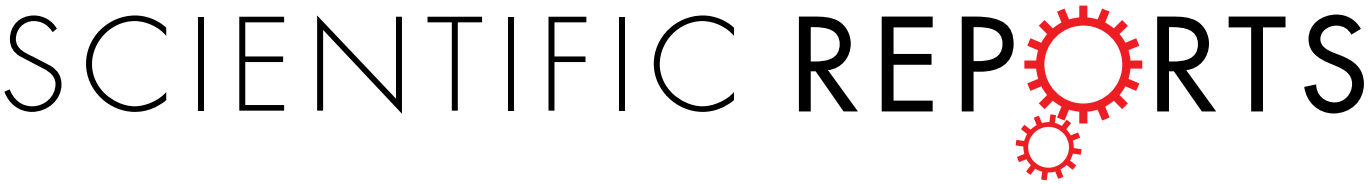

OPEN

A Cluster-then-label Semi-supervised Learning Approach for Pathology Image Classification

Received: 15 November 2017
Accepted: 11 April 2018
Published online: 08 May 2018

Mohammad Peikari[1], Sherine Salama[2], Sharon Nofech-Mozes[2] & Anne L. Martel[1,3]

Completely labeled pathology datasets are often challenging and time-consuming to obtain. Semi-supervised learning (SSL) methods are able to learn from fewer labeled data points with the help of a large number of unlabeled data points. In this paper, we investigated the possibility of using clustering analysis to identify the underlying structure of the data space for SSL. A cluster-then-label method was proposed to identify high-density regions in the data space which were then used to help a supervised SVM in finding the decision boundary. We have compared our method with other supervised and semi-supervised state-of-the-art techniques using two different classification tasks applied to breast pathology datasets. We found that compared with other state-of-the-art supervised and semi-supervised methods, our SSL method is able to improve classification performance when a limited number of labeled data instances are made available. We also showed that it is important to examine the underlying distribution of the data space before applying SSL techniques to ensure semi-supervised learning assumptions are not violated by the data.

Traditionally, there have been two fundamentally different tasks in the spectrum of pattern recognition and machine learning methods. On one side is *supervised learning* in which the goal is to learn a model from labeled data points. The learned model is then applied to an unseen test set and the method is validated based on how successful it was in assigning test data to different classes. The disadvantage of supervised learning techniques is that they are limited to learning from labeled datasets which are often expensive, time consuming, or difficult to generate. If the available labeled dataset is too small and does not represent the true variance of the data space then generalization performance may be poor. This issue is even more decisive in the medical image analysis domain since generating high quality datasets requires the effort of experienced and trained human observers. On the other side of the spectrum are the *unsupervised learning* methods in which unlabeled data points are grouped into clusters that share similar properties. Unlabeled datasets are often easier to acquire and require less human effort to create, however, since the information provided to these techniques is unlabeled, there is no clear way to validate the quality of this approach. In contrast to supervised learning, which only considers labeled data, and unsupervised learning which works only on unlabeled data, semi-supervised learning (SSL) methods work with both labeled and unlabeled data points. Therefore, by using SSL, it is possible to combine the advantages of working with a small labeled dataset to guide the learning process and a larger unlabeled dataset to increase the generalizability of the found solution as shown in Fig. 1[1].

Pathology images are an important source of diagnostic and prognostic information. With the advent of whole slide scanner technologies, pathology slides are being digitized at microscopic resolution making it possible to store and analyze digital pathology images using computer systems. This has led to a rapidly growing field of research into machine learning techniques that can be used to classify images and provide quantitative information. A major difficulty facing researchers is the availability of labeled training data. Whole slide pathology images are orders of magnitude larger than other medical images and they are more complex. Pathologists use a combination of color, texture and morphological information that varies across multiple scales to interpret images and spend many years learning how to cope with enormous variability in the appearance of specific tissue and disease types. This means that it requires an expert to provide ground-truth labels and it also means that for every new application, additional training and validation data is needed; this makes the use of semi-supervised learning particularly relevant for digital pathology.

[1]Medical Biophysics, University of Toronto, Toronto, Canada. [2]Laboratory Medicine and Pathobiology, University of Toronto, Toronto, Canada. [3]Physical Sciences, Sunnybrook Research Institute, Toronto, Canada. Correspondence and requests for materials should be addressed to M.P. (email: mpeikari@sri.utoronto.ca)

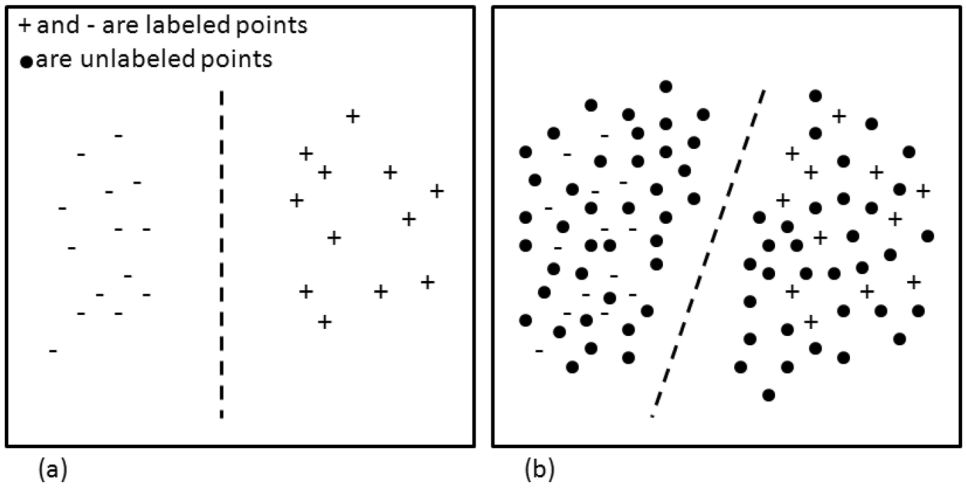

Figure 1. Semi-supervised learning tries to increase the generalization of classification performance by placing the decision boundary through the sparse regions in presence of both labeled and unlabeled data points. (**a**) The decision boundary in presence of labeled data points only, and (**b**) the decision boundary in presence of both labeled and unlabeled data.

In this paper, we present a semi-supervised learning method that analyzes groups of labeled and unlabeled points in multidimensional feature space in order to identify areas of high density and then guides the learning method to place decision boundaries through the regions with low density. We apply this technique to the analysis of digital pathology images of breast cancer.

Related Works

Semi-supervised learning methods are not commonly used in the pathology image analysis field although they have previously been employed in some applications of medical image analysis to improve classification performances on partially labeled datasets[2–5]. In order to make it possible for semi-supervised learning methods to make the most of the labeled and unlabeled data, some assumptions are made for the underlying structure of data space[1]. Among the assumptions, *smoothness* and *cluster assumption* are the basis for most of the state-of-the-art techniques[6]. In the smoothness assumption, it is assumed that points that are located close to each other in data space are more likely to share the same label, and in the cluster assumption, it is assumed that the data points that belong to one class are more likely to form and share a group/cluster of points. Therefore, the core objective of these two assumptions is to ensure that the found decision boundary lies in low density rather than high density regions within data space.

The most basic and easiest SSL method to apply is self-training[7–10], which involves repeatedly training and retraining a statistical model. First, labeled data is used to train an initial model and then this model is applied to the unlabeled data. The unlabeled points for which the model is most confident in assigning labels to, are then added to the pool of labeled points and a new model is trained. This process is repeated until some convergence criterion is met. Another family of methods is based on generative models[11–13], in which some assumptions are made about the underlying probability distribution of data in feature space. The parameters defining the assumed generative model are then found by fitting the model to the data. Graph-based SSL techniques[14–17], attempt to generate an undirected graph on the training data in which every point on the graph is connected by a weighted edge. The weights are assigned to the edges in such a way that closer data points tend to have larger weights and hence they likely share same labels. Labels are assigned to the unlabeled points by propagating labels of labeled points to unlabeled ones through the edges of the graph with the amount dependent on the edge weights. This way unlabeled points can all be labeled even if they are not directly connected to the labeled points.

The support vector machine (SVM) classifier is an efficient and reliable learning method and to date is one of the best classifiers in terms of performance[18] over a wide range of tasks. Semi-supervised SVM techniques expand the idea of traditional SVM to incorporate the ability to use partially labeled datasets to learn reliable models while maintaining accuracy. The idea is to minimize an objective function by examining all possible label combinations of the unlabeled data iteratively in order to find low-density regions in the data space to place the decision boundary through[19–22]. Many implementations of the objective functions have been reported in the literature however these are often time inefficient. The reader is referred to Chapelle *et al.*'s work[23] for a review comparing different methods. Kernel tricks which implement the cluster assumption in SSL have also been proposed[24,25].

Recently, there have been some attempts to replace the lengthy objective function optimization process of semi-supervised SVMs by cluster analysis[6,26,27]. The concept behind these cluster-then-label techniques for semi-supervised learning[28] is to first find point clusters of high density regions in data space and then assign labels to the identified clusters. A supervised learner is then used to find the separating decision boundary that passes through low density regions in data space (i.e. between the clusters). In this study, we propose a novel cluster-then-label semi-supervised learning method and compare its performance with other state-of-the-art techniques for two digital pathology tasks; triaging clinically relevant regions of breast whole mount images[29] and the classification of nuclei figures into lymphocyte, normal epithelial and malignant epithelial objects.

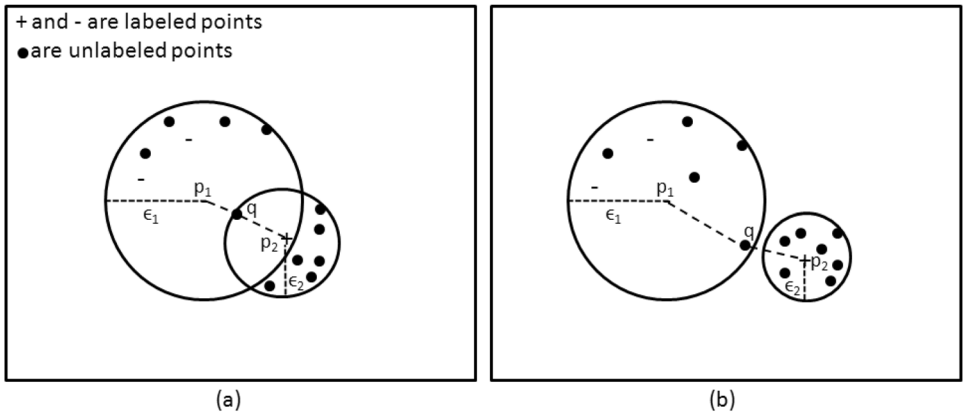

Figure 2. Example showing different scenarios where an unlabeled point q is located with respect to the clusters formed by labeled points p_1 and p_2. (**a**) An unlabeled point q is within the core radii of two labeled points p_1, and p_2. Since $\varepsilon_2 < \varepsilon_1$, q is assigned to p_2 despite the fact that, according to the Euclidean distance, it is actually closer to p_1. (**b**) An unlabeled point q is within the core radius of the labeled point p_1 but not p_2. Since $\varepsilon_1 > ||p_2 - q||_2$, q is again assigned to p_2. ε_1, and ε_2 are the core radii of the labeled points p_1 and p_2 respectively. Please note that the ε_1 and ε_2 are different based on the density of points surrounding them. In this representation, k is set to be 7.

Methodology

Proposed Method. In an earlier work[30], we demonstrated that a semi-supervised cluster-then-label method was able to produce a reliable classification model from small amounts of labeled data. In this study, we propose an improvement of the method proposed in our earlier study[30] and we carry out an extensive experimental comparison with other state-of-the-art semi-supervised techniques on two different pathology image classification tasks.

Clustering Analysis for Semi-supervised learning. Inspired by the work published by Ankerst *et al.*[31], we propose a cluster-then-label based SSL method that works by finding the underlying structure of points (clusters of points forming high density regions) in the data space. A standard supervised SVM is then employed to find the decision boundary using knowledge about the underlying structure of the data space provided by the clustering analysis. In the Ankerst *et al.*'s[31] study, an ordering of points in the data space was found based on how points are spatially located around each other. Therefore, spatially closest points become neighbors in the ordering set. The clustering approach presented by Ankerst *et al.*[31] is unsupervised and finding the clusters from the ordering set is a challenge.

In this paper, our approach in finding spatially closest points in the data space is somewhat similar to the one proposed by Ankerst *et al.*[31], in the sense that points are grouped in such a way that they form clusters of densely populated points separated by regions with sparsely located points (low density). We consider a semi-supervised seeded approach to finding spatially closest points and checking how inclined unlabeled points are toward each of their surrounding labeled points. The algorithm starts by calculating the *core* radii of the labeled points with respect to all points in the data space. A labeled/unlabeled point q is located at a core radius to a labeled point p if it is within a circle/sphere with p as its center and ε as its radius. The value of ε for every point p is defined as the distance from p to the kth closest point located within the neighborhood of p. Parameter k is the minimum number of points located at the neighboring of p that could form a cluster and is specified by the user. Thus, the core radius is low in high density regions and high in low density regions. For this study, the value of k was set to one tenth of the number of points in the data space. This value of k showed a more consistent performance in our preliminary experiments[30].

The core radii of all labeled points with respect to the whole data space are calculated. We define a distance matrix D of the size $l \times u$ where l is the number of labeled points and u is the number of unlabeled points. The Euclidean distance between the labeled point p_i and the unlabeled point q_j is compared to the core distance ε_i, and d_{ij} is defined as the maximum of these two values. Therefore, this could be written as:

$$d_{ij} = max(\varepsilon_i, ||p_i - q_j||_2); \forall_{i=1}^{l} \text{ and } \forall_{j=1}^{u} \quad (1)$$

where ε_i is the core radius of the labeled point p_i, l is the number of labeled points, and u is the number of unlabeled points.

Once the matrix D has been populated using expression (1), it is used to find the closest labeled point for each unlabeled point. Figure 2 shows an example of how core radii are useful in assigning unlabeled points to different clusters and highlights differences between conventional clustering methods. In Fig. 2(a) the unlabeled point q is located within the core radii of both the labeled points p_1 and p_2. Since $\varepsilon_2 < \varepsilon_1$, q is assigned to p_2 despite the fact that, according to the Euclidean distance, it is actually closer to p_1. In Fig. 2(b) the unlabeled point q is within the core radius of p_1 and lies outside of the core radius of p_2 however, since $\varepsilon_1 > ||p_2 - q||_2$, q is again assigned to p_2. In cases where an unlabeled point q is equidistant between two points with different labels, q is assigned a negative label.

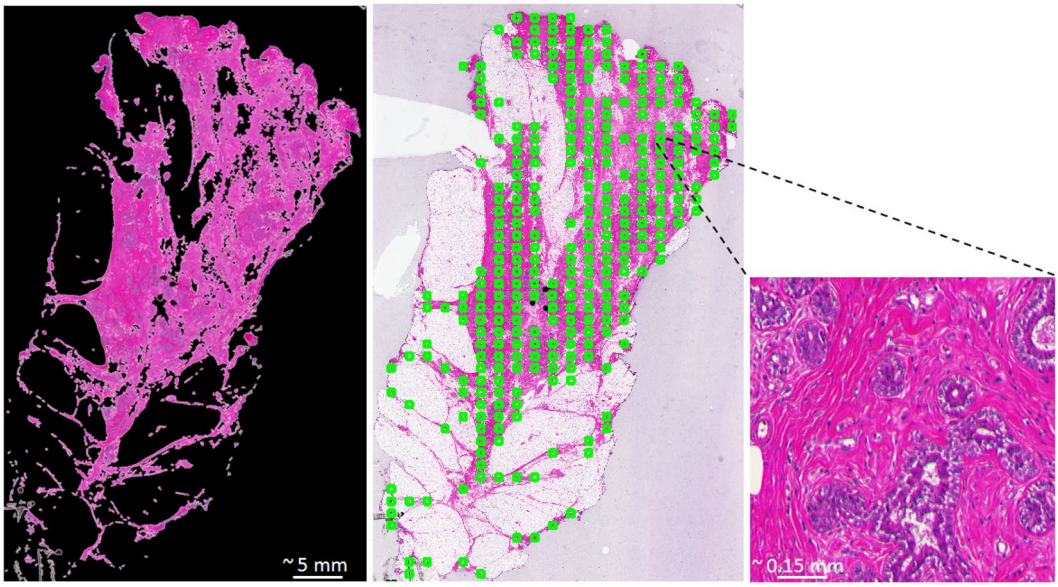

Figure 3. Adaptive thresholding and morphological operations were applied to remove clearly irrelevant structures before patch selection for the data collection process (left), 512 × 512 pixel uniformly spaced box patches (in green) on tissue regions (middle), and an example of a 512 × 512 pixel image patch picked from the WMI (right).

This, however, rarely happens in practice as distances are represented as floating point values. Hence, this method finds and groups the points that match both the smoothness and cluster assumptions in SSL and is referred to as *Semi-supervised Seeded Density-based* (S^3DB) clustering hereafter.

After the groups of points that form clusters are identified and the underlying structure of the data space is learned (using S^3DB), this knowledge is given to an SVM classifier with radial basis function (RBF) as kernel to find the maximum margin boundary that passes through the sparse regions.

Datasets. *Pathology Triaging Image Dataset*. Recently[29], we addressed the problem of triaging digital pathology images and employed a supervised learning method to distinguish between different relevant or irrelevant breast tissue regions. Here, we consider a more extensive dataset and focus on the statistical learning aspect of the problem. The goal is to achieve a high sensitivity of at least 95% in detecting relevant regions while maintaining the highest possible specificity.

Data Collection: To generate a ground-truth dataset, we have used whole-mount images (WMIs)[32] of 30 breast lumpectomy specimens stained with hematoxylin and eosin (H&E) ($n = 150$ WMIs). The slides corresponding to 28 of the patients were scanned at 5X magnification (135 WMIs, 2 μm/pixel) and 2 of them were scanned at 10X (15 WMIs, 1 μm/pixel). Patches of 512 × 512 pixels (1 mm^2 for 5X and 0.25 mm^2 for 10X images) were cropped from each WMI at the highest magnification by overlaying a grid of uniformly spaced squares on the previously preprocessed (adaptive thresholding and morphological operations) tissue regions (Fig. 3). The collaborating pathologist then labeled patches from the 2 patients scanned at 10X magnification (15 WMIs, 2849 patches) and 8 patients scanned at 5X magnification(115 WMIs, 2302 patches labeled, 2100 patches unlabeled). For each patch, the pathologist evaluated the presence of diagnostically relevant information corresponding to each tissue type. According to the pathologist's annotations, diagnostically relevant features include cancers, atypias, microcalcifications and lymphovascular invasion, and irrelevant features include fat, stroma, normal ducts and lobules. To assess inter-observer variability when labeling the triaging ground-truth set, a random subset of 1500 patches were evaluated by a second pathologist. The Kappa agreement coefficient between the two pathologists was $\kappa = 0.77$.

Figure 4 shows a subset of this ground-truth set. We have also added 1500 unlabeled image patches from the remaining 20 patients scanned at 5X (20 WMIs). This set of unlabeled patches was used to improve the generalization performance of the learning models as mentioned in section 3.3.1.

Texture Feature Extraction from Patches: To retrieve texture features from image patches, they were converted from RGB to Lab colorspace and the normalized luminance channel was divided into smaller non-overlapping tiles of size 32 × 32 pixels. Root filter set (RFS)[33] texture filters were used to highlight different textures from image tiles. First order statistical measures (mean, mode, standard deviation, skewness and kurtosis) were calculated from the maximal filter responses along all filter orientations of each scales to combine the texture information. To regroup all extracted information from individual tiles and form one numerical representative per image patch, the bag of words (BoW)[34] technique was used with a previously found optimum dictionary size of 100[29]. The calculated 100-dimensional histograms of words per individual image patch were used to train and evaluate the statistical learning techniques presented in this paper. We used the RBF kernel of the SVM classifier implemented in libsvm library[35] to find the best separating hyperplane between the two classes.

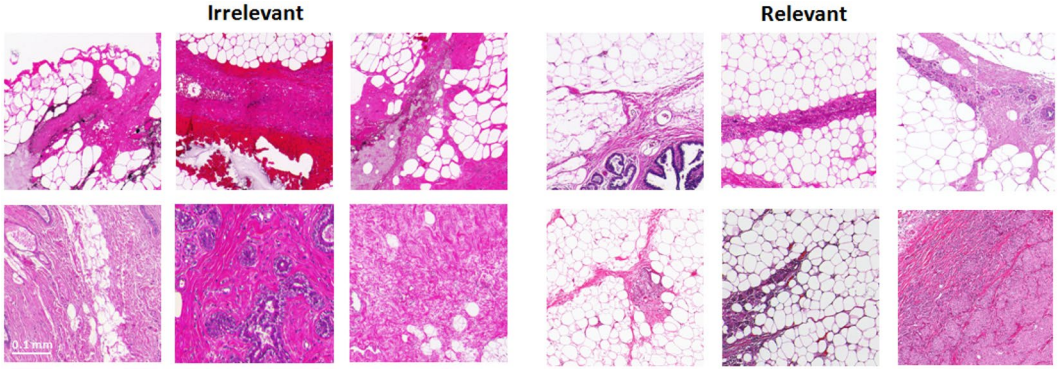

Figure 4. A subset of cropped image patches used as ground-truth in the triaging dataset used for this study with their labels. It is clear that the clinically relevant information may have covered different portions of the patches in the dataset since they were randomly picked from different areas within the tissue region.

	Dataset	
	Triaging Image Dataset	Nuclei Figure Classification
# of labeled instances in training set	2,302	13,821
# of unlabeled instances in training set	2,100 + 1,500	49,000
# of instances in test set	2,849	7,958
Dimensionality of feature vectors	100	125

Table 1. Summary of the data proportions used in each validation stage for the two pathology datasets used in this study.

Nuclei Figure Classification Dataset. Recently[36], we developed an automated method to assess cancer cellularity in breast tissue removed after neoadjuvant chemotherapy (NAT). As a part of the pipeline, we developed a method to classify nuclei figures into three classes of lymphocyte (L), benign epithelial (B) and malignant epithelial (M) figures from a dataset of image patches annotated by a pathologist. Here, we used the same dataset to validate the proposed SSL technique.

Data Collection: H&E stained sections from 92 post-NAT lumpectomy specimens were scanned at 20X magnification (0.5 μm/pixel). The whole slide images (WSIs) were annotated by an expert pathologist using the Sedeen Viewer[37] (Pathcore, Toronto, Canada). A total of $n = 166$ rectangular regions of interest (ROIs) were defined on the 92 WSIs and, within these ROIs, the centers of nuclei figures were labeled as either lymphocyte, benign epithelial, or malignant epithelial. Nuclei that were out of focus, out of plane, or could not be categorized, were not marked by the pathologist. More than 30,000 nuclei figures ($n = 3,868$ lymphocyte, $n = 10,407$ benign epithelial, and $n = 16,419$ malignant epithelial figures) were marked from all 116 ROI patches.

Nuclei Feature Extraction from Patches: In order to train the proposed SSL method, the nuclei have to be segmented first. We have developed a segmentation method[38] that works by manipulating the original RGB colorspace of the image patches to better identify foreground nuclei figures. Multilevel thresholding and marker controlled watershed algorithms were then used to extract nuclei regions and divide overlapping nuclei figures. The nuclei segmentation method achieved an F1-score of 0.9 when tested against a publicly available dataset of 7931 nuclei from 36 images[39]. The effect of color variation from the image patches was reduced by standardizing their color to a reference image as explained in our recent study[36]. The segmentation method was able to segment more than 72% of the ground-truth nuclei figures ($n = 21,779$) from the 166 ROI patches. The segmented figures were used to extract 125-dimensional feature vectors from individual nuclei figures based on intensity, morphological, textural, and spatial properties describing their differences among the three classes[36]. Table 1 summarizes the datasets used to validate and compare performances of the supervised and semi-supervised learning methods described in section 3.1.

Experimental Setup

Comparison with State-of-the-art Methods. We compare our proposed SSL method ($S^3DB + SVM$) with a range of successful supervised and semi-supervised methods in the literature.

Method for supervised learning. The standard supervised SVM technique implemented in the libsvm[35] library was considered to find the separating decision hyperplane. Here we used the RBF kernel and a similar parameter optimization approach to other methods described in this paper was followed, as explained in the subsequent section. Let $X = x_1, x_2, \ldots, x_l$ be the set of d-dimensional labeled points with $Y = y_1, y_2, \ldots, y_l$ be their labels. The SVM technique works by minimizing the optimization function presented in equation (2) to find the maximum margin hyperplane parameters dividing the two classes.

$$min_{\vec{w},b,\xi,C}\frac{1}{2}||\vec{w}||^2 + C\sum_{i=1}^{l}\xi_i \tag{2}$$

subject to:

$$\forall_{i=1}^{l}: y_i[\vec{w}.\ \Phi(\vec{x_i}) + b] \geq 1 - \xi_i$$
$$\forall_{i=1}^{l}: \xi_i > 0$$

where $\vec{w}$ and b are the parameters defining the maximum margin hyperplane, $\Phi(.)$ is the kernel function, C is the parameter defining the trade-off between the margin size and misclassified examples, and ξ is the slack variable.

Methods for semi-supervised learning. *a) semi-supervised Fuzzy c-mean (ssFCM) clustering + SVM*: this method has been previously employed for semi-supervised learning[2,27,40,41]. The idea is to first apply the semi-supervised clustering to both labeled and unlabeled data to find the underlying structure of the space (hard labeling) and then a supervised classifier is trained on the labeled data. The semi-supervised version of the original unsupervised FCM in particular is useful to provide a prior knowledge on the structure of the space in the form of labels[27]. Therefore, in the following optimization problem, the first term is to discover the data space structure of the labeled data and the second term takes care of the unlabeled data. Let $X = x_1, x_2, \ldots, x_l$ be the set of d-dimensional labeled points with $Y = y_1, y_2, \ldots, y_l$ be their labels and $X^* = x_1^*, x_2^*, \ldots, x_u^*$ be the set of unlabeled points, then the ssFCM objective function can be written as:

$$min\rho_m = \sum_{i=1}^{c}\sum_{j=1}^{l} u_{ij}^m ||X_j - V_j||_2 + \sum_{i=1}^{c}\sum_{j=1}^{u} u_{ij}^{*m} ||X_j^* - V_j||_2 \tag{3}$$

where c is the number of classes ($c = 2$ for binary classification), m is the degree of fuzziness (we set $m = 2$), V represents the set of prototypes corresponding with each class, U and U^* are matrices that define the fuzzy membership values for the labeled and unlabeled data points respectively and u_{ij} is the probability that the jth labeled data point belongs to class i. The maximum number of iterations for the experiments using this method was set to 1000 rounds.

We have used the RBF-SVM classifier in conjunction with the semi-supervised FCM method similar to the one presented by Gan *et al.*[27]. The parameters of the SVM classifier was optimized using a similar strategy to other methods described in this paper as explained in section 3.3.

b) TSVM[19,42]: this SSL method is one of the most successful implementations of the semi-supervised SVM technique in terms of performance[23]. The algorithm starts by learning a partially complete model using the labeled data only and then applies this to the unlabeled data. The method then improves the initial solution by switching the labels assigned to the unlabeled data to decrease the objective function after each iteration. The label switching mechanism is important to ensure the balancing constraints between the two classes are maintained. Therefore, the main objective function to be minimized in this method is as follow:

$$min_{y_1^*, \ldots, y_u^*, \vec{w}, b, \xi, \xi^*, C, C^*}\frac{1}{2}||\vec{w}||^2 + C\sum_{i=1}^{l}\xi_i + C^*\sum_{j=1}^{u}\xi_j^* \tag{4}$$

subject to:

$$\forall_{i=1}^{l}: y_i[\vec{w}.\ \Phi(\vec{x_i}) + b] \geq 1 - \xi_i$$
$$\forall_{j=1}^{u}: y_j^*[\vec{w}.\ \Phi(\vec{x_j^*}) + b] \geq 1 - \xi_j^*$$
$$\forall_{i=1}^{l}: \xi_i > 0$$
$$\forall_{j=1}^{u}: \xi_j^* > 0$$

Visualization and Cluster Separability. To visualize the underlying distributions of data spaces used in this study in lower dimensions, t-distributed Stochastic Neighbor Embedding (t-SNE) was used[43]. It is an iterative method which maps data points into lower dimensional space in such a way that the distances between points correspond to their similarity. Also, we have used Fisher Discriminant Ratio (FDR)[44–46], as a measure of cluster separability. FDR measures the cluster separability by calculating the square of the difference between means of points in each cluster divided by the sum of square of their standard deviations:

$$FDR = \frac{(\bar{m}_1 - \bar{m}_2)^2}{\bar{s}_1^2 + \bar{s}_2^2} \tag{5}$$

where $\bar{m}_1$ and $\bar{m}_2$ are the mean of points; and $\bar{s}_1$ and $\bar{s}_2$ are the standard deviations of points in clusters 1 and 2 respectively.

Experimental Design. In order to evaluate the performance of the proposed SSL technique and compare with state-of-the-art methods, the following validation steps were taken.

Triaging Image Dataset. Here, the dataset described in section 2.2.1 was divided into a training set containing the patches scanned at 5X magnification, and a testing set containing the 2 patient datasets scanned at 10X ($n = 2849$ labeled image patches). The training data was further subdivided into two components; one part contained the labeled and unlabeled patches from the 8 patients reviewed by the pathologist ($n = 4402$ patches, 307 relevant, 1995 irrelevant, and 2100 unlabeled image patches with the mean and standard deviation of 283 ± 90 labeled patches and 267 ± 243 unlabeled patches per patient) and the second part consisted of the of 1500 unlabeled image patches taken from the remaining patients' WMIs (section 2.2.1).

An 8-fold patient-wise cross-validation scheme was used to train and validate the performance of learning methods. The optimum SVM-RBF parameters were chosen by examining a range of possible SVM trade-off parameter (C) and the kernel width (γ) values on the training set. The additional set of 1500 unlabeled image patches (section 2.2.1) was included in all folds of the cross-validation scheme.

Validation step for semi-supervised methods: For every fold in semi-supervised learning methods, one or more of the patient datasets were randomly selected to be the labeled set (unlabeled images of chosen patients were kept unlabeled) and labels of the rest of patients were kept hidden (unlabeled set).

Validation step for supervised method: Similarly, for the supervised learning method, in every fold one or more patient datasets were randomly selected to be the labeled set (unlabeled images of chosen patients were discarded) and rest of the patients data were also discarded.

The randomly selected patients, dictionary of words, and histograms of words were kept the same to form paired labeled sets in every fold of each experiment. To do a fair comparison between different methods, we defined the optimum SVM-RBF parameter set by first identifying all sets that produced a sensitivity of 95%; from which the set with maximum specificity was chosen.

Validation using an unseen set: To compare the generalization performance of the methods, the median of the optimized parameters found in all 8-folds of the cross-validation was considered to train an overall model using all training images. For semi-supervised methods, one or more of patients were randomly chosen to form the labeled set and the labels of the rest of patients were kept hidden. For the supervised method also, one or more of patients data were chosen to form the labeled set and the rest of patients data were discarded. The overall performance of the trained model was assessed using the the two unseen patient cases in the test set. To match our previously trained models on 5X magnified images, the test image patches, which were scanned at 10X magnification, were down-sampled.

Nuclei Figure Classification Dataset. The aim of this experiment was to see whether adding many unlabeled instances to an already large set of labeled instances improved the classification performance when comparing an SSL technique with a supervised learning method. A cascaded learning approach was used to first train a classifier to distinguish between lymphocyte versus epithelial figures (L vs. BM) and then to distinguish between benign versus malignant classes (B vs. M).

The supervised SVM was trained using $n = 13{,}821$ labeled nuclei figures ($n = 2{,}260$ Lymphocytes, $n = 3{,}157$ Benign epithelial, and $n = 8404$ Malignant epithelial figures). Both the labeled figures and an additional $n = 49{,}000$ unlabeled figures were used by the semi-supervised methods.

For both supervised and semi-supervised training, a 5-fold cross-validation was performed to assess the performance of the learning methods. In this experiment, the best parameters were chosen in such a way as to maximize the accuracy.

Once the best parameters had been selected a final model was trained on the whole training dataset using the median of the best parameters in all 5 folds and this was applied to an unseen test set of $n = 7{,}958$ nuclei figures to evaluate the generalizability of the trained models.

Results

Comparing Classification Performances. Mean accuracy of the subject-wise cross-validated experiments are shown in Fig. 5 for different number of patients chosen to be the labeled set from the pathology triaging image dataset. As can be seen from Fig. 5, our clustering-based SSL technique (S^3DB + SVM) achieved a superior performance compared to the other state-of-the-art supervised and semi-supervised methods.

Table 2 summarizes the cross-validated performance of different methods on triaging image dataset at an operating point of 95% sensitivity. A pairwise Wilcoxon Signed-Rank test using 8 cross-validated accuracy values was used to compare each method at a given number of labeled patients to that of our proposed SSL technique. For each method tested, we had 6 comparisons, therefore for a two-tailed test with a 5% type I error the Bonferroni adjusted α-value = 0.004. Although none of the comparisons achieved statistical significance against this adjusted threshold, there is an increasing trend in the classification performance of our proposed method compared to other techniques. It is clear by looking at the specificity column that in most cases our method maintained higher specificity values at 95% sensitivity compared to the other methods at all individual experiments (except when number of labeled patients = 1). It is also interesting to note that the average train time of our method is significantly lower than the TSVM technique, which requires a heavy optimization on a 64-bit Intel(R) Xeon(R) CPU (at 3.50 GHz) machine. The supervised SVM method however, took the least amount of time to train a model using the labeled data made available to it. Table 3 summarizes the performance of the four methods using an overall model trained in the first validation phase using number of labeled patients as 3 on a totally unseen test set. The test set consisted of two patient cases scanned at 10X magnification. It is clear that our method consistently performs better in classifying image patches compared to the other supervised and semi-supervised techniques in a totally unseen test set.

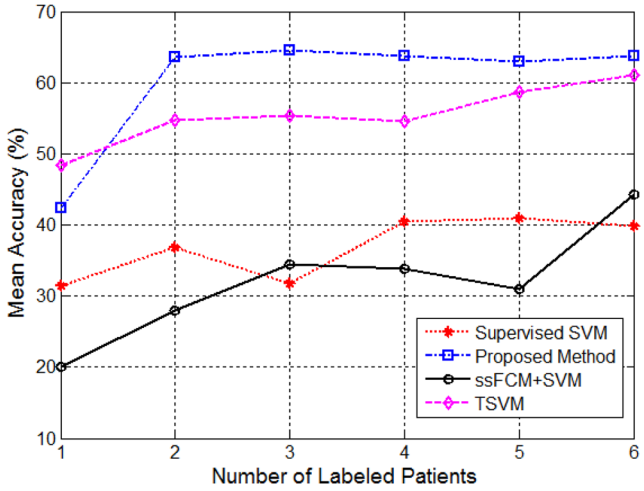

Figure 5. Comparison of mean classification accuracy for 8-fold subject-wise cross-validation of the four supervised and semi-supervised methods discussed in this paper for the breast pathology dataset at different labeled portions. To have a fair comparison, results are reported at an operating point of 95% sensitivity.

Method	# of Labeled Patients	AUC	Accuracy (SD) (%)	95% CI	Specificity (%)	p-value	Train Time (min)
SVM	1	0.71	31.4 (17.2)	[17.0, 45.8]	19.9	0.253	1 ± 0.1
	2	0.76	36.8 (23.8)	[16.9, 56.7]	27.8	0.015	2 ± 0.2
	3	0.75	31.8 (18.3)	[16.5, 47.1]	21.4	0.007	4 ± 0.2
	4	0.76	40.4 (22.6)	[21.5, 59.3]	32.4	0.007	6 ± 0.5
	5	0.78	41.0 (24.1)	[20.8, 61.1]	33.6	0.039	8 ± 0.8
	6	0.79	39.8 (23.6)	[20.1, 59.5]	31.9	0.007	13 ± 1
ssFCM + SVM	1	0.62	20.1 (15.7)	[7.0, 33.2]	5.2	0.007	97 ± 11
	2	0.70	28.0 (28.2)	[4.4, 51.6]	18.1	0.007	120 ± 13
	3	0.72	34.4 (27.4)	[11.5, 57.3]	25.3	0.007	112 ± 16
	4	0.73	33.9 (26.5)	[11.7, 56.6]	24.1	0.007	126 ± 16
	5	0.74	31.0 (27.1)	[8.3, 53.6]	22.5	0.007	131 ± 13
	6	0.78	44.3 (21.7)	[26.1, 62.4]	34.2	0.039	138 ± 17
TSVM	1	0.81	48.4 (18.3)	[33.1, 63.7]	39.2	0.583	6178 ± 3408
	2	0.83	54.7 (17.8)	[39.8, 69.6]	46.4	0.546	7100 ± 3663
	3	0.83	55.4 (18.3)	[40.1, 70.7]	46.8	0.541	8008 ± 4238
	4	0.83	54.6 (15.9)	[41.3, 67.9]	47.8	0.148	7093 ± 4113
	5	0.84	58.7 (17.5)	[44.1, 73.3]	52.8	0.296	7432 ± 4259
	6	0.85	61.0 (16.1)	[47.5, 74.5]	54.4	0.541	6551 ± 3536
S^3DB + SVM	1	0.69	42.3 (16.9)	[28.2, 56.4]	33.1	—	105 ± 21
	2	0.83	63.6 (18.4)	[48.2, 79.0]	59.1	—	88 ± 11
	3	0.84	64.5 (19.9)	[47.9, 81.1]	60.2	—	114 ± 11
	4	0.84	63.8 (17.8)	[48.9, 78.7]	59.1	—	85 ± 11
	5	0.84	62.9 (17.2)	[48.5, 77.3]	57.7	—	137 ± 18
	6	0.84	63.7 (18.1)	[48.6, 78.8]	59.6	—	81 ± 9

Table 2. Results comparing the mean performance of the 8-fold subject-wise cross-validated methods on triaging image dataset. Results are reported at an operating point of 95% sensitivity. A pairwise Wilcoxon Signed-Rank test was used to check for statistical significance in accuracy performances of each method compared with our proposed method. No statistically significant difference was observed between the pairs performances after adjusting for multiple testing using the Bonferroni method (adjusted α-value = 0.004).

Table 4 summarizes the mean performance of the 5 fold cross-validation on $n = 13{,}821$ labeled nuclei figures combined with $n = 49{,}000$ unlabeled objects using our proposed SSL technique compared with supervised SVM method trained on the labeled portion only. No statistically significant difference in performance was observed between the accuracy pairs of the supervised SVM and our proposed SSL method in Table 4 using a pairwise Wilcoxon Signed-Rank test.

Method	AUC	Accuracy (%)	Sensitivity (%)	Specificity (%)
Supervised SVM	0.81	42	93	38
ssFCM + SVM	0.59	29	91	24
TSVM	0.85	49	94	45
S^3DB + SVM	0.86	53	94	49

Table 3. Results comparing the performance of the four methods using an overall trained model from triaging image dataset using 3 patients data only as labeled set on a totally unseen test set of 2 patient scanned at 10X magnification.

Method	Task	AUC	ACC (%)	Sens. (%)	Spec. (%)
Supervised SVM	L vs. BM	0.97	95 (±0.1)	79	99
	B vs. M	0.86	83 (±0.5)	57	92
S^3DB + SVM	L vs. BM	0.95	93 (±0.4)	74	97
	B vs. M	0.73	74 (±0.5)	7	99

Table 4. Mean performance of 5-fold cross-validation on nuclei figure classification dataset using S^3DB + SVM semi-supervised method ($n = 13821$ labeled nuclei objects and $n = 49000$ unlabeled ones) and supervised SVM method ($n = 13821$ labeled nuclei objects).

Method	Class	ACC (%)	Sens. (%)	Spec. (%)
Supervised SVM	L	92	80	94
	B	75	50	92
	M	77	91	63
S^3DB + SVM	L	87	67	91
	B	60	3	99
	M	60	97	20

Table 5. Performance of applying the generated model from dataset used in Table 4 on $n = 7958$ nuclei objects from an independent testing set using S^3DB + SVM and supervised SVM methods.

Dataset	Classes	Fisher Discriminant Ratio (FDR)
Triaging Image Dataset	Relevant vs. Irrelevant	0.39
Nuclei Figure Classification	L vs. BM	0.27
	B vs M	0.02

Table 6. Fisher Discriminant Ratio (FDR) measures for different classes from the datasets used in this study.

Table 5 summarizes the performance of applying the models generated from both supervised and semi-supervised methods on the training part on an independent testing set of $n = 7,958$ nuclei figures. From Tables 4 and 5 it is clear that the proposed SSL method did not fit well for this dataset and the performance is poor compared to the supervised SVM method.

Comparing Cluster Separability Measures. In order to have a sense of how separable the clusters of each class are with respect to each other, FDR measures are summarized in Table 6. From Table 6 we can see that as the FDR measure increases (relevant vs. Irrelevant and L vs. BM) the classes in each dataset tend to form separable clusters while in case of B vs. M where separable clusters are not formed the FDR measure is low. Furthermore, in order to visually compare the distribution of different class labels in feature spaces of both datasets, their dimensions were reduced using the t-SNE method[43]. Figure 6 shows the dimensionality-reduced data space of the triaging image dataset with every point representing an image patch. Similarly, Fig. 7 shows the data space of the nuclei figure classification dataset with every point representing a nuclei figure. Considering Fig. 6, it can be seen that the relevant and irrelevant classes form separable clusters in the feature space while considering Fig. 7, it can be observed that lymphocyte class is better separated compared to the other two classes. Comparing benign versus malignant epithelial classes in the same figure we see that they do not tend to form separable clusters of points thus violating the cluster assumption of SSL.

There is a slight imbalance between the classes in both triaging image dataset and nuclei figure dataset. To determine whether this affected performance, we repeated the training and testing to check the consistency of the classification performance for both datasets, the weight values of the SVM models for minority class were

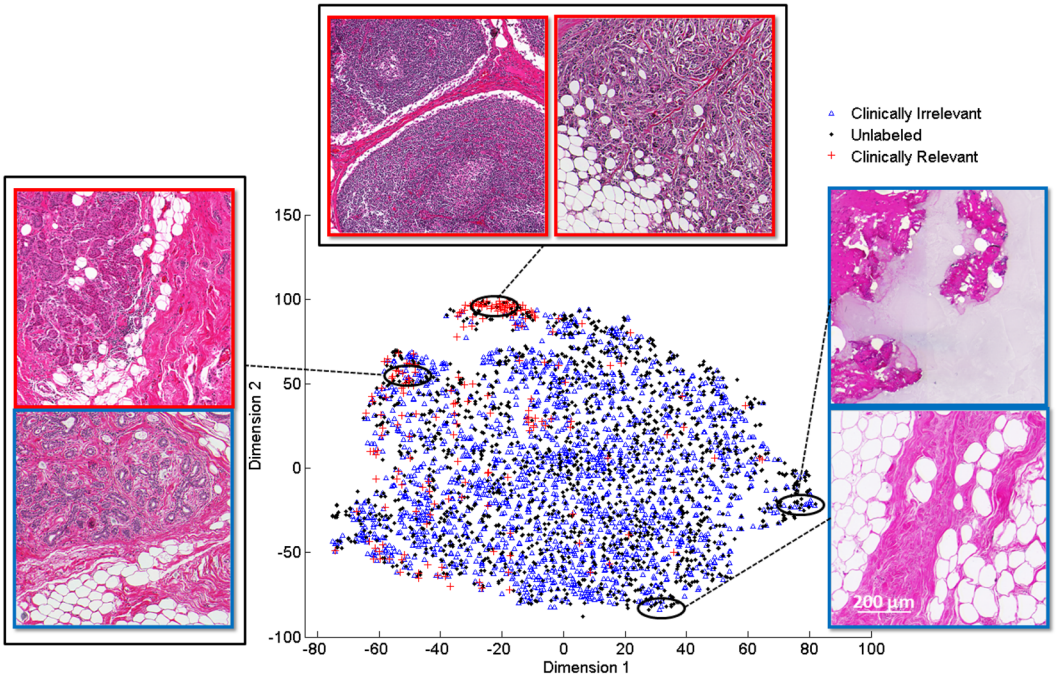

Figure 6. 2D visualization of the triaging image dataset feature space using t-SNE[43]. Every point in this plot represents an image patch from the dataset. As can be seen, relevant versus irrelevant images form separable clusters in this visualization.

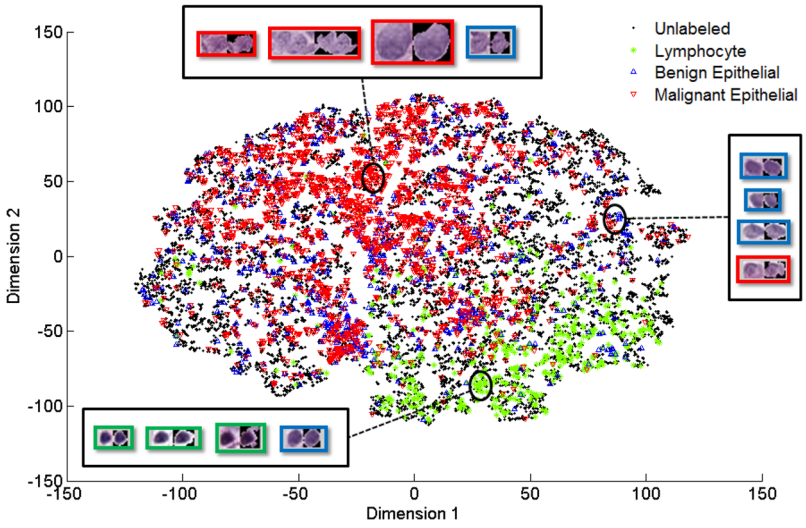

Figure 7. 2D visualization of the feature space from nuclei figure classification dataset using t-SNE[43]. Every point in this plot represents a nuclei figure from the dataset. As can be seen, lymphocyte figures versus the other two classes are better separated while benign versus malignant epithelial figures do not form separable clusters.

set to the ratio of the number of data samples in the majority class to that of the minority class. The classification performance was found to be consistent for both the datasets suggesting that the class imbalance did not affect the classification performance.

Discussion and Conclusion

In this study, we proposed a cluster-then-label based semi-supervised technique to find the underlying structure of the data space and provide this knowledge to train a reliable model. We have compared and validated this technique with other state-of-the-art supervised and semi-supervised methods for triaging breast digital pathology image patches and classifying nuclei figures. We found that when the method is used for the appropriate dataset the classification performance is superior and training time is much lower compared to the other semi-supervised methods.

Our proposed method did not perform as well as TSVM when only 1 patient data was made available as the labeled set for triaging image dataset (Table 2). This is most likely due to the failure in the clustering method because an insufficient number of labeled points were made available to it. The method improved as the number of labeled points increased. Surprisingly, for the triaging image dataset, using the ssFCM method did not add any improvements to the classification performance compared to the supervised SVM method. This may be because of the fact that ssFCM assumes the underlying shape of clusters come from a Gaussian distribution; this leads to incorrect label assignment (in the cases where cluster shapes do not come from a Gaussian distribution) which in turn produces an incorrect decision boundary. Furthermore, one reason for a better performance of our method compared to the other cluster-then-label based techniques is that no assumptions are made about the underlying probability distribution of the clusters and so it can cope with clusters of any shape and form.

Although the improvements in accuracy of our proposed method compared with other techniques in Table 2 were not statistically significant after applying a Bonferonni correction, the effect size was large, with improvements of 20.4% and 25.3% in accuracy over the supervised SVM and the ssFCM methods respectively. The improvement in accuracy compared with the TSVM method was about 4% but the TSVM was very computationally expensive with each model taking more than 4 days to train.

Looking at Tables 4 and 5 for nuclei classification task, our method has a poor performance compared to the supervised SVM. The reason for this poor performance could be because the underlying structure of the data points in these datasets does not form proper clusters. This is supported by the FDR measures reported in Table 6 and t-SNE plot in Fig. 7. As shown in Table 6, the FDR measures for the relevant vs. irrelevant data has a higher value compared to L vs. BM and B vs. M data. Furthermore, from Figs 6 and 7 we can observe that relevant vs. irrelevant and L vs. BM tend to form clusters of points in the dimensionality-reduced t-SNE plots while B vs. M data does not form detectable clusters of points thus violating the cluster assumption of SSL. It is also important to note that semi-supervised learning methods are traditionally suitable for applications where only limited labeled data are available. This means that SSL methods may not work as well as supervised methods when large amounts of labeled data are present[1].

In our preliminary experiments[30], we systematically examined the effect of k, which controls the number of points that lie within the neighborhood of a labeled point, on a subset of our dataset. We found that the performance of our method was stable when a sufficiently large value was chosen for k. The best performance was achieved for $k =$ one tenth of the number of points in the dataset.

Although TSVM is one of the top performing implementations of semi-supervised SVM, its performance was not found to be better on small-sized synthetic datasets when compared to the Branch and Bound (BB) technique[21,23]. The BB method seems to find the globally optimal solution for semi-supervised learning since it efficiently looks through all label combinations in the data space. However, due to its growing search tree basis for finding the solution, its train time is reported to be even slower than TSVM making it infeasible to apply on datasets with more than 200 data points[21].

Our proposed semi-supervised cluster-then-label method showed improved performance over other methods for the triaging task, however, it did not perform well in the nuclei classification task. This suggests that although semi-supervised approaches may be useful in digital pathology where generating sufficiently large labeled datasets is a challenge, additional work is needed to identify whether the clustering assumptions are valid for specific tasks.

References

1. Chapelle, O. & Schölkopf, B. *Semi-Supervised Learning* (The MIT Press, 2006).
2. Helmi, H., Teck, D., Lai, C. & Garibaldi, J. M. Semi-Supervised Techniques in Breast Cancer Classification. In *12th Annual Workshop on Computational Intelligence* (*UKCI*) (2012).
3. Shi, M. & Zhang, B. Semi-supervised learning improves gene expression-based prediction of cancer recurrence. *Bioinforma. (Oxford, England)* **27**, 3017–23, https://doi.org/10.1093/bioinformatics/btr502 (2011).
4. Batmanghelich, K., Ye, D.H.; Pohl, K. & Taskar, B. Disease Classification and Prediction via Semi-supervised Dimensionality Reduction. In *International Symposium on Biomedical Imaging: From Nano to Macro*, 1086–1090 (2011).
5. Moradi, E., Gaser, C., Huttunen, H. & Tohka, J. MRI based dementia classification using semi-supervised learning and domain adaptation. In *MICCAI 2014 Workshop Proceedings*, *Challange on Computer-Aided Diagnosis of Dementia*, *based on Structural MRI Data* (2014).
6. Chapelle, O. & Zien, A. Semi-Supervised Classification by Low Density Separation. In *Tenth International Workshop on Artificial Intelligence and Statistics* (*AISTAT 2005*) (2005).
7. Yarowsky, D. Unsupervised word sense disambiguation rivaling supervised methods. In *ACL '95 Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, 189–196 (1995).
8. Rosenberg, C., Hebert, M. & Schneiderman, H. Semi-Supervised Self-Training of Object Detection Models. In *Seventh IEEE Workshops on Applications of Computer Vision* (*WACV/MOTION'05*), vol. 1, 29–36 (IEEE, 2005).
9. McClosky, D., Charnia, E. &Johnson, M. Effective self-training for parsing. In *HLT-NAACL '06 Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, 152–159 (2006).
10. Tanha, J., van Someren, M. & Afsarmanesh, H. Semi-supervised self-training for decision tree classifiers. *Int. J. Mach. Learn. Cybern.*, https://doi.org/10.1007/s13042-015-0328-7 (2015).
11. Callison-burch, C., Talbot, D. & Osborne, M. Statistical machine translation with word- and sentence-aligned parallel corpora. In *Proceedings of the ACL*, 175–182 (2004).
12. Fujino, A., Ueda, N. & Saito, K. Semisupervised Learning for a Hybrid Generative/Discriminative Classifier based on the Maximum Entropy Principle. *IEEE Transactions on Pattern Analysis and Mach. Intell.* **30**, 424–437 (2008).
13. Nigam, K. & Ghani, R. Analyzing the effectiveness and applicability of co-training. In *Proceedings of the ninth international conference on Information and knowledge management*, 86–93 (2000).
14. He, J., Carbonell, J. & Liu, Y. Graph-Based Semi-Supervised Learning as a Generative Model. In *IJCAI'07 Proceedings of the 20th international joint conference on Artifical intelligence*, 2492–2497 (2007).

15. Talukdar, P. P. & Pereira, F. Experiments in Graph-based Semi-Supervised Learning Methods for Class-Instance Acquisition. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, July, 1473–1481 (2010).
16. Liu, B. W., Wang, J. & Chang, S.-f Robust and Scalable Semisupervised Learning. *Proc. IEEE* **100**, 2624–2638 (2012).
17. Chang, K. C.-C. & Lauw, H. W. Graph-based Semi-supervised Learning: Realizing Pointwise Smoothness Probabilistically. In *Proceedings of the 31st International Conference on Machine Learning*, vol. **32** (2014).
18. Fern, M. & Cernadas, E. Do we Need Hundreds of Classifiers to Solve Real World Classification Problems? *J. Mach. Learn. Res.* **15**, 3133–3181 (2014).
19. Joachims, T. Transductive Inference for Text Classification using Support Vector Machines. In *Proceedings of the International Conference on Machine Learning* (*ICML*) (1999).
20. Yuille, A. L. & Rangarajan, A. The Concave-Convex Procedure (CCCP). *Neural Comput.* **15**, 915–936 (2003).
21. Chapelle, O., Sindhwani, V. & Keerthi, S. Branch and Bound for Semi-Supervised Support Vector Machines. In *Advances in neural information processing systems* (*NIPS*) (2006).
22. Chapelle, O. & Zien, A. A Continuation Method for Semi-Supervised SVMs. In *International Conference on Machine Learning* (2006).
23. Chapelle, O., Sindhwani, V. & Keerthi, S. Optimization Techniques for Semi-Supervised Support Vector Machines. *J. Mach. Learn. Res.* **9**, 203–233 (2008).
24. Chapelle, O., Weston, J. & Scholkopf, B. Cluster Kernels for Semi-Supervised Learning. In *Advances in Neural Information Processing Systems* **15**, 601–608 (2003).
25. Weston, J. *et al*. Semi-supervised protein classification using cluster kernels. *Bioinform. (Oxford, England)* **21**, 3241–7, https://doi.org/10.1093/bioinformatics/bti497 (2005).
26. Dara, R., Kremer, S. & Stacey, D. Clustering unlabeled data with SOMs improves classification of labeled real-world data. In *International Joint Conference on Neural Networks*, 2002. *IJCNN '02. Proceedings of the 2002*, 2237–2242 (2002).
27. Gan, H., Sang, N., Huang, R., Tong, X. & Dan, Z. Using clustering analysis to improve semi-supervised classification. *Neurocomputing* **101**, 290–298, https://doi.org/10.1016/j.neucom.2012.08.020 (2013).
28. Goldberg, A. B. *New Directions in Semi-supervised Learning*. Ph.D. thesis, University of Wisconsin-Madison (2010).
29. Peikari, M., Gangeh, M., Zubovits, J., Clarke, G. & Martel, A. Triaging Diagnostically Relevant Regions from Pathology Whole Slides of Breast Cancer: A Texture Based Approach. *IEEE Transactions on Med. Imaging* (2015).
30. Peikari, M., Zubovits, J. T., Clarke, G. M. & Martel, A. L. Clustering Analysis for Semi-supervised Learning Improves Classification Performance of Digital Pathology. In *Machine Learning in Medical Imaging - 6th International Workshop {MLMI} 2015, Held in Conjunction with {MICCAI} 2015, Munich, Germany, October 5, 2015, Proceedings*, 263–270 (2015).
31. Ankerst, M., Breunig, M. M. & Kriegel, H.-p. OPTICS: Ordering Points To Identify the Clustering Structure. In *SIGMOD '99 Proceedings of the 1999 ACM SIGMOD international conference on Management of data*, 49–60 (1999).
32. Clarke, G. M. *et al*. Increasing specimen coverage using digital whole-mount breast pathology: implementation, clinical feas ibility and application in research. *Comput. Medical Imaging Graphics: Official Journal Comput. Medical Imaging Soc.* **35**, 531–41 (2011).
33. Geusebroek, J.-M., Smeulders, A. W. M. & van de Weijer, J. Fast anisotropic Gauss filtering. *IEEE Transactions on Image Processing: A Publication of the IEEE Signal Processing Society* **12**, 938–43, https://doi.org/10.1109/TIP.2003.812429 (2003).
34. Varma, M. & Zisserman, A. A Statistical Approach to Texture Classification from Single Images. *Int. Journal Computer Vision* **62**, 61–81 (2005).
35. Chang, C.-C. & Lin, C.-J. LIBSVM: A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems Technology* **2**, 27:1–27: 27 (2011).
36. Peikari, M., Salama, S., Nofech-mozes, S. & Martel, L. Automatic Cellularity Assessment from Post-treated Breast Surgical Specimens. *Cytom. A* (in press), 1–30, https://doi.org/10.1002/cyto.a.23244 (2017).
37. Martel, A. L., Hosseinzadeh, D., Senaras, C., Madabhushi, A. & Gurcan, M. N. An Image Analysis Resource for Cancer Research: PIIP—Pathology Image Informatics Platform for Visualization, Analysis, and Management. *Cancer Res.* **77**, e83–e87 (2017).
38. Peikari, M. & Martel, A. L. Automatic cell detection and segmentation from H and E stained pathology slides using colorspace decorrelation stretching. In *SPIE Medical Imaging* (2016).
39. Wienert, S. *et al*. Detection and segmentation of cell nuclei in virtual microscopy images: a minimum-model approach. *Sci. Reports* **2**, 503, https://doi.org/10.1038/srep00503 (2012).
40. Gan, H. *et al*. Discussion of FCM algorithm with partial supervision. In *Proceedings of the Eighth International Symposium on Distributed Computing and Applications to Business, Engineering and Science*, 27–31 (2009).
41. Lai, D. T. C. & Garibaldi, J. M. A Preliminary Study on Automatic Breast Cancer Data Classification using Semi-supervised Fuzzy c-Means. *Int. J. Biomed. Eng. Technol. SI: MEDSIP 2012 Inf. Process.* **13**, 303–322 (2013).
42. Joachims, T. Making Large Scale SVM Learning Practical. In *Support Vector Learning*, 169–184 (1999).
43. Maaten, L. V. D. Accelerating t-SNE using Tree-Based Algorithms. *J. Mach. Learn. Res.* **15**, 1–21 (2014).
44. Theodoridis, S. & Koutroumbas, K. *Pattern Recognition* (Academic Press, New York, 1998).
45. Lin, T.-H., Li, H.-T. & Tsai, K.-C. Implementing the Fisher's discriminant ratio in a k-means clustering algorithm for feature selection and data set trimming. *J. Chemical Information Computer Sciences* **44**, 76–87 (2004).
46. Wang, S., Li, D., Song, X., Wei, Y. & Li, H. A feature selection method based on improved fisher's discriminant ratio for text sentiment classification. *Expert. Syst. with Appl.* **38**, 8696–8702 (2011).

Acknowledgements

This research is funded by Canadian Cancer Society (grant # 703006). We would like to thank Dr. Gina Clarke for providing us with the triaging image dataset that was used as a part of our experiments in this study. We would also like to sincerely thank Dr. Judit Zubovits for her kind support in reviewing the triaging image dataset used to train and validate methods presented in this work.

Author Contributions

Experimental design, analysis of data, and assembling manuscript figures: M.P. Annotating pathology slides: S.N., S.S. Supervision of project: A.L.M. Manuscript writing: M.P., A.L.M. All authors reviewed manuscript.

Additional Information

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.