INFORMATICA, 2018, Vol. 29, No. 1, 75–90

DOI: http://dx.doi.org/10.15388/Informatica.2018.158

Machine Learning Based Classification of Colorectal Cancer Tumour Tissue in Whole-Slide Images

Mindaugas MORKŪNAS[1,2*], Povilas TREIGYS[1], Jolita BERNATAVIČIENĖ[1], Arvydas LAURINAVIČIUS[2], Gražina KORVEL[1]

[1]*Institute of Data Science and Digital Technologies, Vilnius University Akademijos str. 4, LT-08663 Vilnius, Lithuania*

[2]*National Center of Pathology, Affiliate of Vilnius University Hospital Santaros klinikos, P. Baublio str. 5, LT-08406 Vilnius, Lithuania*

e-mail: mindaugas.morkunas@mii.vu.lt, povilas.treigys@mii.vu.lt, jolita.bernataviciene@mii.vu.lt, arvydas.laurinavicius@vpc.lt, grazina.korvel@mii.vu.lt

Received: September 2017; accepted: February 2018

Abstract. The recent introduction of whole-slide scanning systems enabled accumulation of high-quality pathology images into large collections, thus opening new perspectives in cancer research, as well as new analysis challenges. Automated identification of tumour tissue in the whole-slide image enables further use of developed grading systems that classify tumour cell abnormalities and predict tumour developments. In this article, we describe several possibilities to achieve epithelium-stroma classification of tumour tissues in digital pathology images by employing annotated superpixels to train machine learning algorithms. We emphasize that annotating superpixels rather than manually outlining tissue classes in raw images is less time consuming, and more effective way of producing ground truth for computational pathology pipelines. In our approach feature space for supervised learning is created from tissue class assigned superpixels by extracting colour and texture parameters, and applying dimensionality reduction methods. Alternatively, to train convolutional neural network, labelled superpixels are used to generate square image patches by moving fixed size window around each superpixel centroid. The proposed method simplifies the process of ground truth data collection and should minimize the time spent by a skilled expert to perform manual annotation of whole-slide images. We evaluate our method on a private data set of colorectal cancer images. Obtained results confirm that a method produces accurate reference data suitable for the use of different machine learning based classification algorithms.

Key words: tumour, whole-slide image, machine learning, superpixel, ground truth, colour and texture features, convolutional neural network.

1. Introduction

Most often solid tumour cancers are diagnosed by medical pathologist visually inspecting tissue slides. Pathology slides contain important features – spatial information of

*Corresponding author.

tumour cell morphology and tumour microenvironment that can't be captured by other routinely used diagnostic methods. Precise quantification of parameters such as lengths, surface areas, object counts, perimeter, and distance to nearest neighbour is often required. Based on these observations multiple tumour grading systems that classify tumour cell abnormalities and predict tumour developments are available (Lakhani *et al.*, 2012). Confirmation of disease presence, outcome prediction, and choice of therapy explicitly rely on information present in pathology slides. Better cancer diagnostics and advance in treatment evolve with increasing understanding of cancer biology and are driven by research. Visual analysis of pathology tissue slide is time-consuming, and high-level expertise is demanding procedure since pathologist has to recognize subtle changes in cell and tissue patterns. The recent introduction of whole-slide imaging systems, active development of computer vision field greatly changed pathology analysis. Because of the constant discovery of new tumour tissue biomarkers, there is a substantial interest in advanced computational pathology algorithms that would accomplish highly specific tasks of research. Digitized pathology slides (whole-slide images or WSIs) are commonly subjected to computational pathology pipelines to precisely detect, classify, quantify, and segment multiple types of histological objects. Technology has already been shown to be useful for diagnostic tasks effectively discriminating malignant tissue from premalignant, detecting metastasis (Bejnordi *et al.*, 2016; Litjens *et al.*, 2016), and grading tumours (Ertosun and Rubin, 2015).

Qualitative and quantitative analysis of histology objects in a typical pathology image is a complex task that most simply may be viewed as consisting of image segmentation step, feature measurement, and machine learning based classification of segmented image primitives. The latter can be subdivided into traditional (e.g. decision tree learning, support vector machines) and deep learning methods (convolutional neural networks or CNNs). Both approaches as an input take large amounts of labelled data to learn features with the certain degree of interpretability (such as texture or colour) and adapt model parameters according to the distance between the produced and the desired outputs. Finally, predictions on new instances of the same data type have to be made. One of the first computational tasks and a common intermediate goal in comprehensive pathology image analysis is a classification of malignant (or tumour) tissue into epithelium and stromal compartments. The reasoning behind this specific task is that it helps to build a picture of where and to what extent a particular cancer biomarker is present in the tissue. Modern methods of prognostic and predictive stratification of cancer patients evaluate biomarker positive cells distributions in each tissue compartment (McLaughlin *et al.*, 2016; Dunne *et al.*, 2017). Also, for certain types of cancer, tumour epithelium-stroma ratio alone is recognized as the independent prognostic indicator (Panayiotou *et al.*, 2015). The very recent rise of cancer immunotherapy research also requires precise tumour microenvironment compartmentalization algorithms to identify further, and analyse tumour tissue infiltrating immune cells that are known to kill cancer cells (Emens, 2017).

In this paper, we focus on epithelium-stroma classification of tumour tissue. Classification is performed on superpixels, but more importantly, we effectively employ superpixels to collect ground truth data. We review related works on tumour epithelium-stroma

classification in Section 2 and specifically highlight difficulties of current state-of-the-art methods in obtaining ground truth data. In Section 3 we describe the data. In Section 4 we present ways superpixels were processed to be suitable for application in machine learning experiments. Section 4 also describes machine learning approaches that were applied in our study. Section 5 presents experimental results and comparative analysis. Section 6 concludes the paper.

2. Related Work

The majority of publications in the field are focused on breast or prostate cancers. Multiple works (Ahammer *et al.*, 2009; Linder *et al.*, 2012; Bianconi *et al.*, 2015; Bunyak *et al.*, 2015; Sethi *et al.*, 2016; Nava *et al.*, 2016) explore traditional machine learning methods to achieve classification based on handcrafted features extracted from pathology images. In these studies, local binary patterns, Gabor filters, Haralick texture, Tshebichef moments, and colour/intensity parameters were used to build feature space. Feature extraction for epithelium-stroma classification is performed on segmented image primitives. Pathology image segmentation is performed by employing rectangular image blocks (Ahammer *et al.*, 2009), overlapping square patches (Linder *et al.*, 2012), multi-resolution square image blocks (Bianconi *et al.*, 2015), various superpixel approaches – Normalized Cut, Simple Linear Iterative Clustering, Hierarchical Fuzzy C-Means, and also Multiresolution Superpixels (Bunyak *et al.*, 2015; Sethi *et al.*, 2016). For several recent years, solutions for epithelium-stroma classification problem were evolving towards increasing number and complexity of features extracted. Sethi *et al.* (2016) reports the use of Wndchrm (originally by Shamir *et al.*, 2008) software to build 93 features space to discriminate between epithelium and stroma compartments of prostate cancer tumours. As described in the same study, Wndchrm can automatically extract between 1000 and 3000 predefined image features. In a slightly different tumour classification task, relating to normal-malignant breast cancer tissue classification Bejnordi *et al.* (2016) describes 256 handcrafted features extracted per superpixel.

Recently, methods employed to classify tumour tissue shifted from traditional machine learning approaches to deep convolutional neural networks (Huang *et al.*, 2015; Sethi *et al.*, 2016; Litjens *et al.*, 2016; Xu *et al.*, 2016) – an approach that doesn't require supervised feature extraction, but rather relies on end-to-end feature learning from small, square, fixed size, overlapping image patches.

Both traditional and deep learning classification approaches require labelled data, whether to build training data sets or evaluate classifier performance by comparing algorithm predictions to human output. Nearly all published methods are to some extent limited by complicated annotation of WSIs since they rely on pixel-level ground truth images. Annotating pathology images is extremely labour intensive task involving manual delineation of tumour compartments. Skilled experts are not always available for the time needed to perform precise pixel-level annotations. Major obstacles arising from this reason are – lack of precision due to hast, or due to lower magnification images used; incorrect annotations (both false positives and false negatives). This problem is well outlined in Cruz-Roa *et al.* (2014). Authors reached 79% accuracy with handcrafted features

and classical machine learning approaches, and 84% accuracy with deep learning model. Authors also identify the reason of misclassification to be mainly due to the pathologist's not detailed annotations. The process of generating ground truth data is clearly neglected in research papers. Only scarce descriptions are often given, commonly referring manual work done by a skilled specialist. In the paper of Ahammer *et al.* (2009) class labels were assigned per single field of view area at 400× magnification (768 × 589 pixels). Linder *et al.* (2012) only describes that annotation was carried out by one of the researchers and verified by a pathologist. Bianconi *et al.* (2015) states the use of histologically-verified, well-defined various size and resolution images. As deep classification algorithms require significantly larger data sets to efficiently train, it can be noted that with increasing use of latter methods ground truth data mining draws more attention. Sethi *et al.* (2016) refers commercial software used to delineate regions serving as ground truth for both training and testing classifiers. Authors also deal with "impractically tedious" annotation of entire images. Bunyak *et al.* (2015) indicates that the process is also subjective because of the variance either in judgment or in expert dexterity. Authors developed KolamGT software to assist expert annotations. Very recent paper by Huang *et al.* (2015) highlights the problem early in the abstract, and the research is focused on taking advantage of using unlabel data in a training data set via transfer learning.

As a common task epithelium-stroma classification is often fine-tuned to adapt to highly specific end-goals of comprehensive pathology research. Whether a new, unexplored cancer type is concerned, or emerging biological tumour properties need to be evaluated in the field it holds capacity for developing new algorithms to precisely classify tumour tissue.

3. Description of Data Sets

3.1. *Whole-Slide Images*

The method presented in this paper not only employs superpixels, and superpixel derived square image patches classification by different machine learning algorithms but also uses superpixels to annotate ground truth in highest magnification images. Our goal was to show that annotated superpixels retain enough annotation precision, and carry enough colour and texture information to achieve classification accuracy comparable to the results reported by other teams. To demonstrate our approach we have used 3 WSIs of hematoxylin and eosin stained colorectal cancer tissues (see Fig. 1). Specimens were produced in the National Center of Pathology, Lithuania. Tumour-containing zones were manually marked by a pathologist on a tissue slides directly prior to scanning. Digital images were captured using the Aperio ScanScope XT Slide Scanner (Aperio Technologies, Vista, CA, USA) under 20× objective magnification.

3.2. *Superpixels*

From each tumour zone in WSIs 284 RGB tiles of 1000 × 2000 pixels from highest magnification layers were manually selected for analysis. Consecutively each tile was segmented

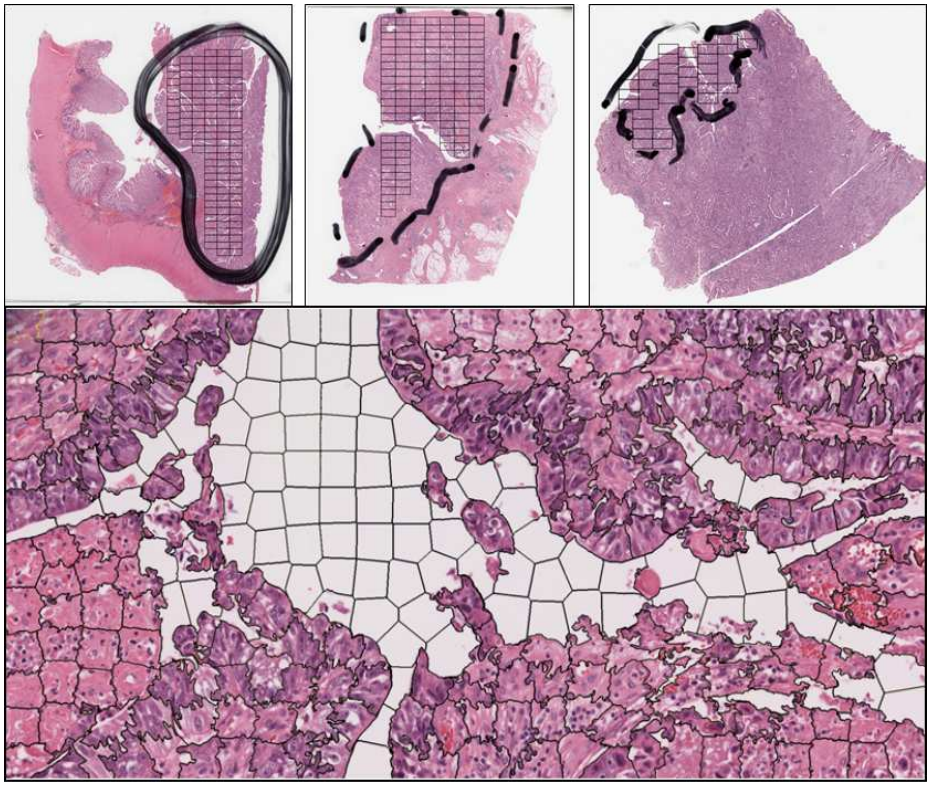

Fig. 1. Top row. Macro-scale annotated WSIs. Tiles selected for analysis are shown inside tumour containing areas marked by a pathologist. Bottom row. A single segmented tile is representing 1000 $\times$ 2000 pixels region in a WSI.

using Simple Linear Iterative Clustering (SLIC) algorithm (Achanta *et al.*, 2012) into approximately 350 superpixels (see bottom row in Fig. 1).

Segmented tiles with boundaries between resulting superpixels highlighted were subjected to micro-scale annotation by the pathologist. Annotation was performed by assigning each superpixel to the tumour (epithelium), stroma or background classes. The whole process of tissue annotation was strictly limited to superpixel selection and label assignment. No manual outlining/boundary drawing was required. A total of 70997 superpixels were assigned whether tumour (epithelium) or stroma class. Background (or otherwise called "glass") is non-informative part of WSI, so it was removed from all data sets. All superpixels mainly representing background could be effectively filtered by mean and dispersion of pixel values in a green colour channel of a superpixel (background superpixels identified with mean pixel value greater than 190, and standard deviation less than 30 in a green colour channel). Also, any multi-class or no-class assigned superpixels were excluded, and remaining superpixels were used to construct 4 data sets (see Table 1). Equal class proportions were maintained in training and validation subsets, and the testing subset was composed to represent real-life data composition (i.e. disbalanced). "General" data set was produced by combining all superpixels from all smaller data sets.

Table 1
Data sets of superpixels.

Data	No. of	Training subset		Validation subset		Testing subset	
sets	tiles	Tumour	Stroma	Tumour	Stroma	Tumour	Stroma
Sample1	125	10000	10000	750	750	9416	1205
Sample2	125	10000	10000	750	750	533	8699
Sample3	34	2200	2200	150	150	3269	175
General	284	28000	28000	2000	2000	7068	3929

3.3. *Colour Descriptors*

For each 2D superpixel (e.g. a grayscale superpixel) two statistics were calculated – mean and standard deviation of the pixel value. For a 3D superpixel (e.g. an RGB superpixel) we calculated eight statistics – mean and standard deviation of the pixel value for each colour channel separately and also per RGB superpixel.

3.4. *Texture Descriptors*

The texture of a 2D superpixel (e.g. a grayscale superpixel) has been measured by calculating spatial gray-level co-occurrence matrix for four directions, or 13 directions for a 3D superpixel (e.g. an RGB superpixel) and 1px displacement vector. From each of the resulting co-occurrence matrices we calculated 13 parameters as follows: angular second moment, contrast, correlation, sum of squares, inverse difference moment, sum average, sum variance, sum entropy, entropy, difference variance, difference entropy, information measure of correlation 1, and information measure of correlation 2 (originally described in Haralick, 1979). For each parameter mean value was obtained from all directions, thus resulting in final 13 descriptors.

3.5. *Dimensionality Reduction and Feature Space*

Each data set prepared according to experimental conditions (described in more detail in subsection 4.3) was subjected to principal component analysis (PCA). For further use of colour and texture descriptors in classification tasks factor analysis (FA) was performed. Number of components to be obtained in FA was selected from PCA results and was defined as a number of components explaining more than 99% of the variance in particular data set. Resulting factor scores were normalized to have zero mean and unit variance per data set per component. Normalized factor scores resulting from colour descriptors are referred here as “colour features”. Similarly, normalized factor scores resulting from texture descriptors are referred as “texture features”. We produced “combined features” by combining colour and texture descriptors and applying similar factorization procedure. Normalized factor scores (features) were used as an input data for classification using traditional machine learning approaches (look in Section 4.1).

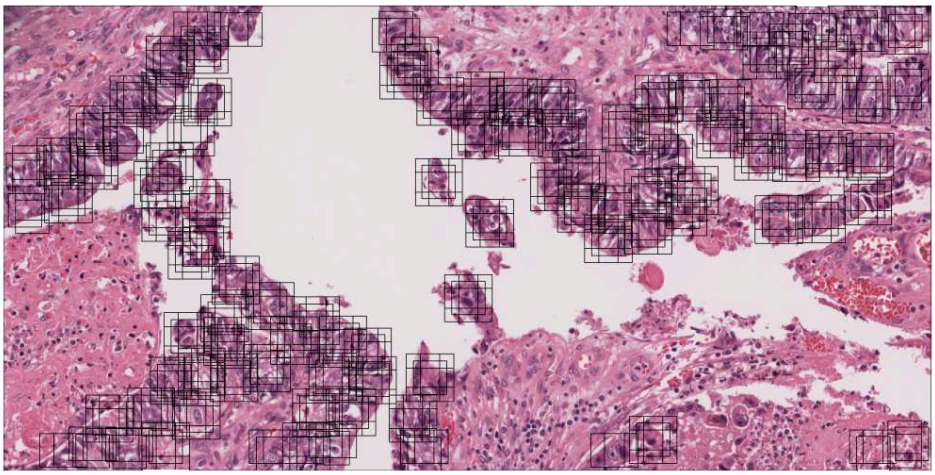

Fig. 2. Tumour-class patches selection for CNN training.

Table 2
Data set used to train CNN classifier.

	Tumour class	Stroma class
Training subset	40000	40000
Validation subset	25000	5000
Testing subset	26000	6000

3.6. *Image Patches*

Annotated superpixels from "Sample1" data set were used to generate patches for CNN based classifier training. Five overlapping 72×72 pixels RGB patches were selected in the original image around each superpixel mass centre by shifting the frame in each direction by 15 pixels (Fig. 2 and Table 2).

Training subset for CNN was built with equal proportions of "Tumour" and "Stroma" classes (Table 2).

4. Methodology

4.1. *Traditional Machine Learning Models*

Since background containing superpixels were removed from all data sets (described in 3.2), thus epithelium-stroma classification problem can be defined as two-class. Support Vector Machines (SVM), Random Decision Forest (RDF) and Multilayer Perceptron (MLP) as implemented by the scikit-learn Python package (Pedregosa *et al.*, 2011) as well as Deep Learning (DL) algorithm implemented by the TensorFlow Python package (Abadi *et al.*, 2016) were employed to solve classification problem.

C-Support Vector Classification algorithm (also called type 1 classification SVM) with Radial Basis Function kernel (RBF, an implementation in scikit-learn Python package assumes gamma parameter is equal to 1/number of features) was used (Cortes and Vapnik, 1995; Pedregosa *et al.*, 2011). With SVM classification is achieved by constructing hyperplane in multidimensional space to separate instances of different classes. Error minimization is employed in an iterative algorithm to find optimal hyperplane. Nonlinear classification is achieved by using kernel functions of which RBF is the most common choice (Chang *et al.*, 2010).

A Random decision forest classifier is a perturb-and-combine technique specifically designed for classification trees (Breiman, 2001). The principle is based on randomization – a bootstrap sample is drawn from a training set to build each of the trees, and the split on the node is chosen as best split among a random subset of the features. A predictor of RDF can be selected in a voting procedure or by averaging individual tree probabilistic predictions. Methods that use averaging are known to have relatively high variance due to feature subsetting at the base-level. Thus, after classification with RDF, a separate calibration of predicted probabilities was performed as post-processing using the sigmoid method and validation data sets (disjoint data used for training and calibration). In an RDF classifier, we empirically selected 25 estimators (base-level trees) and defined the maximum number of features to be used with an individual tree to be integer square root of a total number of features. The nodes were expanded until all leaves had at least two samples. All the input samples had equal weight.

A multilayer perceptron is a feedforward (a network whose neurons are connected only in a forward manner to form layers) neural network that is trained by backpropagation (Rumelhart *et al.*, 1986). Neurons in the network consist of the linear combiner and an activation function. In our model rectifier activation function was used – a function whose output is equal to the input if the input is positive, otherwise, the output is set to 0. Our network was constituted of one input layer containing inputs corresponding to the features in the data set, of two hidden layers with 1000 and 40 neurons respectively, and of an output layer corresponding to our two classes, "epithelium" and "stroma". A number of neurons in hidden layers was selected empirically. We have tested network performance while increasing number of neurons in both hidden layers from 100 and 10 up to 1000 and 40 respectively. Classifier performance was stable among all tested architectures with a mean variation of AUC being less than 0.0003. Connections between layers are defined by weights assigned in a learning process, and each neuron in a layer is connected to every neuron in the next layer. In a training phase, an output of the network (a probability matrix instead of discrete predictions) in an iterative manner is compared to ground truth labels of the input data, and a cross-entropy loss is computed as a negative log-likelihood of a classifier. The model uses limited-memory Broyden–Fletcher–Goldfarb–Shanno (LBFGS) method to optimize log-loss function with respect to the network parameters. LBFGS is an optimizer in the family of quasi-Newton methods (Malouf, 2002). The training phase is ended if the log-loss is not improving by at least a predefined amount (in our case 0.0001) for two consecutive iterations, or when a predefined number of iterations (in our case 200) had been reached.

Deep learning neural network here should be understood similarly to MLP – it is a feedforward neural network with multiple hidden layers between the input and output layers that is trained by backpropagation. The basic difference between DL and MLP used in our study is that DL network is constructed of more hidden layers with more neurons. We constructed DL network with three fully-connected layers (with the corresponding number of neurons – 1024, 512, 128). Neurons in the DL network use rectifier activation function (the same as in MLP). Between the last hidden layer and an output layer, we added a dropout layer. Dropout is a regularization technique for reducing overfitting in the model, by omitting specified proportion (50% in our case) of random neurons output. An output layer has two neurons that use softmax activation function – a function that given a vector outputs a probability distribution over all possible outcomes (in our case 2 – corresponding to two tissue classes). A number of neurons in all three hidden layers was selected empirically. We have tested network performance while increasing number of neurons in hidden layers from 1024, 512 and 128 up to 5120, 2560 and 640 respectively. Classifier performance was stable among all tested architectures with a mean variation of AUC being less than 0.002, which allowed us to select network architecture with the lowest time needed for computation. Log-loss of the model is computed as described in MLP. The network is trained using Adam (adaptive moment estimation, Kingma and Ba (2015) method with adjusted learning rate (0.000001) to minimize cross-entropy loss function.

4.2. *Deep Convolutional Neural Network Model*

Convolutional neural networks are a variation of deep learning neural networks. They are designed to process data that is given as arrays. Thus a local connectivity pattern and weight-sharing between neurons of adjacent layers become possible. Every convolutional layer of a CNN transforms one volume of activations to another through a differentiable function. Thus, neurons in a layer are only connected to a small region of the layer before it through a set of weights (a filter) and are distributed along the depth of an array. In the filtering operation filter is slid over an array with a fixed size stride. Filtering inside a layer is a discrete convolution. Thus the spatial output decreases with each layer. To reduce the number of parameters in the network pooling operation is employed. Pooling combines the outputs of a group of neurons in one layer into a single neuron in the next layer. The dense layer or fully-connected layer is similar to a hidden layer of an MLP or DL. We constructed a simple convolutional neural network to classify image patches prepared as described in Section 3.6. The model used was originally proposed in Litjens *et al.* (2016). We applied slight modifications to the original model. Our model consists of four convolutional layers (conv2d-1, conv2d-2, conv2d-3, conv2d-4), one densely connected layer (dense-1) and an output layer. We adjust filter sizes and strides to accept 72×72 pixel patches, we added batch normalization layers after each conv2d layer, the number of fully-connected layers was reduced to two (including the output layer). Adam method with adjusted learning rate (0.000001) was used to minimize cross-entropy loss function. The network's width (number of neurons in hidden layers) was optimized empirically (see Table 5 for results), and the model was built using the TensorFlow implementation in Python (Abadi *et al.*, 2016).

Table 3
Number of components retained in FA.

	Colour features	Texture features	Combined features
Experiment 1	4	8	8
Experiment 2	4	8	8
Experiment 3	2	8	8

4.3. *Description of Experiments*

Three experimental designs were considered regarding how superpixels were processed before feature extraction. In each experiment descriptors calculated for each superpixel were combined to form an array in which rows correspond to superpixels and columns to descriptors. Dimensionality reduction was applied to each data set as described in Section 3.5. Table 3 summarizes the output of dimensionality reduction.

Experiment 1. Superpixels in each data set were linearized (thus, superpixel shape information was lost), keeping the three colour channels. In this experimental design 4 colour features, 8 texture features, and 8 combined features were extracted (as described in Sections 3.3, 3.4 and 3.5) for every superpixel in all data sets (data sets as described in Table 1), and stored in a matrix with the shape of 20 columns, and number of rows corresponding to the numbers of superpixels per particular data set.

Experiment 2. Superpixels in each data set were bound in a rectangle minimum bounding box, keeping the three colour channels. Pixels inside bounding box, but outside the superpixel area were assigned zero values in each colour channel. Extracted features are the same as in Experiment 1, and are stored similarly.

Experiment 3. Superpixels in each data set were bound in a rectangle minimum bounding box. We assign zero values to the pixels inside bounding box, but outside the superpixel area, and convert bound superpixels to grayscale. We have chosen a colour-to-grayscale transformation that retains *Luminance* of the original image. A detailed review of colour-to-grayscale transformations frequently used in computer vision is given in Kanan and Cottrell (2012). Authors emphasize that *Luminance* is considered a good choice for texture recognition. Eight texture features were extracted as in previous designs. Two colour descriptors were extracted from grayscale images – mean pixel value and standard deviation per bound superpixel. Eight combined features were extracted as described in Section 3.5.

4.4. *Classifier Performance Metrics*

Classifier prediction typically is a numeric value on a continuous scale. A class label is produced by comparing a prediction (or a score) against a threshold value. To study the output of a binary classifier receiver operating characteristic (ROC) curves are often used (Fawcett, 2006). ROC curves are created by plotting the true positive rate against the false positive rate at a range of threshold values. After training the classifiers on a Training

subset, the Testing subset was used to measure each classifier performance by calculating the area under the receiver operating characteristic curve (AUC). Ten iterations were used to train and test the classifiers each time randomly assigning the data to the Training and Testing subsets with exactly the same proportions of class labels. Mean AUC and the variance (standard deviation) of AUC were calculated for each of the classifiers.

5. Results

5.1. *Results of Classification Using Traditional Machine Learning Approaches*

To demonstrate that in the process of ground truth annotation in extremely large WSIs careful, precise, and time-expensive manual outlining of tissue classes can be effectively replaced with simple procedure of annotating automatically generated precise and informative superpixels we attempted classification of all superpixels resulting from three experiments described in 4.3 by the four classifiers described in 4.1. The whole workflow can be summarized as follows:

1. Selected regions inside pathologist macro-annotated WSIs were segmented into superpixels (see Fig. 1).
2. Superpixels were visualized in the WSIs and subjected to pathologist micro-annotation.
3. Superpixels were processed using three techniques (Experiments 1, 2 and 3), and subjected to colour and texture feature extraction (see Sections 3.3, 3.4, and 3.5).
4. Classification using all methods described in 4.1 was attempted on colour features and texture features resulting from Experiments 1, 2, and 3.
5. Classifiers performance was compared with regard to the way superpixels were processed, and the features used to build feature space.

We have tested different types of features to be used in classification task, as well as several ways to extract them from superpixels. A general overview of superpixel classification results is presented in Table 4. Results for Experiment 3 differ from the first two quite substantially. For this experimental design test, AUC achieved were lowest (maximum AUC 0.9443), while with other strategies highest AUC values were above 0.97.

Results for experimental designs 1 and 2 can be reviewed together since they show little difference. Under these experimental conditions classifier performance as measured by area under the ROC curve is ranging from 0.9095 up to 0.9772 in single sample data sets, with SVM being the most accurate classifier. The highest test accuracy achieved on "General" data set was 0.9668 by MLP classifier. Contrary to what was expected, classifier performance is minimally affected by the data set size. Classification accuracies on "Sample 3" data set are comparable to the results of larger data sets (e.g. "Sample 1" data set is nearly four-fold larger than "Sample 3"). With smallest data set best classification results were produced by SVM classifier (AUC $= 0.9554$). No clear tendency towards the type of features extracted could be seen under the conditions of Experiments 1 and 2 for all data sets. In general, it was quite unexpected to find that features extracted from linearized

Table 4
Classifier performance results given as mean AUC and standard deviation (sd) of AUC acquired in 10 repetitions.

	Data set	Features	RDF	sd RDF	SVM	sd SVM	MLP	sd MLP	DL	sd DL
Experiment 1	General	Colour	0.9569	0.0018	0.9611	0.0025	0.9668	0.0024	0.9608	0.0020
		Combined	0.9577	0.0025	0.9613	0.0027	0.9660	0.0026	0.9580	0.0034
		Texture	0.9367	0.0022	0.9498	0.0022	0.9591	0.0034	0.9397	0.0062
	Sample 1	Colour	0.9420	0.0061	0.9504	0.0042	0.9476	0.0036	0.9532	0.0070
		Combined	0.9584	0.0051	0.9664	0.0036	0.9500	0.0038	0.9647	0.0058
		Texture	0.9301	0.0036	0.9401	0.0052	0.9281	0.0035	0.9357	0.0037
	Sample 2	Colour	0.9640	0.0055	0.9688	0.0030	0.9703	0.0034	0.9697	0.0183
		Combined	0.9728	0.0060	0.9772	0.0024	0.9684	0.0044	0.9744	0.0140
		Texture	0.9642	0.0055	0.9732	0.0027	0.9714	0.0041	0.9673	0.0205
	Sample 3	Colour	0.9421	0.0062	0.9425	0.0062	0.9222	0.0068	0.9376	0.0070
		Combined	0.9461	0.0094	0.9554	0.0051	0.9300	0.0055	0.9430	0.0106
		Texture	0.9343	0.0081	0.9431	0.0052	0.9341	0.0080	0.9272	0.0090
Experiment 2	General	Colour	0.9466	0.0023	0.9515	0.0014	0.9577	0.0031	0.9556	0.0026
		Combined	0.9533	0.0022	0.9620	0.0024	0.9666	0.0026	0.9575	0.0019
		Texture	0.9385	0.0035	0.9488	0.0021	0.9542	0.0031	0.9449	0.0040
	Sample 1	Colour	0.9453	0.0046	0.9534	0.0040	0.9526	0.0050	0.9535	0.0049
		Combined	0.9413	0.0044	0.9505	0.0045	0.9387	0.0043	0.9480	0.0042
		Texture	0.9095	0.0051	0.9265	0.0035	0.9107	0.0060	0.9105	0.0055
	Sample 2	Colour	0.9609	0.0049	0.9695	0.0042	0.9683	0.0034	0.9701	0.0114
		Combined	0.9725	0.0058	0.9769	0.0025	0.9713	0.0039	0.9764	0.0148
		Texture	0.9548	0.0077	0.9620	0.0029	0.9586	0.0040	0.9629	0.0145
	Sample 3	Colour	0.9343	0.0057	0.9441	0.0073	0.9217	0.0071	0.9320	0.0060
		Combined	0.9473	0.0067	0.9534	0.0066	0.9333	0.0051	0.9392	0.0080
		Texture	0.9427	0.0055	0.9483	0.0048	0.9282	0.0049	0.9345	0.0070
Experiment 3	General	Colour	0.7083	0.0027	0.7817	0.0023	0.7967	0.0029	0.7967	0.0030
		Combined	0.8553	0.0036	0.8688	0.0033	0.8765	0.0056	0.8694	0.0043
		Texture	0.8694	0.0033	0.8861	0.0033	0.8885	0.0034	0.8777	0.0039
	Sample 1	Colour	0.5533	0.0050	0.6457	0.0069	0.6507	0.0094	0.6507	0.0042
		Combined	0.7748	0.0042	0.7958	0.0060	0.7614	0.0079	0.7923	0.0053
		Texture	0.7984	0.0032	0.8160	0.0052	0.7884	0.0054	0.8049	0.0062
	Sample 2	Colour	0.7991	0.0040	0.8466	0.0039	0.8586	0.0036	0.8575	0.0125
		Combined	0.9267	0.0094	0.9268	0.0021	0.9258	0.0061	0.9261	0.0133
		Texture	0.9257	0.0084	0.9351	0.0019	0.9275	0.0036	0.9273	0.0222
	Sample 3	Colour	0.7690	0.0092	0.8588	0.0059	0.8639	0.0147	0.8668	0.0107
		Combined	0.9317	0.0061	0.9366	0.0048	0.9146	0.0043	0.9322	0.0068
		Texture	0.9371	0.0062	0.9443	0.0043	0.9140	0.0041	0.9321	0.0074

superpixels were equally informative as the ones extracted from two-dimensional RGB superpixels. This finding is in good concordance with the fact that pixel colour information extracted from two-dimensional RGB superpixels alone allows classification accuracies very close to highest test accuracies achieved in this study. Keeping in mind that texture context is only meaningful for two or more dimensional arrays and the fact that the second dimension in linearized superpixels is that of the three colour channels, features extracted from pixel value co-occurrence matrixes of linearized superpixels reflect both colour and texture information. Therefore, we acknowledge that features here referred to as "texture" features in the context of linear superpixels are somewhat ambiguous.

The effect of type of features extracted on classification accuracy is more obvious under "Experiment 3" conditions. In the process of pathology tissue slide production, often complicated procedures involving manual human work are applied. Some amount of variation in slide staining quality was expected. Bearing that fact in mind, in order to reduce expected inter-slide colour variability gray-scaled superpixels experimental design was considered. Surprisingly extracted colour features proved to be exclusively valuable for tumour tissue classification. Under generalization conditions, gray-scaled superpixels could not be effectively classified by any method tested (AUC < 0.9). Obviously, gray-scaled superpixels carry less colour information. Therefore, colour features alone allow

Table 5
Parameters of tested CNN models. Each row represents separate model configuration.

Model	Number of neurons in hidden layers					Performance metrics	
No.	conv2d-1	conv2d-2	conv2d-3	conv2d-4	dense-1	Mean AUC	sdAUC
1	32	32	64	64	384	0.9704	0.0007
2	32	32	64	64	1024	0.9667	0.0005
3	64	64	128	128	384	0.9726	0.0004
4	64	64	128	128	1024	0.9687	0.0008
5	**96**	**96**	**192**	**192**	**384**	**0.9745**	**0.0003**
6	96	96	192	192	1024	0.9721	0.0003

less accurate classification than texture features (in certain cases less than 0.6, see Table 4). Maximum AUC achieved with colour and texture features in single sample data sets was 0.8668 and 0.9443 respectively. Best performance on "General" data set was observed with MLP and texture features (AUC = 0.8885), which is substantially lower when compared to the best results of other experimental strategies.

Both colour and texture features, as well as superpixel linearization approach, were suitable to achieve good classification results with most machine learning methods used (AUC $>$ 0.95). We achieved most effective performance under generalization conditions with MLP on combined texture and colour features (AUC = 0.9668). Our results align well with results published by other groups using traditional machine learning approaches AUC 0.97–0.99 (Linder *et al.*, 2012); AUC = 0.9847 (Nava *et al.*, 2016) both targeting colorectal cancer, and AUC = 0.977 (Sethi *et al.*, 2015) on prostate cancer images.

5.2. *Results of Classification Using Convolutional Neural Network*

To demonstrate that annotated superpixels can also be effectively employed to build reliable ground truth data sets for experiments involving deep convolutional neural networks we attempted a classification of image patches (prepared as described in Section 3.6) using a CNN described in Section 4.2. The workflow can be summarized as follows:

1. Selected regions inside pathologist macro-annotated WSIs were segmented into superpixels (see Fig. 1).
2. Superpixels were visualized in the WSIs and subjected to pathologist micro-annotation.
3. Five overlapping 72 $\times$ 72 pixels RGB patches were selected in the original image around each annotated superpixel mass centre by shifting the frame in each direction by 15 pixels (please refer to Fig. 2 and Section 3.6).
4. Classification using CNN described in Section 4.2 was performed.
5. Classifier performance was evaluated.
6. A set of model parameters was tested to select best performing architecture (see Table 5).

We applied convolutional neural network approach to a limited data set, with no possibility to test for generalization. CNN model was tested on the set of patches generated

from "Sample 1", and reached maximum AUC = 0.9745. This is slightly higher accuracy than maximum AUC = 0.9664 achieved with SVM on combined colour and texture features extracted from superpixels of the same sample. Our results are comparable to results achieved by other groups on different types of cancer – breast cancer (AUC = 0.9316 in Xu *et al.*, 2016), and prostate cancer (AUC = 0.965 in Sethi *et al.*, 2016, and AUC = 0.99 in Litjens *et al.*, 2016).

6. Conclusions

The use of superpixels in tumour tissue classification task is not a new approach. It has been widely applied by many authors before (Bunyak *et al.*, 2015; Sethi *et al.*, 2016). In this paper, we propose a method that allows annotating superpixels rather than manually outlining tissue compartments in raw images. Using this approach we were able to produce accurate reference data, and minimize the most effort-intensive part of digital pathology workflow. We demonstrate that ground truth obtained from annotated superpixels can be used to train traditional machine learning based classifiers. We have also tested the ways to process superpixels and extract a set of colour and texture based handcrafted features to achieve high tumour tissue classification accuracy. Finally, by using annotated superpixel centroids, we produced a data set of tumour tissue image patches that allowed us to train deep convolutional neural networks to achieve similar performance regarding classification accuracy as the state-of-the-art methods.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I.J., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mane, D., Monga, R., Moore, S., Murray, D.G., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P.A., Vanhoucke, V., Vasudevan, V., Viegas, F.B., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X. (2016). TensorFlow: large-scale machine learning on heterogeneous distributed systems. arXiv preprint, 1603.04467. arxiv.org/abs/1603.04467. Software available from tensorflow.org.

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Süsstrunk, S. (2012). SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11), 2274–2282.

Ahammer, H., Kriipfl, J., Hackl, C., Sedivy, R. (2009). Image statistics and data mining of anal intraepithelial neoplasia. *Pattern Recognition Letters*, 29, 2189–2196.

Bejnordi, B. E., Balkenhol, M., Litjens, G., Holland, R., Bult, P., Karssemeijer, N., van der Laak, J.A. (2016). Automated detection of DCIS in whole-slide H&E stained breast histopathology images. *IEEE Transactions on Medical Imaging*, 35(9), 2141–2150.

Bianconi, F., Alvarez-Larran, A., Fernandez, A. (2015). Discrimination between tumour epithelium and stroma via perception-based features. *Neurocomputing*, 154, 119–126.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.

Bunyak, F., Hafiane, A., Al-Milaji, Z., Ersoy, I., Haridas, A., Palaniappan, K. (2015). A segmentation-based multi-scale framework for the classification of epithelial and stromal tissues in H&E images. In: *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. doi:10.1109/BIBM.2015.7359726.

Chang, Y.W., Hsieh, C.J., Chang, K.W., Ringgaard, M., Lin, C.J. (2010). Training and testing low-degree polynomial data mappings via linear SVM. *Journal of Machine Learning Research*, 11, 1471–1490.

Cortes, C., Vapnik, V. (1995). Support-vector networks. *Mach Learn*, 20, 273.

Cruz-Roa, A., Basavanhally, A., Gonzalez, F., Gilmore, H., Feldman, M., Ganesan, S., Shih, N., Tomaszewski, J., Madabhushi, A. (2014). Automatic detection of invasive ductal carcinoma in whole slide images with Convolutional Neural Networks. In: *Proc. of SPIE*, 9041(904103).

Dunne, M.R., Michielsen, A.J., O'Sullivan, K.E., Cathcart, M.C., Feighery, R., Doyle, B., Watson, J.A., O'Farrell, N.J., Ravi, N., Kay, E., Reynolds, J.V., Ryan, E.J., O'Sullivan, J. (2017). HLA-DR expression in tumor epithelium is an independent prognostic indicator in esophageal adenocarcinoma patients. *Cancer Immunology, Immunotherapy*, 66(7), 841–850. http://doi.org/10.1007/s00262-017-1983-1.

Emens, L.A. (2017). Breast cancer immunotherapy: facts and hopes. *Clinical Cancer Research*. doi:10.1158/1078-0432.CCR-16-3001.

Ertosun, M.G., Rubin, D.L. (2015) Automated grading of gliomas using deep learning in digital pathology images: a modular approach with ensemble of convolutional neural networks. In: *AMIA Annual Symposium Proceedings*, pp. 1899–1908.

Fawcett, T. (2006). An introduction to ROC analysis (PDF). *Pattern Recognition Letters*, 27(8), 861–874.

Haralick, R.M. (1979). Statistical and structural approaches to texture. *Proceedings of the IEEE*, 67(5), 786–804.

Huang, Y., Zheng, H., Liu, C., Ding, X., Rohde, G. (2015). Epithelium-stroma classification via convolutional neural networks and unsupervised domain adaptation in histopathological images. *IEEE Journal of Biomedical and Health Informatics*. doi:10.1109/JBHI.2017.2691738.

Kanan, C., Cottrell, G.W. (2012). Color-to-grayscale: does the method matter in image recognition? *PLoS ONE*, 7(1), e29740. https://doi.org/10.1371/journal.pone.0029740.

Kingma, D.P., Ba, J.L. (2015). Adam: a method for stochastic optimization. In: *International Conference on Learning Representations*, pp. 1–13.

Lakhani, S.R., Ellis, I.O., Schnitt, S.J., Tan, P.H., van de Vijver, M.J. (2012). *WHO Classification of Tumours*, 4th edition, Vol. 4. IARC.

Linder, N., Konsti, J., Turkki, R., Rahtu, E., Lundin, M., Nordling, S., Haglund, C., Ahonen, T., Pietikäinen, M., Lundin, J. (2012). Identification of tumor epithelium and stroma in tissue microarrays using texture analysis. *Diagnostic Pathology*, 7, 22. http://doi.org/10.1186/1746-1596-7-22.

Litjens, G., Sanchez, C.I., Timofeeva, N., Hermsen, M., Nagtegaal, I., Kovacs, I., Hulsbergen - van de Kaa, C., Bult, P., van Ginneken, B., van der Laak, J. (2016). Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. *Scientific Reports*, 6, 26286. doi:10.1038/srep26286.

Malouf, R. (2002). A comparison of algorithms for maximum entropy parameter estimation. In: *Proceedings of the Sixth Conference on Natural Language Learning (CoNLL)*, pp. 49–55.

McLaughlin, J., Han, G., Schalper, K.A., Carvajal-Hausdorf, D., Pelakanou, V., Rehman, J., Velcheti, V., Herbst, R., LoRusso, P., and Rimm, D.L. (2016). Quantitative assessment of the heterogeneity of PD-L1 expression in non-small cell lung cancer (NSCLC). *JAMA Oncology*, 2(1), 46–54. http://doi.org/10.1001/jamaoncol.2015.3638.

Nava R., Gonzalez G., Kybic J., Escalante-Ramirez B. (2016). Classification of tumor epithelium and stroma in colorectal cancer based on discrete tchebichef moments. In: Oyarzun L.C., et al. (Eds.), *Clinical Image-Based Procedures. Translational Research in Medical Imaging. CLIP 2015. Lecture Notes in Computer Science*, Vol. 9401. Springer, Cham.

Panayiotou, H., Orsi, N.M., Thygesen, H.H., Wright, A.I., Winder, M., Hutson, R., Cummings, M. (2015). The prognostic significance of tumour-stroma ratio in endometrial carcinoma. *BMC Cancer*, 15, 955. http://doi.org/10.1186/s12885-015-1981-7.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E. (2011). Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

Rumelhart, D.E., Hinton, G.E., Williams, R.J. (1986). Learning internal representations by backpropagating errors. *Nature*, 323, 533–x 536.

Sethi, A., Sha, L., Vahadane, A.R., Deaton, R.J., Kumar, N., Macias, V., Gann, P.H. (2016). Empirical comparison of color normalization methods for epithelial-stromal classification in H and E images. *Journal of Pathology Informatics*, 7, 17. http://doi.org/10.4103/2153-3539.179984.

Shamir, L., Orlov, N., Eckley, D. M., Macura, T., Johnston J., Goldberg I.G. (2008). Wndchrm – an open source utility for biological image analysis. *Source Code for Biology and Medicine*, 3, 13. https://doi.org/10.1186/1751-0473-3-13.

Xu, J., Luo, X., Wang, G., Gilmore, H., Madabhushi, A. (2016). A deep convolutional neural network for segmenting and classifying epithelial and stromal regions in histopathological images. *Neurocomputing*, 191, 214–223. http://doi.org/10.1016/j.neucom.2016.01.034.

M. Morkūnas graduated from the Vilnius Gediminas Technical University, Lithuania, in 2002. In 2016 he started PhD studies in informatics engineering at the Institute of Data Science and Digital Technologies, Vilnius University, Lithuania. His interests include bioinformatics, cancer biology, image analysis, machine learning, artificial neural networks.

P. Treigys graduated from the Vilnius Gediminas Technical University, Lithuania, in 2005. In 2010 he received the doctoral degree in computer science (PhD) from Institute of Mathematics and Informatics jointly with Vilnius Gediminas Technical University. He is a member of the Lithuanian Society for biomedical engineering. His interests include image analysis, detection and object's feature extraction in image processing, automated image objects segmentation, optimization methods, artificial neural networks, and software engineering.

J. Bernatavičienė graduated from the Vilnius Pedagogical University in 2004 and received a master's degree in informatics. In 2008, she received the doctoral degree in computer science (PhD) from Institute of Mathematics and Informatics jointly with Vilnius Gediminas Technical University. She is a researcher at the Cognitive Computing Group of Vilnius University, Institute of Data Science and Digital Technologies. Her research interests include databases, data mining, neural networks, image analysis, visualization, decision support systems and Internet technologies.

A. Laurinavičius, MD, PhD fulltime professor at Vilnius University, Department of Pathology, Forensic Medicine and Pharmacology. Director and consultant pathologist at National Center of Pathology. Chair, and board member of multiple international professional societies. Fields of interest: renal pathology, digital pathology image analysis, pathology informatics, health information systems, standards, testing of cancer biomarkers in tissue, multi-resolution analysis of biomarkers.

G. Korvel received her BS degree in mathematics and MS degree in informatics (with honors) from Lithuanian University of Educational Sciences, in 2007 and 2009, respectively. She received the doctoral degree from Vilnius University Institute of Data Science and Digital Technologies (former Institute of Mathematics and Informatics) in 2013. Currently she works in this institution. Her research interests include speech signal processing, developing of mathematical models, applications of soft computing and computational intelligence.