Original Article

A Methodology for Texture Feature-based Quality Assessment in Nucleus Segmentation of Histopathology Image

Si Wen[1], Tahsin M. Kurc[2], Yi Gao[2], Tianhao Zhao[2], Joel H. Saltz[2], Wei Zhu[1]

Departments of [1]Applied Mathematics and Statistics and [2]Biomedical Informatics, State University of New York at Stony Brook, Stony Brook, NY, USA

Received: 19 May 2017 Accepted: 11 July 2017 Published: 07 September 2017

Abstract

Context: Image segmentation pipelines often are sensitive to algorithm input parameters. Algorithm parameters optimized for a set of images do not necessarily produce good-quality-segmentation results for other images. Even within an image, some regions may not be well segmented due to a number of factors, including multiple pieces of tissue with distinct characteristics, differences in staining of the tissue, normal versus tumor regions, and tumor heterogeneity. Evaluation of quality of segmentation results is an important step in image analysis. It is very labor intensive to do quality assessment manually with large image datasets because a whole-slide tissue image may have hundreds of thousands of nuclei. Semi-automatic mechanisms are needed to assist researchers and application developers to detect image regions with bad segmentations efficiently. **Aims:** Our goal is to develop and evaluate a machine-learning-based semi-automated workflow to assess quality of nucleus segmentation results in a large set of whole-slide tissue images. **Methods:** We propose a quality control methodology, in which machine-learning algorithms are trained with image intensity and texture features to produce a classification model. This model is applied to image patches in a whole-slide tissue image to predict the quality of nucleus segmentation in each patch. The training step of our methodology involves the selection and labeling of regions by a pathologist in a set of images to create the training dataset. The image regions are partitioned into patches. A set of intensity and texture features is computed for each patch. A classifier is trained with the features and the labels assigned by the pathologist. At the end of this process, a classification model is generated. The classification step applies the classification model to unlabeled test images. Each test image is partitioned into patches. The classification model is applied to each patch to predict the patch's label. **Results:** The proposed methodology has been evaluated by assessing the segmentation quality of a segmentation method applied to images from two cancer types in The Cancer Genome Atlas; WHO Grade II lower grade glioma (LGG) and lung adenocarcinoma (LUAD). The results show that our method performs well in predicting patches with good-quality segmentations and achieves F1 scores 84.7% for LGG and 75.43% for LUAD. **Conclusions:** As image scanning technologies advance, large volumes of whole-slide tissue images will be available for research and clinical use. Efficient approaches for the assessment of quality and robustness of output from computerized image analysis workflows will become increasingly critical to extracting useful quantitative information from tissue images. Our work demonstrates the feasibility of machine-learning-based semi-automated techniques to assist researchers and algorithm developers in this process.

Keywords: Classification, nuclei segmentation quality assessment, texture feature

Introduction

Whole-slide tissue specimens have long been used to examine how the disease manifests itself at the subcellular level and modifies tissue morphology. By examining glass tissue slides under high-power microscopes, pathologists evaluate changes in tissue morphology and render diagnosis about a patient's state. Advances in digital pathology imaging have made it feasible to capture high-resolution whole-slide tissue images rapidly. Coupled with decreasing storage and computation costs, digital slides have enabled new opportunities for research. Research groups have developed techniques for quantitative analysis of histopathology images and demonstrated the application of tissue imaging in disease research.[1-13]

Access this article online

Quick Response Code:

Website: www.jpathinformatics.org

DOI: 10.4103/jpi.jpi_43_17

Address for correspondence: Ms. Si Wen, Department of Applied Mathematics and Statistics, State University of New York at Stony Brook, Stony Brook, NY 11794-3600, USA. E-mail: siwen.statistics@gmail.com

How to cite this article: Wen S, Kurc TM, Gao Y, Zhao T, Saltz JH, Zhu W. A methodology for texture feature-based quality assessment in nucleus segmentation of histopathology image. J Pathol Inform 2017;8:38.

Available FREE in open access from: http://www.jpathinformatics.org/text.asp?2017/8/1/38/214169

Nucleus/cell detection and segmentation are the common methodologies in tissue image analysis.[14] Over the past decade, researchers have developed a variety of nucleus segmentation methods.[1,14,15] Nucleus segmentation pipelines process images to detect the locations of nuclei and extract their boundaries. After the boundaries of nuclei are determined, imaging features (such as size, intensity, shape, and texture features) can be computed for each segmented nucleus and used in downstream analyses for mining and classification. Achieving accurate and robust segmentation results remains a difficult problem because of image noise, such as image acquisition artifacts, differences in staining, and variability in nuclear morphology within and across tissue specimens. It is not uncommon that a segmentation pipeline optimized for a tissue type will produce bad segmentations in images from other tissue types and even in different regions of the same image. Figure 1 shows sample patches with good-segmentation results (good-quality-segmentation) and sample patches with two categories of bad segmentations (under-segmented and over-segmented) from the same segmentation algorithm. The "under-segmented" patches in the figure refer to the cases, in which some nuclei were missed due to poor contrast between the nuclei and the tissue. The "over-segmented" patches, on the other hand, have nonnuclear material segmented as nuclei or sets of single nuclei segmented as multiple nuclei.

It is necessary to have a quality control stage to assess the quality of segmentation results before the results are used in downstream analyses for knowledge discovery and scientific interpretation. It is labor intensive to manually check every image and every segmented nucleus in an image. A typical whole-slide tissue image contains a few hundred thousand to over a million nuclei. This data problem is compounded by the fact that datasets with thousands of images are becoming common in image analysis projects with the help of advanced tissue slide scanners and increased storage capacity of modern computing platforms. (Semi-) automated error checking workflows are needed that can help researchers and algorithm developers detect bad segmentation results quickly and reliably.

Bamford and Lovell[16] have proposed a nucleus segmentation method with a confidence measure in segmentation output. Since the confidence measure is related to a specific parameter in the particular segmentation method, this quality control method cannot easily be expanded to other segmentation algorithms. Cukierski *et al*.[17] assigned a numeric value to each segmented object. Probability is calculated from a logistic regression built on the morphological, texture, and contextual features of the segmented object. By ranking the segmentations based on their probabilities, well-segmented objects were selected. In another recent work,[18] an artificial neural network was trained to classify accurately segmented nuclei and other segmented objects using the shape, intensity, and texture feature of the segmented objects. An experimental evaluation showed that this selection procedure can help increase the precision of segmented objects from 17% to 71.5%. Brinker *et al*.[19] trained a support vector machine (SVM) classifier with the appearance-based features (area, circularity, and solidity) and shape-based features (intensity variance and entropy) of a segmented object. The trained classifier is then used to differentiate correct and incorrect cell segmentations in the preparation for automatic segmentation correction. The previous work on segmentation quality assessment and improvement has developed methods that work at the object level. The methods aim to assess the correct segmentation of individual objects. This process can become computationally very expensive in high-resolution images with millions of nuclei and may not scale to large datasets.

In this paper, we propose a novel quality control workflow that uses patch-level intensity and texture features to evaluate nucleus segmentation results in high-resolution whole-slide tissue images. This approach is motivated by the observation that image regions with similar intensity and texture features tend to have comparable segmentation quality given a segmentation algorithm and a set of segmentation parameter

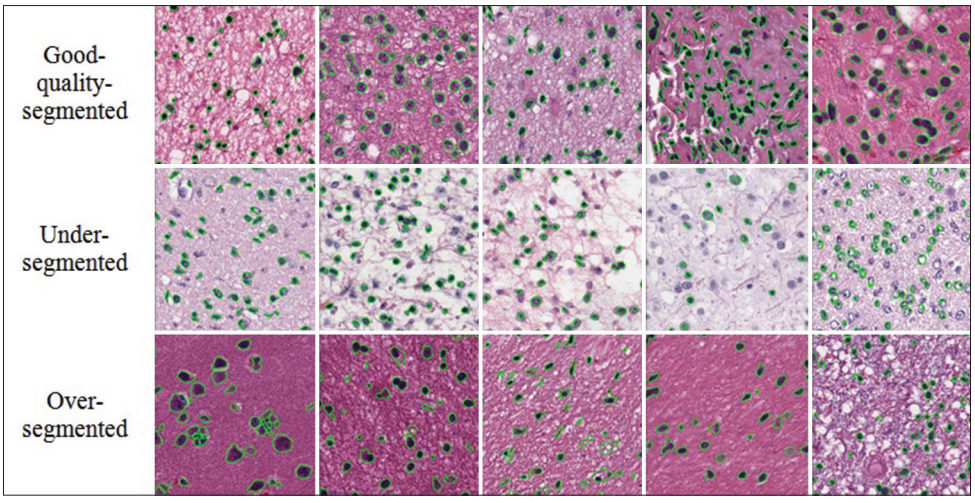

Figure 1: Sample patches of WHO Grade II lower grade glioma for each of the three categories: good-quality-segmentation, under- and over-segmented

values. In our current implementation, segmentation results are used only for labeling image regions in the training phase but not used in the prediction (or classification) phase. To scale millions of nuclei and large numbers of images, our approach assesses the segmentation quality of image patches, instead of the accuracy of pixel-level boundary delineation[20] or the probability that an object has been segmented well.[17] The proposed approach is executed as follows with segmentation results obtained from a collection of whole-slide tissue images. In the training phase, a sample image set is randomly selected by a pathologist from the collection of images. The pathologist examines the segmentation results in the sample images and selects representative regions in each sample image. S/he then classifies them into regions with good-segmentation results (good-quality segmentation) and regions with bad-segmentation results (under-segmented or over-segmented). The selected regions are partitioned into equal-sized patches, and a set of intensity and texture features is computed for each patch. A machine-learning model is trained using the features and labels of each patch in the training set. In the classification phase, test images are partitioned into patches (of the same size as the patches in the training set), the same set of intensity and texture features is computed, and each patch is classified using the features and the trained model. The classification model must be retrained for results obtained from a different segmentation algorithm or a different set of algorithm parameter values. In that case, the training phase will use the same set of intensity and texture features, but the set of patches and their labels may be different. We plan to explore the utilization of segmentation results and morphological features (such as size and shape of a segmented nucleus) in the training and classification phases in the future work.

We have experimentally evaluated our methodology with two different cancer types: WHO Grade II lower grade gliomas (LGGs) and lung adenocarcinoma (LUAD) cases from The Cancer Genome Atlas (TCGA) project.[21] For each of the cancer types, we segmented images with a segmentation algorithm which discriminates between background tissue and target nuclei through a threshold parameter.[22] Threshold parameters are used in many nucleus segmentation algorithms to delineate the boundaries of target objects. The choice of threshold parameter values leads to under-segmentation or over-segmentation of an image. Our approach not only can predict the segmentation quality based on the image information but also can provide suggestions as to which direction the threshold value should be adjusted to obtain better segmentation results. Please see the Segmentation Algorithm subsection in the Results and Discussion section for an example of how predicting if a patch is under-segmented or over-segmented can be used to guide the selection of algorithm parameters to improve segmentation results.

The rest of the paper is organized as follows. The Methods section outlines the construction of the proposed segmentation quality assessment pipeline, including generation of labeled sample patches, patch-level texture feature extraction, and classification. An experimental evaluation of the pipeline is presented in the Results and Discussion section.

Methods

Our approach consists of a pipeline of training and classification steps as is illustrated in Figure 2. In the training phase, image regions in a sample set of images are selected and labeled by a pathologist to create the training set. The image regions are then partitioned into image patches, and a set of intensity and texture features is computed for each patch. The last step in this phase is to train a classifier using the labels of the image regions and the computed features. In the classification step, the classification model is applied to the test data to assess the quality of segmentations in image patches extracted from images in the test dataset.

Training phase

A subset of segmented whole-slide images in the target dataset is randomly chosen. The pathologist marks up regions in each selected image and assigns a classification label to each region. There are three classification labels; region with good-segmentation results (good-quality-segmentation), region with under-segmented nuclei (under-segmented), and region with over-segmentation (over-segmented). If a region is labeled under-segmented, it means that the segmentation algorithm has missed some nuclei in the region and/or segmented single nuclei as multiple nuclei. If a region is labeled over-segmented, it means that the segmentation algorithm has segmented more objects than there are actual nuclei.

The stratified sampling method[23] is used to select a subset of images for the training set. If the images belong to some natural strata, images are randomly selected from each group based on the number of images in each group. In our experiments, images are grouped based on their tissue source site i ($i = 1,\ldots n$). The tissue source site indicates from which institution the tissue was obtained. Grouping images based on tissue source site is performed to accommodate for variability in images due to differences in tissue preparation and image

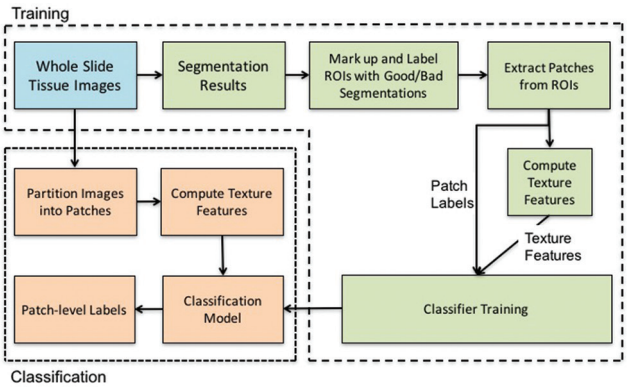

Figure 2: Workflow of the nucleus segmentation quality assessment pipeline

acquisition across source sites. To select N images for the training set, we compute the ratio $p_i (\sum p_i = 1)$ of images from each site i to the total number of images. The relative size of the sample images from site i would be $N \times p_i$. To ensure that our training set has images from all the source sites, we select one image randomly from the sites with $N \times p_i \leq 1$. For the rest of the sites $i \in \{i \mid N \times p_i > 1\}$, we recompute the ratio $p_i = \frac{p_i}{\sum_{\{i|N \times p_i > 1\}} p_i}$ and randomly choose $(N - \sum 1_{\{i|N \times p_i \leq 1\}}) \times p'_i$ images from those source sites. $\sum 1_{\{i|N \times p_i \leq 1\}}$ indicates the number of source sites from which only one image has been selected.

Image regions selected by the pathologist can be of any shape and size. Each region is partitioned into nonoverlapping patches of the same size and shape. The number of patches in each region will depend on the region's shape and size. All of the patches in a region are assigned the same label as that of the region. Two sets of intensity and texture features are calculated for each region – note that the features are computed at the patch level, not for nuclei segmented in the patch. The first set contains 16 features from three groups (pixel statistics, gradient statistics, and edge). A total of 32 features are computed from this set; 16 for the red channel and 16 for the blue channel. These features are listed in Table 1. The second set of features consists of the mean and standard deviation of the intensity values of the red, green, and blue channels.

To avoid collinearity among the features and to select the more informative features, stepwise variable selection in logistic regression is applied. Variable selection is an essential preprocedure and has many benefits for classifiers, such as reducing the measurement requirements, reducing training and utilizing times, and alleviating the curse of dimensionality to improve prediction performance.[24] Stepwise variable selection in logistic regression is one of the commonly used variable selection methods.

Two sets are created for the variable selection step. One set contains the patches that have good segmentation results and the patches that are under-segmented. The other set is composed of the patches that have good-segmentation results and the patches that are over-segmented. The variable selection process is applied to the two sets independently. The label of each patch is treated as a binary response variable. The computed features are added to or removed from the feature set at each iterative step to achieve a smaller Bayesian information criterion until no more action can be done to reduce the criterion. In this way, the selected features are the smallest subset of the input features with sufficient information to differentiate the two categories (good vs. under-segmented or good vs. over-segmented). For a different set of sample patches, the selected features might be different based on their distinct texture characters. We use the implementation of stepwise selection for generalized linear regression model in Statistics and Machine Learning Toolbox™ in MATLAB to carry out the variable selection step.

Table 1: List of the patch level texture features for red and blue channel

Category	Name	Brief description
Pixel statistics	IntensityMean	Average of raw pixel value
	IntensityMax	Maximum of raw pixel value
	IntensityMin	Minimum of raw pixel value
	IntensityStd	SD of raw pixel value
	IntensityEntropy	Entropy of the normalized co-occurrence matrix of pixel value
	IntensityEnergy	Sum of squared elements in the normalized co-occurrence matrix of pixel value
	IntensitySkewness	Skewness of the normalized pixel value
	IntensityKurtosis	Kurtosis of the normalized pixel value
Gradient statistics	GradientMean	Average of gradient channel value
	GradientStd	SD of gradient channel value
	GradientEntropy	Entropy of the normalized co-occurrence matrix of gradient channel value
	GradientEnergy	Sum of squared elements in the normalized co-occurrence matrix of gradient channel value
	GradientSkewness	Skewness of the normalized gradient channel value
	GradientKurtosis	Kurtosis of the normalized gradient channel value
Edge	CannyNonZero	Number of pixel with nonzero canny value
	CannyMean	Average of canny value

SD: Standard deviation

Classification models

The features selected for good versus under-segmented may not be able to differentiate over-segmented patches from patches with good-segmentation results, and similarly, the features for good versus over-segmented may not be able to separate under-segmented patches from patches with good-segmentation results. The proposed approach trains two classification models. One model is trained using the set of patches with good-segmentation results and under-segmented patches. The second model is trained using the set of patches with good-segmentation results and over-segmented patches. These two models are applied to a test patch to predict the test patch's label as we shall describe in the next section.

Test phase

When a new patch with no labels goes through the classification process, it will get two labels, one from each classification model. One label indicates whether the patch is under-segmented or not-under-segmented. The other label classifies whether the patch is over-segmented or not-over-segmented. The two classification results are combined to make a final decision about the segmentation quality of the patch. This is illustrated in Figure 3. In Figure 3, we refer one of the models as "under remover" which labels a patch under-segmented or

not-under-segmented and the other model as "over remover" which labels a patch over-segmented or not-over-segmented. If the under remover labels a patch under-segmented and the over remover labels the patch not-over-segmented, the final label of the patch will be under-segmented. This decision is based on the expectation that the over remover will not be able to differentiate between a patch with good-segmentation results and an under-segmented patch and that the under remover will be more accurate with under-segmented patches. Similarly, if the over remover labels a patch over-segmented and the under remover labels the patch not-under-segmented, the final label of the patch will be over-segmented. If the over remover and the under remover label a patch not-over-segmented and not-under-segmented, respectively, the final label of the patch is chosen to be "patch with good-segmentation results." If the over remover labels a patch over-segmented and the under remover labels the patch under-segmented, we can only conclude that the patch has bad-segmentation results; however, we cannot tell whether the patch is under- or over-segmented.

Results and Discussions

We have evaluated the proposed pipeline using two sets of whole-slide images obtained from two different cancer types: WHO Grade II LGG and LUAD. The whole-slide images were downloaded from TCGA data set.

Segmentation algorithm

For each cancer type, a computerized nucleus segmentation method[22] was applied. The method segments nuclei in H&E-stained whole-slide tissue images. It applies color normalization in the L*a*b color space on input images using a properly stained template image. It then extracts the hematoxylin (stained on nuclei mainly) channel through a color decomposition process. A localized region-based level set method with a user-defined threshold value determines the contour of each nucleus. In cases where several nuclei are clumped together, a hierarchical mean shift algorithm is employed to separate the clump into individual nuclei.

The threshold parameter in the level set method significantly affects the quality of segmentation. Figure 4 shows the segmentation results generated by two different threshold values. The blue polygons in the images are the segmentation results obtained with a small threshold value, while the red polygons show the results using a large threshold value. In the images in the first row, some light-colored nuclei have been missed with the low threshold value. In the areas highlighted with a yellow circle, only six nuclei were segmented with the low threshold value. There are actually 10 nuclei in that area. The blue result in this case represents a bad result with under-segmentation. After increasing the threshold value, the result (red result) can be considered good-quality segmentation. The images in the second row show an example of over segmentation; the large threshold value (the red result) would lead to segmentation of nonnuclear material. By decreasing the threshold parameter value, the segmentation result (the blue result) is much better. Therefore, if our quality assessment pipeline predicts whether a patch is under-segmented or over-segmented, this information can be used to guide the selection of algorithm parameters to improve segmentation results.

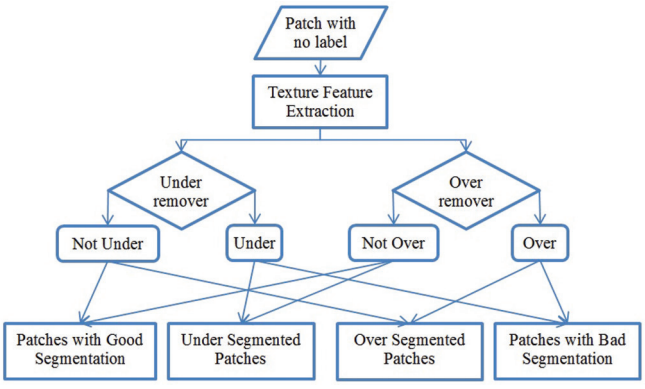

Figure 3: Framework of decision-making strategy for patches with no labels

Classification methods

In our experiments, we generated the classification models using two classification methods and compared their results: random forest and SVM. Generally, both random forest and SVM have their own pros and cons. Whether one method is better than the other depends on the problem and data set.[25]

Random forest is an ensemble learning method for classification that works by bagging multiple decision trees and outputting the classification label by taking the majority vote. Each of the decision trees is built on a bootstrap sample of the training data using a randomly selected subset of variables.[26] It has the distinct advantage that decision trees' habit of overfitting to their training set can be avoided. Furthermore, since there are no parameters to be tuned, the runtime for training a

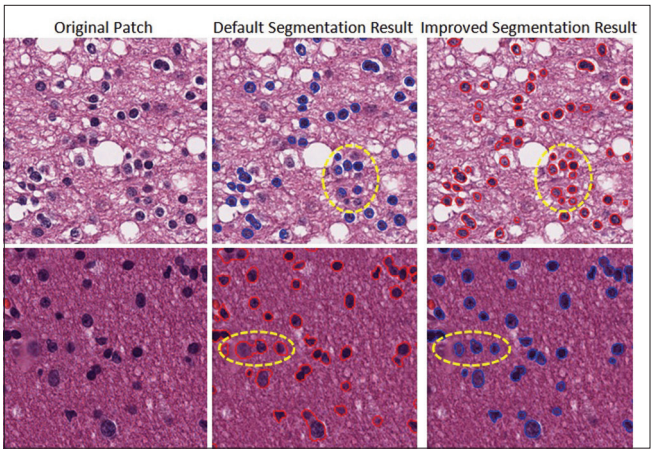

Figure 4: Comparison of segmentation result with different threshold parameter values (the example in the first row shows that segmentation method with small threshold parameter [blue polygons] would neglect some light-colored nuclei; while the example in the second row indicates that segmentation method with large threshold parameter [red polygons] would segment nonnuclei area)

random forest is usually short. We use the implementation of random forest in Statistics and Machine Learning Toolbox™ in MATLAB. In our experiments, we set the number of trees as 1000. We also evaluated a random forest with 10,000 trees. The results showed that the performance of the random forest did not improve much.

SVM is supervised learning method. It finds hyperplanes in a high-dimensional space that maximizes the distance to the nearest training data point of any classes. In addition to performing linear classification, SVM can efficiently perform a nonlinear classification using a kernel which implicitly maps inputs to high-dimensional feature spaces. This is the reason why SVM can achieve better performance than traditional linear classifiers, such as linear discriminant analysis and logistic regression. We used the MATLAB version of LIBSVM[27] with the radial basis function kernel. The kernel parameter gamma and cost were selected by 5-fold cross-validation.

Generation of sample patches

Forty images segmented by the segmentation method were randomly selected from each cancer type for our collaborating pathologist to manually label regions (also referred to as regions of interest [ROIs] in the rest of the paper). Since the whole-slide images in TCGA were collected from different tissue source sites, we applied the stratified sampling step as presented in the Methods section to create the training dataset.

The pathologist reviewed the segmented images using our QuIP application and manually labeled regions good-quality-segmentations, under-segmented, and over-segmented in each image in the training set. QuIP is a web-based suite of tools and services that are designed to support analysis, management, and query of pathology image data and image analysis results (http://www.quip1.bmi.stonybrook.edu; https://www.github.com/SBU-BMI/quip_distro.git). It provides web applications and interfaces for users to interact with and visualize images and analysis results. The web interfaces and applications are backed by a database, which indexes metadata about whole-slide tissue images and results from segmentation and feature computation pipelines. Users can view high-resolution whole-slide tissue images and segmentation results using the caMicroscope application.[28] A user can request an image and select analysis results from an analysis algorithm and view the analysis results overlaid on the image as polygons. The user can pan and zoom in the image, markup regions using rectangular or freehand drawing tools, annotate the regions with a label, and save the results in the database. Figure 5 shows the main image viewing and markup interface with segmentation results [Figure 5a] and regions marked up by the pathologist [Figure 5b] in a whole-slide tissue image.

In the regions with good-segmentation results, most of the nuclei were segmented correctly. In the under-segmented regions, some nuclei were not segmented or single nuclei were segmented as multiple nuclei. In the over-segmented regions, some none-nucleus material was delineated as nuclei. Regions with no tissue (e.g., regions containing slide background and image acquisition artifacts) were not taken into consideration. Some regions, such as necrosis regions or blood clots, may have tissue but no nuclei. If there are segmented objects in such regions, they are labeled over-segmented regions; otherwise, they are labeled "region with good-segmentation results." In our experiments, some regions with no nuclei were over-segmented, but there were few regions with no segmentation results.

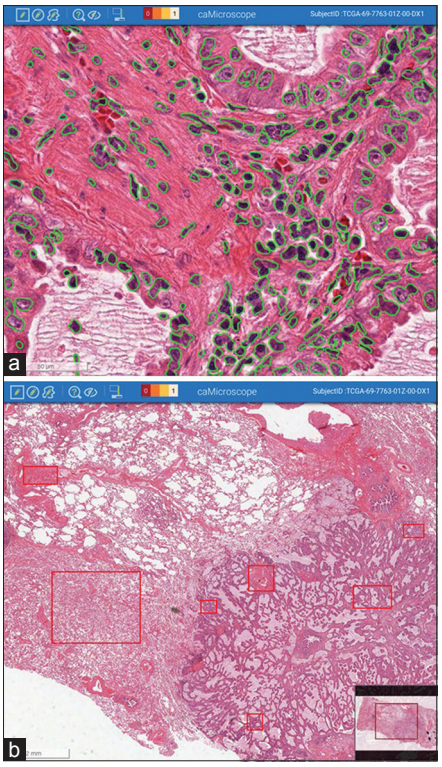

Figure 5: CaMicroscope application in QuIP to view images, segmentation results, and markup regions. (a) Segmentation results displayed as polygons overlaid on a whole-slide tissue image, (b) regions marked up a pathologist to create a training set

After the manually labeled regions were obtained, nonoverlapping patches of 512 × 512 pixels were extracted from each of the regions. Some sample patches of WHO Grade II LGG with segmentation results for each category are shown in Figure 1.

Table 2 lists the number of manually labeled ROIs and the number of patches for each cancer type and each segmentation quality category. The number of patches for an ROI was determined by the size of the ROI. Larger ROIs generated more patches. There are fewer patches generated for LUAD even though there are more regions labeled for LUAD. This is because tissue images from the WHO Grade II LGG cases have usually more uniform texture than images from the LUAD cases. Hence, larger regions were selected in WHO Grade II LGG images.

Computation and comparison of region texture features

For each good-quality-segmentation, under-segmented, or over-segmented patch, 38 texture features were calculated. Thirty-two of them are the features listed in Table 1 for both the red and the blue channels. The remaining six features are the mean and standard deviation of the raw pixel value in each of the red, green, and blue channels (feature names: meanR, meanG, meanB, stdR, stdG, and stdB). To remove redundant features, stepwise selection of logistic regression was carried out for the good-quality-segmentation versus under-segmented patches and good-quality-segmentation versus over-segmented patches for each cancer type. The selected significant features are listed in Table 3. "r_" or "b_" prefix in a feature name means that the feature was computed for the red channel or the blue channel, respectively. We can see from the table that different cancer types have different sets of significant features. Comparing with the features selected for WHO Grade II LGG, more features were selected for LUAD. This means that it was harder to discriminate between good- and under-segmented patches or between good-quality-segmentation and over-segmented patches in LUAD than those in WHO Grade II LGG. As a result, more features were needed to be selected.

There are 38 texture features in total. For WHO Grade II LGG good versus under, 11 features were selected; for WHO Grade II LGG good versus over, 8 features were selected; for LUAD good versus under, 22 features were selected; and for LUAD good versus over, 21 features were selected.

Even for the same cancer type, different texture features were selected in different segmentation quality comparisons. As is seen in Table 3, r_GradientMean (the average of gradient value for the red channel) was selected for WHO Grade II LGG in the good versus under comparison, but not in the good versus over comparison. On the other hand, stdG (representing standard deviation of the raw pixel value for green channel) was selected in the good versus over comparison but not in the good versus under comparison.

Table 2: List of information for manual labeled regions of interest

	Number of images	Number of manual labeled ROI (number of patches)		
		Good-quality-segmented	Under-segmented	Over-segmented
WHO Grade II LGG	40	34 (5819)	24 (6122)	17 (4718)
LUAD	40	28 (3992)	39 (3087)	27 (3121)

ROI: Regions of interest, LGG: Lower grade glioma, LUAD: Lung adenocarcinoma

Table 3: Texture features selected to each cancer and category comparison

Feature name	WHO Grade II LGG		LUAD		Feature name	WHO Grade II LGG		LUAD	
	Good versus under	Good versus over	Good versus under	Good versus over		Good versus under	Good versus over	Good versus under	Good versus over
r_IntensityMean		✓	✓		b_IntensityMean	✓			
r_IntensityMax					b_IntensityMax			✓	✓
r_IntensityMin	✓		✓	✓	b_IntensityMin				
r_IntensityStd		✓	✓	✓	b_IntensityStd	✓	✓		
r_IntensityEntropy			✓	✓	b_IntensityEntropy			✓	✓
r_IntensityEnergy		✓	✓	✓	b_IntensityEnergy			✓	
r_IntensitySkewness	✓		✓	✓	b_IntensitySkewness	✓	✓	✓	✓
r_IntensityKurtosis	✓			✓	b_IntensityKurtosis	✓			✓
r_GradientMean	✓		✓		b_GradientMean			✓	
r_GradientStd		✓	✓	✓	b_GradientStd				✓
r_GradientEntropy			✓		b_GradientEntropy			✓	✓
r_GradientEnergy				✓	b_GradientEnergy			✓	✓
r_GradientSkewness	✓				b_GradientSkewness				✓
r_GradientKurtosis					b_GradientKurtosis			✓	✓
r_CannyNonZero				✓	b_CannyNonZero			✓	✓
r_CannyMean		✓	✓		b_CannyMean				
meanR					stdR	✓			
meanG	✓		✓	✓	stdG		✓	✓	
meanB			✓		stdB				✓

There are 38 texture features in total. For WHO Grade II LGG good versus under, 11 features were selected; for WHO Grade II LGG good versus over, 8 features were selected; for LUAD good versus under, 22 features were selected; and for LUAD good versus over, 21 features were selected. LGG: Lower grade glioma, LUAD: Lung adenocarcinoma '✓' means the feature is selected for the category comparison

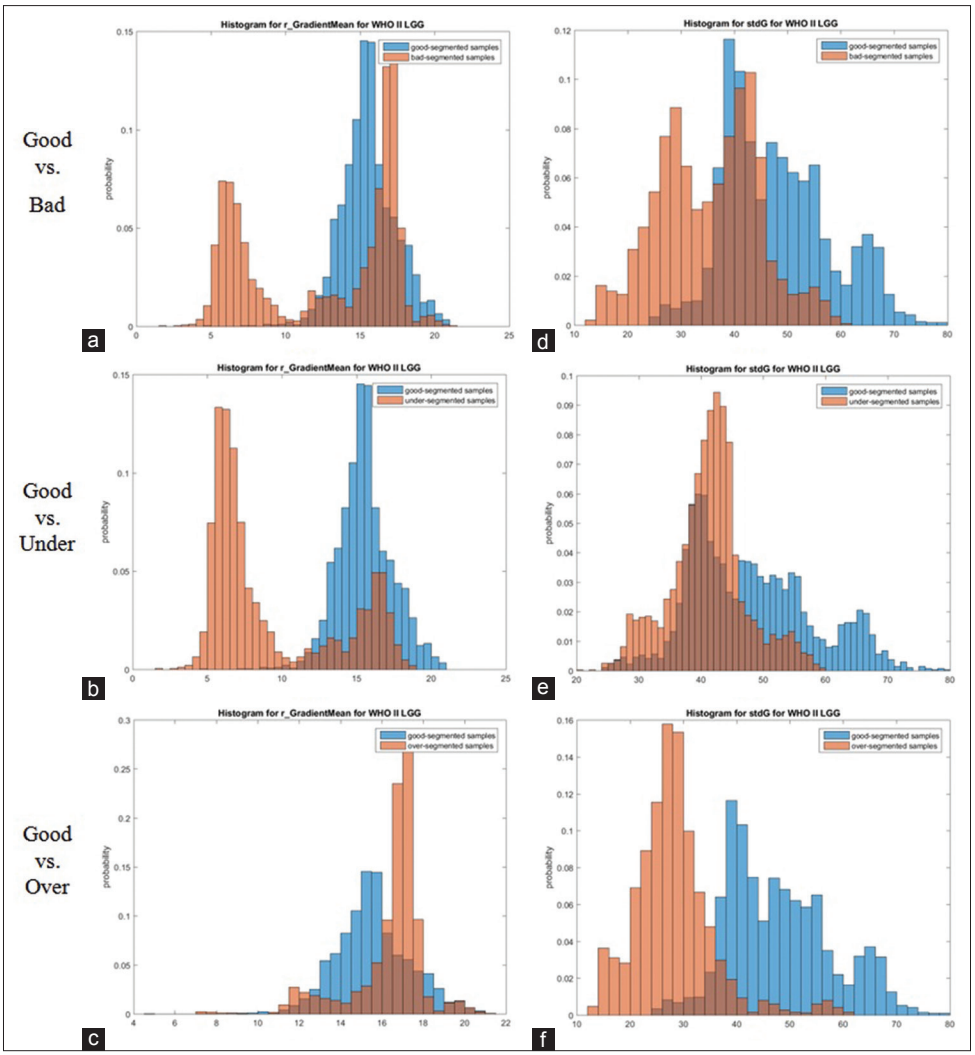

Figure 6: Histogram of texture features for WHO II lower grade glioma (histogram of r_GradientMean of [a] good-quality-segmentation group versus bad-quality-segmentation group, [b] good-quality-segmentation group versus under-segmented group, [c] good-quality-segmentation group versus over-segmented group, and histogram of stdG of [d] good-quality-segmentation group versus bad quality segmentation group, [e] good-quality-segmentation group versus under-segmented group, [f] good-quality-segmentation group versus over-segmented group)

The histograms in Figure 6 explain why the two features were selected in only one of the two category comparisons and why we need to divide the bad-quality-segmentation group into two subgroups. It can be seen in Figure 6b that patches with r_GradientMean <10 are all under-segmented patches and all the patches with good-quality segmentation have r_GradientMean >10. That is, patches with smaller variation in the red channel tended to be under-segmented patches. There is still a small portion of under-segmented patches having large variation in the red channel, which is caused by the texture variation across different slides. Feature r_GradientMean was able to distinguish most of the under-segmented patches from patches with good-quality segmentation. However, as seen in Figure 6c, r_GradientMean could not differentiate between over-segmented patches and patches with good-quality segmentation. If we combined under-segmented patches with over-segmented patches and grouped them together as patches with bad-quality segmentations, as shown in Figure 6a, there would be significant overlap between the histogram of the patches with bad-quality segmentations and that of the patches with good-quality segmentations. By dividing the bad-quality-segmentation group into two subgroups, we can more easily differentiate one particular subgroup of bad-quality-segmentation patches from good-quality-segmentation patches. Similarly, feature stdG provided more information when distinguishing good-quality-segmentation patches from over-segmented ones but was less informative when dividing the patches into good-quality-segmentation group and under-segmented group or bad-quality-segmentation group.

Classification result

For each cancer type, SVM and random forest for good quality segmentation group and under/over-segmented group were trained using the selected significant texture features. Since the primary purpose of our pipeline is to pick up

good-quality segmentation patches in a whole-slide image, we need to separate the good-quality segmentation group from all the subgroups of bad segmentations (the under-segmented subgroup and the over-segmented subgroup). The training accuracy, which is computed as the average of 10-fold cross validation accuracy, and the test accuracy are listed in Table 4. The accuracy is defined as the sum of true positive and true negative divided by the total number of cases in the two categories of each training or test set. SVM achieved higher training accuracy than random forest. However, random forest achieved better test accuracy in most of the cases, especially for the test of LUAD under remover. This indicates that the SVM is easier to be overfitted than random forest.

Although the training accuracies of the under remover models (from the SVM and random forest) with the LUAD images were above 90%, their test accuracies were low, especially that of the model implemented by the SVM. This is unsurprising given the wide variety of cell appearances in the LUAD images and the similarity of texture feature for good-quality-segmentation patches and under-segmented patches. Figure 7 shows some sample patches: patches in the first row are good-quality-segmentation patches that were classified as under-segmented and the ones in the second row were under-segmented patches but labeled good-quality segmentation. Compared with the LGG samples shown in Figure 1, the size and shape of LUAD nuclei are quite different, across the slides or even within the same slide. For the patches generated from a single ROI, their texture features are quite similar. They share the same label with the label given to the ROI. However, the texture appearance of different ROIs circled from a particular slide or from different slides may be very different. As seen in Figure 7, the patches in the first row were all labeled as good-quality segmentation, but they look significantly different from each other. This fact also increases the difficulty of accurately classify patches into these two quality categories for LUAD.

Table 5 lists the precision, recall, and F1 score of the three categories: good, under, and over using different methods to classify WHO Grade II LGG and LUAD patches. For both of the classification methods, the under-segmented category achieved the highest F1 score in the LGG dataset. This is because patches in this category have consistent and differentiable texture features than the other two categories (as shown in Figure 1, under-segmented sample patches are all light-colored and quite different from patches in the other two categories). By applying our segmentation quality control pipeline to the other WHO Grade II LGG whole-slide images, we can find out the regions in which the segmentation method with a given set of parameters may fail to detect all the nuclei. A possible further step would be to increase the threshold parameter value for those regions to get more accurate segmentation results. Due to more heterogeneous collection of nuclei in the LUAD images, the performances of the two classification methods were worse with the LUAD segmentation results than with the LGG segmentation results. Among the three categories, the good-quality-segmentation category achieved the highest F1 score. This can help in finding regions that are better segmented when processing an LUAD whole-slide image through our quality control workflow.

We have also combined the two subgroups of patches with bad segmentations together to form the bad-quality-segmentation group and trained the classifiers for each cancer type by the two classification methods to differentiate good-quality-segmentation patches from the bad-quality-segmentation ones. The performance of the good versus bad classifiers is given in Table 6. The good versus bad classifier has better performance in detecting good-quality-segmentation areas in the LGG images, while it has much worse performance with the LUAD images compared with the performances of the classification models for good-versus-under and good-versus-over. These results show that dividing the bad-quality-segmentation group into two subgroups not only provides us with information about how to improve the segmentation quality for the bad-quality-segmentation areas but also achieves better classification performance with cancer types, in which morphology of nuclei has higher variance.

In general, compared with SVM, random forest has better performance in the LUAD dataset. It achieves higher training accuracy and F1 score. Therefore, we recommend applying

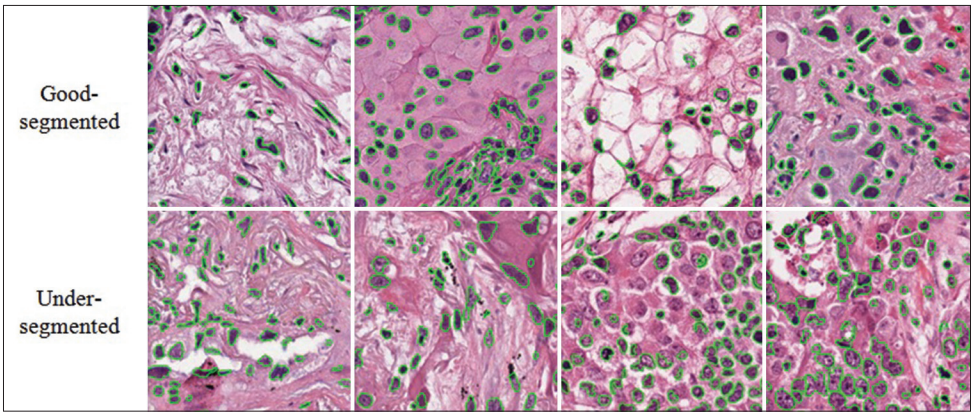

Figure 7: Samples of misclassified good-quaility and under-segmented lung adenocarcinoma patches

Table 4: Performance of support vector machine and random forest for each cancer type and category comparison

	Random forest		SVM	
	Training accuracy (%)	Test accuracy (%)	Training accuracy (%)	Test accuracy (%)
WHO Grade II LGG under remover	97.23	97.56	99.60	93.48
WHO Grade II LGG over remover	96.59	82.43	98.21	84.40
LUAD under remover	92.61	83.77	90.57	66.43
(a) LUAD over remover	86.75	83.89	91.43	84.01
WHO Grade II LGG	**Random forest Under remover**		**SVM Under remover**	
	Tested as not under-segmented	Tested as under-segmented	Tested as not under-segmented	Tested as under-segmented
Good/over	5286	150	5029	407
Under	2	3039	0	3041
WHO Grade II LGG	**Random forest Over remover**		**SVM Over remover**	
	Tested as not over-segmented	Tested as over-segmented	Tested as not over-segmented	Tested as over-segmented
Good/under	6064	174	6146	92
Over	918	1321	756	1483
LUAD	**Random forest Under remover**		**SVM Under remover**	
	Tested as not under-segmented	Tested as under-segmented	Tested as not under-segmented	Tested as under-segmented
Good/over	2644	870	2308	1206
Under	295	1335	445	1185
LUAD	**Random forest Over remover**		**SVM Over remover**	
	Tested as not over-segmented	Tested as over-segmented	Tested as not over-segmented	Tested as over-segmented
Good/under	2705	851	1611	945
(b) Over	365	1223	395	1193

The table shows the training accuracy (average of 10-fold cross validation accuracy) and test accuracy of each test. The accuracy is defined as the sum of true positive and true negative divided by the total number of cases in the two categories of each training or test set. The tables present the confusion matrices of each test for overall test set. LGG: Lower grade glioma, LUAD: Lung adenocarcinoma, SVM: Support vector machine (a) shows the training accuracy (average of 10-fold cross validation accuracy) and test accuracy of each test. The accuracy is defined as the sum of true positive and true negative divided by the total number of cases in the two categories of each training or test set. (b) present the confusion matrices of each test for overall test set.

random forest to create classification model. When comparing the results for the two cancer types, we observe that the classification results for LGG are better than those for LUAD. This indicates that our pipeline will likely work better for cancer types with more consistent texture features. By dividing the bad-quality-segmentation group into under- and over-segmented subgroups, the F1 score for good-quality-segmentation group has increased. In addition, the subgroup labels can also be used to guide algorithm parameter selection to improve segmentation results as described earlier in the segmentation algorithm section.

Conclusions

Effective utilization of whole-slide tissue images in biomedical research hinges on robust methods that extract useful information accurately and reliably. Nucleus segmentation is one of the core image analysis methods employed by researchers to turn image data into knowledge. While computer algorithms generally produce more reproducible results, they often are sensitive to input data and may not produce consistently good segmentation results across image datasets. In this paper, a quality assessment pipeline is proposed that evaluates nucleus segmentation quality based on patch-level texture features. The decision to use patch-level texture features is based on our experience that images with similar texture features in tissue usually have comparable segmentation quality given a set of segmentation parameters.

Our experimental results show that the proposed approach is able to predict segmentation quality at the patch level and performs well across a range of images. Since the texture appearance for different cancer types may vary

Table 5: Classification performance for each cancer type in each category

	Random forest			SVM		
	Precision (%)	Recall (%)	F1 score (%)	Precision (%)	Recall (%)	F1 score (%)
WHO Grade II LGG						
Good	76.68	94.62	84.71	78.53	86.49	82.32
Under	95.56	95.43	95.49	89.94	100.00	94.71
Over	98.36	59.00	73.76	98.34	66.24	79.16
LUAD						
Good	74.76	76.12	75.43	65.09	58.00	61.34
Under	66.19	45.03	53.60	47.44	37.55	41.92
Over	76.79	47.29	58.54	74.45	48.62	58.82

LGG: Lower grade glioma, LUAD: Lung adenocarcinoma, SVM: Support vector machine

Table 6: Classification performance for each cancer type by using good-quality-segmented versus bad-quality-segmented

	Random forest			SVM		
	Precision (%)	Recall (%)	F1 score (%)	Precision (%)	Recall (%)	F1 score (%)
WHO Grade II LGG						
Good	85.16	87.05	86.09	80.60	88.65	84.43
Bad	92.05	90.81	91.43	92.68	87.08	89.80
LUAD						
Good	67.64	69.89	68.74	44.00	32.76	37.56
Bad	81.61	79.99	80.79	65.09	75.05	69.71

LGG: Lower grade glioma, LUAD: Lung adenocarcinoma, SVM: Support vector machine

significantly, different texture features would be selected for the classification model and the training phase would result in a different classification model. While the model trained in our experimental evaluation cannot be directly applied to other cancer types, our pipeline can be generalized to other cancer types by training new classification models for other cancer types. By comparing the results of applying our pipeline on two different cancer types, we found that our pipeline works better when applying to the images with less heterogeneity of nuclei appearance and more consistent texture characteristics. The proposed approach can be applied to segmentation methods that make use of the texture characteristics of tissue to detect and delineate nuclear boundaries. However, if a segmentation method adapts its parameters across images, our method may not be suitable to check for segmentation quality since similar texture features may have different segmentation results in that case. We are in the process of involving the size, shape, and texture features of the segmented objects in our segmentation quality assessment pipeline and will report on our findings in the future work.

Financial support and sponsorship

Nil.

Conflicts of interest

There are no conflicts of interest.

References

1. Gurcan MN, Boucheron LE, Can A, Madabhushi A, Rajpoot NM, Yener B, *et al.* Histopathological image analysis: A review. IEEE Rev Biomed Eng 2009;2:147-71.
2. Chen XS, Wu JY, Huang O, Chen CM, Wu J, Lu JS, *et al.* Molecular subtype can predict the response and outcome of chinese locally advanced breast cancer patients treated with preoperative therapy. Oncol Rep 2010;23:1213-20.
3. Leo P, Lee G, Madabhushi A. Evaluating stability of histomorphometric features across scanner and staining variations: Predicting biochemical recurrence from prostate cancer whole slide images. In: SPIE Medical Imaging. Bellingham, Washington USA: International Society for Optics and Photonics; 2016. p. 97910I.
4. Ojansivu V, Linder N, Rahtu E, Pietikäinen M, Lundin M, Joensuu H, *et al*. Automated classification of breast cancer morphology in histopathological images. Diagn Pathol 2013;8:S29.
5. Romo-Bucheli D, Janowczyk A, Romero E, Gilmore H, Madabhushi A. Automated tubule nuclei quantification and correlation with oncotype DX risk categories in ER+ breast cancer whole slide images. In: SPIE Medical Imaging. Bellingham, Washington USA: International Society for Optics and Photonics; 2016. p. 979106.
6. Yu KH, Zhang C, Berry GJ, Altman RB, Ré C, Rubin DL, *et al.* Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. Nat Commun 2016;7:12474.
7. Chen W, Chu V, Hu J, Yang L, Wang F, Kurc T, *et al*. ImageMiner: A medical image analysis and image management UML data model. APIII: Advancing Practice. Pittsburgh, PA: Instruction & Innovation Through Informatics; 2009.
8. Foran DJ, Yang L, Chen W, Hu J, Goodell LA, Reiss M, *et al.* ImageMiner: A software system for comparative analysis of tissue microarrays using content-based image retrieval, high-performance computing, and grid technology. J Am Med Inform Assoc 2011;18:403-15.
9. Ren J, Sadimin E, Foran DJ, Qi X. Computer aided analysis of prostate histopathology images to support a refined Gleason grading system. In: SPIE Medical Imaging. Bellingham, Washington USA: International Society for Optics and Photonics; 2017. p. 101331V.
10. Yang L, Tuzel O, Chen W, Meer P, Salaru G, Goodell LA, *et al.* PathMiner: A web-based tool for computer-assisted diagnostics in pathology. IEEE Trans Inf Technol Biomed 2009;13:291-9.
11. Gurcan MN, Kong J, Sertel O, Cambazoglu BB, Saltz JH, Çatalyürek ÜV. Computerized pathological image analysis for neuroblastoma prognosis.

In: AMIA. Oxford Univerisity Press: Citeseer; 2007.
12. Kong J, Sertel O, Boyer KL, Saltz JH, Gurcan MN, Shimada H, *et al.* Computer-assisted grading of neuroblastic differentiation. Arch Pathol Lab Med 2008;132:903-4.
13. Sertel O, Kong J, Catalyurek UV, Lozanski G, Saltz JH, Gurcan MN. Histopathological image analysis using model-based intermediate representations and color texture: Follicular lymphoma grading. J Signal Process Syst 2009;55:169.
14. Irshad H, Veillard A, Roux L, Racoceanu D. Methods for nuclei detection, segmentation, and classification in digital histopathology: A review current status and future potential. IEEE Rev Biomed Eng 2014;7:97-114.
15. Xing F, Yang L. Robust nucleus/cell detection and segmentation in digital pathology and microscopy images: A comprehensive review. IEEE Rev Biomed Eng 2016;9:234-63.
16. Bamford P, Lovell B. A methodology for quality control in cell nucleus segmentation. DICTA 1999;1:21-5.
17. Cukierski WJ, Nandy K, Gudla P, Meaburn KJ, Misteli T, Foran DJ, *et al.* Ranked retrieval of segmented nuclei for objective assessment of cancer gene repositioning. BMC Bioinformatics 2012;13:232.
18. Nandy K, Gudla PR, Amundsen R, Meaburn KJ, Misteli T, Lockett SJ, *et al.* Automatic segmentation and supervised learning-based selection of nuclei in cancer tissue images. Cytometry A 2012;81:743-54.
19. Brinker A, Fredrikson A, Zhang X, Sourvenir R, Zhang S. Toward consistent cell segmentation: Quality assessment of cell segments via appearance and geometry features. In: SPIE Medical Imaging. Bellingham, Washington USA: International Society for Optics and Photonics; 2015. p. 942000.
20. Zhang H, Fritts JE, Goldman SA. Image segmentation evaluation: A survey of unsupervised methods. Comput Vis Image Underst 2008;110:260-80.
21. Network TC. Comprehensive molecular portraits of human breast tumours. Nature 2012;490:61.
22. Gao Y, Ratner V, Zhu L, Diprima T, Kurc T, Tannenbaum A, *et al*. Hierarchical nucleus segmentation in digital pathology images. In: SPIE Medical Imaging.Bellingham, Washington USA: International Society for Optics and Photonics; 2016. p. 979117.
23. Esfahani MS, Dougherty ER. Effect of separate sampling on classification accuracy. Bioinformatics 2013;30:242-50.
24. Guyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res 2003;3:1157-82.
25. Caruana R, Niculescu-Mizil A. An Empirical Comparison of Supervised Learning Algorithms. In: Proceedings of the 23rd International Conference on Machine Learning. ACM; 2006. p. 161-8.
26. Breiman L. Random forests. Mach Learn 2001;45:5-32.
27. Chang CC, Lin CJ. LIBSVM: A library for support vector machines. ACM Trans Intell Syst Technol 2011;2:1-27.
28. Sharma A, Kazerouni A, Saghar YN, Commean P, Tarbox L, Prior F. Framework for Data Management and Visualization of the National Lung Screening Trial Pathology Images. Pathology Informatics Summit; 2014.