An automated pattern recognition system for the quantification of inflammatory cells in hepatitis C infected liver biopsies

Simon Hodgson, Robert F. Harrison and Simon S. Cross

Department of Automatic Control and Systems Engineering
The University of Sheffield
Mappin Street
S1 3JD UK

Research Report No 834

August 2003

An automated pattern recognition system for the quantification of inflammatory cells in hepatitis C infected liver biopsies

Simon Hodgson*†, Robert F. Harrison* and Simon S. Cross**

*Department of Automatic Control and Systems Engineering, University of Sheffield, Sheffield S1 3JD, UK. **Academic Unit of Pathology, Division of Genomic Medicine, School of Medicine and Biomedical Sciences, University of Sheffield, Sheffield S10 2RX, UK.
e-mail: {cop01sh, r.f.harrison, s.s.cross}@shef.ac.uk

Research Report No. 834
August 2003

Abstract

Hepatitis C is a common viral infection of the liver. The degree of inflammation associated with the infection is normally estimated manually from a liver biopsy, by considering the quantity and nature of inflammatory cells. This paper presents an automated pattern recognition system for the quantification of inflammatory cells in liver biopsies. Initially, images are corrected for colour variation. Features are then extracted from colour biopsy images at positions of interest identified by adaptive thresholding and clump decomposition. A sequential floating search method and principal component analysis are used to reduce the dimensionality of the feature vector. Manually annotated training images allow supervised training by providing the class membership for each position of interest. Gaussian parametric and gaussian mixture model density estimation methods are compared, and are used to classify cells as either inflammatory or healthy via Bayes' theorem. The system is optimised using a response surface method, where the response or system performance is derived from the area under the receiver operating characteristic curve. The optimised system is then tested on test images previously ranked by a number of observers with varying levels of pathology experience. The observers results are compared to the automated system using Spearman rank correlation. Results show that this system can rank 15 test images, with varying degrees of inflammation, in strong agreement with five expert pathologists.

1 Introduction

> "It is called the silent epidemic and it has infected an estimated 500,000 people in Britain, and 170 million worldwide, without attracting the kind of tabloid headlines devoted to HIV/Aids. It can lay dormant in a carrier for up to 25 years, and health professionals fear it could be a timebomb ticking away with no one knowing when it might explode. It is the hepatitis C virus, or HCV [16]. "
>
> — Paul Humphries, The Guardian, Wednesday March 6,2002.

Although the above newspaper extract is sensational, the figures quoted can be substantiated. The World Health Organisation (WHO) estimates that 170 million people, 3% of the world's population, are currently infected with the hepatitis C virus (HCV) [37]. This virus is usually transmitted by exposure to the blood or blood products of an infected person. In the majority of cases infected people do not develop symptoms for a number of years, leaving them totally unaware of their situation [38]. Liver damage is not caused by the virus itself but by the body's immune response to the attack. This damage can be extremely serious, resulting in liver failure and death of the patient. The current treatment for HCV, according to the UK clinical guidelines, is with a combination therapy of two drugs, Interferon–alpha and Ribavirin [5]. A major factor in prescribing combination therapy is that both drugs produce side effects in most people [5]. The cost of combination therapy is between £3000 and £12000 per patient per year [5]. It is generally thought that treating patients with expensive drugs with potentially serious side-effects may be inappropriate unless there is evidence of disease activity [1] . A liver biopsy is currently the only method available to assess HCV activity. The biopsy, involves removing a small core of tissue, approximately 15mm in length by 2-3mm in diameter, as shown in figure 1. This core is then processed in paraffin wax, cut into slices along its length and then stained. At this stage a trained histopathologist [2] will examine the samples under a light microscope and use his/her experience, combined with a detailed definition, to assess the level of damage. The damage can normally be categorised into two types and it is common to assign a numerical score relative to the level of damage for each type. One of the most widely used scoring method is the Ishak system [17], which can be summarised as

(1) Inflammation: assigned a necroinflammatory [3] (activity) score from 0–18.
(2) Scarring: assigned a fibrosis [4] (stage) score in the range 0–6.

Scarring is an indication of long-term disease activity and as a result remains relatively constant. For this reason, it is the assessment of inflammation that is normally the determining factor for a patient to receive treatment. The scoring process is time con-

[1] HCV is particularly likely to be associated with chronic disease[28]; for 20% of people with this form, liver disease will slowly progress to cirrhosis of the liver during the first 10 to 20 years.

[2] A person who studies the tissue changes associated with disease.

[3] Cell death caused by the body's inflammatory response.

[4] The formation of fibrous tissue.

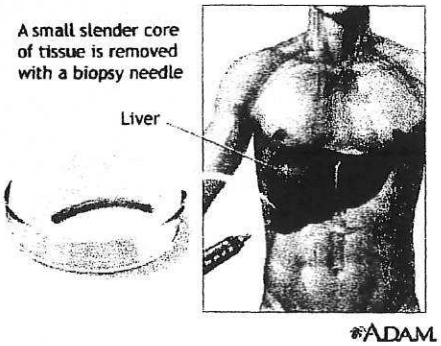


Fig. 1. An illustration of the liver biopsy procedure [1].

suming and requires highly experienced and qualified personnel. Studies have shown that it is often difficult for observers to agree on disease activity and stage scores when evaluating the same samples, and it is common for the same observer to assign different scores at a later date [11]. This inter- and intra-observer variability has been studied in depth by [11], who found that observer agreement was far better for the assessment of fibrosis (stage) than for inflammation (activity). This finding, together with emphasis on inflammatory activity when considering treatment, stresses the urgent need for improved reliability in the assessment of inflammation. It is proposed that an automated system could be developed using image processing and pattern recognition techniques, to assess, systematically, the level of inflammation in liver biopsies.

This paper presents research on the design, optimisation and testing of an automated pattern recognition system, to quantify, reliably, the amount of liver inflammation. Initially, the liver biopsy is examined in more detail with particular consideration given to the colour variation in biopsy samples. Next, previous approaches to this problem are discussed and a new pattern recognition system is presented. A method of system optimisation is then outlined. Finally, the optimised system is tested using images previously evaluated by human observers.

2 Liver Biopsy Interpretation

This section introduces the image characteristics of an HCV infected liver biopsy and discusses the colour variation between biopsy samples. To understand this investigation in more detail it is first necessary to consider the histopathological elements of a normal liver biopsy and the different forms of damage. A microscopic view of a standard liver biopsy from a healthy person shows liver cells (hepatocytes) forming interconnecting walls created by the close contact of cell membranes [34], as shown in figure 2. The nucleus is the dark mass located at the centre of each cell. The array of hepatocytes is only interrupted by other structural elements of the liver, such as

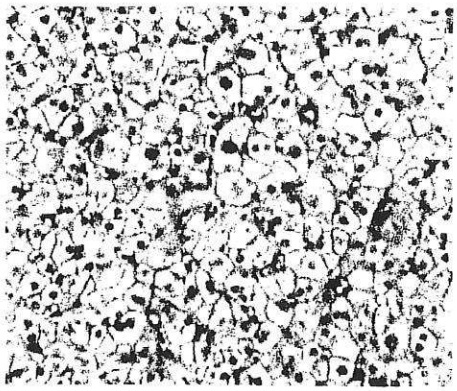


Fig. 2. Microscopic (10× objective) view of a normal liver biopsy.

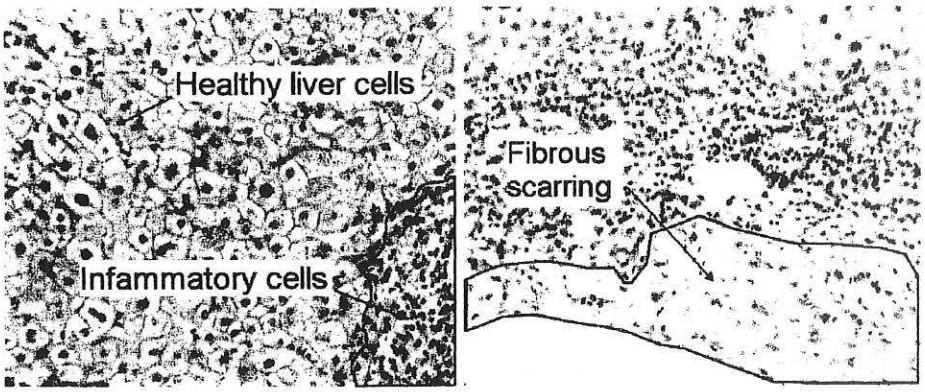
portal tracts[5], hepatic veins[6] and bile ducts[7] (not shown). The damage caused by HCV alters this structure and can normally be categorised into two types:

(1) Inflammation - Cell death (necrosis), caused by the viral attack, evokes an inflammatory response which is manifested by the appearance of inflammatory cells. The majority of these inflammatory cells are lymphocytes [29]. Figure 3(a) shows a region of lymphocyte cells. The lymphocyte cells are generally smaller, with the cell nuclei smaller and darker than those of hepatocyte cells.
(2) Fibrosis - The death of small groups of hepatocytes may leave the reticulum (cell membrane system) intact and the resulting regeneration will repair the damage. However if the reticulum is damaged, healing can only occur by scar and will lead to fibrosis. If scars are produced throughout the liver the lack of blood circulation leads to cirrhosis [29]. Figure 3(b) shows an example of scarring. The remainder of the cells are lymphocytes.

As explained in section 1, the focus of this work is to measure the degree of inflammation relative to the amount of other tissue, not including the background. This means the main task of this system is to group the cells into two classes, inflammatory (C_1) and healthy (C_2). Scar tissue will therefore be classified as 'healthy' for our purpose.

The biopsies used in this study are all stained using haematoxylin and eosin. This usually causes lymphocyte nuclei to appear dark purple, the hepatocyte nuclei to appear light purple and the background to appear white. Haematoxylin and eosin stain is commonly used by many pathology departments. This method can produce high colour variability across different samples as the stain mixture varies at different hospitals and laboratories. Another factor producing image variability is the illumination at the time of image capture. A system must be robust to these factors in order to interpret, adequately, new images.

[5] A tract of the portal system of the liver, which is a network of veins that begin and end in capillaries.
[6] Blood vessel in the liver that returns blood to the heart.
[7] Pathway for the transportation of bile from the liver to the gallbladder.



(a) (b)

Fig. 3. Microscopic (10× objective) view of a liver biopsy. (a) Healthy and inflammatory liver cells (b) A region of fibrous scarring.

2.1 Colour Correction

Cardei et al [7] propose a method of colour correction to counteract illumination variability. This involves using the difference between the background of images viewed under different illumination to colour correct the whole image. This simple technique can be expanded to correct colour variation in tissue caused by stain and illumination change. In brief, a reference image is selected by eye, using the natural human ability to determine mid-range colour attributes. A raw image requiring colour correction is also selected. Q_{raw} is the raw RGB image reshaped into an $N \times 3$ matrix, where N is the total number of pixels in the image. Similarly, Q_{cc} is a matrix containing values of the colour corrected image. Applying the diagonal model of illumination change [7] shows that

$$Q_{cc} = Q_{raw}.M \tag{1}$$

where

$$M = \text{diag}\left(\frac{\lambda^R_{ref}}{\lambda^R_{raw}}, \frac{\lambda^G_{ref}}{\lambda^G_{raw}}, \frac{\lambda^B_{ref}}{\lambda^B_{raw}}\right). \tag{2}$$

Thresholding each image to remove the background leaves only the tissue portion, which is defined by a region, R^i_j, where $i \in \{R, G, B\}$ and $j \in \{ref, raw\}$. The mean values, λ^i_j, are then derived from the tissue portion by

$$\lambda_j^i = \frac{1}{M_j} \sum_{m=1}^{M_j} (R_j^i)^n \tag{3}$$

where M_j is the number of pixels in each region. Figure 5 shows the result of colour correction on the images presented in figure 4. Qualitatively, the colour corrected images can be seen to be more similar that the raw images.

3 Pattern Recognition System

In this section we first discuss previous approaches to the cell classification problem. Next, details of the images used during the training process are presented and finally, a new pattern recognition system for cell classification is introduced.

3.1 Previous Approaches

The development of pattern recognition systems for cell classification is an active research field. Although no previous work as been identified which directly addresses the tasks outlined in section 1, a number of related studies have been completed.

Zhang et al [40] propose an image analysis system for the quantification of stellate cells in rat liver. This system is simpler than that required for inflammatory cells, because stellate cells have an auto-fluorescence property which allows them to be clearly identified using a fluorescence microscope. The resulting homogeneous objects are then simply segmented by thresholding. This work was conducted using macros written for an existing image analysis suite (computer-NIH) [27]. Quantification occurs by pixel counting. As the stellate cells are the only cells to fluoresce, no classification is required. This work is primarily of interest because of its use of the image analysis suite. This software is commonly used by researchers in this field for pattern recognition tasks.

Masseroli et al [23] investigate the quantification of liver *fibrosis*. The authors propose a novel image analysis system, 'FibroQuant', to segment semi-automatically regions of fibrosis tissue. The samples are first corrected for colour variation by a background subtraction technique. The system uses adaptive thresholding and area measurements to segment and classify different types of fibrosis. Results are then compared to the semi-quantitative scoring methods discussed in section 1 and show good correlation. Although the usefulness of fibrosis quantification for the assessment of HCV progression is discounted in section 1, this work is useful in demonstrating the importance of validation to existing manual methods.

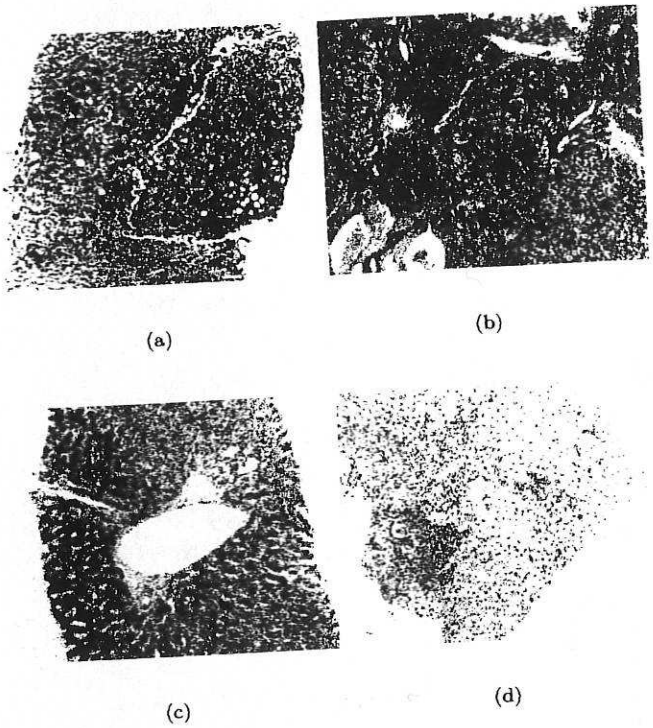


Fig. 4. Before colour correction

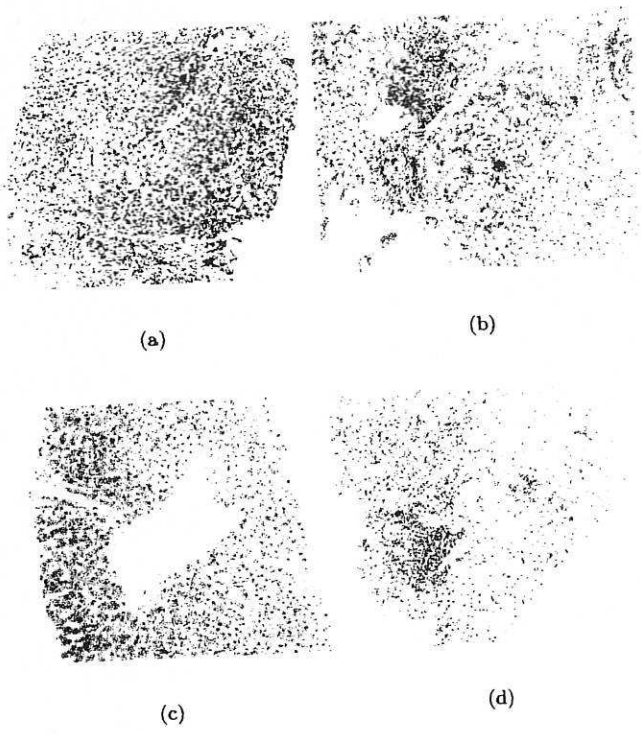


Fig. 5. After colour correction

Lake-Bakaar et al [21] address the very relevant problem of hepatocyte quantification. Hepatocytes are the cells crudely referred to as 'healthy cells' in section 1. This study prefers a histological approach to cell segmentation with the use of high contrasting stains, rather than pattern recognition techniques. The non-conventional staining allows much simpler segmentation than that required for the assessment of inflammation.

Santos et al [32] propose a more complete pattern recognition study for the detection of cellular necrosis (cell death) in cell cultures from swine. The feature vector is produced from 12 parameters derived from the local histogram and co-occurrence [14] matrix of a sliding window. No segmentation is performed. The authors use Fisher linear discriminant analysis to classify regions as either 'alive', 'dead' or 'background'. The number of cells is approximated by dividing the total cell area by a user defined size. The automatic method is then compared to manual identification using contingency table analysis.

3.2 Training Images

To train the system, two sets of 86 colour images of liver biopsies are used. Set 1 contains the raw images and set 2 contains an annotated version of the raw images. Annotated images show regions of inflammation, as demonstrated in figure 6(b). To simplify the time consuming manual annotation process, inflammation was only annotated for close groups of six or more inflammatory (lymphocyte) cells. Each image is a 1000 $\times$ 1280 pixels bitmap of red/green/blue (RGB) layers, taken at 10$\times$ objective magnification and shows only a part of the whole liver biopsy. The images have been specially selected to show a cross section of inflammatory and healthy cells, with variation in stain and illumination. The liver biopsy images were supplied and annotated by Dr Simon S. Cross (SSC), a consultant pathologist in the Academic Unit of Pathology, at the University of Sheffield, UK. During preprocessing, the closed annotated regions shown in figure 6(b) are converted into binary masks, as demonstrated in figure 6(c). This is later overlaid on the raw image to provide supervised cell classification during the training process.

3.3 New Approach

After completing the preprocessing steps of colour correction and the creation of the binary masks, the system is trained using the steps detailed in sections 3.3.1 to 3.3.4. The evaluation of new images using the trained system is then discussed in section 3.3.5.

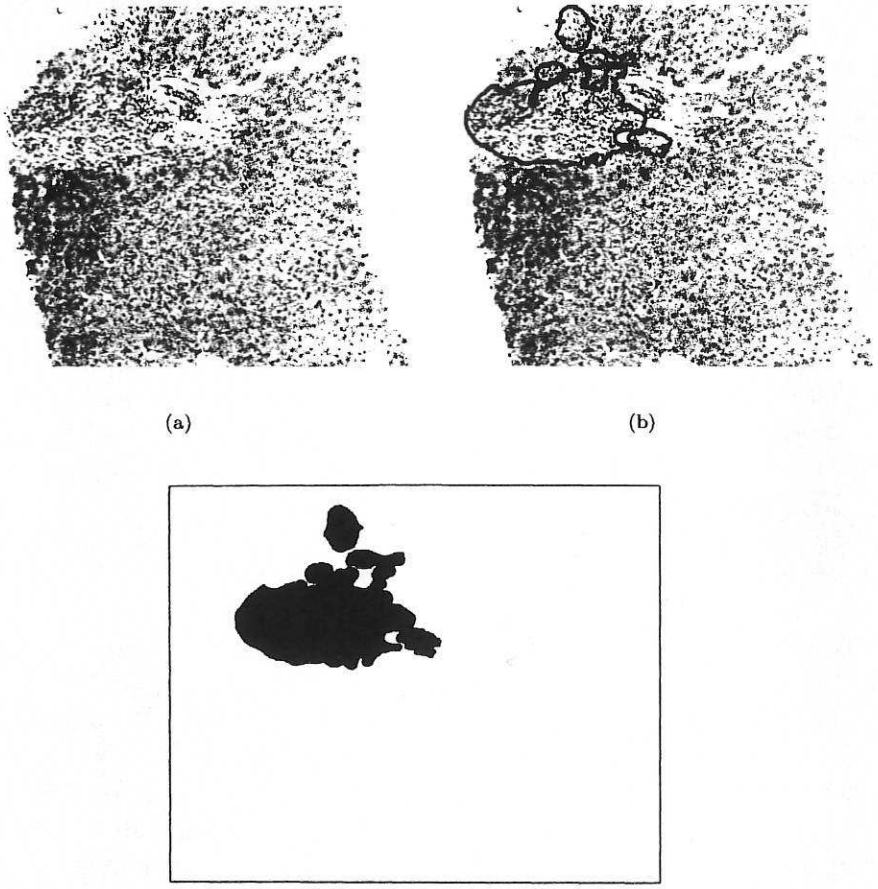


Fig. 6. Method to generate the binary mask images for supervised training. (a)A raw liver biopsy image. (b) Annotation showing regions of inflammation. Portal tract inflammation is shown in black and non-portal tract inflammation in blue. For this study, black and blue regions are considered of equal interest.(c) The binary mask image derived from the annotated image.

3.3.1 Thresholding

The image is thresholded to highlight the points of interest (POI) within each region e.g. all the cell nuclei. The method of thresholding uses histogram analysis. To describe this method in more detail, it is necessary to consider a greyscale representation of the raw RGB image. The histogram of grey levels taken across the whole image is either unimodal or bimodal, depending on the amount of background included in the original image, as shown in figure 7. This method uses the histogram lobe corresponding to the tissue region to calculate the threshold level. Therefore, it is first necessary to identify the tissue lobe and tissue lobe maximum. This is done by hill climbing a smoothed version of the original greyscale histogram, starting at the zero greyscale value (left side of figures 7(a) and 7(b)). Once a peak is found, local checks are performed to

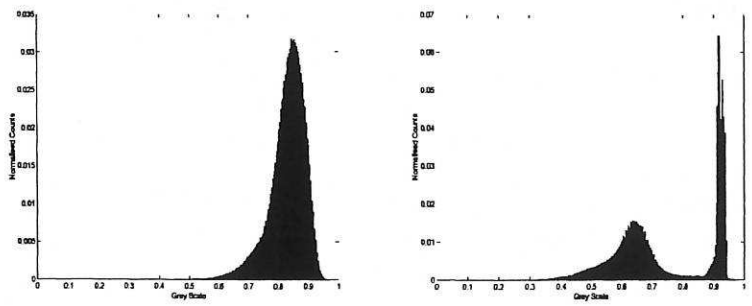
ensure this is the true lobe maximum. With the tissue lobe identified, thresholding at a suitable value within the lobe allows the darker cell nuclei to be segmented from the other tissue. Through experimentation, it was found that thresholding at 1.2 standard deviations (σ) below the tissue lobe maximum produces the best segmentation of cell nuclei across all training images. This method is illustrated in figure 8. Because of the non-gaussian form of the original histogram, σ is calculated by mirroring the lower half of the tissue lobe about the tissue lobe maximum and assuming a gaussian distribution.

3.3.2 Clump decomposition

Thresholding produces a binary representation of the cell nuclei. These nuclei are often touching or merged. This prevents identification of the true cell centroid [8] and makes it impossible to accurately quantify the number of cells. Clump decomposition is a technique to separate merged parts by the morphology of the combined or clumped parts [36]. In this study, a method of uniform recursive erosion is implemented based on the well known *watershed* [31] technique. This method is used to identify merged or marginally touching nuclei by splitting clumps at narrow points within the component. This method is demonstrated in figure 9 and discussed in more detail below:

(1) Initially, the morphological *opening* [33] operator is used to remove noise from the thresholded image (see figure 9(b)). An *opening* consists of an erosion [33] followed by a dilation [33].
(2) Each component (or clump) is then labelled using connected component analysis (CCA) [15]. This means each pixel within the image is allocated to an individual component and each component centroid is identified.

[8] The centre of mass of the region.



(a) (b)

Fig. 7. Two examples of greyscale biopsy histograms. (a) dark image with no background (b) lighter image with background.

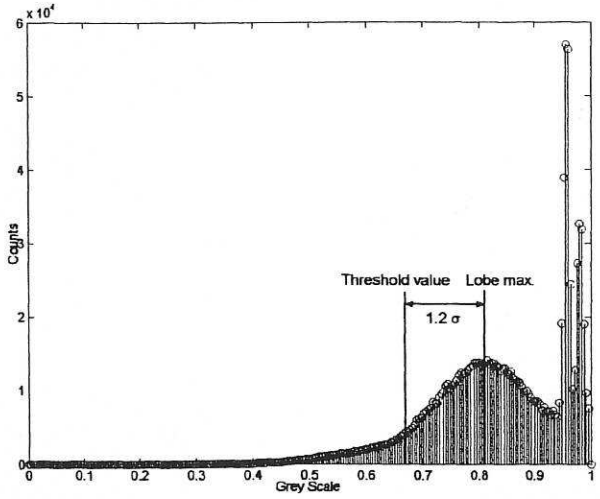


Fig. 8. Experiments show that thresholding at 1.2 standard deviations below the tissue lobe maximum produce the optimum thresholding results.

(3) The binary image is then uniformly eroded using an array structuring element of 3×3 pixels. The erosion splits the components at narrow sections which correspond to marginally touching cells. If a component splits by this process, CCA is used to calculate the centroids of the newly created 'child' components. If a component does not split, the original 'parent' centroid remains.
(4) The image is then recursively eroded, using the methods described in step 3, until no more components remain (see figure 9(d)).
(5) The final list of component centroids, containing the resulting mixture of 'parent' and 'child' details, is superimposed onto the image produced in step 1. Each pixel in this binary image is then re-allocated to the nearest component centroid, thus creating the patch work effect illustrated in figure 9(c).

Although more complex clump decomposition methods are now available (see [36,39,22]), this study has found the erosion technique effective and computationally efficient.

3.3.3 Extracting Image Features

Outputs from the clump decomposition process (the component centroids and the patch work of cells) are used to identify the POI within the colour biopsy image. Features are then calculated from image data extracted about these POI, using one of two methods: (1) $k \times k \times 3$ blocks of image data centred on each component centroid. (2) all pixels belonging to the region supplied by the patch-work of cells. Features can be defined as the measurements or attributes describing an object of interest. For this study, a collection of image features is used to generate a D-dimensional feature vector to discriminate between inflammatory and healthy cells. A full list of the features used is presented in Appendix A. Common sense would suggest that the greater the

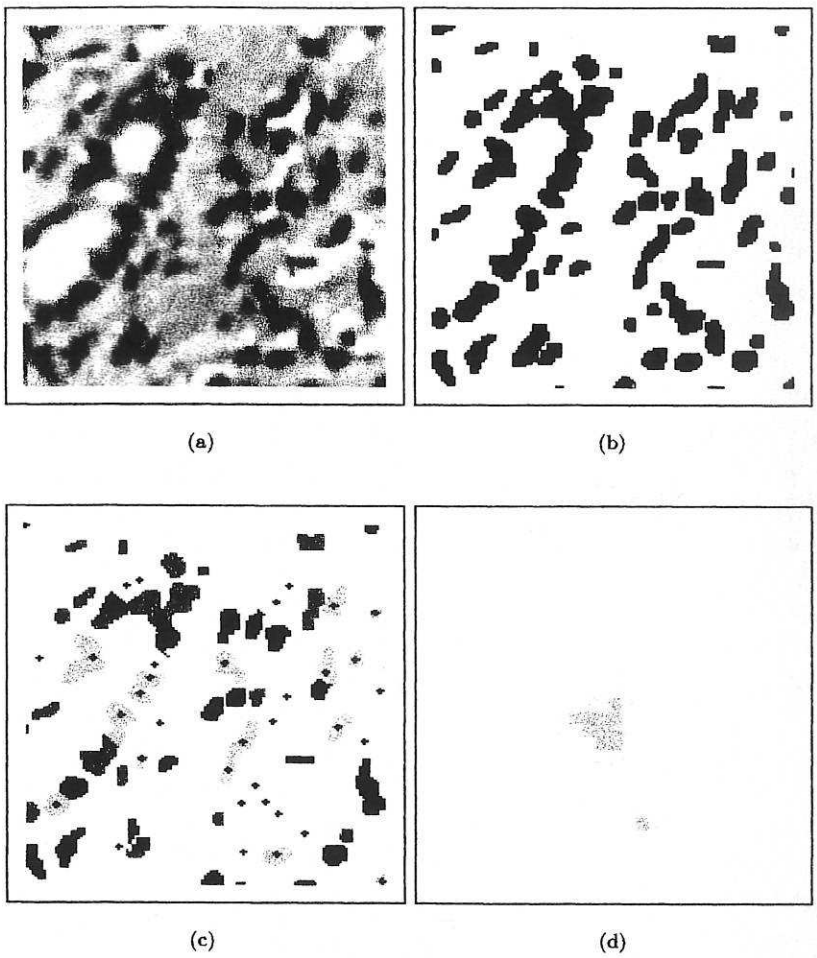


Fig. 9. Illustration of the clump decomposition technique. (a) Raw image. (b) Thresholded image after opening. (c) After clump decomposition. The markers show the new cell centroids and the patch-work effect illustrates the pixel allocation about those centroids. (d) A close-up of the erosion technique. The clump of cells is eroded until the component disappears. Any resulting break-up of the component is used to generate new or 'child' components. Using these new 'child' components, pixels from the original thresholded image are then re-allocated to the closest centroid.

number of features the easier it becomes to classify an object. However, in reality classification accuracy can decrease as a result of having too many features, owing to a phenomenon commonly called the 'curse of dimensionality' [35]. The term 'curse of dimensionality' was first coined by Bellman [2] in 1961 and describes a characteristic of high-dimensional data sets [3]. In brief, the hyper-volume of the feature space increases exponentially as a function of dimensionality. As most problems are dealing with a limited amount of data this rapidly leads to sparsely populated high-dimensional spaces which are difficult to characterise [12]. For this reason, dimensionality reduction is seen as a key step in any pattern recognition system. This system uses two main approaches to dimensionality reduction and these are discussed below.

3.3.3.1 Feature Selection In feature selection a subset of input features is selected for their suitability to a classification problem. This reduces dimensionality and the computational cost of feature gathering. The only guaranteed method of finding an optimal subset of d features from an original D-dimensional feature vector, is to perform an exhaustive search of all $\frac{D!}{d!(D-d)!}$ subsets of the reduced feature vector [19]. However, this is impractical because the number of subsets grows combinatorially. To demonstrate this effect, the analysis of approximately 1.4 million subsets would be required in order to generate an optimal 12-dimensional feature vector from an original 23-dimensional set. A number of suboptimal selection methods are available which are discussed in [19]. Of these, Jain et al [18] found that the sequential forward floating search (SFFS) [30] method produced the best results, performing close to the optimal, and demanding lower computational resources than other feature selection methods. The SFFS method is a bottom up search procedure, where the term *floating* identifies that the number of features dynamically changes, with one feature included and/or excluded, at each iteration. The SFFS method is used for feature selection in this system.

To summarise this method, $X_d = \{x_i | i = 1, 2, \ldots, d, x_i \in Y\}$ is a subset of d features taken from a set $Y = \{y_j | j = 1, 2, \ldots, D\}$ of D available features. $J(X_d)$ is the criterion function used to evaluate the effectiveness of X_d. For this study J is chosen to be the area under the receiver operating characteristic (ROC) curve, a commonly used test of classifier performance [35]. The method of ROC curves is discussed in more detail in section 4.1. The algorithm is initialised with an empty feature subset $X_0 = 0$. The most significant feature from Y ($\max J(Y)$) is then added to the subset X_0. This step is then repeated once more, taking the most significant feature from the remaining available features $Y - X_1$. The following steps are then performed:

(1) The most significant feature from $Y - X_d$ is added to the current subset, X_d.
(2) The least significant feature ($\min J(X_d - \{x_i\})$) is conditionally excluded from the current subset, X_d. If the newly added feature is the least significant or joint least significant with another feature, then step 1 is repeated. Otherwise, the least significant feature from the current subset, X_d, is excluded and step 3 is performed.
(3) This step is a continuation of the conditional exclusion in step 2. Once again the least significant feature, x_i, from X_d is located. If the resulting subset $X_d - \{x_i\}$ is better than the previous best subset of the same cardinality, then feature, x_i, is excluded from X_d and step 3 is repeated. Otherwise, the feature is retained and step 1 is repeated.

If the cardinality of X_d returns to 2 at either exclusion step (2 or 3), then the algorithm goes to step 1. The algorithm terminates when the required cardinality is achieved. Through experimentation, a final cardinality not exceeding 12 provides the best results here. Table 1 demonstrates the progression of the SFFS algorithm and presents the final reduce subset of features in bold-face.

Iteration	Feature subset
1	{4}
2	{4, 11}
3	{4, 11, 3}
4	{4, 11, 3, 16}
5	{4, 11, 3, 16, 1}
6	{4, 11, 3, 16, 1, 13}
7	{4, 11, 3, 16, 1, 13, 21}
8	{4, 11, 3, 16, 1, 13, 21, 22}
9	{4, 11, 3, 16, 1, 13, 21, 22, 6}
10	{4, 11, 3, 16, 1, 13, 21, 22, 6, 18}
11	{4, 11, 3, 16, 1, 13, 21, 22, 6, 18, 14}
12	{4, 11, 3, 16, 1, 13, 21, 22, 6, 18, 14, 7}
13	{4, 11, 3, 16, 1, 13, 21, 22, 6, 14, 7}
14	{4, 11, 3, 16, 1, 13, 21, 22, 6, 14, 7, 2}
15	{4, 11, 3, 16, 1, 13, 21, 22, 14, 7, 2}
16	{4, 11, 3, 16, 1, 13, 21, 14, 7, 2}
17	{4, 11, 3, 16, 1, 13, 14, 7, 2}
18	{4, 11, 3, 1, 13, 14, 7, 2}
19	{4, 11, 3, 1, 13, 14, 7, 2, 16}
20	{4, 11, 3, 1, 13, 14, 7, 2, 16, 21}
21	{4, 11, 3, 1, 13, 14, 7, 2, 16, 21, 22}
22	**{4,11,3,1,13,14,7,2,16,21,22,20}**

Table 1
The feature subsets considered by the SFFS method. The final subset is in bold-face, the previous subsets demonstrate the floating nature of the SFFS technique. Features corresponding to the numbers shown, are defined in Appendix A.

3.3.3.2 Feature Extraction Although, there are many feature extraction techniques available (see [19] for a review), this study implements principal component analysis (PCA) [10], one of the most widely used methods. The dimensionality of the d-dimensional ($d = 12$) feature vector derived from feature selection is further reduced using PCA. PCA or the Karhunen-Loève transform is an unsupervised linear transformation technique, which seeks to project the high-dimensional input data into lower dimensional space [3]. In simple terms this means that new features are created from a transformation of the input features. The feature vector, $\mathbf{x} = (x_1, \ldots, x_d)^{\mathbf{T}}$, is first normalised for all N data points using

$$\mathbf{y}^n = \Phi^{-1}(\mathbf{x}^n - \overline{\mathbf{x}}), \quad (n = 1, \ldots, N) \tag{4}$$

where

$$\overline{\mathbf{x}} = (\overline{x}_1, \ldots, \overline{x}_d) \tag{5}$$

$$\overline{x}_i = \frac{1}{N}\sum_{n=1}^{N} x_i^n \tag{6}$$

$$\Phi = \mathrm{diag}(\sigma_1, \ldots, \sigma_d) \tag{7}$$

$$\sigma_i^2 = \frac{1}{N-1}\sum_{n=1}^{N}(x_i^n - \overline{x}_i)^2 \tag{8}$$

This normalisation is intended to counter the intolerance of PCA to data with different orders of magnitude(p.298 [3]). For PCA, the mean vector, $\overline{\mathbf{y}}$, and covariance matrix, Σ, are then computed for the normalised feature vector, $\mathbf{y}^n$.

$$\overline{\mathbf{y}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{y}^n \tag{9}$$

$$\Sigma = \frac{1}{N-1}\sum_{n=1}^{N}(\mathbf{y}^n - \overline{\mathbf{y}})(\mathbf{y}^n - \overline{\mathbf{y}})^{\mathrm{T}} \tag{10}$$

The eigen-decomposition of the covariance matrix

$$\Sigma\mathbf{u}_j = \lambda_j\mathbf{u}_j \tag{11}$$

is then calculated and sorted according to decreasing eigenvalue. Because Σ is a covariance matrix its eigenvalues are real and non-negative [3]. In most cases a small number of eigenvalues will dominate, indicating the inherent dimensionality of the data [10]. By forming a matrix, $\mathbf{U}$, whose columns are the $pc < d$ eigenvectors corresponding to the pc largest eigenvalues

$$\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_{pc}), \tag{12}$$

it is possible to define

$$\mathbf{z} = \mathbf{U}^{\mathrm{T}}(\mathbf{y} - \overline{\mathbf{y}}) \tag{13}$$

a pc-dimensional vector of linearly transformed variables. The optimum number of principal components, pc, will be discussed in section 4. Although, in principle, PCA should provide optimal dimensionality reduction without feature selection. The prohibitive cost of generating large numbers of features makes the inclusion of feature selection desirable for this study.

3.3.4 Probability Density Estimation

The conclusion of training process is to derive the class-conditional probability densities, $p(\mathbf{z}|C_j)$. The density estimate can then be used for the Bayesian classification discussed in section 3.3.5.1. The binary masks (see figure 6) can be overlayed onto the output from the clump decomposition process to provide the class (C_j, $j = 1, 2$) labels for each transformed feature vector $\mathbf{z}_j$. As a result, we can approximate the required probability distribution for each class. In this study two methods of density estimation are compared:

(1) Gaussian parametric model (GPM)—This is a parametric method where a fixed gaussian functional form of density estimation is assumed. This technique is easy to compute and simple to implement. For the multivariate case, the density estimation takes the form

$$p(\mathbf{z}|C_j) = \frac{1}{(2\pi)^{d/2}|\Sigma_j|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{z} - \overline{\mathbf{z}}_j)^{\mathrm{T}}\Sigma_j^{-1}(\mathbf{z} - \overline{\mathbf{z}}_j)\right\} \tag{14}$$

where the class covariance matrix Σ_j and mean vector $\overline{\mathbf{z}}_j$ are derived from the transformed feature vectors of the training set.

(2) Gaussian mixture model (GMM)—This is a semi-parametric [3] method where mixtures of gaussians are used to build more complex density models e.g. multimodal [10]. For the multivariate case, the probability density function for each class is estimated by a linear combination of K_j ($j = 1, 2$) gaussian basis functions of the form

$$p(\mathbf{z}|C_j) = \sum_{k=1}^{K_j} P_{jk} p_j(\mathbf{z}|k), \quad (\mathbf{z} \in C_j) \tag{15}$$

where

$$p_j(\mathbf{z}|k) = \frac{1}{(2\pi)^{d/2}|\Sigma_{jk}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{z} - \overline{\mathbf{z}}_{jk})^{\mathrm{T}}\Sigma_{jk}^{-1}(\mathbf{z} - \overline{\mathbf{z}}_{jk})\right\} \tag{16}$$

where Σ_{jk} is the covariance matrix for the k^{th} gaussian for the j^{th} class and $\overline{\mathbf{z}}_{jk}$ is the mean vector for the k^{th} gaussian for the j^{th} class. Typically, the GMM parameters are determined using the *expectation-maximisation* (EM) algorithm [26].

The performance of each estimator is evaluated during the optimisation process outlined in section 4.

3.3.5 Evaluating new images

When a new image is presented to the system, it is first necessary to perform the following, previously discussed, steps:

- Colour correct the new image (using the original reference image).
- Threshold the image.
- Apply clump decomposition.
- Extract the feature vector $\mathbf{x}$ (the reduced subset of features) from each POI.
- Apply PCA to produce the projected feature vector $\mathbf{z}$.

The projected feature vector $\mathbf{z}$ can then be applied to one of the density estimation techniques (GPM or GMM) detailed in section 3.3.4, to give the likelihood of the cell at each POI belonging to a particular class. Bayes theorem (17) can then be used to calculate the posterior probability of class membership, which allows a decision to be made regarding class membership of the individual nuclei. This method is discussed in more detail below.

3.3.5.1 Classification Bayesian decision theory is the fundamental approach to the classification problem [3]. Considering this approach for supervised classification, Bayes theorem (17) permits the posterior probability, $P(C_j|\mathbf{z})$, to be expressed in terms of the prior probability, $P(C_j)$, the likelihood, $p(\mathbf{z}|C_j)$, and a normalisation factor, $p(\mathbf{z})$ [10].

$$P(C_j|\mathbf{z}) = \frac{p(\mathbf{z}|C_j)P(C_j)}{p(\mathbf{z})} \tag{17}$$

$P(C_j)$ is the probability of each class occurring based on *a priori* knowledge of the training set. $p(\mathbf{z}|C_j)$ is the class-conditional probability density function. In practice, an estimate of the probability density function for each class is required, as discussed in section 3.3.4. By assuming that new nuclei belong to one of the two classes C_1–inflammatory or C_2–healthy, then the posterior probabilities obey

$$P(C_1|\mathbf{z}) = 1 - P(C_2|\mathbf{z}) \tag{18}$$

Each nucleus may then be assigned class membership according to a user defined classification threshold T, as follows

$$\begin{aligned} &P(C_1|\mathbf{z}) > T, \text{ then assign to } C_1 \\ &P(C_1|\mathbf{z}) < T, \text{ then assign to } C_2 \end{aligned} \tag{19}$$

where $0 < T < 1$. The method of selecting a suitable classification threshold is discussed in section 4.1.

4 Optimisation

The role of optimisation in this study is to select a good set of the adjustable system parameters. To determine the optimum system performance it is necessary to evaluate images with pre-classified cells. As the training images discussed in section 3.2 already provide pre-classified cells, the system is optimised by the m-fold cross validation ($m = 10$) of these training images. This simply means the training images are randomly divided into m equally sized subsets [10]. The system then evaluates one subset, with the remainder used for training. This operation is then repeated until all subsets have been evaluated. It can be shown [10] that applying m performance measures gives an estimate of the true system performance. A method of quantifying system performance from the results of m-fold cross validation is discussed in section 4.1. The optimisation method and the final optimised system parameters are presented in section 4.2.

4.1 The Receiver Operating Characteristic

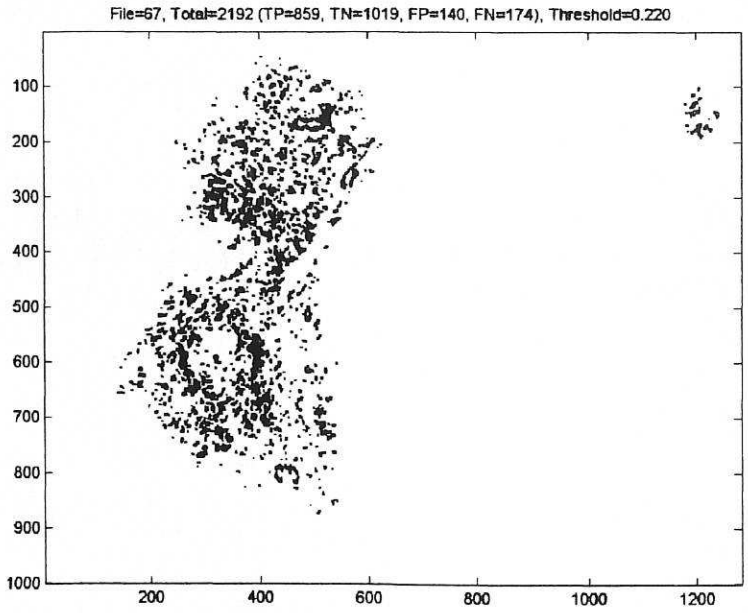
System performance may be evaluated using contingency table data derived from the m-fold cross validation of the training set. The contingency table is defined in table 2 and an example of the results obtained from cross validation of the training set is shown in figure 10.

		Test results	
		Class 1	Class 2
Observer Results	Class 1	True Positive (TP)	False Negative (FN)
	Class 2	False Positive (FP)	True Negative (TN)

Table 2
Contingency table definition. The observer results are derived from the annotated test images and the test results are derived from cell classification at a particular classification threshold.

For a two class problem it is possible to evaluate the system more robustly by plotting the receiver operating characteristic (ROC) curve, a technique commonly used in medical imaging [13]. The ROC curve is constructed from contingency table data by plotting *sensitivity* $\left(\frac{TP}{TP+FN}\right)$ against one minus *specificity* $\left(\frac{FP}{TN+FP}\right)$ as the classifier threshold varies from 0 to 1. The area under the ROC curve (AUROC) can be considered to be a measure of the overall quality of the classification model [35].

(a) Raw image



(b) Experimental result

Fig. 10. Experimental result for training image RAW00067 generated from 10-fold cross validation of the training set. Inflamed cells are shown in black, healthy cells are grey and the annotated regions provided by SSC are in blue. This image shows good correlation between automatic and observed cell classification.

Maximising the area by changing important system variables should lead to the optimum cell classification. As each point lying on the curve corresponds to a different threshold, the final classification threshold may be chosen based on desired levels of sensitivity and/or specificity. Using the Neyman-Pearson criterion (NPC) for this purpose [35], a maximum false positive rate (one minus specificity) is specified by the user, as shown in figure 12. A final classification threshold is then selected with the highest false positive rate, which is less than the NPC. For the purpose of illustration in this study, the maximum permitted false-positive rate is set at 0.1.

4.2 Response Surface Methodology

Response surface methodology (RSM) is a technique to reduce the cost of optimisation by searching for combinations of variables (factors) that maximise the performance of the system [25]. In this study RSM is used to maximise the AUROC by searching for the optimum value of key system factors. The first stage of this technique is to develop a strategy for gathering experimental data, known as the 'design of experiments' (DoE) [6]. This involves identifying factors which have a significant effect on the response of the system, a procedure which is normally carried out by screening out insignificant factors during the development process. Once identified, the factors are constrained to an allowable range by identifying suitable upper and lower limits for each factor. The range is then discretised at equal spacing to generate *levels* within the allowable range. A common approach, when considering a small number of input factors (less than five), is to evaluate the system at all combinations of factors and corresponding levels. This approach is known as a *full factorial* design [6]. The next stage of RSM is to develop a model of the system response. This model can then be searched to find the maximum predicted response and thus the optimal factor values. For an example with two input factors, the *full factorial* design provides a 2-dimensional grid of system evaluation points. The model can then be visualised as a response surface constructed on the grid. Considering a system more formally

$$y = f(\mathbf{v}) \tag{20}$$

where y is the system response, f is an unknown function and $\mathbf{v} = (v_1, v_2, \ldots, v_q)$ is a vector of q independent factors. It is common to construct a model of this system by fitting a low-order (either linear, quadratic or cubic) polynomial to the experimental data. For a cubic polynomial with p combinations of factors and levels, this takes the form

$$y^n = b_0 + \sum_{i=1}^{q} b_i v_i^n + \ldots \sum_{i=1}^{q}\sum_{j=1}^{q} b_{ij} v_i^n v_j^n + \ldots \sum_{i=1}^{q}\sum_{j=1}^{q}\sum_{k=1}^{q} b_{ijk} v_i^n v_j^n v_k^n \tag{21}$$

where b_0, b_i, b_{ij} and b_{ijk} are the unknown polynomial coefficients. y^n is the experimental observed response value and $n = (1, \ldots, p)$. This model can be written in matrix notation as

$$\mathbf{y} = \mathbf{V}\mathbf{b} \tag{22}$$

where

$$\begin{aligned} \mathbf{y} &= (y^1, y^2, \ldots, y^p), \\ \mathbf{b} &= (b_0, b_1, \ldots, b_q, b_{11}, \ldots, b_{qq}, b_{111}, \ldots, b_{qqq}) \end{aligned} \tag{23}$$

and $\mathbf{V}$ is the experimental design matrix constructed from p rows of $\hat{\mathbf{v}}^n$, a vector corresponding to the factor terms in 21, of the form

$$\hat{\mathbf{v}}^n = (1, v_1^n, \ldots, v_q^n, v_1^n v_1^n, \ldots, v_q^n v_q^n, v_1^n v_1^n v_1^n, \ldots, v_q^n v_q^n v_q^n) \tag{24}$$

The coefficients $\mathbf{b}$ can then be estimated using the least squares method

$$\mathbf{b} = \mathbf{V}^\dagger \mathbf{y} \tag{25}$$

where $\mathbf{V}^\dagger$ is the *pseudo-inverse* of $\mathbf{V}$ [3]. The value of the input variables that provide the maximum system response can then be derived from the polynomial given by $\mathbf{V}\mathbf{b}$. Considering the DoE for this study, the following system factors have a significant effect on the system response.

(1) Block size (k)—Features 1–6 defined in Appendix A rely on square blocks of data extracted from around each cell centroid. This factor represents the block size and is constrained between 1 and 81 pixels. The discretised levels of this allowable range are $\{21, 41, 61, 81\}$.
(2) Number of Principal components (pc)—Principal component analysis is used for dimensionality reduction and discussed in section 3.3.3.2. This factor governs the number of dimensions that the reduced feature vector is mapped to and is constrained between 1 and 12 (levels = $\{3, 6, 9, 12\}$).
(3) Density estimation method—Either GPM or GMM, as discussed in section 3.3.4. For GMM only, the following extra factor requires optimisation:
    (a) Number of basis functions (bf)—The number of gaussian basis functions to fit the data. This is constrained to be between 1 and 8 (levels = $\{2, 4, 6, 8\}$).

The final cardinality of the feature selection method, discussed in section 3.3.3.1, should also be treated as a factor. However, this is impractical because of the high computational cost of combining the SFFS technique with the full factorial design. To

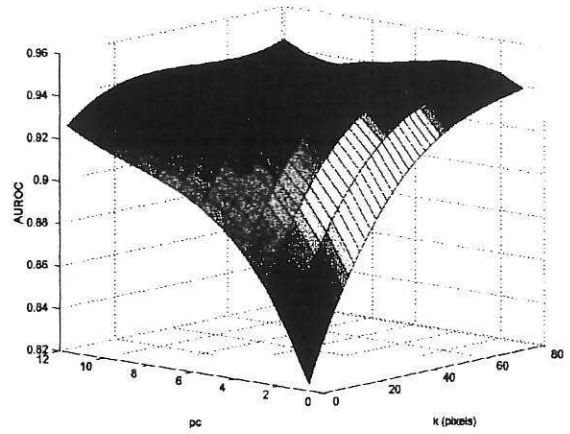


Fig. 11. The theoretical response surface derived by RSM for GPM estimation. This surface is generated by least squares fitting a cubic polynomial to experimental data.

compare the GPM and GMM density estimation (DE) methods discussed in section 3.3.4, two response surfaces are generated. This allows the system to be independently optimised for each DE method. The optimised systems are then compared for the maximum system response by re-evaluating the AUROC. A cubic polynomial is used to model the response in both cases. With this in mind, the two forms of DE will now be considered.

- GPM—only variables k and pc are applicable. Evaluating these variables using a full factorial design requires 16 evaluations of the AUROC. The response surface generated from this data is shown in figure 11.
- GMM—all the above variables (k, pc, bf) are applicable. Evaluating this set using a full factorial design requires 64 evaluations of the AUROC.

As all the factors under discussion can take only integer values within the constraints discussed previously, the maximum predicted response may be derived by evaluating the model at all variable combinations between the upper and lower bounds for each variable. Although this is a combinatorial problem, the task of evaluating the model is computationally trivial in comparison to evaluating the AUROC. Searching each model for the maximum response using this method provides the factor values listed in table 3. Evaluating the AUROC at these parameters shows that GMM provides the optimum DE method. However the improvement gained by using the GMM method rather than the computationally more efficient GPM method is only marginal (0.65%). As the intended end-users of this system are pathologists, it is thought that adopting the conceptually simpler GPM method will aid the understanding and trust of this system by medical professionals who may not be familiar with pattern recognition theory. Therefore the GPM method is applied in this study. The ROC curve for the optimum GPM configuration is illustrated in figure 12. By applying the NPC technique discussed in section 4.1, the final classification threshold of 0.22 can be derived from the curve.

	Variable	Value (GPM)	Value (GMM)
1	Block size (k)	55	53
2	Number of principal components (pc)	5	6
3	Number of gaussian basis functions (bf)	N/A	5
	AUROC	0.9559	0.9619

Table 3
Table showing the optimised system factors for both GPM and GMM density estimation methods and the corresponding AUROC.

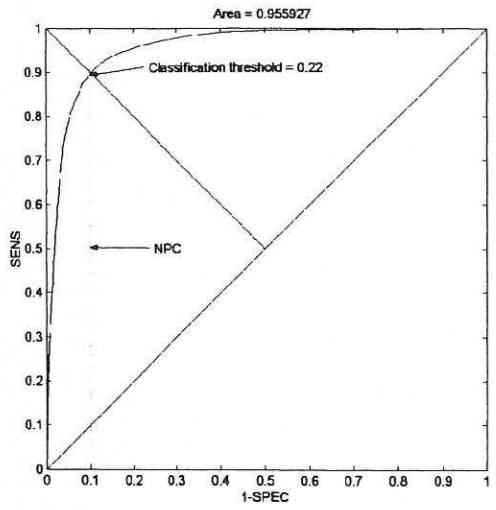


Fig. 12. ROC curve for the optimised system. For the purpose of illustration the maximum false-positive rate is set at 0.1.

5 Testing

It is common to test a system of this type against a 'gold standard', a set of universally accepted test images where the amount of inflammation is accurately quantified. However, no 'gold standard' exists for liver biopsy inflammation. The closest alternative is the Ishak scoring system [17] discussed in section 1, but as previously shown, this suffers from high inter- and intra-observer variability. With this in mind, our system is tested using a separate group of 15 test images previously evaluated in a study by Cross at al [9]. In this study, 25 observers (including 5 consultant pathologists, 4 trainee pathologists and 16 control observers) were asked to compare 15 liver biopsy images with varying degrees of inflammation. It can be assumed that consultant pathologists have the most experience in identifying cell inflammation, followed by the trainee pathologists and finally the 16 control observers. The images are named '*mild1*, ..., *mild5*, *mod1*, ..., *mod5*, *sev1*, ..., *sev5*'. Each of the 15 images is compared to each other image producing 105 pairs. The observers are then asked to identify the image containing the most inflammation from each pair. A rank order of images is then produced for each observer using a ranking algorithm normally

used to rank competitive chess players [8]. Finally, Spearman rank correlation (SRC) [41] is used to assess, statistically, the level of agreement between the observers. As the results show good inter-observer agreement using this comparison technique, it is proposed that image ranks from each observer can be used to test the automated system.

Initially, the optimised system is trained on all 86 training images. Each test image is then processed using the methods outlined in section 3.3.5 and using the classification threshold determined by NPC (see figure 12). The test images are the same size and magnification as the training images. Quantification of the inflammation is carried out by counting the number of cells classified either C_1–inflammatory or C_2–healthy. The percentage of inflammatory cells is then computed. The results from processing all 15 test images are shown in Appendix B. Quantifying the inflammation in percentage form, allows the images to be placed in rank order of severity and compared to the image ranks from the previous study [9]. The results are shown in Appendix C, tables 1 and 2. As in [9], SRC can then be used to assess the level of agreement between the image rank produced by this automated system and the image ranks produced by the 25 observers. SRC is defined by

$$r_s = 1 - \frac{6\sum_{n=1}^{N} d_n^2}{N(N^2-1)} \tag{26}$$

where d_n is the difference between each pair of ranks and N is the number of paired observations. The resulting values of r_s are presented in Appendix C, tables 3 and 4. They show good agreement between the observers and the automated system. Using a null hypothesis that there is no correlation between any of the ranks presented in Appendix C, tables 1 and 2, the *significance* of each r_s value can be determined by calculating the probability (P-value) that this hypothesis is true [41]. For $N > 10$ ($N = 15$ in this case), r_s has a Normal distribution with a mean of zero and a variance of $\frac{1}{(n-1)}$ [4]. To test the significance of r_s the z value is first calculated as follows [41]

$$z = \frac{r_s}{\sqrt{\frac{1}{(n-1)}}} = r_s\sqrt{(n-1)} \tag{27}$$

The P-value is then determined from z, using tables of the area under the Normal distribution curve [4]. Focusing on the relationship between consultant pathologists and this automated system, table 4 shows a sample of the r_s values given in Appendix C, and the corresponding probability that the null hypothesis is true in each case. P-values were also calculated for all other observers (not shown). Table 4 shows that $P < 10^{-3}$ is the maximum probability that the null hypothesis is true when considering the correlation between consultants and this system. This is also the maximum when considering the correlation between our system and the trainee pathologists. However the P-value rises to $P < 10^{-2}$, when considering the correlation between control observers and our system. Historically $P < 10^{-2}$ or a 'one percent probability level'

suggests the null hypothesis may be rejected [20]. Although, this means that the null hypothesis may be rejected for all observer groups, the low P-values seen in table 4, suggest a strong correlation with consultants. It can also be shown that consultants have the lowest inter-observer variability with each other. This means our system can rank 15 test images in correlation to five consultant pathologists, who in turn strongly agree with each other, indicating this system has an expert capability in this test.

For completeness the above tests were also conducted using the GMM density estimation method discussed in sections 3.3.4 and 4. Although the GMM method produces a better system performance (greater AUROC) when considering 10-fold cross validation of the training set, results here show a similar performance to the GPM method when considering the correlation between our system and the three observer groups. This confirms the decision, made in section 4.2, to use GPM density estimation.

6 Conclusions

An effective and systematic method of evaluating the liver biopsies of patients with hepatitis C will become increasingly important owing to the large number of people currently infected with the disease. Previous approaches to similar cell classification problems do not adequately address the specific issues associated with the automatic segmentation and classification of inflammatory cells in HCV infected liver biopsies. The system outlined in this study, offers a fully automatic pattern recognition solution to quantify inflammatory cells. The simplicity of the pattern recognition techniques used aids the understanding and trust of this system by pathologists and facilitate the implementation of this system on a standard desktop PC in a pathology laboratory. Important steps forward have been made in: (1) colour correcting images for

	Consultant 1	Consultant 2	Consultant 3	Consultant 4	Consultant 5	Computer
Consultant 1	1.000	0.946 ($P < 10^{-3}$)	0.936 ($P < 10^{-3}$)	0.936 ($P < 10^{-3}$)	0.971 ($P < 10^{-3}$)	0.943 ($P < 10^{-3}$)
Consultant 2	0.946 ($P < 10^{-3}$)	1.000	0.950 ($P < 10^{-3}$)	0.968 ($P < 10^{-3}$)	0.971 ($P < 10^{-3}$)	0.989 ($P < 10^{-3}$)
Consultant 3	0.936 ($P < 10^{-3}$)	0.950 ($P < 10^{-3}$)	1.000	0.911 ($P < 10^{-3}$)	0.943 ($P < 10^{-3}$)	0.971 ($P < 10^{-3}$)
Consultant 4	0.936 ($P < 10^{-3}$)	0.968 ($P < 10^{-3}$)	0.911 ($P < 10^{-3}$)	1.000	0.975 ($P < 10^{-3}$)	0.964 ($P < 10^{-3}$)
Consultant 5	0.971 ($P < 10^{-3}$)	0.971 ($P < 10^{-3}$)	0.943 ($P < 10^{-3}$)	0.975 ($P < 10^{-3}$)	1.000	0.971 ($P < 10^{-3}$)
Computer	0.943 ($P < 10^{-3}$)	0.989 ($P < 10^{-3}$)	0.971 ($P < 10^{-3}$)	0.964 ($P < 10^{-3}$)	0.971 ($P < 10^{-3}$)	1.000

Table 4
SRC between the automated system (computer) and five consultant pathologists. The *significance* (P-value) of each result is shown parentheses.

stain and illumination variability, (2) the segmentation of individual cells via clump decomposition and (3) the application of feature selection and extraction methods to reduce the cost of feature gathering and the dimensionality of the feature vector.

The comparison of two commonly used density estimation methods, GPM and GMM, shows that the simpler GPM technique provides an equivalent system performance when considering novel images. The implementation of the GPM method shows the system can rank a set of 15 previously unseen test images in correlation to five consultant pathologists and four trainee pathologists with a level of significance of $P < 10^{-3}$. Although the level of significance is reduced when considering the relationship between this system and the control observers ($P < 10^{-2}$), the consultants have the lowest inter-observer variability and have the most experience of interpreting biopsy images. Therefore, the correlation between consultants can be considered a 'gold standard' for this test. Although results show an expert capability of the system, equivalent to consultants, improvements can be made to its robustness. First, a major weakness is that the training images are annotated by a single pathologist (SSC). This means the system is trained by single expert who may or may not suffer from the inter- and intra-observer variability found by [11]. This will be corrected in future studies by using training images annotated by a number of experts. Second, as the effectiveness of the system is reduced by poor quality images, it is planned to screen unsuitable images presented for evaluation. This means that *outliers*, images with colour characteristics which do not match the majority of the training images, will be referred back to the user. Finally, it is also intended that future work in this area will use biopsy images taken at a higher magnification (40× objective). This will allow a greater number and more appropriate features to be extracted by combining the low magnification techniques discussed in this study and a high magnification examination of each cell. In particular it is hoped to gather textural information from each cell nuclei.

Acknowledgements

† The author would like to thank the UK EPSRC for their financial support of this work through a doctoral training award.

References

[1] ADAM. *Illustration of liver biopsy procedure.* World Wide Web, http://www.adam.com/, 2002.

[2] R.E. Bellman. *Adaptive Control Processes.* Princeton University Press, Princeton,NJ, 1961.

[3] C.M. Bishop. *Neural Networks for Pattern Recognition.* Oxford University Press, Oxford, 1995.

[4] M. Bland. *An Introduction to Medical Statistics.* Oxford Medical Publications, Oxford, 2nd edition, 1995.

[5] J.C.L. Booth, J. O'Grady, and J Neuberger. Clinical guidelines on the management of hepatitis c. *Gut*, 49(Supplement 1):i1–21, 2001.

[6] G.E.P. Box and N.R. Draper. *Empirical Model-Building and Response Surfaces.* Wiley, New York, 1987.

[7] V.C. Cardei, B. Funt, and M. Brockington. Issues in color correcting digital images of unknown origin. In *CSCS 12.* Bucharest, 1996.

[8] Internet Chess Club. *ICC Help File: RATINGS.* World Wide Web, http://www.chessclub.com/help/ratings, 2002.

[9] S.S. Cross, N. Bashir, P. Hempshall, S. Hodgson, and R.F. Harrison. Estimating inflammation in liver biopsies in chronic hepatitis - reproducibility is far higher comparing paired rather than single images. *Journal of Clinical Pathology*, 2002. In Press.

[10] R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern Classification.* John Wiley and sons, New York, 2nd edition, 2001.

[11] R.D. Goldin, J.G. Goldin, and A.D. Burt et al. Intra-observer and inter-observer variation in the histopathological assessment of chronic viral hepatitis. *Journal of Hepatology*, 25(5):649–654, 1996.

[12] J.S. Hallinan. *Detection of Malignancy Associated Changes in Cervical Cells Using Statistical and Evolutionary Computational Techniques.* PhD thesis, Cytometrics Project, Centre for Sensor Signal and Information Processing, University of Queensland, Australia, 1999.

[13] J.A. Hanley and B.J. McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143:29–36, 1982.

[14] R.M. Haralick, K. Shanmugam, and I. Dinstein. Textural features for image classification. *IEEE Transaction on System, Man and Cybernetic*, 3:610–621, 1973.

[15] R.M. Haralick and L.G. Shapiro. *Computer and Robot Vision*, volume I. Addison-Wesley, Reading,MA, 1992.

[16] P. Humphries. An explosive situation. *The Guardian*, March 6 2002.

[17] K. Ishak, A. Baptista, and L. Bianchi et al. Histological grading and staging of chronic hepatitis. *Journal of Hepatology*, 22:696–699, 1995.

[18] A. Jain and D. Zongker. Feature selection: Evaluation, application, and small sample performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19:153–158, 1997.

[19] A.K. Jain, R.P.W. Duin, and J. Mao. Statistical pattern recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1):4–37, 2000.

[20] M. Kendall and J.D. Gibbons. *Rank correlation methods.* Edward Arnold, London, 1990.

[21] G. Lake-Bakaar, V. Mazzoccoli, and L. Ruffini. Digital image analysis of the distribution of proliferating cell nuclear antigen in hepatitis c virus-related chronic hepatitis, cirrhosis, and hepatocellular carcinoma. *Digestive Diseases and Sciences*, 47(7):1644–1648, 2002.

[22] L. Liu and S. Sclaroff. Deformable shape detection and description via model-based grouping. Technical Report 98-017, Computer Science, Boston University, November 1998.

[23] M. Masseroli, T. Caballero, and F. O'Valle et al. Automatic quantification of liver fibrosis: design and validation of a new image analysis method: comparison with semi-quantitative indexes of fibrosis. *Journal of Hepatology*, 32:453–464, 2000.

[24] MathWorks. *Matlab: Image Processing Toolbox User's Guide.* The MathWorks, Inc., MA, 2002.

[25] R.H. Myers and D.C. Montgomery. *Response surface methodology: process and product optimization using designed experiments.* Wiley-Interscience, New York, 2002.

[26] I.T. Nabney. *NETLAB: Algorithms for Pattern Recognition.* Springer-Verlag, 2002.

[27] National Institutes of Health (NIH). *Public domain image processing and analysis program.* World Wide Web, http://rsb.info.nih.gov/nih-image/, 2002.

[28] Chief Medical Officer. *Annual Report of the Chief Medical Officer 2001.* Department of Health, London, 2001.

[29] E. Orfei. *Review of pathology of the Liver.* Department of Pathology, Stritch School of Medicine, Loyola University of Chicago, http://www.meddean.luc.edu/, 2003.

[30] P. Pudil, J. Novovicova, and J. Kittler. Floating search methods in feature selection. *Pattern Recognition Letters*, 15:1119–1125, 1994.

[31] J.C. Russ. *The image processing handbook.* CRC Press, London, 1992.

[32] A. Santos, C. Ramiro, and M. Desco et al. Automatic detection of cellular necrosis in epithelial cell cultures. *Proceedings of SPIE*, 4322:1836–1844, 2001.

[33] R.J. Schalkoff. *Digital image processing and computer vision.* John Wiley & Sons, Inc., 1989.

[34] P.J. Scheuer and J.H. Lefkowitch. *Liver Biopsy Interpretation.* W.B. Saunders Company Ltd, London, 5th edition, 1994.

[35] M.J.J. Scott, M. Niranjan, and R.W. Prager. Parcel: feature subset selection in variable cost domains. Technical Report CUED/F-INFENG/TR. 323, Cambridge University Engineering Department, May 1998.

[36] T.T.E. Teo, X.C. Jin, S.H. Ong, Jayasooriah, and R. Sinniah. Clump splitting through concavity analysis. *Pattern Recognition Letters*, 15:1013–1018, 1994.

[37] WHO. Hepatitis c global prevalence (update). *Weekly Epidemiological Record (World Health Organisation)*, 74:421–428, 1999.

[38] WHO. *Hepatitis C (Fact Sheet No. 164).* World Health Organisation, Geneva, 2000.

[39] C. Xu and J.L Prince. Gradient vector flow: A new external force for sankes. In *Conference on Computer Vision and Pattern Recognition*, pages 66–71. IEEE, 1997.

[40] X.Y. Zhang, C.K. Sun, and A.M. Wheatley. A novel approach to the quantification of hepatic stellate cells in intravital fluorescence microscopy of the liver using a computerized image analysis system. *Microvascular Research*, 60:232–240, 2000.

[41] F.H. Zuwaylif. *Applied General Statistics.* Mass:Addison-Wesley Pub. Co, Reading, 3rd edition, 1979.

Appendix A: Features

Notation

S = A region of RGB data defined by a block of $k \times k \times 3$ pixels centred on each position of interest (POI).

L = A region of RGB data defined by all pixels belonging to each component, segmented by clump decomposition (see section 3.3.2).

$i \in \{R, G, B\}$ = each layer of RGB data.

$\text{cell\_density}(O, r)$ = The number of cells contained with in a circle of radius r pixels, centred at the centroid of O.

$|O|$ = number of pixels in O.

$O(x)$ = The value of O at x.

$\text{sort}(O)$ = A sorted version of O

$CH(O)$ = A binary image representing the convex hull of O. The convex hull is the smallest convex polygon that fully surrounds the region [24].

$\text{greyscale}(O)$ = A grey scale representation of O.

$ELL(O)$ = An ellipse with the same second-moments as region O.

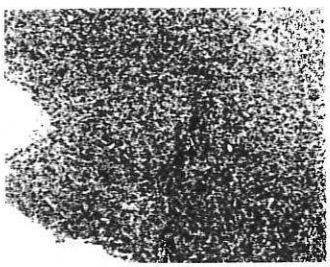
$\text{major\_axis}(E)$ = The length in pixels of the major axis of ellipse E.

$\text{minor\_axis}(E)$ = The length in pixels of the minor axis of ellipse E.

Feature Definitions

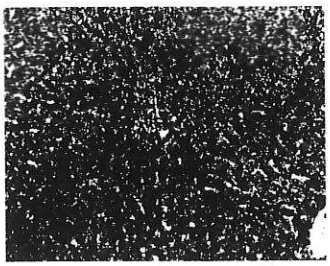
Feature No.	Definition	Description
1–3	$\mu_i^{block} = \frac{1}{M}\sum_{m=1}^{M} S_i^m$, where $M = k \times k$	Block mean
4–6	$\sigma_i^{block} = \frac{1}{M-1}\sum_{m=1}^{M}(S_i^m - \mu_i^{block})^2$	Block standard deviation
7	$A = \frac{\lvert L\rvert}{3}$	Cell area
8	$ECC = \frac{\text{major\_axis}(ELL(L))}{\text{minor\_axis}(ELL(L))}$	Cell eccentricity
9	$ED = \sqrt{\frac{4A}{\pi}}$	Equivalent circle diameter
10	$SOL = \frac{A}{CH(L)}$	Solidity
11	$CD = \text{cell\_density}(L, 120)$	Cell density
12–14	$\mu_i^{cell} = \frac{1}{N}\sum_{n=1}^{N} L_i^n$, where $N = A$	Cell mean
15–17	$\theta_i^{cell} = \text{sort}(L_i)(h)$, where $h = (N+1)/2$ for odd N and $h = \frac{(N/2)+((N+1)/2)}{2}$ for even N	Cell median
18–20	$\sigma_i^{cell} = \frac{1}{N-1}\sum_{n=1}^{N}(L_i^n - \mu_i^{cell})^2$	Cell standard deviation
21	$\mu_i^{grey} = \frac{1}{N}\sum_{n=1}^{N}\text{greyscale}(L)^n$	grey scale mean
22	$\theta_i^{grey} = \text{sort}(\text{greyscale}(L))(h)$, where $h = (N+1)/2$ for odd N and $h = \frac{(N/2)+((N+1)/2)}{2}$ for even N	Grey scale median
23	$\sigma_i^{grey} = \frac{1}{N-1}\sum_{n=1}^{N}(\text{greyscale}(L)^n - \mu^{grey})^2$	Grey scale standard deviation

Appendix B: Test Results

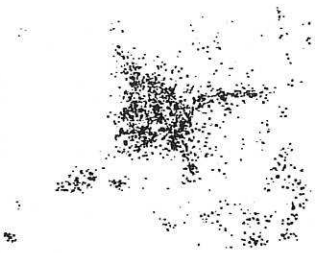


(a) Raw image ('mod 1')

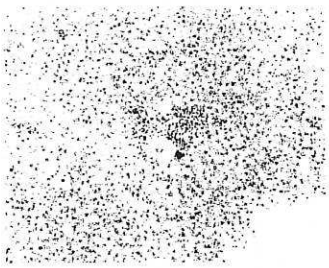
(b) Test result ('mod 1'). Showing 22.3% inflammation.



(c) Raw image ('mod 2')



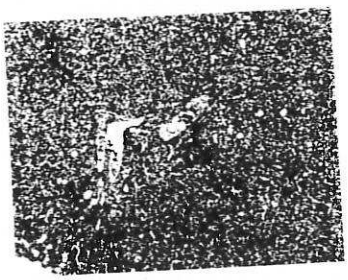
(d) Test result ('mod 2'). Showing 45.9% inflammation.



(e) Raw image ('mod 3')

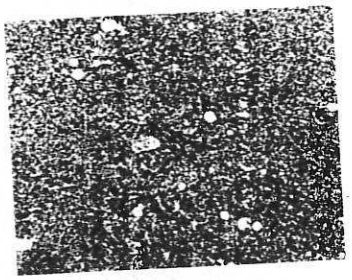
(f) Test result ('mod 3'). Showing 19.3% inflammation.

Fig. .1. Experimental test results. Inflamed cells are shown in black and healthy cells are grey.



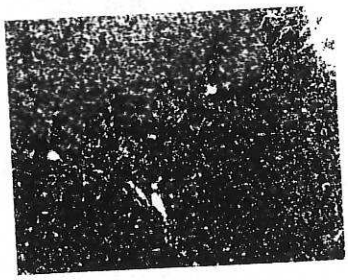
(a) Raw image ('mod 4')

(b) Test result ('mod 4'). Showing 30.8% inflammation.

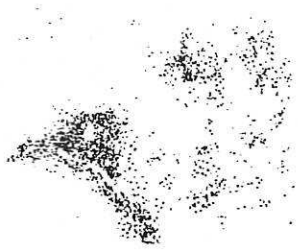


(c) Raw image ('mod 5')

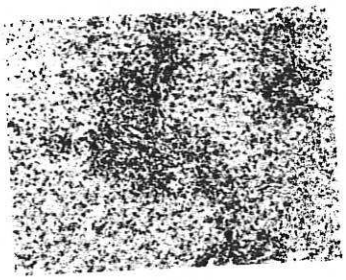
(d) Test result ('mod 5'). Showing 14.6% inflammation.



(e) Raw image ('mild 1')



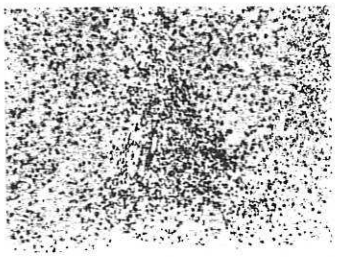
(f) Test result ('mild 1'). Showing 49.4% inflammation.



(g) Raw image ('mild 2')

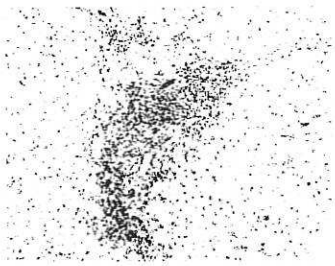
(h) Test result ('mild 2'). Showing 62.9% inflammation.

Fig. .2. Experimental test results. Inflamed cells are shown in black and healthy cells are grey.



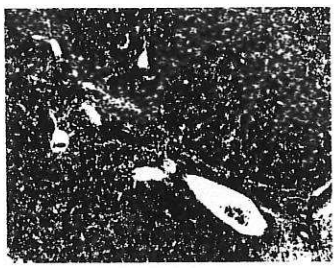
(a) Raw image ('mild 3')

(b) Test result ('mild 3'). Showing 61.6% inflammation.



(c) Raw image ('mild 4')

(d) Test result ('mild 4'). Showing 57.1% inflammation.



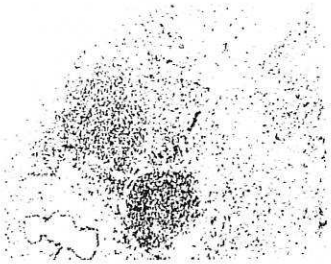
(e) Raw image ('mild 5')

(f) Test result ('mild 5'). Showing 66.2% inflammation.

(g) Raw image ('sev 1')

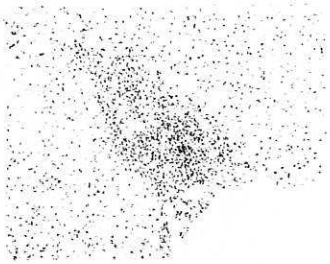
(h) Test result ('sev 1'). Showing 88.1% inflammation.

Fig. .3. Experimental test results. Inflamed cells are shown in black and healthy cells are grey.



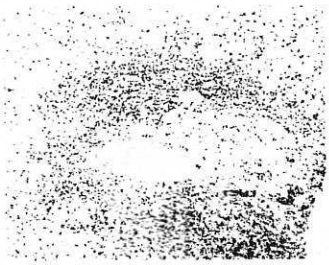
(a) Raw image ('sev 2')

(b) Test result ('sev 2'). Showing 78.9% inflammation.

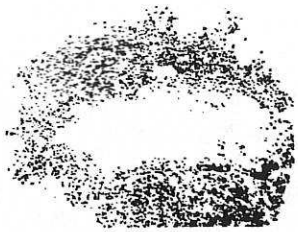


(c) Raw image ('sev 3')

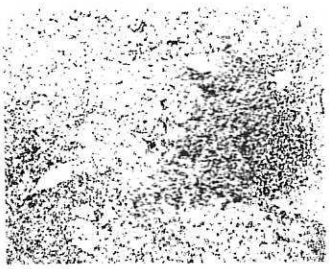
(d) Test result ('sev 3'). Showing 53.2% inflammation.



(e) Raw image ('sev 4')



(f) Test result ('sev 4'). Showing 75.9% inflammation.



(g) Raw image ('sev 5')

(h) Test result ('sev 5'). Showing 72.8% inflammation.

Fig. .4. Experimental test results. Inflamed cells are shown in black and healthy cells are grey.

Appendix C: Rank Correlation

Image	Cns1	Cns2	Cns3	Cns4	Cns5	Trn1	Trn2	Trn3	Trn4	Cnt1	Cnt2	Cnt3	Cnt4
mild 1	12	12	13	13	13	12	13	12	13	12	13	12	13
mild 2	10	11	12	10	10	11	11	11	11	11	11	11	11
mild 3	13	14	14	12	12	13	12	14	12	14	14	14	14
mild 4	15	13	10	14	14	14	14	13	14	13	12	13	12
mild 5	14	15	15	15	15	15	15	15	15	15	15	15	15
mod 1	11	10	11	11	11	10	10	10	8	10	10	10	10
mod 2	4	7	4	7	5	5	6	6	7	5	5	7	7
mod 3	5	6	6	6	7	8	7	8	10	4	6	5	6
mod 4	8	9	8	8	9	9	8	7	9	9	8	9	9
mod 5	7	5	7	5	6	6	5	5	4	8	7	4	5
sev 1	1	1	1	1	1	1	1	1	1	2	1	2	1
sev 2	3	2	2	4	3	2	3	3	2	3	2	1	2
sev 3	9	8	9	9	8	7	9	9	6	6	9	8	8
sev 4	2	3	3	3	2	4	2	2	5	1	4	6	4
sev 5	6	4	5	2	4	3	4	4	3	7	3	3	3

Table 1. Part A: Table showing the rank order of test images per observer. The observer names are abbreviated (Cns = Consultant, Trn = Trainee and Cnt = Control).

Image	Cnt5	Cnt6	Cnt7	Cnt8	Cnt9	Cnt10	Cnt11	Cnt12	Cnt13	Cnt14	Cnt15	W.B.r	Comp
mild 1	13	13	13	11	12	13	14	14	13	13	14	13	13
mild 2	11	11	11	12	11	11	13	12	11	10	11	11	11
mild 3	14	14	14	13	13	14	12	9	14	14	13	14	14
mild 4	12	12	12	15	14	12	11	13	12	12	12	12	12
mild 5	15	15	15	14	15	15	15	15	15	15	15	15	15
mod 1	10	10	10	10	10	10	10	10	10	11	10	10	10
mod 2	9	9	6	5	9	7	8	8	3	5	6	3	6
mod 3	5	6	4	9	8	8	5	7	7	7	7	7	7
mod 4	7	7	8	8	7	9	6	6	8	9	8	8	8
mod 5	8	8	9	6	6	6	9	11	6	2	5	6	5
sev 1	1	1	1	2	1	1	4	2	1	1	1	1	1
sev 2	2	2	5	4	2	3	3	1	4	4	3	2	2
sev 3	6	5	7	7	4	5	7	5	9	8	9	9	9
sev 4	4	3	2	3	5	2	1	4	2	6	4	4	3
sev 5	3	4	3	1	3	4	2	3	5	3	2	5	4

Table 2. Part B: Table showing the rank order of test images per observer. The automated system is labelled 'Comp', the other observer names are also abbreviated (Cns = Consultant, Trn = Trainee and Cnt = Control).

Cmpd to	Cns1	Cns2	Cns3	Cns4	Cns5	Trn1	Trn2	Trn3	Trn4	Cnt1	Cnt2	Cnt3	Cnt4
Cns1	1.000	0.946	0.936	0.936	0.971	0.939	0.961	0.946	0.846	0.957	0.946	0.896	0.921
Cns2	0.946	1.000	0.950	0.968	0.971	0.975	0.979	0.979	0.936	0.936	0.975	0.975	0.993
Cns3	0.936	0.950	1.000	0.911	0.943	0.936	0.939	0.950	0.854	0.932	0.979	0.914	0.954
Cns4	0.936	0.968	0.911	1.000	0.975	0.957	0.982	0.964	0.918	0.882	0.957	0.943	0.968
Cns5	0.971	0.971	0.943	0.975	1.000	0.979	0.989	0.971	0.925	0.936	0.964	0.925	0.961
Trn1	0.939	0.975	0.936	0.957	0.979	1.000	0.971	0.968	0.964	0.911	0.971	0.954	0.971
Trn2	0.961	0.979	0.939	0.982	0.989	0.971	1.000	0.986	0.936	0.918	0.964	0.936	0.968
Trn3	0.946	0.979	0.950	0.964	0.971	0.968	0.986	1.000	0.929	0.911	0.968	0.932	0.968
Trn4	0.846	0.936	0.854	0.918	0.925	0.964	0.936	0.929	1.000	0.821	0.907	0.925	0.939
Cnt1	0.957	0.936	0.932	0.882	0.936	0.911	0.918	0.911	0.821	1.000	0.921	0.875	0.911
Cnt2	0.946	0.975	0.979	0.957	0.964	0.971	0.964	0.968	0.907	0.921	1.000	0.957	0.982
Cnt3	0.896	0.975	0.914	0.943	0.925	0.954	0.936	0.932	0.925	0.875	0.957	1.000	0.982
Cnt4	0.921	0.993	0.954	0.968	0.961	0.971	0.968	0.968	0.939	0.911	0.982	0.982	1.000
Cnt5	0.886	0.954	0.914	0.929	0.914	0.929	0.918	0.921	0.889	0.911	0.950	0.936	0.961
Cnt6	0.886	0.950	0.911	0.914	0.918	0.929	0.918	0.925	0.896	0.925	0.936	0.914	0.950
Cnt7	0.929	0.936	0.936	0.932	0.936	0.918	0.925	0.921	0.832	0.950	0.954	0.886	0.936
Cnt8	0.900	0.932	0.879	0.943	0.946	0.971	0.946	0.950	0.936	0.857	0.929	0.900	0.925
Cnt9	0.850	0.936	0.850	0.914	0.907	0.946	0.911	0.914	0.954	0.857	0.904	0.925	0.936
Cnt10	0.904	0.968	0.929	0.932	0.957	0.964	0.950	0.957	0.943	0.932	0.943	0.918	0.964
Cnt11	0.846	0.889	0.893	0.893	0.882	0.868	0.893	0.886	0.811	0.882	0.907	0.846	0.896
Cnt12	0.814	0.839	0.829	0.846	0.861	0.871	0.854	0.825	0.843	0.829	0.864	0.811	0.846
Cnt13	0.961	0.950	0.975	0.932	0.968	0.950	0.964	0.971	0.879	0.936	0.968	0.896	0.943
Cnt14	0.875	0.943	0.904	0.939	0.925	0.939	0.929	0.932	0.921	0.829	0.932	0.954	0.957
Cnt15	0.914	0.971	0.950	0.979	0.964	0.968	0.975	0.971	0.939	0.875	0.982	0.957	0.986
W.B.r	0.954	0.957	0.982	0.925	0.961	0.964	0.957	0.964	0.900	0.921	0.982	0.932	0.957
Comp	0.943	0.989	0.971	0.964	0.971	0.971	0.982	0.989	0.932	0.918	0.986	0.961	0.989

Table 3. Part A: Spearman rank correlation between the automated system and the 25 observers of the comparison study[9] (-1 = systematic disagreement, 0 = no connection, 1 = strong agreement). The observer names are abbreviated (Cns = Consultant, Trn = Trainee and Cnt = Control).

Cmpd to	Cnt5	Cnt6	Cnt7	Cnt8	Cnt9	Cnt10	Cnt11	Cnt12	Cnt13	Cnt14	Cnt15	W.B.r	Comp
Cns1	0.886	0.886	0.929	0.900	0.850	0.904	0.846	0.814	0.961	0.875	0.914	0.954	0.943
Cns2	0.954	0.950	0.936	0.932	0.936	0.968	0.889	0.839	0.950	0.943	0.971	0.957	0.989
Cns3	0.914	0.911	0.936	0.879	0.850	0.929	0.893	0.829	0.975	0.904	0.950	0.982	0.971
Cns4	0.929	0.914	0.932	0.943	0.914	0.932	0.893	0.846	0.932	0.939	0.979	0.925	0.964
Cns5	0.914	0.918	0.936	0.946	0.907	0.957	0.882	0.861	0.968	0.925	0.964	0.961	0.971
Trn1	0.929	0.929	0.918	0.971	0.946	0.964	0.868	0.871	0.950	0.939	0.968	0.964	0.971
Trn2	0.918	0.918	0.925	0.946	0.911	0.950	0.893	0.854	0.964	0.929	0.975	0.957	0.982
Trn3	0.921	0.925	0.921	0.950	0.914	0.957	0.886	0.825	0.971	0.932	0.971	0.964	0.989
Trn4	0.889	0.896	0.832	0.936	0.954	0.943	0.811	0.843	0.879	0.921	0.939	0.900	0.932
Cnt1	0.911	0.925	0.950	0.857	0.857	0.932	0.882	0.829	0.936	0.829	0.875	0.921	0.918
Cnt2	0.950	0.936	0.954	0.929	0.904	0.943	0.907	0.864	0.968	0.932	0.982	0.982	0.986
Cnt3	0.936	0.914	0.886	0.900	0.925	0.918	0.846	0.811	0.896	0.954	0.957	0.932	0.961
Cnt4	0.961	0.950	0.936	0.925	0.936	0.964	0.896	0.846	0.943	0.957	0.986	0.957	0.989
Cnt5	1.000	0.993	0.954	0.886	0.957	0.950	0.939	0.918	0.882	0.868	0.936	0.896	0.939
Cnt6	0.993	1.000	0.946	0.886	0.964	0.968	0.936	0.921	0.886	0.854	0.921	0.893	0.936
Cnt7	0.954	0.946	1.000	0.896	0.879	0.936	0.939	0.868	0.936	0.857	0.929	0.914	0.929
Cnt8	0.886	0.886	0.896	1.000	0.921	0.932	0.864	0.839	0.918	0.900	0.939	0.911	0.929
Cnt9	0.957	0.964	0.879	0.921	1.000	0.954	0.868	0.904	0.846	0.882	0.914	0.868	0.914
Cnt10	0.950	0.968	0.936	0.932	0.954	1.000	0.900	0.868	0.936	0.911	0.946	0.929	0.961
Cnt11	0.939	0.936	0.939	0.864	0.868	0.900	1.000	0.921	0.864	0.779	0.900	0.850	0.893
Cnt12	0.918	0.921	0.868	0.839	0.904	0.868	0.921	1.000	0.793	0.725	0.850	0.814	0.836
Cnt13	0.882	0.886	0.936	0.918	0.846	0.936	0.864	0.793	1.000	0.921	0.954	0.986	0.971
Cnt14	0.868	0.854	0.857	0.900	0.882	0.911	0.779	0.725	0.921	1.000	0.961	0.936	0.950
Cnt15	0.936	0.921	0.929	0.939	0.914	0.946	0.900	0.850	0.954	0.961	1.000	0.961	0.986
W.B.r	0.896	0.893	0.914	0.911	0.868	0.929	0.850	0.814	0.986	0.936	0.961	1.000	0.979
Comp	0.939	0.936	0.929	0.929	0.914	0.961	0.893	0.836	0.971	0.950	0.986	0.979	1.000

Table 4. Part B: Spearman rank correlation between the automated system and the 25 observers of the comparison study[9] (-1 = systematic disagreement, 0 = no connection, 1 = strong agreement). The automated system is labelled 'Comp', the other observer names are also abbreviated (Cns = Consultant, Trn = Trainee and Cnt = Control).