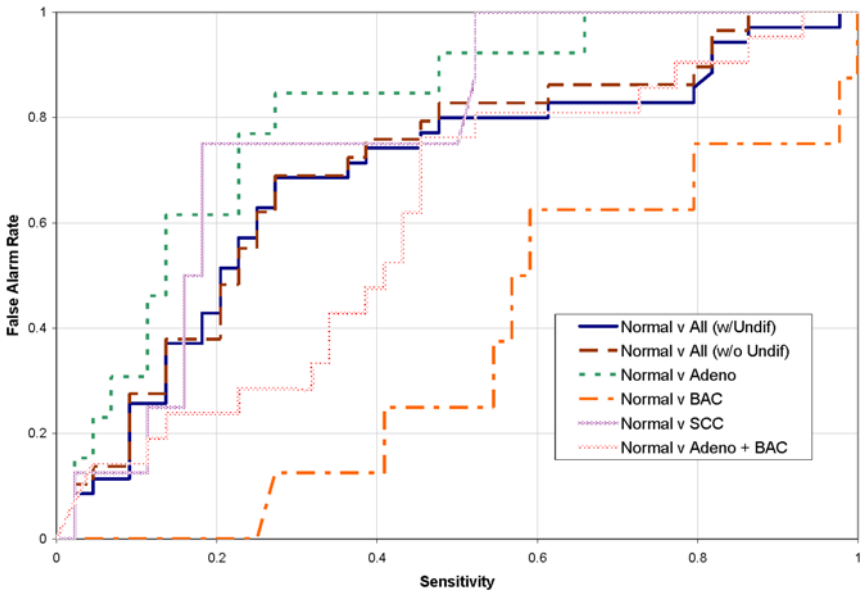
5.1.3.1 Normal tissue vs. combined adenocarcinomas and bronchioalveolar carcinomas

Finally, when classifying the normal images against the combined adenocarcinomas and bronchioalveolar carcinomas from the training set, the maximum classification accuracy was 69% (missing 20 out of 65 images). The maximum A_Z achieved, 0.601, fell between the maximum values for these two cancer types when evaluated individually, which is expected, given the difficulty we experienced in classifying the bronchioalveolar carcinomas from the test set. Table 4 summarises the kernel parameters used to achieve these results. Figure 7 shows the actual ROC curves for the best performer (based on ROC A_Z) in each combination.

Table 4 Best performing kernels for normal vs. combined adenocarcinoma and bronchioalveolar carcinoma images

Kernel used	*Regularisation parameter* (C)	*Classification accuracy*	*ROC* A_Z
RBF, $\sigma = 1$	2	49/68 (72%)	0.738
RBF, $\sigma = 2$	5	47/68 (69%)	0.727

Figure 7 ROC curves for best performing classifiers evaluated on the test data (see online version for colours)



5.1.4 Comparison of training dataset and test dataset results

With the exception of the bronchioalveolar carcinomas, the performance on the test data exceeded the performance using one-hold-out cross-validation on the training data for each of the groupings (see Table 5). This indicates that the classifier is generalising well – not only among the training cases, but also with respect to cases it has not seen. Although more training data is always desirable, the performance on the test data also argues that the training data used is sufficient to generate a reasonably accurate model, and that the performance achieved in this study is an accurate baseline for the process's potential performance.

Table 5 Comparison of best performers on the training data and the test data

Grouping	*Training performance*	*Test performance*
Normal vs. all (w/Undif)	N/A	0.687
Normal vs. all (w/o Undif)	0.610	0.704
Normal vs. adenocarcinomas	0.767	0.806
Normal vs. bronchioalveolar carcinomas	0.650	0.357
Normal vs. combined adenocarcinomas and bronchioalveolar carcinomas	0.738	0.601
Normal vs. squamous cell carcinomas	0.594	0.768

5.1.5 Discussion of the Fuzzy FCM-SVM approach

The primary original contributions of the study are threefold. The first is a complete end-to-end process – other research is currently working on individual pieces of a process

such as this, but a complete end-to-end process is unique to our work. Additionally, very few, if any, other researchers are working with cytological diagnosis based on the type of source data used in this study (large field microscopy images). The second contribution is the kernelised extension to FCM – although FCM has been in widespread use over the past 25 years, we have recast this framework in terms of kernel-based classifiers. This allows the clustering done by FCM to be based on the kernel's assessment of the similarity between data points. The third contribution is the identification of an expended list of cytological features which improves the accuracy of the classification process. One of the key advantages of this integrated end-to-end process over a manual assessment of the image is the objectivity of the automated evaluation. Rather than requiring a technician to make subjective observations about a particular tissue sample, which may vary from one technician to the next (inter-observer variability), or may even be scored differently by the same technician on different occasions (intra-observer variability), the tissue image can be evaluated directly by this process, providing a higher degree of consistency and reproducibility in the results.

5.2 Results using Thiran and Macq's feature vector

Thiran and Macq (1996) describe four morphological cell features used as a basis for lung cancer classification in their study. The first feature is the nucleocytoplasmic ratio, or the ratio of nucleus area to cytoplasm area. This is exactly the same measure as Metric 5 used in this study.

The second feature is a measure of anisonucleosis, or variation in nucleus size. Thiran and Macq estimate the size of each nucleus by taking radius measurements in each of eight cardinal directions and averaging those measurements (an additional step is taken to remove from the calculation any measurements which are greater than the distance to the next nearest nucleus centre, since Thiran and Macq's segmentation did not separate joined nuclei – since our segmentation separates each nucleus into its own contiguous area, we need take no such precaution). The final feature is computed by taking the variance of the distribution of cell measurements and dividing by the square of the mean nucleus size. We measure this feature in Metric 2, the standard deviation of the nucleus area, and also in Metric 14, standard deviation of the distribution of nucleus radius averages.

The third measure is nuclear deformity, or how much the nucleus varies from being perfectly round. The radius measurements are used again (once again with the accommodation for joined nuclei), but the variance is measured for the radius measurements for a particular nucleus. The feature is then calculated as the mean of the distribution of nucleus radius measurement variances, once again divided by the mean nucleus size. In this study, nuclear deformity is measured through Metric 13, the average of the distribution of standard deviations from individual nucleus radius measurements, and also through Metric 15, which is the standard deviation corresponding to the distribution whose average is reported in Metric 13.

The final measure is hyperchromasia, or the average contribution of the small spikes in pixel intensity which are responsible for the granular appearance of the nuclei. We measure this feature in Metrics 7 and 9, the standard deviation of the nucleus pixel intensities across the entire image, and the average of the distribution of standard deviations of nucleus pixel intensities when each nucleus is evaluated individually.

Each of these four features was extracted from each of the 83 segmented training images, and this four-element feature vector evaluated using one-hold-out cross-validation and the same battery of classification tests performed with the training data. Table 6 summarises the best results (based on ROC A_Z) using Thiran and Macq's feature vector for each arrangement of the classes, and compares those results to the best results obtained with the full feature vector used in this study.

Table 6 Comparison of best performers on the training data using Thiran and Macq's (1996) feature vector and the feature vector developed in this study

Grouping	*A_Z using Thiran and Macq's feature vector*	*A_Z w/our feature vector*	*Percentage improvement*
Normal vs. all	0.604	0.610	0.9
Normal vs. adenocarcinomas	0.642	0.767	19.5
Normal vs. bronchioalveolar carcinomas	0.639	0.650	1.7
Normal vs. combined adenocarcinomas and bronchioalveolar carcinomas	0.622	0.738	18.6
Normal vs. squamous cell carcinomas	0.547	0.594	8.6

The above table shows that the performance of our feature vector as good as or better than Thiran and Macq's in every case. With the exception of bronchioalveolar carcinomas, each individual cancer classification task experienced at least a modest improvement, with two groupings showing nearly 20% improvement in classification accuracy due to our feature vector.

6 Conclusions

The accuracy obtained by this process, both on the training data alone, and when evaluated with an independent test set, demonstrates the promise of this end-to-end approach to the problem. The test results for a single type of cancer (adenocarcinoma) are particularly good for a classification problem based on challenging real-world data, especially taking into account the amount of additional optimisation possible throughout the process. Comparisons with other related research also show the superiority of this approach. While additional research is needed to maximise the accuracy of such a system, the components in this study form the basis for a robust, self-adapting process that can quickly and accurately generate reproducible classifications for these images.

The accuracy of the process on the test data supports the hypothesis that an accurate predictive model can be generated from the training images. The fact that the performance of the process on the independent test data set is comparable to the one-hold-out performance on the training data alone also supports the hypothesis that the performance achieved in this study is an accurate baseline for the process's potential performance against much larger quantities of data.

Specifically, we demonstrate that the performance of our feature vector is as good as or better than Thiran and Macq's in every case. With the exception of bronchioalveolar carcinomas, each individual cancer classification task experienced at least a modest improvement, with two groupings each showing nearly 20% improvement in classification accuracy due to our feature vector.

Acknowledgements

This work was supported in part by the New York Health Science Board under Grant No. C017959.

References

American Cancer Society (2007a) *Cancer Facts and Figures 2007*, American Cancer Society, Atlanta.

American Cancer Society (2007b) *What are the Key Statistics for Lung Cancer?*, Revised 10/25/2006, Atlanta, http://www.cancer.org/docroot/CRI/content/CRI_2_4_1x_What_Are_the_Key_Statistics_About_Lung_Cancer_15.asp?rnav=cri, Accessed 9 October.

American Cancer Society (2007c) *What is Non-small Cell Lung Cancer?*, Revised 10/25/2006, Atlanta, http://www.cancer.org/docroot/CRI/content/CRI_2_4_1x_What_Is_Non-Small_Cell_Lung_Cancer. asp?rnav=cri, Accessed 9 October.

Bengtsson, E., Wählby, C. and Linblad, J. (2004) 'Robust cell image segmentation methods', *Pattern Recognition and Image Analysis*, Vol. 14, No. 2, pp.157–167.

Bezdek, J. (1981) *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press, New York.

Burges, C.J.C. (1998) 'A tutorial on Support Vector Machines for pattern recognition', *Data Mining and Knowledge Discovery*, Vol. 2, No. 2, pp.121–167.

Cotran, R.S., Kumar, V. and Collins, T. (1999) *Robbins Pathologic Basis of Disease*, 6th ed., W.B. Saunders Company, Philadelphia.

Cristianini, N. and Shawe-Taylor, J. (2000) *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*, Cambridge University Press, New York.

Demir, C. and Yener, B. (2002) *Automated Cancer Diagnosis Based on Histopathological Images: A Systematic Survey*, Technical Report, Rensseler Polytechnic Institute, TR-05-09.

Demir, C. and Yener, B. (2005) *Automated Cancer Diagnosis Based on Histopathilogical Images: A Systematic Survey*, Technical Report, Rensselaer Polytechnic Institute, Department of Computer Science, TR-05-09.

Einstein, A.J., Gil, J., Wallenstein, S., Bodian, C.A., Sanchez, M., Burstein, D.E., Wu, H.S. and Liu, Z. (1997) 'Reproducibility and accuracy of interactive segmentation procedures for image analysis in cytology', *Journal of Microscopy*, Vol. 188, pp.136–148.

Fogel, D.B., Wasson, E.C. and Boughton, E.M. (1995) 'Evolving neural networks for detecting breast cancer', *Cancer Letters*, Vol. 96, pp.49–53.

Fogel, D.B., Wasson, E.C., Boughton, E.M. and Porto, V.W. (1997) 'A step toward computer-assisted mammography using evolutionary programming and neural networks', *Cancer Letters*, Vol. 119, pp.93–97.

Fogel, D.B., Wasson, E.C., Boughton, E.M. and Porto, V.W. (1998a) 'Evolving artificial neural networks for screening features from mammograms', *Artificial Intelligence in Medicine*, Vol. 14, pp.317–326(10).

Fogel, D.B., Wasson, E.C., Boughton, E.M., Porto, V.W. and Angeline, P.J. (1998b) 'Linear and neural models for classifying breast masses', *IEEE Trans. Medical Imaging*, Vol. 17, No. 3, pp.485–488.

Garbay, C., Chassery, J. and Brugal, G. (1986) 'An iterative region-growing process for cell image segmentation based on local color similarity and global shape criteria', *Analytical and Quantitative Cytology and Histology*, Vol. 8, No. 1, pp.25–34.

Graham, R.M. (1972) *The Cytologic Diagnosis of Cancer*, 3rd ed., W.B. Saunders Company, Philadelphia.

Gunduz, C., Yener, B. and Gultekin, S.H. (2004) 'The cell graphs of cancer', *Bioinformatics*, Vol. 20, pp.i145–i151.

Gunn, S. (1998) 'Support Vector Machines for classification and regression', *ISIS Technical Report, Image Speech and Intelligent Systems Group*, University of Southampton, Southhampton, UK.

Haykin, S. (1999) *Neural Networks*, Prentice-Hall, New Jersey.

Jiang, T. and Yang, F. (2002) 'An evolutionary Tabu search for cell image segmentation', *IEEE Transactions on Systems, Man, and Cybernetics – Part B-Cybernetics*, Vol. 32, No. 5, pp.675–678.

Koss, L.G. (1979) *Diagnostic Cytology and Its Histopathologic Bases*, 3rd ed., J.B. Lippincott Company, Lippincott Williams & Wilkins, 4th ed., (April 1992) Philadelphia, Vol. 2.

Land Jr., W.H., Bryden, M., Lo, J.Y., McKee, D.W. and Anderson, F.R. (2002a) 'Performance tradeoff between Evolutionary Computation (EC)/Adaptive Boosting (AB) hybrid and Support Vector Machine breast cancer classification paradigms', *Proc. 2002 Congress on Evolutionary Computation, 2002 (CEC '02)*, Vol. 1, pp.187–192.

Land Jr., W.H., Masters, T. Lo, J.Y. and McKee, D. (2000) 'Using evolutionary computation to develop neural network breast cancer benign/malignant classification models', *Proc. 4th World Conference on Systemics, Cybernetics and Informatics (SCI2000)*, July 23–26, Orlando, FA, USA Vol. 10, pp.343–347.

Land Jr., W.H., Masters, T. Lo, J.Y. and McKee, D. (2001a) 'Application of evolutionary computation and neural network hybrids for breast cancer classification using mammogram and history data', *Proc. 2001 Congress on Evolutionary Computation, 2001*, Vol. 2, pp.1147–1154.

Land Jr., W.H., Masters, T. Lo, J.Y. and McKee, D. (2001b) 'Application of adaptive boosting to EP-derived Multi-Layer Feedforward Neural Networks (MLFNs) to improve benign/malignant breast cancer classification', in Sonka, M. and Hanson, K.M. (Eds.): *Proc. SPIE, Medical Imaging 2001: Image Processing*, Vol. 4322, pp.1717–1724.

Land Jr., W.H., Masters, T., Lo, J.Y., McKee, D.W. and Anderson, F.R. (2001c) 'Performance analysis of Evolutionary Computation (EC)/Adaptive Boosting (AB) hybrids for breast cancer classification', *Proc. 6th International Conference on Information Systems Analysis and Synthesis (ISAS2001)*, Orlando, FL, July, pp.22–25.

Land Jr., W.H., Masters, T., Lo, J.Y., McKee, D.W. and Anderson, F.R. (2001d) 'New results in breast cancer classification obtained from an evolutionary computation/adaptive boosting hybrid using mammogram and history data', *Proc. 2001 IEEE Mountain Workshop on Soft Computing in Industrial Applications, 2001, (SMCia/01)*, pp.47–52.

Land Jr., W.H., McKee, D.W. Anderson, F.R. and Lo, J.Y. (2003a) 'Improving the predictive value of mammography using a specialized evolutionary programming hybrid and fitness functions', in Sonka, M. and Fitzpatrick, J.M. (Eds.): *Proc. SPIE, Medical Imaging 2003: Image Processing*, Vol. 5032, pp.898–907.

Land Jr., W.H., McKee, D.W. Anderson, F.R. and Lo, J.Y. (2004a) 'Breast cancer classification improvements using a new kernel function with evolutionary-programming-configured Support Vector Machines', in Fitzpatrick, J.M. and Sonka, M. (Eds.): *Proc. SPIE, Medical Imaging 2004: Image Processing*, Vol. 5370, pp.880–887.

Land Jr., W.H., McKee, D.W. Lo, J.Y. and Anderson, F.R. (2002b) 'Improving mammogram screening using a bank of Support Vector Machines (SVMs)', in Dagli, C.H., Buczak, A.L. Ghosh, J., Embrechts, M.J., Ersoy, O. and Kercel, S.W. (Eds.): *Intelligent Engineering Systems Through Artificial Neural Networks (ANNIE 2002)*, ASME Press, Fairfield, NJ, Vol. 12, pp.779–784.

Land Jr., W.H., Wong, L., McKee, D. Masters, T., Anderson, F., Raturi, A. and Lo, J.Y. (2004c) 'New results in Computer Aided Diagnosis (CAD) of breast cancer using a recently developed SVM/GRNN oracle hybrid', in Fitzpatrick, J.M. and Sonka, M. (Eds.): *Proc. SPIE, Medical Imaging 2004: Image Processing*, Vol. 5370, pp.777–784.

Land Jr., W.H., Wong, L., McKee, D., Embrechts, M., Salih, R. and Anderson, F. (2004b) 'Applying Support Vector Machines to breast cancer diagnosis using screen film mammogram data', *Proc. 17th IEEE Symposium on Computer-Based Medical Systems, 2004 (CBMS 2004)*, pp.224–228.

Land Jr., W.H., Wong, L., McKee, D., Masters, T., Anderson, F. and Sarvaiya, S. (2004d) 'Data fusion of several support-vector-machine breast-cancer diagnostic paradigms using a GRNN oracle', in Dasarathy, B.V. (Ed.): *Proc. SPIE, Multisensor, Multisource Information Fusion: Architectures, Algorithms, and Applications*, Vol. 5434, pp.423–430.

Land Jr., W.H., Wong, L., McKee, D.W., Masters, T. and Anderson, F.R. (2003c) 'Breast cancer Computer Aided Diagnosis (CAD) using a recently developed SVM/GRNN oracle hybrid', *IEEE International Conference on Systems, Man and Cybernetics*, Vol. 5, pp.4705–4711.

Land, W.H., McKee, D.W., Velazquez, R., Wong, L., Lo, J.Y. and Anderson, F.R. (2003b) 'Application of Support Vector Machines to breast cancer screening using mammogram and clinical history data', in Sonka, M. and Fitzpatrick, J.M. (Eds.): *Proc. SPIE, Medical Imaging 2003: Image Processing*, Vol. 5032, pp.546–556.

Lee, K-M. and Street, W.N. (1999) 'A fast and robust approach for automated segmentation of breast cancer nuclei', *Proceedings of the Second IASTED International Conference on Computer Graphics and Imaging*, Palm Springs, CA, pp.42–47.

Lo, J.Y., Baker, J.A., Kornguth, P.J. and Floyd Jr., C.E. (1999) 'Effect of patient history data on the prediction of breast cancer from mammographic findings with artificial neural networks', *Academic Radiology*, Vol. 6, pp.10–15.

Lo, J.Y., Baker, J.A., Kornguth, P.J., Iglehart, J.D. and Floyd, C.E. (1997) 'Predicting breast cancer invasion with artificial neural networks on the basis of mammographic features', *Radiology*, Vol. 203, pp.159–163.

McKee, D.W. (2001) *Boosting Evolved Artificial Neural Networks to Improve Breast Cancer Classification Accuracy*, Master's Thesis, State University of New York, Binghamton, NY.

Morrison, M. and Attikiouzel, Y. (1992) 'A probabilistic neural network based image segmentation network for magnetic resonance images', *Proceedings of the International Joint Conference on Neural Networks*, Vol. 3, pp.60–65.

Nedzved, A., Ablameyko, S. and Pitas, I. (2000) 'Morphological segmentation of histology cell images', *Proceedings of 15th International Conference on Pattern Recognition, 2000, (ICPR '00)*, pp.500–503.

Nielsen, B., Albregtsen, F. and Danielsen, H.E. (1999) 'The use of fractal features from the periphery of cell nuclei as a classification tool', *Analytical Cellular Pathology*, Vol. 19, pp.21–37.

Pham, D.L. (2003) 'Unsupervised tissue classification in medical images using edge-adaptive clustering', *Proc. 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2003*, Vol. 1, pp.634–637.

Platt, J. (1998) 'Fast training of Support Vector Machines using sequential minimal optimization', in Scholkopf, B., Burges, C. and Smola, A. (Eds.): *Advances in Kernel Methods: Support Vector Training*, MIT Press, Cambridge, MA, pp.185–209.

Sammouda, M., Sammouda, R., Niki, N., Yamaguchi, N. and Moriyama, N. (2002) 'Cancerous nuclei detection on digitized pathological lung color images', *Journal of Biomedical Informatics*, Vol. 35, pp.92–98.

Sanei, S. and Lee, T.K.M. (2003) 'Cell recognition based on PCA and bayesian classification', *4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003)*, Nara, Japan, April.

Schnorrenberg, F., Pattichis, C.S., Kyriacou, K.C. and Schizas, C.N. (1997) 'Computer-aided detection of breast cancer nuclei', *IEEE Transactions on Information Technology in Biomedicine*, Vol. 1, No. 2, pp.128–140.

Takahashi, M. (1971) *Color atlas of Cancer Cytology*, J.B. Lippincott Company, Philadelphia.

Thiran, J. and Macq, B. (1996) 'Morphological feature extraction for the classification of digital images of cancerous tissues', *IEEE Transactions on Biomedical Engineering*, Vol. 43, No. 10, pp.1011–1020.

Vapnik, V.N. (1995) *The Nature of Statistical Learning Theory*, Wiley, New York.

Wählby, C., Sintorn, I-M., Erlandsson, F., Borgefors, G. and Bengtsson, E. (2004) 'Combining intensity, edge and shape information for 2D and 3D segmentation of cell nuclei in tissue sections', *Journal of Microscopy*, Vol. 215, No. 1, pp.67–76.

Zhou, Z-H., Jiang, Y., Yang, Y-B. and Chen, S-F. (2002) 'Lung cancer cell identification based on artificial neural network ensembles', *Artificial Intelligence in Medicine*, Vol. 24, No. 1, pp.25–36.