The University of Saskatchewan
Department of Computer Science

Technical Report #2012-04

UNIVERSITY OF SASKATCHEWAN

Classification of Four Types of Odontogenic Cysts

A. Frydenlund* M. Eramian† T. Daley‡

Abstract

Odontogenic cysts originate from remnants of the tooth forming epithelium in the jaws and gingiva. There are various kinds of such cysts with different biological behaviours that carry different patient risks and require different treatment plans. Types of odontogenic cysts can be distinguished by the properties of their epithelial layers in H&E stained samples. Herein we detail a set of image features for automatically distinguishing between four types of odontogenic cysts in digital micrographs and evaluate their effectiveness using two statistical classifiers – a support vector machine (SVM) and bagging with logistic regression as the base learner (BLR). Cyst type was correctly predicted from among four classes of odontogenic cysts between 83.8% and 92.3% of the time with an SVM and between $90 \pm 0.92\%$ and $95.4 \pm 1.94\%$ with a BLR. One particular cyst type was associated with the majority of misclassifications. Omission of this cyst type from the data set improved the classification rate for the remaining three cyst types to about 96.2% for both SVM and BLR.

1 Introduction

Tooth formation involves a complex interaction of the enamel organ and surrounding connective tissues. Occasionally, the remnants of the enamel organ can give rise to a variety of odontogenic tumours and cysts. The rarity of these pathologies mean that proper diagnosis is often the responsibility of subspecialists. Proper diagnosis of odontogenic cysts is crucial because different biological behaviours of the various types of cysts require different treatment plans, and present significantly different risks to patients.

As was argued in Eramian et al. (2011), there is a need for an accurate computer assisted diagnostic protocols to reduce the workload of oral pathologists and to potentially reduce the costs of multiple expert diagnostic opinions. Many odontogenic cysts can be distinguished by histologic examination of characteristics of their epithelial layer. As a first step towards a computer assisted diagnostic system for classifying odontogenic cysts, we presented in Eramian et al. (2011) an automated epithelial segmentation algorithm for hematoxylin and eosin (H&E) stained digital microscopy images of odontogenic cysts. The algorithm was evaluated using a data set consisting of four types of odontogenic cysts: dentigerous cysts (DCyst), odontogenic keratocysts (OKC), lateral periodontal cysts (LPC), and glandular odontogenic cysts (GOC). Eramian et al. (2011) presented a summary of the details of these pathologies, their prevalence, and their radiographic appearance.

In the current study, we make the assumption that the epithelial regions of such digital images as studied in Eramian et al. (2011) can be perfectly segmented. We propose a set of image features that can be computed from such epithelial regions to form descriptions of the regions and show that the proposed region descriptions can be used to distinguish between and accurately classify samples of the four types of odontogenic cysts mentioned above using standard classification algorithms.

*Department of Computer Science, University of Saskatchewan, aaf955@mail.usask.ca

†Department of Computer Science, University of Saskatchewan, eramian@cs.usask.ca

‡Department of Pathology, Schulich School of Medicine and Dentistry, University of Western Ontario, Tom.Daley@schulich.uwo.ca

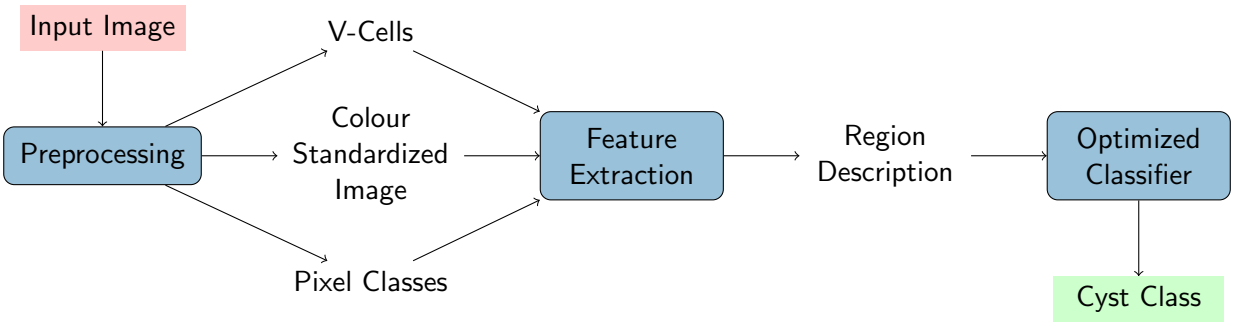


Figure 1: Cyst Classification System Overview

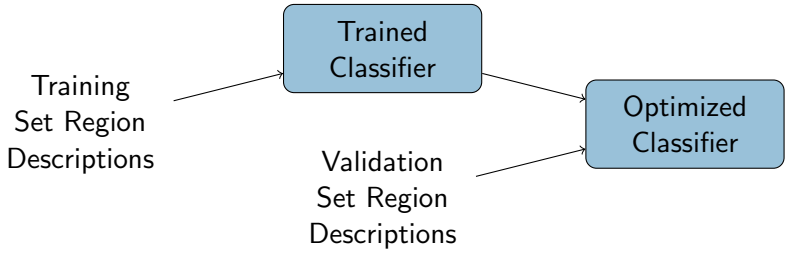


Figure 2: Classifier Training Overview. Region descriptions were formed by the same process outlined in Figure 1.

1.1 System Overview

The high-level architecture of our system is depicted in Figure 1. The cyst classification system is fairly standard in that it follows the stages of input preprocessing, feature extraction, and classification. Input images are preprocessed to reduce stain colour hue variations caused by differences in sample thickness, sample age, amount of stain, etc. which can adversely affect features based on colour. We also extract other information about virtual cells (V-Cells) and whether each pixel is predominantly hematoxylin-stained (nucleus), eosin-stained (other tissue), or unstained (non-tissue); we call these the pixel classes. This information is used to extract features which form a region description (feature vector) that is fed to the trained cyst classifier which outputs its prediction of the input images cyst type.

The classifier was trained from a set of training images. For each training image a region description was produced in the same manner as depicted in Figure 1. The set of all such region descriptions was used to train two standard classifiers. The parameters of the classifiers were then optimized using epithelial region descriptions extracted from a validation set of images, disjoint from the training set. This process is depicted in Figure 2. The optimized classifier was then evaluated using a third, disjoint set of test images.

2 Materials and Methods

2.1 Image Data Sets

Three different sets of images which were acquired at different times with slightly different acquisition parameters were used in our work.

2.1.1 Training Set

Epithelial region descriptions (feature vectors) from the images in the training set were used to train classifiers with examples of descriptions from each cyst class and to construct colour histogram models of

	Subset 1	Subset 2	Class Total
DCyst	10	10	20
LPC	10	10	20
OKC	10	10	20
GOC	8	5	13
Total	38	35	73
Resolution (px)	1300×1030	1080×734	
Scale (px/μm)	3	3	

Table 1: Summary of training data set composition (73 images total) denoting counts of samples for Dentigerous cysts (DCyst), odontogenic keratocysts (OKC), lateral periodontal cysts (LPC) and glandular odontogenic cysts (GOC). Two groups of training data were gathered at different times with the same equipment, but slightly different resolution.

	Validation Set	Test Set
DCyst	10	10
LPC	10	10
OKC	10	11
GOC	7	8
Total	37	39
Resolution (px)	1300×1030	1300×1030
Scale (px/μm)	2.6	2.6

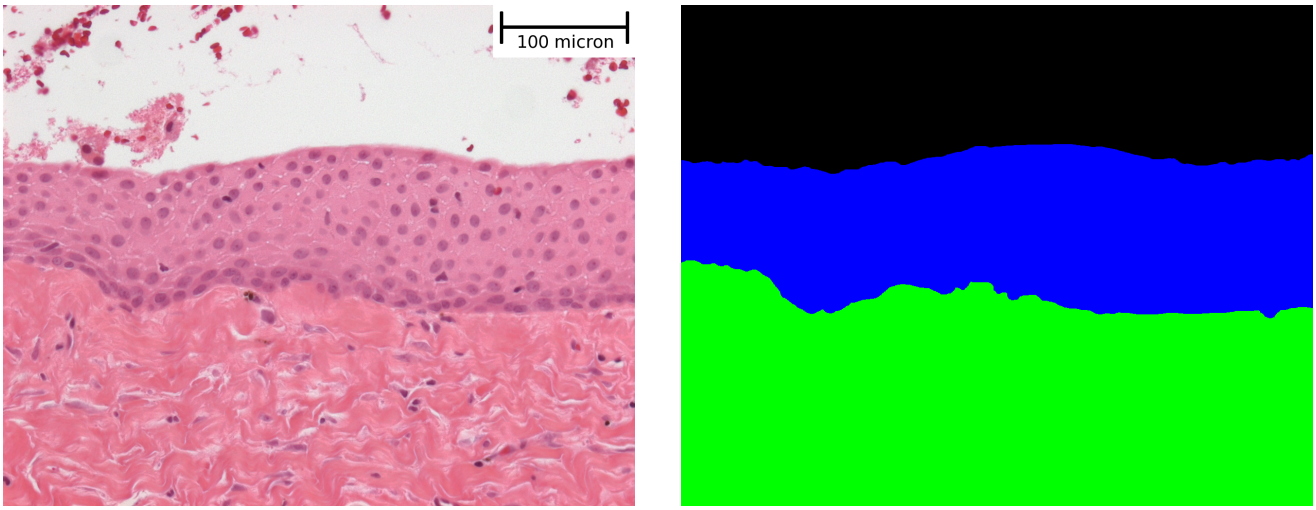
Table 2: Summary of validation and test sets. The two sets were gathered with the same equipment and at the same time.

epithelial and stromal regions in the colour standardization process (Section 2.3).

A total of 73 images were obtained using a Zeiss Axiocam MRc at 25X objective magnification. Of these, 38 images were obtained at a resolution of 1300×1030 pixels with an inter-pixel distance of $1/3\mu$m (3 pixels per micron), consisting of 10 dentigerous cysts (DCyst), 10 lateral periodontal cysts (LPC), ten odontogenic keratocysts (OKC), and 8 glandular odontogenic cysts (GOC). The remaining 35 images were obtained at a resolution of 1080×734 pixels with an inter pixel distance of $1/3\mu$m (3 pixels per micron), consisting of 10 DCyst, 10 LPC, 10 OKC, and 5 GOC. Composition of the training set is summarized in Table 1. The inter-pixel distance (scale) for the images was determined by measuring the diameter in pixels of red blood cells incidentally occurring in the images which are known to be 7 microns in diameter. The mean diameter in pixels of 60 red blood cells sampled from image subsets sharing the same acquisition parameters was measured and divided by 7 to determine the image's scale in pixels per micron (px/μm).

2.1.2 Validation and Test Sets

The validation and test sets were collected at the same time using the same camera as the training set with a 20X objective magnification. These images have size of 1300×1030 pixels and an inter-pixel distance of 0.385μm (2.6 pixels per micron). Scale was determined in the same manner as that for the training set, using 92 sample blood cells. The number of images of each type of cyst contained in each set are summarized in Table 2. The validation set was used on the trained classifiers to tune classifier parameters and produced the optimized classifiers.



(a) (b)

Figure 3: An image and its ground truth. Left: A dentigerous cyst. Right: The epithelial region is blue, the stroma region is green, and the non-tissue region is black.

2.2 Image Scale

Some of the image features extracted to form descriptors of the epithelial regions include measurements of distance within the image. Because the training set has a different scale than the validation and test sets, such features were scaled to the same scale as the training images. The scaling factor for the features from the training set was therefore $\phi_{\text{Train}} = 1.0$. For the validation and test sets, $\phi_{\text{Test}} = \phi_{\text{Validation}} = 3.0/2.6$.

2.2.1 Manual Epithelial Regions (Ground Truth)

Ground truths were established for all images in all data sets by the same observer. Images were manually segmented into three types of region: epithelium, stroma (other tissue), and non-tissue. See Figure 3 for an image with its ground truth. Let $G_{\text{epi}}(I)$, $G_{\text{str}}(I)$, and $G_{\text{nt}}(I)$ be the sets of pixels from image I labeled as epithelium, stroma, and non-tissue, respectively.

The cyst images in our data sets all contain a region of epithelial tissue which forms the inner lining of the cyst (e.g. the middle darker pink band in Figure 3(a)). Typically this region is an irregular band or strip bordered on one side by the *lumen*, the formerly fluid-filled cavity of the cyst, which, after preparation of the samples for imaging, appears a largely tissue-free region (top band of Figure 3(a)). The opposite border of the epithelial region is adjacent to deeper tissues relative to the lumen; it is mainly collagen and connective tissue, referred to collectively as *stroma* (brighter pink band at the bottom of Figure 3(a)).

In the manually segmented ground truth, the epithelial region was segmented from the adjacent lumen and stroma regions. In some cases, particularly for glandular cysts, the epithelial region contained significant holes, either due to the presence of gland-like structures, or other irregularities. These regions were labeled as non-tissue, even when they contained some stained material, which was quite frequent. This was done because the material within the gland-like structures is not epithelial tissue and would not be expected to be grouped with the epithelial region by an automated segmentation algorithm. Similarly, stray stained cellular material in the lumen region was grouped with the lumen region (e.g. the black region in Figure 3(b)) because it was shown in Eramian et al. (2011) that the lumen region could be segmented extremely reliably despite these stray bits of material.

Nevertheless, on occasion we will wish to consider the region formed by the union of the pixels in $G_{\text{epi}}(I)$ and the holes in this region. We denote this set of pixels by $G_{\text{epi}}^{\text{holes}}(I)$.

2.3 Colour Standardization

Our training, validation, and test data sets were colour standardized using the algorithm in Eramian et al. (2011). This standardization process reduces luminance and chrominance variation between images in the data sets that arise from variations in the amount of dye used, sample thickness, and lighting conditions. The colour histogram models required by the algorithm were established using the training images, and this same histogram was used to standardize all three of our image sets. Some of the image features used for classification, described in Section 2.6, were derived from the standardized images. The result of this step is denoted "Colour Standardized Image" in Figure 1.

2.4 Pixel Classification

In this preprocessing step, each pixel in an image's epithelial region is classified as either "nucleus", "non-nucleus", or "non-tissue"; this roughly corresponds to pixels where the predominant stain is respectively eosin, hematoxylin, or none. In turn, this corresponds to classifying each pixel as either "purple", "pink", or "white", which was achieved using k-means clustering (MacQueen, 1967, Spath, 1985).

Each image I was smoothed by processing with a 3×3 vector median filter (Astola, 1990) in RGB colour space. The vector median filter changes the colour vector of the pixel in the center of the neighbourhood N to the colour vector x_m from the neighborhood which minimizes the summed Manhattan distance between x_m and the colour vectors of all other pixels in the neighbourhood. The average colour $\bar{x}_{nt}$ of those pixels in the smoothed image which correspond to a non-tissue pixel was computed; each non-tissue pixel contributing to this average was then set to the colour $\bar{x}_{nt}$. This replacement was to avoid a few cases where variation in the colour and luminance of non-tissue pixels within a single image was large enough to cause some non-tissue pixels to be classified as "non-nucleus". The resulting image is denoted $I^{\bar{x}_{nt}}_{\mathrm{VMF}}$.

The RGB colour vectors of the epithelial and non-tissue pixels (determined by the ground truth) in $I^{\bar{x}_{nt}}_{\mathrm{VMF}}$ were clustered using k-means clustering ($k = 3$); the non-tissue pixels were included to avoid situations where there were insufficiently many non-tissue samples to be identified as a unique cluster. Three random vectors were used for the initial cluster centers. Clustering was repeated 10 times. The replicant with the minimum summed distances between samples and their cluster centers was selected as the final result.

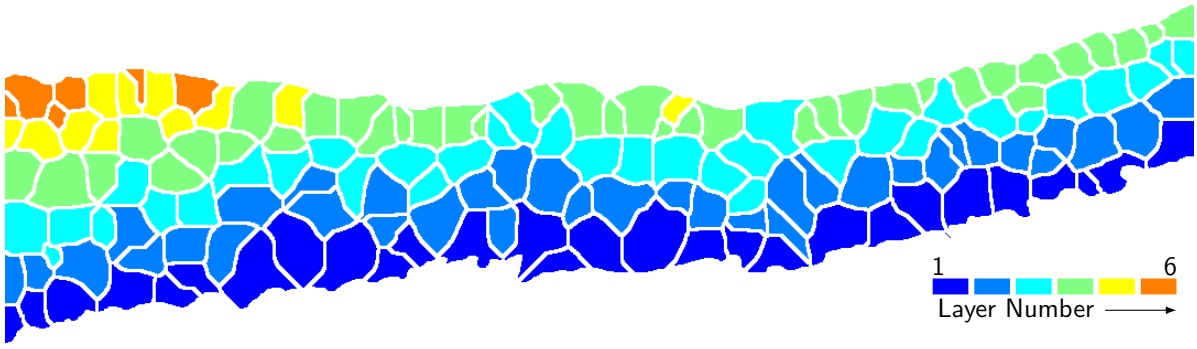
The output of this operation was a label image where pixels with the same label belong to the same cluster. To account for the possibility that cluster labels across different images did not necessarily represent the same class, the class identities of the cluster centers were inferred from the BT.601 Luma ($0.299R + 0.587G + 0.114B$) values of the colour represented by the cluster centers and were made consistent across all images. Let the resulting sets of nucleus, non-nucleus and non-tissue pixels be denoted $P_{\mathrm{nuc}}(I)$, $P_{\mathrm{nonnuc}}(I)$, and $P_{\mathrm{nt}}(I)$, respectively; these are the "Pixel Classes" denoted in Figure 1.

2.5 V-Cell Segmentation and Layering

The epithelial region of each image I was segmented into cell regions called V-cells. Each V-cell corresponds to a theoretical extent of an epithelial cell profile. The result of this preprocessing step is denoted "V-Cells" in Figure 1.

V-cells were found within the region of the smoothed I defined by $G_{\mathrm{epi}}(I)$ using a slightly modified version of the procedure introduced by Landini and Othman (2003). The modification was that the set of nuclei pixels $P_{\mathrm{nuc}}(I)$ resulting from the k-means pixel classification process was used to refine the "domes" computed by Landini's method. Dome regions that did not share at least one pixel with $P_{\mathrm{nuc}}(I)$ were removed, complementing the methods already used in Landini and Othman (2003) to establish which domes correspond to true cell nuclei. Figure 4(a) shows a smoothed original image with V-cell boundaries overlaid. Figure 4(b) shows an example result of V-cell computation where V-cell boundaries are white and the colour of each cell represents its layer. Determining the cell layer is the last step in Landini's algorithm. The layer of each V-Cell with respect to the lumen is determined iteratively. The first level of V-cells are those that are adjacent to the lumen. The second layer of V-cells are those that are adjacent

(a) Original image after vector median filtering with V-cell boundaries overlaid.



(b) V-Cells of the epithelial layer. The white lines denote the V-Cell boundaries found using a modified version of Landini and Othman (2003). The colour of each V-cell indicates the layer of the cell with respect to the lumen. The boundaries between V-cells have been thickened for presentation.

Figure 4: Examples V-cells and Cell Layer computation.

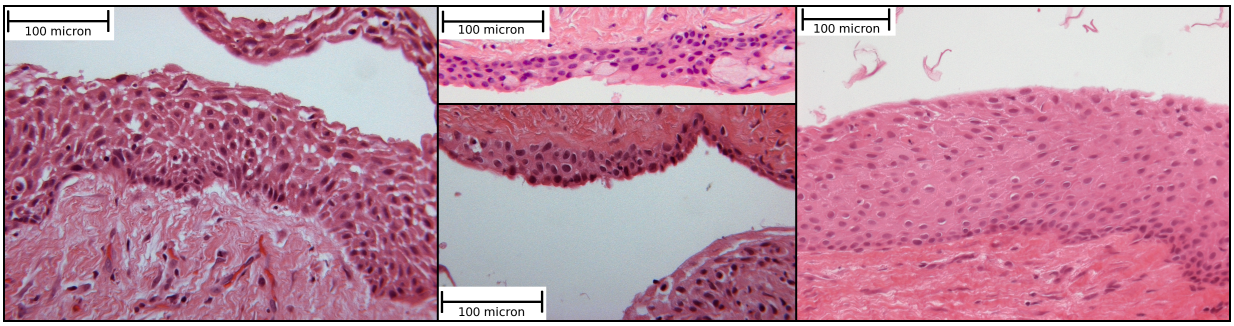


Figure 5: Examples of dentigerous cysts. The top center cyst has mucous cells, which are similar in appearance to the GOC gland-like structures. The left image has a large variance of thickness where the right has little. In the bottom middle image, the colour of the epithelium near the lumen resembles that of parakeratin which is normally a property of OKC.

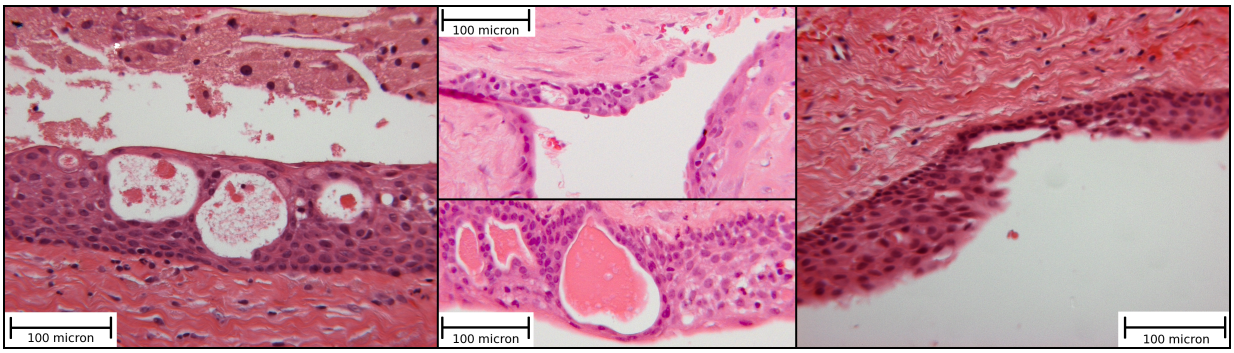


Figure 6: Examples of glandular odontogenic cysts. The thickness of the epithelial region can be quite variable across samples. The size, shape and opacity of the gland-like structures are also highly variable, as is the distribution and shape of nuclei.

to a cell in the first layer, and so on. We also determined the cell layers with respect to the stroma region.

2.6 Feature Extraction

In this section we describe the image features we used to form descriptions of epithelial regions, denoted "Region Descriptions" in Figure 1. Features are summarized in Table 3.

The distinguishing characteristics of the four cyst types being considered were described in some detail in Eramian et al. (2011). The main characteristics of the microscopic appearance of the epithelia of these cysts are briefly reviewed here since they are the attributes we need our image-derived features to capture. The strongest discriminators of cyst type are thickness of the epithelial layer, shape of epithelial cells, and the presence or absence of gland-like structures and parakeratin.

DCyst typically have a variably thick epithelium with stratified squamous epithelial cells. GOC are distinguished by the presence of gland-like structures in the epithelium; they also exhibit a relatively thick epithelial layer. OKC generally have a uniformly thick epithelial layer, 5-10 cells thick with basal cells ranging from cuboidal to columnar in shape, and exhibit a thin, superficial layer of parakeratin at the epithelium/lumen interface which stains a slightly darker pink than other non-nuclear components. Nuclei in OKC are often more densely distributed within the basal layers. LPC have a much thinner epithelial layer, 1-3 cells thick, and may have scattered plaques of cells with clear cytoplasm. Features were selected for their ability to capture information about these general characteristics of the appearance of each type

of cyst.

Despite these simple sounding characterizations, there is a high degree of variation in appearance within all four classes of cyst. Figures 5 and 6 show examples which illustrate this for DCyst and GOC cyst types, respectively.

2.6.1 Preliminary Definitions

The following definitions are used in the explanation of the features used in the epithelial region descriptions.

The *area* of an image region is the number of pixels that make up the region. *Scaled area of a region*, in units of square μm, is its area multiplied by ϕ_x^2, where ϕ_x is ϕ_{Train}, $\phi_{\text{Validation}}$, or ϕ_{Test}, depending on the origin of the image.

The *perimeter* of a region is defined as the total length of its inner boundary in pixels, including inner boundaries of holes in the region. *Scaled perimeter* is the perimeter of a region multiplied by ϕ_x.

Compactness of a region is the ratio of the square of the region's perimeter to its area; it is a measure of "circularity". The most compact region is a circle; larger values of compactness indicate greater deviation from a circular shape.

The *major and minor axes* of a region are defined as the major and minor axes of the ellipse that has the same normalized second central moments as the region. The orientation vectors of the axes coincide with the eigenvectors of the covariance matrix C of the population of ordered pairs of (x, y) coordinates of the pixels that comprise the region. The lengths of the axes are twice the magnitude of the eigenvalues of C. The *scaled* major and minor axes lengths are the axes lengths scaled by ϕ_x. The *eccentricity* of a region is defined as the ratio of the distance between foci of the ellipse and the major axis length. Eccentricity characterizes how elongated a region is and takes on values between 0, for a circular region, and 1, in the case of a line segment.

2.6.2 Epithelial Region Description Features

In the following sections the features used in the descriptions of the epithelial regions of an image I are enumerated and explained.

Compactness of the Epithelial Region (f_1): The first feature is the compactness of the epithelial region, that is, the region defined by the set of pixels $G_{\text{epi}}(I)$. Thicker epithelial regions are more compact than thinner regions.

Pixel Class Proportion Features (f_2–f_5): Features f_2 through f_4 are the proportion of pixels in the epithelium and its holes of each pixel class. Formally

$$f_2 = \left|P_{\text{nuc}}(I) \cap G_{\text{epi}}^{\text{holes}}(I)\right| / \left|G_{\text{epi}}^{\text{holes}}(I)\right|, \tag{1}$$

$$f_3 = \left|P_{\text{nonnuc}}(I) \cap G_{\text{epi}}^{\text{holes}}(I)\right| / \left|G_{\text{epi}}^{\text{holes}}(I)\right|, \tag{2}$$

$$f_3 = \left|P_{\text{nt}}(I) \cap G_{\text{epi}}^{\text{holes}}(I)\right| / \left|G_{\text{epi}}^{\text{holes}}(I)\right|, \tag{3}$$

where $|\cdot|$ denotes the cardinality of a set. Feature f_5 is the proportion of nucleus pixels from among only tissue pixels in the epithelium:

$$f_5 = \frac{\left|P_{\text{nuc}}(I) \cap G_{\text{epi}}^{\text{holes}}(I)\right|}{\left|P_{\text{nuc}}(I) \cap G_{\text{epi}}^{\text{holes}}(I)\right| + \left|P_{\text{nonnuc}}(I) \cap G_{\text{epi}}^{\text{holes}}(I)\right|} \tag{4}$$

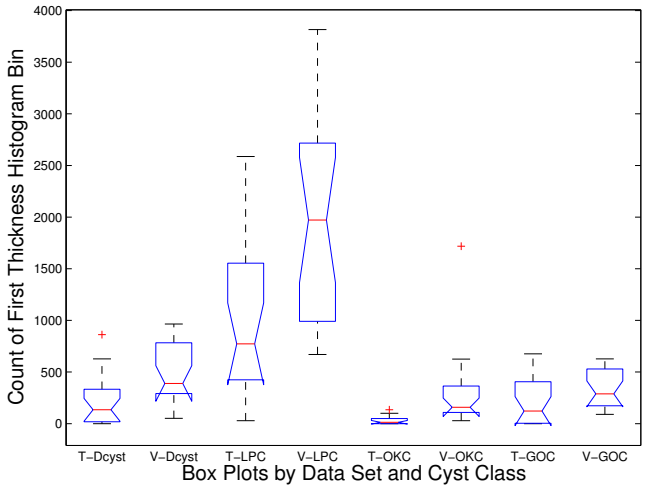


Figure 7: The box plot of the first bin of the thickness histogram (f_7). Plots for each class for both the training set (T prefix) and the validation set (V prefix) are shown. We see a larger count for LPC's than the other classes as expected.

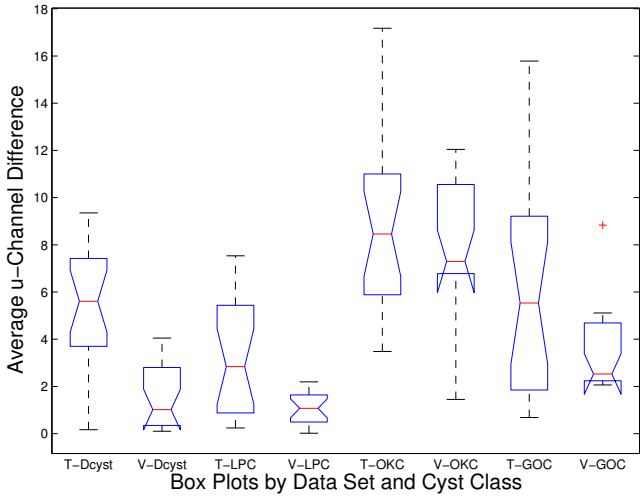


Figure 8: Box plot of the absolute difference between the average u-channel value in the contour fringe and average u-channel value of the rest of the epithelial region (f_{12}). Plots for each class for both the training set (T prefix) and the validation set (V prefix) are shown. Generally, OKC exhibit a larger difference, as expected.

Thickness Features (f_6–f_8): Since thickness of the epithelial region is a distinguishing characteristic of cyst type, multiple features were defined to capture thickness and variations in thickness of the epithelial region. The skeleton of the region formed by $G_{\text{epi}}^{\text{holes}}(I)$ was computed using the method of Cychosz (1994). The shortest distances from each point on the boundary of the region formed by $G_{\text{epi}}^{\text{holes}}(I)$ to the skeleton were found. These distances were doubled to correctly reflect thickness since regions are locally symmetric about their skeleton, and scaled by ϕ_x; denote this set of scaled thicknesses T. Feature $f_6 = \text{var}(T)$, that is, the variance of the thicknesses in T. The real-valued interval $[0, 331\phi_x]$ was divided into seven non-overlapping sub-intervals of equal length. The upper bound of this interval was the largest thickness found in the training set; thickness values outside the interval arising from the validation and training sets (which, in practice, were very few) were not included in the histogram. A histogram of thickness values was computed by counting the number of elements in T falling within each sub-interval. Ultimately only the first bin's value was used as a feature, since the other bins did not generalize well to the validation set. Thus f_7 = value of thickness histogram bin 1. The feature f_8 was defined as the difference between the mean of the smallest 10% of the elements in T and the mean of the largest 10% of elements in T. Thus, f_8 characterizes the difference in thickness between the thinnest and thickest parts of the epithelial region. The distributions for f_7 for the different cyst types in the training and validation set are shown in Figure 7 and illustrate the feature's potential fitness to distinguish between LPC and other classes.

Co-occurrence Features (f_9–f_{10}): We elected to consider some texture features since these could potentially capture valuable information not so easily discerned by the eye, for example, variations in density of nuclei, which could manifest as variations in texture features due to variation in the number of strong edges. The statistical features of Haralick et al. (1973) are established general purpose features for characterizing texture. Asymmetric co-occurrence matrices were computed for the epithelial region of red channel of the unfiltered (no VMF) standardized image (only pixels in the red channel corresponding to $G_{\text{epi}}(I)$ contributed to the matrices). Eight matrices were computed, one for each of the eight horizontal and vertical directions at a distance of $1\phi_x$px. To tackle the implementation issue of a potentially fractional

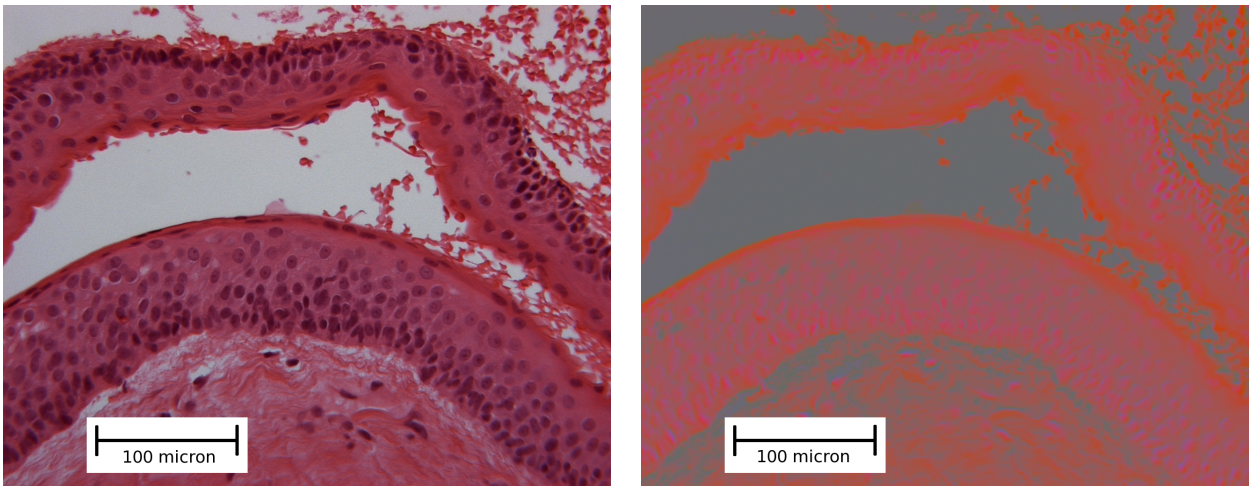


Figure 9: Parakeratin in a regular and a standardized image. Parakeratin stains with a more orange hue compared to the rest of the epithelium

displacement distance, the images were scaled (re-sampled) by a factor of ϕ_x using bilinear interpolation so that a displacement distance of 1 could be used in all cases. The red channel was used since it would contain most of the colour differences between nuclei tissue, non-nuclei tissue, and parakeratin. The average correlation and and entropy features of the eight co-occurrence matrices were used for features f_9 and f_{10}, respectively.

Contour Fringe Features (f_{11}–f_{12}): The purpose of these feature is to gauge whether parakeratin is present at the epithelium/lumen interface, which is a distinguishing feature of OKC. Figure 9 shows how parakeratin stains a slightly more orange hue than the rest of the epithelium. Both f_{11} and f_{12} are derived from the average colour value of all epithelial pixels in $G_{\text{epi}}(I)$ within $12\phi_x$ pixels of the lumen. We call these pixels the *fringe* of the epithelial region. The colour-standardized and 3×3 vector-median smoothed I was transformed from the RGB colour space to the CIEL*u*v* colour space. The average u-channel value $\bar{u}_{\text{fringe}}$ was computed over all pixels in the fringe. The average u-channel value $\bar{u}_{\text{non-fringe}}$ was also computed over all pixels in $G_{\text{epi}}(I)$ not in the fringe. The features f_{11} and f_{12} are defined as:

$$f_{11} = \bar{u}_{\text{fringe}} \quad (5)$$

$$f_{12} = |\bar{u}_{\text{fringe}} - \bar{u}_{\text{non-fringe}}| \quad (6)$$

Figure 8 illustrates the distributions of f_{12} for the four cyst classes and the potential for f_{12} to distinguish between OKC and other cysts.

Eccentricity of Nuclei Feature (f_{13}–f_{17}): These features were designed to capture the aforementioned variations in the shape of the epithelial cells between cyst classes. For example, eccentricity of columnar cells will be larger than that of squamous or cuboidal cells on account of their more elongated shape. The shape of the nuclei approximates the shape of the cell therefore the eccentricity of every nucleus in the epithelial region of I was computed. The eccentricity of each maximal contiguous set of pixels from $P_{\text{nuc}}(I)$ was computed and a histogram of the resulting eccentricities was computed by dividing the range $[0.0, 1.0]$ of possible eccentricities into eight sub-intervals of equal size and counting the number of nuclear eccentricities in each interval. Features f_{13} through f_{17} are the bin counts for bins 1, 5, 6, 7, and 8, respectively. Bins 2 through 4 were not used after boxplots of these features computed from the training images (similar to Figures 7 and 8) were consulted and the features were not found to have much potential to discriminate between cyst classes.

Largest Hole and Lumen Component Features (f_{18}–f_{24}): These features were defined with the expectation that they would be representative of the presence or absence of gland-like structures and help to distinguish between GOC and other cysts. The hole regions in the epithelium are defined by $H = G_{\text{epi}}^{\text{holes}}(I) - G_{\text{epi}}(I)$. Each hole in $G_{\text{epi}}(I)$ corresponds to a subset of H of spatially contiguous pixels. Feature f_{18} was defined as the product of ϕ_x^2 and the area (in pixels) of the largest hole. Feature f_{19} was defined to be the compactness of the largest hole. Feature f_{20} was defined as the average hole area (scaled by ϕ_x^2) over all holes. Feature f_{21} was defined as total number of holes. Similar features were computed from the regions formed by the non-tissue pixels within the epithelium, that is, the spatially contiguous regions defined by $NT = P_{\text{nonnuc}}(I) \cap G_{\text{epi}}^{\text{holes}}(I)$. Feature f_{22} was defined as the scaled area of the largest non-tissue region. Feature f_{23} was defined as the eccentricity of the largest non-tissue region. Feature f_{24} was defined as the compactness of the largest non-tissue region. All compactness features were computed from scaled areas and perimeters.

V-Cell Features (f_{25}–f_{31}): These features were used to capture properties of the first layer of V-cells adjacent to the lumen and the first layer adjacent to the stroma since there are specific differences in these areas expected between classes. Feature f_{25} was defined as the number of V-cells in the first layer adjacent to the lumen. Feature f_{26} was defined as the number of V-cells in the first layer adjacent to the stroma. Features f_{27} was defined as the average eccentricity of the cells in the first layer adjacent to the lumen . Feature f_{28} was defined to be the proportion of pixels within the V-cells of the first layer adjacent to the stroma that belong to $P_{\text{nuc}}(I)$. Feature f_{29} was defined to be the average compactness of all V-cells. Feature f_{30} was defined as the average eccentricity of all V-cells. Finally, Feature f_{31} was defined as the average minor axis length of all V-cells, scaled by ϕ_x.

First Layer Ratios and Nuclei Distance Histogram Features (f_{32}–f_{52}): These features are intended to capture differences in the distribution of nuclei across the epithelial region between classes.

Let L_1^{Lumen} denote the portion of the region formed by the union of the V-cells in the first layer adjacent to the lumen. Similarly L_1^{Str} is the region formed by the union of the V-cells in the first layer adjacent to the stroma.

Feature f_{32} was defined as the proportion of nuclei pixels in L_1^{Lumen} to the proportion nuclei pixels in L_1^{Str}.

$$f_{32} = \frac{\left(\left|P_{\text{nuc}}(I) \cap L_1^{\text{Lumen}}\right| / \left|L_1^{\text{Lumen}}\right|\right)}{\left(\left|P_{\text{nuc}}(I) \cap L_1^{\text{Str}}\right| / \left|L_1^{\text{Str}}\right|\right)} \tag{7}$$

The remaining features f_{33}–f_{52} are central moments of histograms of nuclei positions in the epithelium relative to the lumen, stroma and skeleton of the epithelium. The skeleton of the epithelial region $G_{\text{epi}}(I)$ was computed using the algorithm of Cychosz (1994). For each point p in $P_{\text{nuc}}(I) \cap G_{\text{epi}}(I)$ (nucleus pixels in the epithelium) the shortest distance from p to the stroma ($D_{\text{str}}(p)$), lumen ($D_{\text{lumen}}(p)$), and epithelium skeleton ($D_{\text{skel}}(p)$) were found using the distance transform algorithm of Meijster et al. (2002). Figure 10 illustrates how these distances are computed.

For each p, its position between the nucleus and the skeleton, $h(p)$ was computed:

$$h(p) = \begin{cases} D_{\text{lumen}}(p)/2(D_{\text{lumen}}(p)+D_{\text{skel}}(p)), & \text{if } D_{\text{lumen}}(p)<D_{\text{str}}(p) \\ (1-D_{\text{str}}(p)/(D_{\text{str}}(p)+D_{\text{skel}}(p)))/2 + 0.5, & \text{if } D_{\text{lumen}}(p) \geq D_{\text{str}}(p) \end{cases} \tag{8}$$

A 30-bin histogram of $h(p)$ over all p was computed, denoted $H_{\text{lumen}}^{\text{str}}$. Features f_{33}, f_{34}, and f_{35} were defined to be the second, third, and fourth central moments of $H_{\text{lumen}}^{\text{str}}$, respectively. A histogram $J_{\text{lumen}}^{\text{str}}$ was computed in the same fashion but where each p was the centroid of a V-cell region; this was done to take advantage of both of our estimates of nucleus locations. Features f_{36}, f_{37} and f_{38} were defined to be the first, second, and third central moments of $J_{\text{lumen}}^{\text{str}}$.

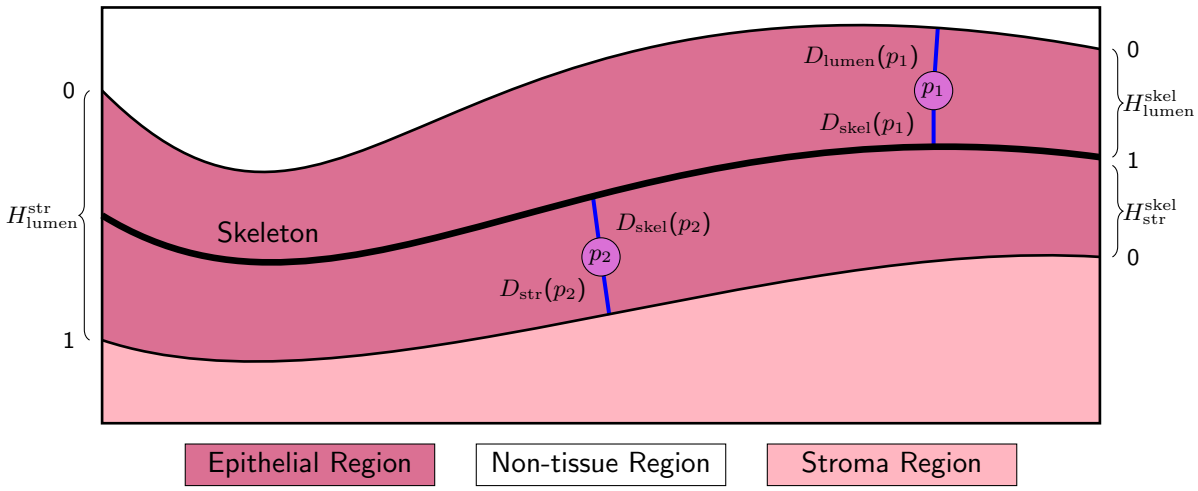


Figure 10: Diagram of the nuclei distance histogram features. p_1, p_2, are examples of nuclei points on either side of the skeleton. Blue lines are the shortest distances from the nuclei location to skeleton, lumen and stroma contours. The zones at the left and right illustrate the positions indicated by $h(p)$ for the three types of histograms.

A histogram $H^{\text{skel}}_{\text{lumen}}$ of relative position of nuclei between the lumen and the skeleton was computed in the same way as $H^{\text{str}}_{\text{lumen}}$ but using $h(p) = D_{\text{lumen}}(p)/(D_{\text{lumen}}(p) + D_{\text{skel}}(p))$. Only nuclei between the skeleton and the lumen contributed to this histogram. Features f_{39}, f_{40}, f_{41}, and f_{42} were defined as the first, second, third, and fourth central moments of $H^{\text{skel}}_{\text{lumen}}$, respectively. Histogram $J^{\text{skel}}_{\text{lumen}}$ was computed in the same way as $H^{\text{skel}}_{\text{lumen}}$, but using the centroids of the V-cell regions as the nuclei locations. Features f_{43}, f_{44}, and f_{45} were defined as the first, third, and fourth central moments of $J^{\text{skel}}_{\text{lumen}}$ (the second moment was omitted after consulting box plots for that feature similar to Figure 7).

Histograms $H^{\text{skel}}_{\text{str}}$ and $J^{\text{skel}}_{\text{str}}$ of relative position of nuclei between the stroma and the skeleton were computed in the same way as $H^{\text{skel}}_{\text{lumen}}$ and $J^{\text{skel}}_{\text{lumen}}$, but using $h(p) = D_{\text{str}}(p)/(D_{\text{str}}(p) + D_{\text{skel}}(p))$. Only nuclei situated between the skeleton and the stroma contributed to these histograms. Features f_{46}, f_{47}, f_{48} and f_{49} were defined as the first, second, third and fourth central moments of $H^{\text{skel}}_{\text{str}}$. Features f_{50}, f_{51} and f_{52} were defined as the second, third, and fourth moments of $J^{\text{skel}}_{\text{str}}$ (first moment was omitted after consulting box plots).

2.6.3 Feature Normalization

Features f_1 through f_{52} were computed for all images in the training, validation and test sets. For the i-th feature, the minimum and maximum values of f_i in the training set were found; denote these f_i^{Tmin} and f_i^{Tmax}. All features from all images were then normalized:

$$\hat{f}_i = (f_i - f_i^{\text{Tmin}})/(f_i^{\text{Tmax}} - f_i^{\text{Tmin}}). \tag{9}$$

Features from the training set take on a value in the interval $[0, 1]$, however, since features from the validation and test sets may not fall within the interval $[f_i^{\text{Tmin}}, f_i^{\text{Tmax}}]$, features from images in these sets do not necessarily fall within $[0, 1]$.

ID	Feature
Compactness Features	
f_1	Compactness of epithelial region
Label Proportion Features	
f_2	Proportion of nuclei in epithelial
f_3	Proportion of non-nuclei tissue in epithelial
f_4	Proportion of non-tissue in epithelial
f_5	Proportion of nuclei over tissue in in epithelial
Thickness Features	
f_6	Thickness variance
f_7	Thickness histogram bin 1 of 7
f_8	Difference between thinnest 10% and thickest 10%
Co-occurrence Features	
f_9	Texture correlation of $G_{\text{epi}}(I)$ region of red channel.
f_{10}	Texture entropy of $G_{\text{epi}}(I)$ region of red channel
Contour Fringe Features	
f_{11}	Contour fringe average
f_{12}	Contour fringe difference of averages
Eccentricity of Nuclei Feature	
f_{13}	Hist of eccentricity of nuclei components bin 1 of 8
f_{14}	Hist of eccentricity of nuclei components bin 5 of 8
f_{15}	Hist of eccentricity of nuclei components bin 6 of 8
f_{16}	Hist of eccentricity of nuclei components bin 7 of 8
f_{17}	Hist of eccentricity of nuclei components bin 8 of 8
Largest Hole and Lumen Component Features	
f_{18}	Scaled area of largest hole
f_{19}	Compactness of largest hole
f_{20}	Average scaled area of holes
f_{21}	Number of holes
f_{22}	Size of the largest non-tissue component
f_{23}	Eccentricity of the largest non-tissue component
f_{24}	Compactness of the largest non-tissue component

ID	Feature
V-cell Features	
f_{25}	Number of V-cells in the first layer from lumen
f_{26}	Number of V-cells in the first layer from stroma
f_{27}	Average eccentricity of first layer V-cells from lumen
f_{28}	Proportion of nuclei in first layer of V-cells from stroma
f_{29}	Average compactness of V-cells
f_{30}	Average eccentricity of V-cells
f_{31}	Average minor axis of V-cells
First Layer Ratios and Nuclei Distance Histogram Features	
f_{32}	Ratio of first layers
f_{33}	Variance of $H^{\text{str}}_{\text{lumen}}$.
f_{34}	Skew of $H^{\text{str}}_{\text{lumen}}$.
f_{35}	Kurtosis $H^{\text{str}}_{\text{lumen}}$.
f_{36}	Mean of $J^{\text{str}}_{\text{lumen}}$.
f_{37}	Variance of $J^{\text{str}}_{\text{lumen}}$.
f_{38}	Skew of $J^{\text{str}}_{\text{lumen}}$.
f_{39}	Mean of $H^{\text{skel}}_{\text{lumen}}$.
f_{40}	Variance of $H^{\text{skel}}_{\text{lumen}}$.
f_{41}	Skew of $H^{\text{skel}}_{\text{lumen}}$.
f_{42}	Kurtosis of $H^{\text{skel}}_{\text{lumen}}$.
f_{43}	Mean of $J^{\text{skel}}_{\text{lumen}}$.
f_{44}	Skew of $J^{\text{skel}}_{\text{lumen}}$.
f_{45}	Kurtosis of $J^{\text{skel}}_{\text{lumen}}$.
f_{46}	Mean of $H^{\text{skel}}_{\text{str}}$.
f_{47}	Variance of $H^{\text{skel}}_{\text{str}}$.
f_{48}	Skew of $H^{\text{skel}}_{\text{str}}$.
f_{49}	Kurtosis of $H^{\text{skel}}_{\text{str}}$.
f_{50}	Variance of $J^{\text{skel}}_{\text{str}}$.
f_{51}	Skew of $J^{\text{skel}}_{\text{str}}$.
f_{52}	Kurtosis of $J^{\text{skel}}_{\text{str}}$.

Table 3: Summary of epithelial region description features.

2.7 Classification

To classify the cysts, two classifiers were used, both from the Weka (Hall et al., 2009) suite of machine learning classifiers. The classifiers were a support vector machine (SVM) and an ensemble method, bagging, using logistic regression (without a regularization term) as its base learner (BLR). The classifier names within Weka are SMO and Bagging using Simple Logistic.

A SVM, given a two class problem, is a classifier which aims to find a linear decision boundary, or margin, such that the margin has maximum separation between the two classes. The WEKA implementation uses pairwise classification for multiple class data. To get around the limitation of linear decision boundaries, SVMs also allow for a potentially non-linear mapping of the feature space by means of a kernel function. This permits non-linear decision boundaries. In practice, classes are often not completely separable, nor does finding a completely separable boundary often generalize to new data. Because of this, the concept of a soft margin was introduced. This allows for a trade off between minimizing misclassifications and maximizing the margin. This trade off, or regularization, is controlled by a parameter C. The margin is a hard margin when $C = \infty$. See Burges (1998) for a tutorial on SVMs.

Bagging, or bootstrap aggregating (Breiman, 1996), is an ensemble method where the training set is sampled randomly with replacement. Each of these sample subsets is the same size. From each of these subsets, a model is trained using a secondary classifier or a base learner, which in our case was logistic regression. Then for classification, each model is used to predict the class. Each model is given equal weight and a majority vote determines the overall class. The two parameters for bagging are the size of the sample, and the number of models created. Because we used logistic regression without a regularization term, the base learner has no parameters.

3 Results

3.1 Experiment 1: Training, Validation and Testing

The SVM and BLR classifiers were trained using the set of feature vectors computed for the images from the training set ("Training Set Region Descriptions" in Figure 2).

The validation set was used to optimize the classifier parameters; this is the portion of Figure 2 where "Trained Classifier" and "Validation Set Region Descriptions" combine to form the "Optimized Classifier". All results given about the validation sets are after optimization. Optimization was done by hand. A number of combinations of parameter values were tried and the best preforming ones selected.

The SVM had two parameters requiring optimization: 1) the regularization term C; and 2) the kernel function. A value of 0.1 for C and a kernel using both quadratic and linear terms resulted in the best classifier performance on the validation set. For the Bagging classifier, there are two parameters to choose when using boosting: 1) the subset size; and 2) the number of models. A value of 86% of the original training data for the subset size was chosen and 40 models were formed. These were selected based on performance of a single run on the validation set. Since no regularization term for logistic regression was used, there were no tuneable parameters for the base learner.

The test set was used to evaluate the performance of the optimized classifier. The classifiers had not previously been exposed to the data in the test set and none of the test set data had any influence on classifier choice, feature selection, or classifier training. The performance of the classifiers on this set are the main results of this paper.

3.1.1 Results for SVM Classifier

Overall, the optimized SVM classifier correctly classified 83.8% and 92.3% of the validation and test sets, respectively (Table 4). Table 4 also shows the kappa statistics for the test and validation sets of 0.782 and 0.897, respectively. The kappa statistic is a chance-corrected measure of the agreement between the classifier predictions and the true classes of samples where 1.0 indicates perfect agreement.

	Validation Set	Test Set
Correctly classified samples	31	36
Incorrectly classified samples	6	3
Total	37	39
Percentage correct:	83.8%	92.3%
Kappa statistic	0.782	0.897

Table 4: Overall SVM classifier performance.

	DCyst	LPC	OKC	GOC	Weighted Avg.
TP Rate (Recall)	0.80	1.0	0.909	1.0	0.923
FP Rate	0.034	0.0	0.036	0.032	0.026
Precision	0.889	1.0	0.909	0.889	0.923
F-Measure	0.842	1.0	0.909	0.941	0.922
ROC Area	0.924	1.0	0.974	0.984	0.970

Table 5: Detailed SVM classifier performance on the test set. For each column, predictions of the indicated class are treated as positives and any other predictions are considered negatives.

Detailed performance statistics for each cyst class were also computed; we briefly define them in this paragraph. A *true positive* (TP) with respect to class ω is a sample from class ω that was labeled as class ω. A *false positive* (FP) with respect to class ω is a sample from some class $\xi \neq \omega$ which was incorrectly labeled as class ω. A *false negative* (FN) with respect to class ω is a sample from class ω which was incorrectly labeled as some other class $\xi \neq \omega$. A *true negative* (TN) with respect to class ω is a sample from class $\xi \neq \omega$ that was correctly not given a label of ω. The *TP Rate* or *recall* with respect to class ω is $TP/(TP + FN)$, the proportion of samples from ω that were correctly classified. The *FP rate* with respect to class ω is $FP/(FP + TN)$, the proportion of samples not in ω that were incorrectly classified as ω. The *precision* is $TP/(TP + FP)$, the proportion of samples labeled as ω that were labeled correctly. The F-measure is $(2 \times \text{recall} \times \text{precision})/(\text{recall} + \text{precision}) = 2TP/(2FP + FP + FN)$, the harmonic mean of precision and recall. The area under the ROC curve reflects the performance of the classifier for a particular class when members of all other classes are considered "negatives". The ROC curve parameter t is the threshold for the probability that a classifier prediction is a "positive". For example, the operating point $t = 0.7$ represents the case where a prediction is considered a "positive" if the probability of the sample belonging to the class equals or exceeds 0.7.

Table 5 gives the detailed classifier performance on the test set on a per-class basis. From the table we can see that overall, performance was worst for dentigerous cysts, with an F-measure of 0.842. LPC were all correctly classified, and no non-LPC sample was incorrectly classified as LPC. Performance for OKC and GOC was between that of LPC and DCyst with F-measures of 0.909 and 0.941, respectively. The *weighted average* column in Table 5 is the average of the measures for each class weighted according to the number of samples in each class.

Table 6 shows the confusion matrices for the validation set and the test set. All classification errors, save one image in the validation set, were either misclassified dentigerous cysts, or non-dentigerous cysts misclassified as dentigerous.

3.1.2 Results for BLR

The sampling of learning sets for bagging in BLR is random so test set results for BLR were taken over 60 runs (using different random seed values). For this reason the BLR results are reported as means and standard deviations over these 60 runs with the exception of the results on the validation which was optimized based on the performance of a single run.

Table 7 contains the overall performance results for the BLR classifier which were slightly better, on

Validation Set				Test Set				
Classified As				Classified As				
a	b	c	d	a	b	c	d	
8	2	0	0	8	0	1	1	a = DCyst
2	8	0	0	0	10	0	0	b = LPC
1	0	9	0	1	0	10	0	c = OKC
0	1	0	7	0	0	0	8	d = GOC

Table 6: Confusion matrices of the SVM classifier for the validation and test sets.

	Validation Set	Test Set
Correctly classified samples	34	37.2 ± 0.756
Incorrectly classified samples	3	1.8 ± 0.756
Total	37	39
Percent Correct	91.9%	$95.4 \pm 1.94\%$
Kappa Statistic	0.891	0.938 ± 0.026

Table 7: Overall BLR classifier performance.

average, than the results for the SVM classifier. For the test set, the coefficient of variation of the percentage of correct classifications over the 60 runs was 2.01%.

Table 8 shows the detailed BLR classifier performance on the test set on a per-class basis. Compared with the SVM classifier, on average, the BLR improves the F-measure for the DCyst class from 0.842 to 0.937 and the F-measure for OKC from 0.909 to 0.972 at the expense of a slight decrease in the F-measure for LPC and GOC. The confusion matrices for BLR are given in Table 9 where there is no clear pattern of confusion.

3.2 Experiment 2: Cross Validation of All Sets

Since we are considering a four-class problem, our data sets are relatively small in that we don't have a very large number of samples from each class. As such, a secondary experiment was conducted which attempts to show how the classifier might perform on larger data sets. The training, validation, and test sets were merged into a single data set on which a leave-one-out cross-validation experiment was performed. Since cross-validation tends to over-fit the data on which it is trained, the same normalization and classifier parameters that were determined in Experiment 1 were used for Experiment 2 in an attempt to ameliorate this problem. As a result, in the cross-validation experiment, the classifiers were optimized to a part of the data, but not the complete data set.

	DCyst	LPC	OKC	GOC	Weighted Avg.
TP Rate (Recall)	0.896 ± 0.02	0.978 ± 0.042	0.951 ± 0.059	1.0 ± 0.0	0.954 ± 0.019
FP Rate	0.005 ± 0.013	0.013 ± 0.017	0.001 ± 0.007	0.039 ± 0.014	0.013 ± 0.006
Precision	0.984 ± 0.038	0.965 ± 0.045	0.996 ± 0.019	0.869 ± 0.037	0.959 ± 0.018
F-Measure	0.937 ± 0.025	0.97 ± 0.029	0.972 ± 0.035	0.93 ± 0.022	0.954 ± 0.019
ROC Area	0.995 ± 0.002	0.999 ± 0.001	1.0 ± 0.002	0.997 ± 0.002	0.998 ± 0.001

Table 8: Detailed BLR classifier performance on the test set. For each column, predictions of the indicated class are treated as positives and any other predictions are considered negatives.

Validation Set				Test Set				
Classified As				Classified As				
a	b	c	d	a	b	c	d	
9	1	0	0	8.96 ± 0.198	0.0 ± 0.0	0.04 ± 0.198	1.0 ± 0.0	a = DCyst
1	9	0	0	0.0 ± 0.0	9.78 ± 0.418	0.0 ± 0.0	0.22 ± 0.418	b = LPC
0	0	10	0	0.16 ± 0.37	0.38 ± 0.49	10.5 ± 0.646	0.0 ± 0.0	c = OKC
0	1	0	6	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	8.0 ± 0.0	d = GOC

Table 9: Confusion matrices of the BLR classifier for the validation and test sets.

	SVM	BLR
Correctly classified samples	131	134.1 ± 1.37
Incorrectly classified samples	18	14.9 ± 1.37
Total	149	149
Percent Correct	87.9%	90.0 ± 0.92%
Kappa Statistic	0.838	0.866 ± 0.012

Table 10: The classification rates for the SVM and BLR when using leave-one-out cross validation on the merged training, validation and test data sets.

3.2.1 Results for Cross Validation of All Sets

Table 10 shows the overall performance of the SVM and BLR classifiers in the cross-validation experiment. The classification rates fall between the results on the validation set and on the test set from Experiment 1.

Table 11 shows the confusion matrices for both classifiers. Once again, dentigerous cysts are the most frequent source of confusion; most classification errors are misclassified DCyst samples, or non-DCyst samples misclassified as DCyst.

3.3 Experiment 3: Three class problem without DCyst

Experiment 1 was repeated after omitting the DCyst samples to determine how well the features distinguish between the remaining classes and to confirm the hypothesis that it is the DCysts which caused most of the misclassifications in Experiment 1.

For Experiment 3, the same features and parameter settings for the SVM and BLR from the first experiment were used. Parameters were reevaluated for both classifiers but the original parameters continued to exhibit the best performance. The results for the BLR test set were averaged over 60 runs.

SVM				BLR				
Classified As				Classified As				
a	b	c	d	a	b	c	d	
34	3	3	0	34.2 ± 1.2	2.96 ± 0.198	2.5 ± 1.11	0.34 ± 0.557	a = DCyst
4	36	0	0	1.82 ± 0.691	35.2 ± 0.708	0.94 ± 0.24	2.02 ± 0.141	b = LPC
4	1	36	0	1.2 ± 0.404	1.0 ± 0.0	38.8 ± 0.404	0.0 ± 0.0	c = OKC
3	0	0	25	1.1 ± 0.364	0.98 ± 0.141	0.0 ± 0.0	25.4 ± 3.67	d = GOC

Table 11: The confusion matrices for the SVM and BLR using leave-one-out cross validation on the merged training, validation and test data sets.

	SVM Validation	SVM Test	BLR Validation	BLR Test
Number Correct	26	29	26	28.1 ± 0.252
Number Incorrect	1	0	1	0.933 ± 0.252
Total	27	30	27	30
Percent Correct	96.2%	100.0	96.2%	96.78 ± 0.867%
Kappa Statistic	0.943	1.0	0.943	0.951 ± 0.013

Table 12: The classification rates for the SVM and BLR for both the validation and test sets for the three class problem.

SVM Validation			SVM Test			
Classified As			Classified As			
a	b	c	a	b	c	
10	0	0	10	0	0	a = LPC
0	10	0	0	11	0	b = OKC
1	0	6	0	0	8	c = GOC

BLR Validation			BLR Test			
Classified As			Classified As			
a	b	c	a	b	c	
10	0	0	10.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	a = LPC
0	10	0	0.933 ± 0.252	10.1 ± 0.252	0.0 ± 0.0	b = OKC
1	0	6	0.0 ± 0.0	0.0 ± 0.0	8.0 ± 0.0	c = GOC

Table 13: The confusion matrices for the SVM and BLR for both the validation and test sets for the three class problem.

3.3.1 Results for Three class problem without DCyst

Table 12 contains classification statistics and Table 13 the confusion matrices for the three-class case. There was a significant increase in overall classification performance for both the test and validation sets for the SVM. Performance of BLR also increased, though less significantly. For both classifiers, only one image in the validation set was incorrectly classified.

4 Discussion

Dentigerous cysts posed the greatest challenge in this project. The other cyst classes all have features which fairly obviously separate them from the others, as confirmed with Experiment 3. The thickness measures separate the LPC from the other classes, the features detecting parakeratin separate the OKC, and the largest holes or component features separate the GOC. There is no such definitive characteristic for DCysts. This is not surprising as the DCyst class seems to have the largest within-class variance. Despite the classification rates on the test set, robust detection of DCyst is lacking. This can be seen in the performance of the SVM on the validation set in Experiment 1, where the confusion matrix in Table 6 clearly shows that DCyst is the major problem. This can be seen again in the confusion matrices for the leave-one-out cross validation in Experiment 2 (Table 11). Finding a feature which separates DCyst from the others is a priority for future work. Another possibility for improving the classification rate for D-Cysts is making use of of radiographic information; in diagnosis of dentigerous cysts, oral pathologists rely heavily on the fact that these cysts present as a non-keratinized cyst around the crown of an impacted tooth.

Normally, it is expected that the performance on the validation set should exceed that of the test set. The fact that this didn't happen can be explained by the size of our data sets. Because they are relatively

small a difference of a few images can lead to a sizeable difference in performance rates. Combined with the previous evidence of a lack of a definitive feature for distinguishing DCysts, the anomalous relative performance of the validation and tests sets can be explained by small, chance variations in the number of DCyst misclassifications.

The performance of the SVM was relatively stable throughout the project. Small changes in features often caused little or no change to the SVM's performance on the training set. In comparison, the performance of logistic regression on the training set was relatively unstable but this instability was mitigated by using bagging with logistic regression.

A significant number of features are based on the skeleton of the epithelial region. Skeletons are often very noisy and tend to contain many branches or unwanted spurs. Pruning these branches should improve the performance of these features.

A number of other features were implemented, but ultimately not used. The first of these was a histogram of the nuclei orientations. Orientation was determined relative to the orientation of the skeleton, which is necessary for rotational invariance. This feature was expected to distinguish between OKC which often have elongated nuclei perpendicular to the epithelial/stoma interface. That these features were not useful speaks to the amount of noise in the k-means classification and the region skeletons.

A second group of poor features were features analogous to f_{28} and f_{32} for V-cell layers other than the first layers, including features similar to f_{32} but which included the first 2, 3, etc. V-cell layers from the lumen and stroma, respectively.

A third group of poor features were alternatives to the Nuclei Distance Histogram Features which make up a significant portion of our features (features f_{33} through f_{52}). In an attempt to eliminate the parameter specifying the number of bins for these histograms, we replaced these features with the central moments computed from the un-binned $h(p)$ values (equivalent to using an infinite number of histogram bins). However, this reduced classifier prediction rates by 3–5%. The histogram moments may have worked better because of the number-of-bins parameter, which can be tuned to fit the training data.

The next step toward an automated cyst classification algorithm is to study the integration of our classifiers with an automatic or semiautomatic epithelium segmentation algorithm (e.g. Eramian et al. (2011)) to determine how tolerant our classifiers are to errors in the segmentation. The success of cyst class prediction also motivates expending the effort to improve the existing epithelium segmentation algorithm to reduce as much as possible the segmentation errors that must be tolerated by the classifier.

5 Conclusion

Based on the minimum and maximum performances observed in Experiments 1 and 2, the type of odontogenic cysts can be predicted from among four classes of odontogenic cyst (dentigerous, lateral periodontal, odontogenic keratocyst, and glandluar) between 83.8% and 92.3% of the time with a support vector machine classifier and between $90 \pm 0.92\%$ and $95.4 \pm 1.94\%$ with bagging with logistic regression as the base learner.

Based on Experiment 3, omission of dentigerous cysts from the data set improves the classification rate of the remaining three cyst types to about 96.2% (only one misclassification) for both the support vector machine and bagging with logistic regression.

6 Acknowledgements

This work was supported by the Natural Sciences and Engineering Research Council of Canada through grant number RGPIN262027-2007 and the Undergraduate Student Research Awards program.

References

Astola, J. (1990). Vector median filters. *Proceedings of the IEEE*, 78(4):678–689.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24:123–140.

Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167.

Cychosz, J. M. (1994). Efficient binary image thinning using neighborhood maps. In *Graphics Gems IV*. Academic Press.

Eramian, M., Daley, M., Neilson, D., and Daley, T. (2011). Segmentation of epithelium in H&E stained odontogenic cysts. *Journal of Microscopy*, 244(3):273–292.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The WEKA data mining software: An update. *SIGKDD Explorations*, 11(1):10–18.

Haralick, R., Shanmugam, K., and Dinstein, I. (1973). Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-3(6):610–621.

Landini, G. and Othman, I. E. (2003). Estimation of tissue layer level by sequential morphological reconstruction. *Journal of Microscopy*, 209(2):118–125.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In Le Cam, L. M. and Neyman, J., editors, *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, volume 1, pages 281–297. University of California Press.

Meijster, A., Roerdink, J., and Hesselink, W. (2002). A general algorithm for computing distance transforms in linear time. *Computational Imaging and Vision*, 18(8):331–340.

Spath, H. (1985). *Cluster dissection and analysis: theory, FORTRAN programs, examples*. Computers and their applications. Chichester Publishing; Halsted Press [distributor], Chichester and New York. Translated by J. Goldschmidt.