Hindawi
Computational and Mathematical Methods in Medicine
Volume 2019, Article ID 3041250, 13 pages
https://doi.org/10.1155/2019/3041250

Research Article

Direct Cellularity Estimation on Breast Cancer Histopathology Images Using Transfer Learning

Ziang Pei, Shuangliang Cao, Lijun Lu, and Wufan Chen

School of Biomedical Engineering and Guangdong Provincial Key Laboratory of Medical Image Processing, Southern Medical University, Guangzhou 510515, China

Correspondence should be addressed to Lijun Lu; ljlubme@gmail.com and Wufan Chen; chenwf@smu.edu.cn

Received 3 February 2019; Accepted 30 April 2019; Published 9 June 2019

Academic Editor: Xiaoqi Zheng

Residual cancer burden (RCB) has been proposed to measure the postneoadjuvant breast cancer response. In the workflow of RCB assessment, estimation of cancer cellularity is a critical task, which is conventionally achieved by manually reviewing the hematoxylin and eosin- (H&E-) stained microscopic slides of cancer sections. In this work, we develop an automatic and direct method to estimate cellularity from histopathological image patches using deep feature representation, tree boosting, and support vector machine (SVM), avoiding the segmentation and classification of nuclei. Using a training set of 2394 patches and a test set of 185 patches, the estimations by our method show strong correlation to those by the human pathologists in terms of intraclass correlation (ICC) (0.94 with 95% CI of (0.93, 0.96)), Kendall's tau (0.83 with 95% CI of (0.79, 0.86)), and the prediction probability (0.93 with 95% CI of (0.91, 0.94)), compared to two other methods (ICC of 0.74 with 95% CI of (0.70, 0.77) and 0.83 with 95% CI of (0.79, 0.86)). Our method improves the accuracy and does not rely on annotations of individual nucleus.

1. Introduction

Breast cancer is the most common malignant cancer occurring in women [1]. Preoperative neoadjuvant therapy (NAT) [2] can reduce the breast tumor size, so as to facilitate the complete resection of tumor and the performance of breast-conserving surgery instead of mastectomy for patients with large tumor. With NAT, a significant reduction of recurrence and metastasis can be achieved [3]. Pathologic complete response (pCR) [4] has been conventionally accepted as the primary end point to evaluate the efficacy of NAT [5]. However, the predictive potential of pCR on long-term prognosis is impaired by its blurry definition (e.g., there is no agreement on whether pCR should be also applied to the axillary lymph node and whether the presence of only noninvasive cancer should be defined as pCR) [4] and the roughness of dichotomizing the tumor response (as complete response or residual disease) [5].

Unlike pCR, the residual cancer burden (RCB) index is improved by measuring both in situ and invasive cancer in residual tumor and the metastasis through lymph nodes [5]. RCB is a new staging system basically devised to continuously quantify the residual breast cancer that ranges from complete response to chemotherapy resistance. It is standardized by defining a pipeline of specimen collection and tumor bed identification and has proved to be a significant indicator of distant relapse-free survival of breast cancer [5].

Clinically, RCB is assessed by checking the histological sections from primary breast tumor site and regional lymph nodes [6]. The RCB index is calculated using six parameters, namely, the primary tumor bed area (length and width), the overall cancer cellularity, the percentage of in situ cancer, the number of positive lymph nodes, and the diameter of the largest metastasis [6]. Among them, the estimation of cancer cellularity is a critical and challenging task. Cancer cellularity is defined as the proportion of cancer within the residual tumor bed. In clinical practice, the largest cross-sectional area of the preidentified tumor bed is divided into multiple slides, which are stained with hematoxylin and eosin (H&E)

and then reviewed with a microscope. In each microscopic field, a pathologist estimates the local cellularity by comparing the proportion of area containing cancer to the standard reference. The top row of Figure 1 presents some of the computer-generated diagrams that illustrate the distribution of cancerous nuclei under different cellularity and can be used as references to assist manual estimation. The overall cellularity is then obtained by averaging these manual estimations over all the fields. However, the reliability of manual assessment on tumor cell percentage is subject to inter-rater variability [7], and the procedure is time-consuming and also requires expertise and experience.

These problems can be probably solved using computerized methods. Digital pathology [8, 9] has enabled software to retrieve useful information from digitized slides that could be further analyzed using advanced statistical learning models. Combined with machine learning, digital pathology has been developed as a powerful tool in various clinical applications such as histological classification [10, 11] and segmentation [12–15], prognosis prediction [16, 17], and cancer diagnosis [18]. Recently, deep learning [19] has gained much attention due to its impressive performances in computer vision tasks. Deep learning methods are generally based on convolutional neural networks (CNNs) that learn features and prediction models with fewer human interventions than conventional machine learning and are also combined with fuzzy learning for robust feature representation [20] and reinforcement learning for self-taught decision-making [21].

As for automatic cancer cellularity assessment, there are two literatures addressing this challenge [22, 23]. In [22], Peikari et al. use a two-stage method consisting of nuclei segmentation and classification. The images extracted from whole slide images (WSIs) are preprocessed with decorrelation stretching for contrast enhancement, and the nuclei are segmented with multilevel Otsu's thresholding [24, 25], morphological operations, and marker-controlled watershed [26]. Then, two support vector machines (SVMs) trained using shape, textural, and spatial features extracted from the annotated nucleus figures are applied to distinguish the lymphocytes from epithelial nuclei and classify the epithelial as benign or malignant. The cellularity is estimated as the proportion of area filled with the cytoplasm of malignant epithelial cells. This procedure is analogous to the workflow of the pathologist; however, it is subject to the accuracy of segmentation and classification. In [23], Akbar et al. fine-tuned two modified Inception [27] networks independently to identify the images as cancerous or non-cancerous and predict the cellularity for those cancerous patches. The method is validated using the same dataset as in [22], and superior performance is achieved. This method is based on end-to-end deep learning models, and the prediction of cellularity is directly given.

In recent literatures [28–30], transfer learning has also been applied to pathological images. Transfer learning is generally defined as the strategy to apply the knowledge obtained from one task to another related task [31, 32]. Practically, transfer learning makes training less dependent on the quantity of data utilizing the prior knowledge integrated in the models pretrained on a large scale dataset. In [29], the authors extract deep features from the top three layers of a pretrained AlexNet and use logistic regression to classify benign and malign tissue from breast cancer pathological slides. Weiss et al. [30] made a comparison on different features extracted from different models and found that the features from lower layers of Xception [33] perform the best. The basic idea of [28] is similar, but the authors used the lower layers of 3 different pretrained models for feature extraction and gradient boosting decision trees for classification. The performance improves compared to end-to-end CNN classifier.

In this work, we propose a novel framework to directly estimate cellularity from breast cancer histopathological slide patches combining deep feature representation, tree boosting, and SVM. Unlike the previous approach proposed by Peikari et al., our methods require for training only the cellularity labels on image patches instead of the annotations on individual nucleus. The following contributions are made:

(1) Our method can directly estimate cellularity from breast cancer slide patches and avoid the segmentation and classification of nuclei, which is similar to the approach in [23].

(2) We validate the transferability to tissue microscopy of the deep features learned from natural images. With the pretrained features, we manage to address the problem of data scarcity.

(3) To tackle the problem of label imbalance, we make our prediction using regression and learn to rank model.

According to the experiments, we show that our methods are robust enough and state-of-the-art performances are achieved.

2. Materials and Methods

2.1. Materials. The data used in our research are acquired from the SPIE-AAPM-NCI BreastPathQ Challenge [34] and are from the same batch as those in [22, 23]. The dataset consists of 69 H&E stained WSI collected from the resection specimens of 37 post-NAT patients with invasive residual breast cancer. The specimens are processed following regular histopathological protocols, and the WSIs are scanned at 20x magnification (0.5 μm/pixel). The tumor beds of these slides are roughly segmented into 4 regions (normal (0), low cellularity (1–30%), medium cellularity (31–70%), and high cellularity (71–100%)), and approximately equal numbers of patches are selected from each of them. In total, 2579 image patches with ROI of 512×512 pixels (about 1 mm^2) are selected and labeled by an expert pathologist with manually estimated cellularity ranging between [0, 100%], as described in [22]. These patches are randomly partitioned by the challenge organizer into training set with 33 patients (2394 patches) and test set with 4 patients (185 patches). In our experiments, the test set is separated from training.

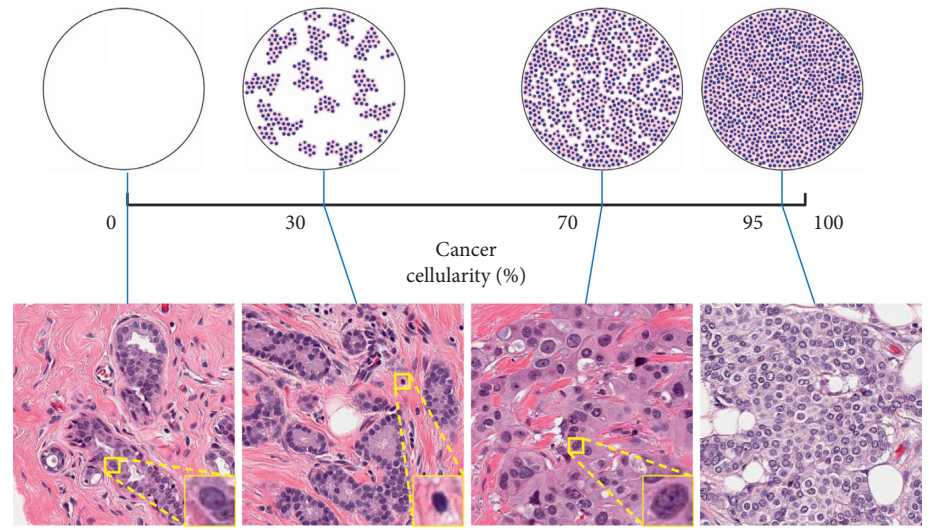

Figure 1: Illustrations of the definition of cellularity. In the top row are 4 computer generated diagrams [6] used as reference for manual estimation. In the bottom row are 4 example patches from the test set labeled with manually estimated cellularity and 3 zooming windows showing the morphology of benign epithelial, lymphocyte, and malignant epithelial nuclei.

The bottom row of Figure 1 shows 4 examples from the test dataset with different cellularity and morphological features of benign epithelial, lymphocyte, and malignant epithelial nuclei. Besides, stromal nuclei are also presented in the images and are largely different in shape from the other nuclei. Nuclei presented in this dataset can be categorized as one of these 4 classes. Cellularity is defined as the percentage of the area of malignant epithelial cell.

The histogram of cellularity value on the whole dataset is presented in Figure 2, from which it is clear that there is heavy imbalance on cellularity distribution. The bin density varies in different intervals: the cellularity difference between neighboring bins is 5% in [10%, 90%], whereas in [0, 10%) ∪ (90%, 100%], it is 1%. Numbers of patches varies among the bins: 705 patches are labeled as 0 cellularity; 14 of the bins contain 50–200 patches; 22 of them contain fewer than 50 patches; some of the bins contain no patches (4%, 6%, 9%, 91%, 94%, 96%).

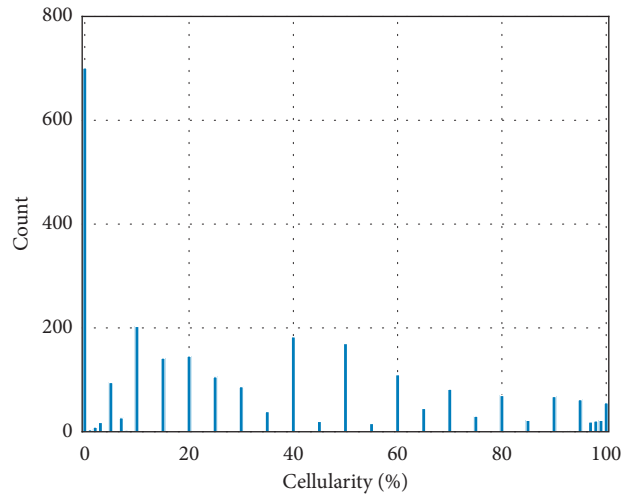

Figure 2: Histogram of cellularity distribution of the whole dataset.

2.2. Methods

2.2.1. Overview. There are two major problems in this project. The first problem is data scarcity, as only less than 3000 samples from 37 patients are included in our dataset. This could be addressed using transfer learning [35]. Transfer learning reduces the dependency on data quantity with the prior knowledge learned from another dataset that has proved transferability and generalizability to other scenarios [31, 32]. Generally, there are two options to apply transfer learning in medical image: (1) to fine-tune the pretrained CNN; (2) to use the pretrained CNN as feature extractor and combine it with conventional machine learning [36]. The second problem, label imbalance, as shown in Figure 2 where cellularity is not uniformly distributed in all the discrete bins, poses a challenge to learning algorithms susceptible to data distribution. In this work, we solve this problem using regression and learn to rank models.

We propose a framework that holds promise to overcome the two problems stated above following the second option of transfer learning. As illustrated in Figure 3, the workflow of our methods consists of the following steps: (1) data preprocessing that includes stain normalization and data augmentation, as detailed in Section 2.2.2; (2) deep feature representation that extracts thousands of features from each patch, as detailed in Section 2.2.3; (3) feature selection via minimum redundancy maximum relevance (mRMR) [37] and dimensionality reduction by principal component analysis (PCA), as detailed in Section 2.2.4; (4) training gradient boosting decision trees (GBDT) classifier using the refined features to distinguish between cancerous and noncancerous patches, as detailed in Section 2.2.5; (5) training GBDT and SVM to predict the cellularity for those cancerous patches, as detailed in Sections 2.2.6 and 2.2.7. The

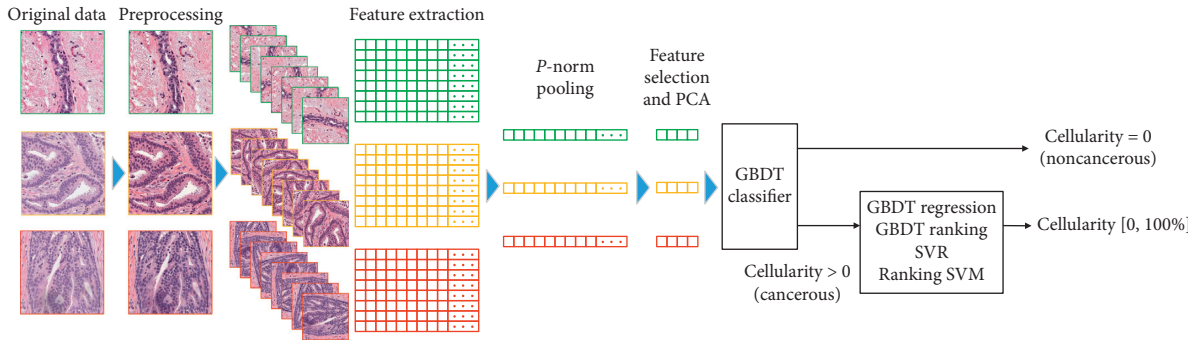

Figure 3: Overview workflow of our methods. The images are first preprocessed to obtain consistent stain. Thousands of features are extracted from each image using the pretrained CNNs and are pooled to enhance rotation invariance. A few hundred features are obtained from mRMR feature selection and PCA and are used to train GBDT and SVM.

metrics to evaluate the performance of our methods are briefly described in Section 2.2.8.

2.2.2. Data Preprocessing. Stain inconsistency of digital WSI due to fixation, embedding, cutting, and staining of tissue sections is common among slides from different microscopes and different staining batches [38]. This contributes to very negative effect on quantitative analysis. To reduce stain variation in our dataset, we normalize the H&E stain on the slide patches with the algorithm proposed in [39]. It first converts the RGB image to optical density (OD) space and finds the optimal stain vectors for H and E by calculating singular value decomposition (SVD) on the OD tuple, a $N \times 3$ matrix, where N is the number of pixels and 3 represents the RGB channels. Then, the image is deconvolved using these stain vectors and normalized using 2 predefined reference vectors.

Generally, data augmentation is indispensable to reduce overfitting in statistical learning systems. Following the method proposed in [28, 40], we first convert the color of the tissue from RGB space to H&E space, then we multiply each channel by a factor randomly and uniformly sampled from range [0.7, 1.3]. Other commonly used augmentations, such as cropping and rescale, are ineligible for this task as rescale causes loss in resolution and cropping varies the exact cellularity when cancerous or noncancerous regions are cropped out.

2.2.3. Deep Feature Representation. In our method, deep features are extracted from the augmented patches with CNN pretrained on ImageNet [41], which is similar to [28]. Three architectures, namely, VGG-16 [42], ResNet-50 [43], and Inception-v3 [27], are applied. As suggested in [44], lower layers of deep learning models are expected to preserve more generic and transferable features. Therefore, we focus on the convolutional layers of these models. Global average pooling is used in each layer to decrease the dimensionality of features. In Figure 4, ResNet is taken as example to show our scheme of feature extraction. In total, 4224, 15168, and 10048 features are to be extracted from VGG, ResNet, and Inception, respectively. With rotation (0, 90°, 180°, and 270°) and flipping, 8 variations are to be generated from each patch, so that 8 different feature vectors are to be extracted. To obtain rotation invariance, we combine them into one vector by applying the p-norm pooling approach on each dimension [45, 46], which can be formulated as

$$f_{\text{pooling}} = \left(\frac{1}{N} \sum_{i=1}^{N} (f_i)^p \right)^{1/p}, \quad (1)$$

where f is the feature dimension to be pooled, $N = 8$ is the number of vectors, and p is the norm, which is set to 3.

2.2.4. Feature Selection and Dimensionality Reduction. We apply the mRMR feature selection method suggested in [37] to search for the fewest features that preserve sufficient information for prediction. The basic idea of mRMR is to select a feature subset where the features are marginally as correlated to the target variable as possible while mutually as uncorrelated to each other as possible. Practically, mRMR is developed to be a filter that iteratively selects features with greedy algorithm based on the mutual information between the target variable and each feature. In this work, mutual information quotient (MIQ) is chosen as the feature filtering criterion. The performances of using different numbers of top ranking features (n_{feat}) are evaluated to determine the optimal n_{feat}.

PCA is used to reduce the dimensionality of the selected features. PCA transforms the features into uncorrelated components, among which those with the smallest variance are considered noisy and are thus abandoned. In general, feature selection and dimensionality reduction eliminate irrelevant information and speedup training.

2.2.5. Tissue Classification Based on GBDT. Boosting is an ensemble method where new models are added serially to minimize the loss of previous models until no further improvements can be made. In this procedure, a series of weak models are combined to make up a stronger one with the primary goal to reduce bias and variance. GBDT [47] is a boosting approach where new decision trees are trained to fit the residuals of the existing trees and then added together to

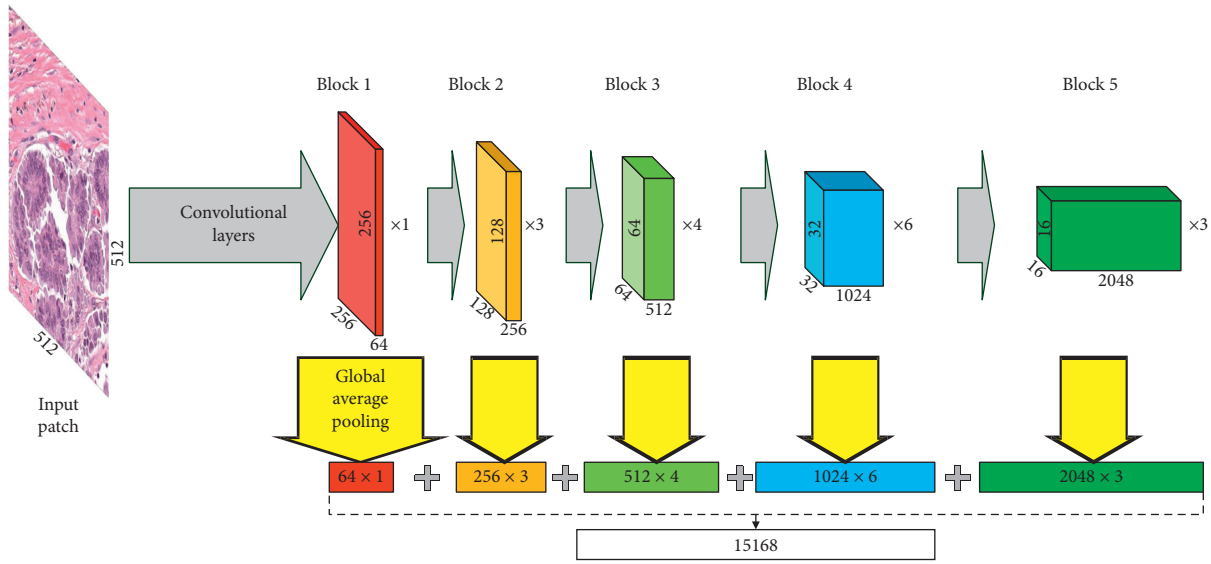

Figure 4: Schematic diagram of feature extraction with ResNet.

optimize the object function. It is a powerful and popular technique that has been used in many leading solutions to data science challenges with success in classification and regression. The name gradient boosting comes from the gradient descent algorithm that is used to minimize the loss when adding new models. XGBoost [48] is an implementation of GBDT developed for faster speed and superior performance. Remarkable features of XGBoost include parallelization of tree construction, distributed computing, out-of-core computing, and cache-aware prefetch. LightGBM [49] is another gradient boosting framework achieving higher efficiency and lower memory usage with the use of histogram-based splitting, leafwise tree growth, and optimization of parallel. Both XGBoost and LightGBM support L1, L2, and model complexity regularization in order to reduce overfitting.

As illustrated in Figure 2, our data are heavily imbalanced. As the patches with zero cellularity, i.e., noncancerous tissue, make up nearly 1/4 of the training samples, sifting them out helps balancing the dataset and improving regression performance. In this section, we describe our GBDT-based method to classify the patches as cancerous or noncancerous. In general, GBDT aims to minimize a predefined loss function. For the target of binary classification, binary cross entropy (BCE) is optimized, which can be formulated as

$$L_{\mathrm{BCE}} = -\sum_{i=1}^{M} y_i \log\left(\sigma\left(F\left(\mathbf{x_i}\right)\right)\right) + \left(1 - y_i\right)\log\left(1 - \sigma\left(F\left(\mathbf{x_i}\right)\right)\right), \tag{2}$$

where $F(\mathbf{x_i})$ is the output of the trees for sample $\mathbf{x_i}$, M is the number of training samples, σ is the sigmoid function, and $y_i \in \{0, 1\}$ is the training label. BCE is a smooth and differentiable function, which is a prerequisite to be optimizable for GBDT. Our model is trained using the selected and reduced features.

2.2.6. Prediction for Cancerous Patches Using GBDT. As described in Section 2.2.5, GBDT is a powerful model that hold promises to solve both classification and regression problems. To predict the cellularity on a continuous scale for those cancerous patches using the selected features, 2 kinds of loss are optimized.

(1) Huber Loss. Huber loss can be formulated as

$$L_{\mathrm{Huber}} = \sum_{i=1}^{M} h_i, \tag{3}$$

where

$$h_i = \begin{cases} \dfrac{1}{2}\left(y_i - F\left(\mathbf{x_i}\right)\right)^2, & \left|y_i - F\left(\mathbf{x_i}\right)\right| \le \delta, \\ \delta\left|y_i - F\left(\mathbf{x_i}\right)\right| - \dfrac{1}{2}\delta^2, & \text{otherwise}, \end{cases} \tag{4}$$

where y_i is the target variable, $F(\mathbf{x_i})$ is the model prediction on sample $\mathbf{x_i}$, M is the number of training samples, and δ is the parameter controlling the steepness of Huber function.

It is clear that Huber loss is actually a combination of L1- and L2-norm loss. In our regression models, the target variable y_i is cellularity.

(2) Ranking Loss. In information retrieval, RankNet [50] is a supportive machine learning model that solves the ranking problems. Generally speaking, a RankNet is intended to output a ranking score indicating the ranking position given an input feature vector.

Analogous to the manual workflow, the cellularity of an object image patch could be estimated by comparing it to the patches of known cellularity. This problem could be probably handled using the pairwise ranking method. In this task,

we apply the pairwise RankNet [50], which trains a ranking model by optimizing the loss function based on pairs of feature vectors in adjacent levels. Prior to training, different cellularities are first translated into different levels. The cellularity 0, 1%, 2%, . . ., 10% are translated to level 0, 1, 2, . . ., 10, respectively. 15%, 20%, 25%, . . ., 90% are translated to 11, 12, 13, . . ., 26 and 91%, 92%, 93%, . . ., 100% to 27, 28, 29, . . ., 36, respectively. The probability that feature vector $\mathbf{x_i}$ should be ranked higher than $\mathbf{x_j}$ (i.e., the cellularity of $\mathbf{x_i}$ is lower than that of $\mathbf{x_j}$) can be formulated as

$$P_{ij} = \sigma\left(k\left(F\left(\mathbf{x_i}\right) - F\left(\mathbf{x_j}\right)\right)\right), \tag{5}$$

where σ is the sigmoid function, F is the ranking model, and k controls the shape of sigmoid. The ranking loss is defined as the sum of BCE of P_{ij}:

$$L_{\text{rank}} = \sum_{i,j\in A} -\overline{P}_{ij}\log P_{ij} - \left(1-\overline{P}_{ij}\right)\log\left(1-P_{ij}\right), \tag{6}$$

where $\overline{P}_{ij} \in \{0, 1\}$ denotes the true ranking between $\mathbf{x_i}$ and $\mathbf{x_j}$ and A denotes the set of any feature vector pairs in adjacent cellularity levels.

The output ranking scores are recalibrated to [0, 100%] using K-nearest neighborhood (KNN) ($k = 30$), where the ranking scores are mapped to cellularity with reference to the scores of the training set. The illustration of our KNN mapping method is presented in Figure 5.

2.2.7. Prediction for Cancerous Patches Using SVM. SVM is a classical model that provides promising solutions to classification and regression problems. Similar to GBDT, we propose methods based on support vector regression (SVR) [51, 52] and ranking SVM [53] for the estimation of cellularity for those cancerous patches.

(1) SVR. SVR is a linear prediction model that tries to minimize the largest margin produced by the training samples by penalizing the outliers using ε-insensitive loss. Those samples that lie beyond the ε margin are called support vectors. Similar to SVM classifier, the training of SVR can be formulated as

$$\begin{aligned} \min_{\mathbf{w},b,\xi_i,\widehat{\xi}_i} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{M}\left(\xi_i + \widehat{\xi}_i\right), \\ \text{s.t.} \quad & f\left(\mathbf{x_i}\right) - y_i \le \varepsilon + \xi_i, \\ & y_i - f\left(\mathbf{x_i}\right) \le \varepsilon + \widehat{\xi}_i, \\ & \xi_i \ge 0, \widehat{\xi}_i \ge 0,\ i = 1, 2, \ldots, M, \end{aligned} \tag{7}$$

where $f(\mathbf{x_i}) = \mathbf{w}^{\mathrm{T}}\mathbf{x_i} + b$ is the linear regression model, y_i is the target variable, ξ_i and $\widehat{\xi}_i$ are the slack variables, C is the tradeoff parameter, and ε controls the width of the margin. This problem can be solved using the classical Lagrange duality and KKT conditions. Kernel tricks can be implemented by replacing $\mathbf{x_i}$ with $\Phi(\mathbf{x_i})$, and the most commonly used kernels are linear kernels and radial basis functions (RBF). By setting ε as 0 and substituting ξ_i^2 and $\widehat{\xi}_i^2$ for ξ_i and $\widehat{\xi}_i$ in the loss function, SVR is equivalent to L2-regularized linear regression. In our method, the target variable is cellularity.

(2) Ranking SVM. As stated in Section 2.2.6, the problem of cellularity estimation could be handled with ranking models. We use ranking SVM [53] for this problem. Ranking SVM tries to build a linear ranking model using weight vector $\mathbf{w}$ and minimizes the pairwise ranking error by maximizing the minimal margin between pairs of ranking scores, as illustrated in Figure 6, where the minimum distance d on the weight vector $\mathbf{w}$ between any two samples is maximized. Based on the idea of SVM classifier, ranking SVM can be formulated as a convex optimization problem:

$$\begin{aligned} \min_{\mathbf{w},\xi_{i,j}} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum \xi_{i,j}, \\ \text{s.t.} \quad & \mathbf{w}^{\mathrm{T}}\mathbf{x_i} - \mathbf{w}^{\mathrm{T}}\mathbf{x_j} \ge 1 - \xi_{i,j}, \\ & \xi_{i,j} \ge 0, \forall\left(\mathbf{x_i} \rhd \mathbf{x_j}\right), \end{aligned} \tag{8}$$

where $\mathbf{w}^{\mathrm{T}}\mathbf{x_i}$ and $\mathbf{w}^{\mathrm{T}}\mathbf{x_j}$ are the ranking scores for $\mathbf{x_i}$ and $\mathbf{x_j}$, C is the tradeoff parameter, $\xi_{i,j}$ is the slack variable, and $\mathbf{x_i} \rhd \mathbf{x_j}$ denotes that $\mathbf{x_i}$ should be ranked higher that $\mathbf{x_j}$. Prior to training, we translate cellularity into different ranking levels, the same as in Section 2.2.6. The ranking scores predicted by SVM are mapped to cellularity using the KNN method described in Section 2.2.6.

2.2.8. Evaluation Metrics. To evaluate the performance of different methods on the test set using the manual labels as reference, 3 metrics are used: intraclass correlation (ICC), Kendall's tau-b (τ_b) [54, 55], and the prediction probability (P_K) [56].

The ICC in this work refers to ICC (A, 1) according to the definitions in [57, 58]. It uses two-way model of analysis of variance (ANOVA) to evaluate the absolute agreement between any two measurements and can be formulated as

$$\text{ICC} = \frac{\text{MS}_R - \text{MS}_E}{\text{MS}_R + (k-1)\text{MS}_E + k/n\left(\text{MS}_C - \text{MS}_E\right)}, \tag{9}$$

where MS_R denotes the mean square for rows (testing samples), MS_E denotes the mean square error, MS_C denotes the mean square for columns (measurements), $k = 2$ is the number of measurements, and $n = 185$ is the number of testing samples.

Kendall rank correlation coefficient is a metric used to evaluate the ordinal correlation between two measurements with nonparametric hypothesis test [55]. It is calculated based on the measurements of pairs of testing samples. In this work, we use Kendall's tau-b [54], which is adjusted for ties and can be formulated as

$$\tau_b = \frac{n_c - n_d}{\sqrt{\left(n - n_{T_x} - n_{T_{xy}}\right)\left(n - n_{T_y} - n_{T_{xy}}\right)}}, \tag{10}$$

where x denotes the automatically estimated cellularity, y denotes the manual measurement, n_c and n_d denote the number of concordance and discordance pairs, respectively,

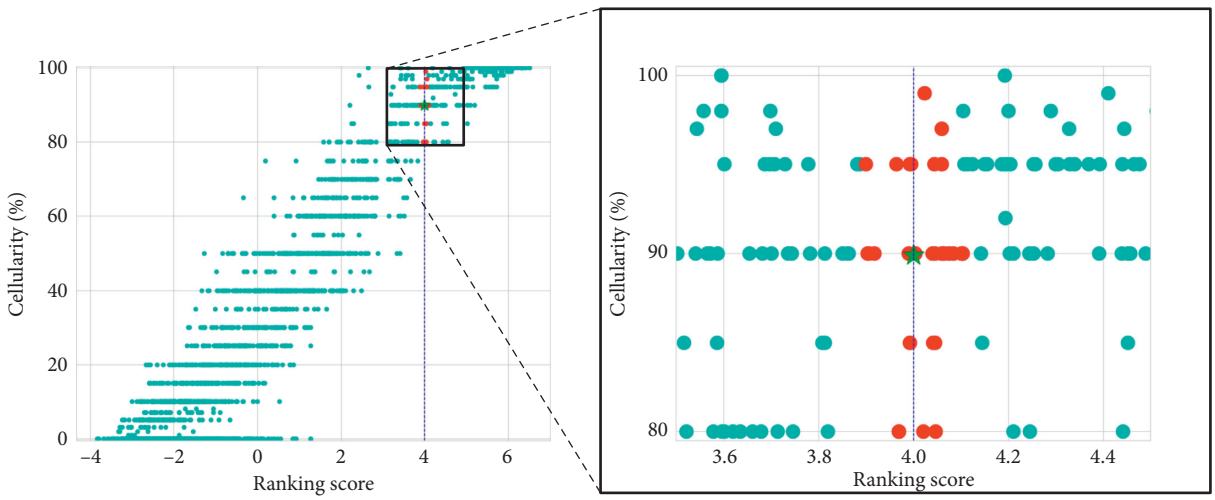

FIGURE 5: Illustration of our KNN method to map the ranking scores to cellularity. The red dots represent the training samples that lie within the 30-nearest neighborhood of a testing samples with ranking score 4.0. The cyan dots represent those outside the neighborhood. The prediction for the testing sample is the mean of cellularity of all the red dotted samples, as denoted by the green star.

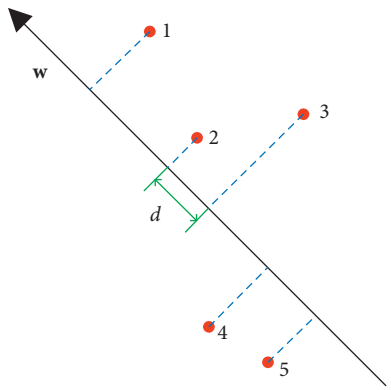

FIGURE 6: Illustration of Ranking SVM using 2D features, where **w** is the weight vector, 1–5 are the training samples, and d is the minimal margin that should be maximized.

n_{T_x} denotes the number of ties only in x, n_{T_y} denotes the number of ties only in y, and $n_{T_{xy}}$ denotes the number of ties in both x and y.

Prediction probability [56] is the metric adopted by the challenge organizer. It also evaluates the ordinal correlation and can be formulated as

$$P_K = \frac{n_c + n_{T_x}/2}{n_c + n_d + n_{T_x}}. \tag{11}$$

As explained by the challenge organizer, the reason for evaluating automatic methods with ordinal correlation metrics is that there exists variability among expert pathologists on cellularity assessment, and thus it is hard to define a calibrated and unbiased cellularity reference. Moreover, the cellularity estimated by the algorithm can be normalized to [0, 100%] without loss in consistency.

3. Experiments and Results

The feature extraction is carried out using Keras with TensorFlow backend, with the support of a NVIDIA GTX 1080 Ti GPU. All parameters of the pretrained models are provided by their respective authors.

3.1. Feature Selection. We perform feature selection using MIQ as criterion and sort the features into a list according to their prediction potential. Then, different numbers of top ranking features are selected to train GBDT models with Huber loss, and the performance is evaluated using P_K with 9-fold cross-validation on the training set. The parameter setting for GBDT is listed in Table 1. The relation between model performances and the number of features is illustrated in Figure 7. If not specified, the number of features selected from VGG, ResNet and Inception are 400, 800, and 1000, respectively, in the following experiments.

3.2. Tissue Classification Based on GBDT. Our GBDT classifier is constructed with the LightGBM package and is trained using the ResNet features that are not processed with mRMR and are processed with PCA where 40 principal components are preserved. Three hundred decision trees with depth of 7 are constructed, with the primary goal to minimize the BCE loss. The feature fraction and bagging fraction are 0.6 and 0.8, respectively, and the learning rate is 0.01. The other parameters are set as default. The ROC curve of our classification on the test set is presented in Figure 8, and the area under curve (AUC) is 0.95. This classifier can detect cancerous images in the test set with sensitivity of 0.98, specificity of 0.84, and accuracy of 0.96.

3.3. Prediction for Cancerous Patches. For tree boosting regression using Huber loss, we construct GBDT with the LightGBM package using the augmented training set. The δ in Huber loss is set to 1. For RankNet, we train the model with the XGBoost package. The parameter settings are summarized in Table 1, and parameters not listed are set as

Table 1: Parameter settings of the GBDT for cancerous patches.

Parameter	Huber	RankNet
Maximum tree depth	4	3
L1 regularization	0.001	0.001
L2 regularization	0	0
Feature fraction	0.6	0.6
Bagging fraction	0.7	0.8
Learning rate	0.01	0.01
Number of trees	1500	1000

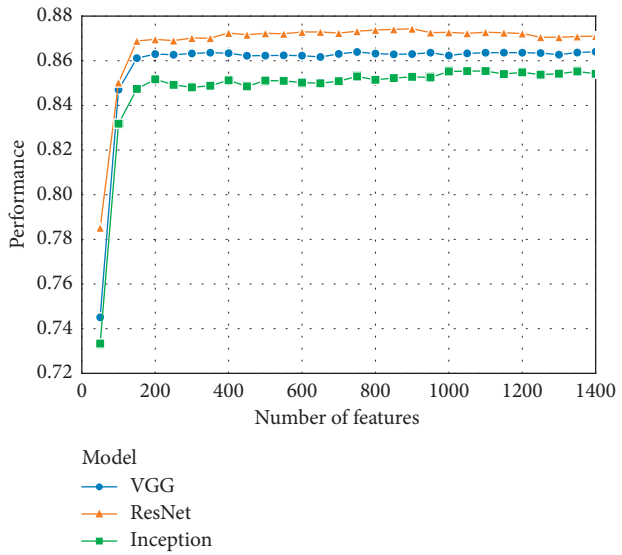

Figure 7: The curves showing the relation between model performances and the number of features.

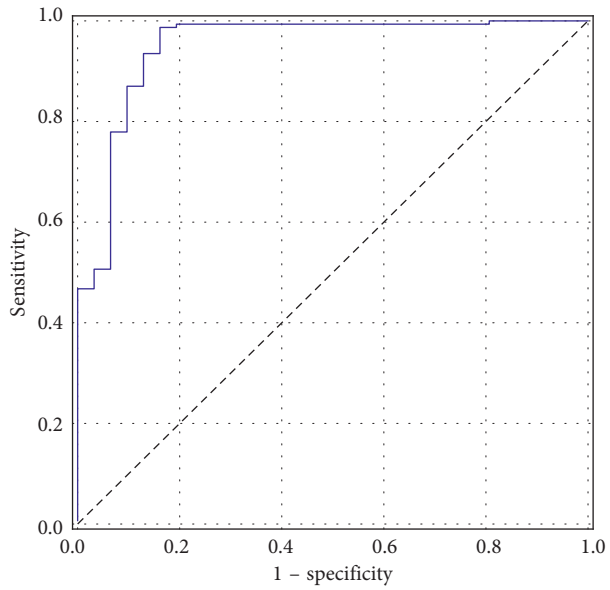

Figure 8: The ROC curve of our GBDT classification.

default. The features used to train the GBDTs are selected using mRMR and are not processed by PCA.

For SVR, we use LIBSVM [59] wrapped in the Scikit-learn [60] package. The top 4000, 8000, and 4000 features are selected from VGG, ResNet, and Inception, respectively, and the number of components for PCA is set as 150. The tradeoff parameter C is set as 200, and RBF with $\gamma = 0.0001$ is used as the kernel for the SVR with $\varepsilon = 0$.

For Ranking SVM, we use the software provided in [53]. The top 50 features are selected from each CNN, and the number of components for PCA is set as 20. Linear kernel is used, and the tradeoff parameter C is set as $1e8$.

Linear correction is used to make the mean and standard deviation of the estimated cellularity equal to those of the manual labels of the training set. All the parameter settings in our experiments are tuned using 9-fold cross-validations on the training set.

The classifier is first used to identify cancerous patches, and regression models combined with different features are applied to further predict the cellularity for them. Complete evaluations of our methods on the test set are listed in Table 2, from which it is clear that SVR combined with ResNet features obtains the best performance. With high correlation scores and tight upper and lower bounds of the 95% confidence interval, this method shows good agreement with the pathologist and produces stable predictions. Overall, the features from ResNet are more robust than the others, and SVR is the best model.

The ICC of the methods proposed by Peikari et al. [22] and Akbar et al. [23] are reported as 0.74 [0.70, 0.77] and 0.83 [0.79, 0.86], significantly lower than those of many of our models. The ICC between some of our methods and the pathologist's estimation is even higher than that between two different pathologists (0.89 [0.70, 0.95]) reported in [22], indicating that our methods can possibly supplant manual estimation in clinical practice.

The scatter plots in Figure 9 show the agreement on cellularity estimation between the pathologist and our automated methods. Note that to make the plots clear, each estimated cellularity is rounded to its nearest discrete bin. As presented in the figure, our predictions are in good agreement with the assessment by the pathologist. Generally, most of the estimations made by our methods are close to those by the pathologist. The black lines show the linear regressions between manual and automated measurements. As can be seen, the slopes of the regression lines of SVR and ranking SVM methods are close to 1, indicating good correlations to manual estimation.

3.4. Comparisons to the Method by Peikari et al. For direct comparisons to the method proposed by Peikari et al. [22] using exactly the same test set, we implement their algorithm. We follow their instructions for nuclei segmentation, labeling and classification, and cellularity computation. The dataset of manually annotated nuclei is provided by them as a part of the challenge. We compute 2 thresholds for each image using the Otsu algorithm and the lower one is used for segmentation. Morphological opening with a disk with radius of 3 is used to smooth nuclei boundaries and separate neighboring objects. Distance transform and local maxima finding are used to locate the centroids of overlapping nuclei, and marker-controlled watershed is applied to separate them. According to our observation and the proportion of

Table 2: Evaluations of all our methods.

Model	Feature	ICC (95% CI)	τ_b (95% CI)	P_K (95% CI)
GBDT with Huber loss	VGG	0.91 [0.87, 0.93]	0.78 [0.69, 0.79]	0.90 [0.86, 0.91]
	ResNet	0.91 [0.87, 0.93]	0.77 [0.73, 0.81]	0.90 [0.87, 0.91]
	Inception	0.87 [0.84, 0.90]	0.71 [0.65, 0.76]	0.86 [0.84, 0.89]
GBDT with RankNet	VGG	0.91 [0.88, 0.93]	0.77 [0.72, 0.81]	0.89 [0.87, 0.91]
	ResNet	0.91 [0.88, 0.94]	0.79 [0.75, 0.82]	0.90 [0.88, 0.92]
	Inception	0.88 [0.85, 0.91]	0.74 [0.70, 0.79]	0.88 [0.86, 0.90]
SVR	VGG	0.94 [0.92, 0.95]	0.80 [0.76, 0.84]	0.91 [0.89, 0.93]
	ResNet	**0.95 [0.93, 0.96]**	**0.83 [0.79, 0.86]**	**0.93 [0.91, 0.94]**
	Inception	0.89 [0.85, 0.92]	0.74, [0.69, 0.79]	0.88 [0.85, 0.90]
Ranking SVM	VGG	0.91 [0.88, 0.93]	0.76 [0.71, 0.80]	0.89 [0.87, 0.91]
	ResNet	0.92 [0.89, 0.94]	0.78 [0.73, 0.82]	0.90 [0.88, 0.92]
	Inception	0.89 [0.86, 0.92]	0.76 [0.71, 0.80]	0.89 [0.86, 0.91]

Bold indicates the best performance in terms of the corresponding metric.

matched nuclei, this parameter setting obtains optimal segmentation performance. Stromal nuclei are eliminated from classification by removing objects with ratio of major to minor axes ≥ 3. In total, 21573 nuclei (close to the 21799 nuclei in [22]) are matched to manual annotations and are partitioned into training set (13945 nuclei) and test set (7628 nuclei). The test set is separated from training and preserved to evaluate the generalization performance of the classifiers. Based on 5-fold cross-validation on the training set, we set $C = 100$ and $\gamma = 0.01$ for the two SVMs.

Our performance of 5-fold cross-validation on the training set of nuclei classification is presented in Figure 10, where the ROC curves of distinguishing lymphocyte from epithelial nuclei (L vs. BM) and classifying benign and malignant epithelial nuclei (B vs. M) are shown. It is clear that the points representing the performances reported by Peikari et al. locate on or under our ROC curves and that our AUCs are 0.99 and 0.94, trivially higher than those reported by them (0.97 and 0.86). The performance of our classifiers on the test set is summarized in Table 3. Comparing it to Table 3 of [22], it can be concluded that our implementation achieves comparable classification performance. We apply the nuclei detection and the postprocessing methods (including KNN correction and morphological dilation) described by Peikari et al. on the 185 testing samples for cellularity estimation and the ICC, τ_b and P_K are 0.76 [0.69, 0.81], 0.57 [0.49, 0.63], and 0.79 [0.75, 0.82], respectively, consistent with the performance reported in [22] and significantly outperformed by our methods.

4. Discussion

Cancer cellularity is an important component in the assessment of RCB, an effective indicator of tumor response. In this work, we propose a novel framework based on pretrained CNN features, GBDT, and SVM for direct cellularity estimation from histopathological slide patches within the tumor bed of breast cancer. Prior to training, feature selection is implemented to eliminate irrelevant features and PCA is used to reduce dimensionality, in order to reduce overfitting and improve training efficiency. Our methods are validated on a dataset consisting of 185 image patches and obtain state-of-the-art performance in terms of agreement with the human pathologist, in comparison to two other methods [22, 23]. The agreements between our methods and the pathologists' are even better than that between different pathologists, thus they are potential substitutes for manual inspection in clinical practice.

The problem of data scarcity is overcome using transfer learning. We extract deep features from multiple layers and manage to construct GBDT and SVM with a limited quantity of data. The results of our method demonstrate the transferability of deep features learned from natural images to specific histopathological data. The validity of general deep features on pathological images is also verified by recent literatures [28–30], where other tasks are handled.

The dataset poses a great challenge of label imbalance: the cellularity is not uniformly distributed in all the discrete bins. Two strategies are used in our framework to solve this problem. First, the noncancerous patches are sifted out using a GBDT classifier, as these patches make up a large part of our data. Second, regression and learn to rank models are used to predict the cellularity for the cancerous patches, as these models are insusceptible to such imbalance.

The approach proposed by Peikari et al. [22] is a two-stage method including segmentation and classification of individual nucleus with hand-crafted features, thus the performance depends heavily on the accuracy of nuclei identification and additional datasets of annotated nuclei are necessary for training. These drawbacks are overcome by the methods proposed by us and Akbar et al., where direct predictions for cellularity are given and only manual labels of cellularity are necessary. In [23], Akbar et al. fine-tuned two modified CNN for classification and regression. It can be concluded that direct learning models are generally more robust than nuclei detection-based estimations. There are two major differences between our methods and that reported by Akbar et al. First, the features extracted by our deep models are processed with more advanced learning algorithms, while in their method, predictions are based on linear models. Second, our methods extract features from multiple levels of CNN while their predictions are purely based on the last layer. The results show that our methods achieve better agreement with the human pathologist.

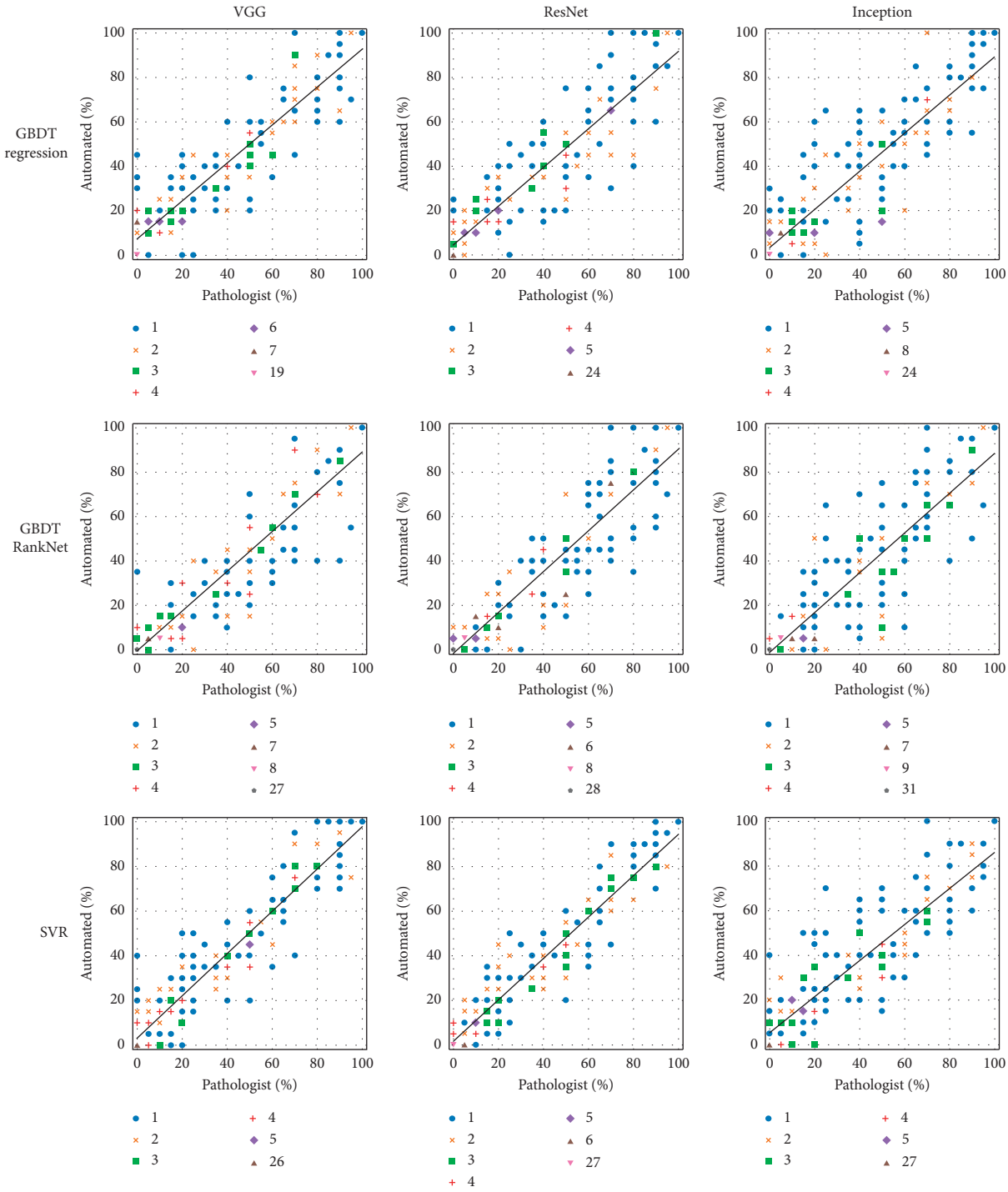

Figure 9: Continued.

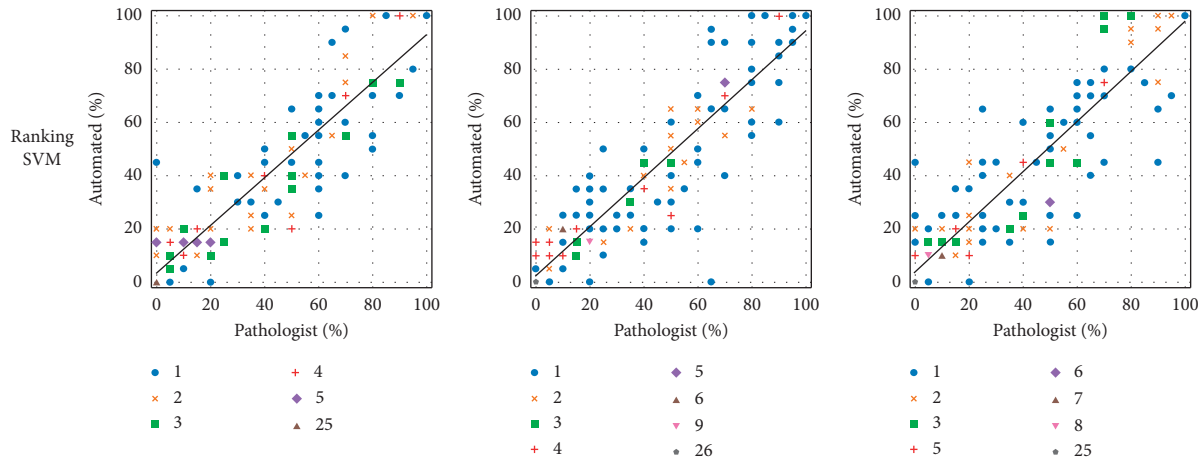

Figure 9: Scatter plots showing the agreement between the human pathologist and the automated approaches. The marker style of each dot indicates the number of patches with the corresponding manually and automatically estimated cellularity.

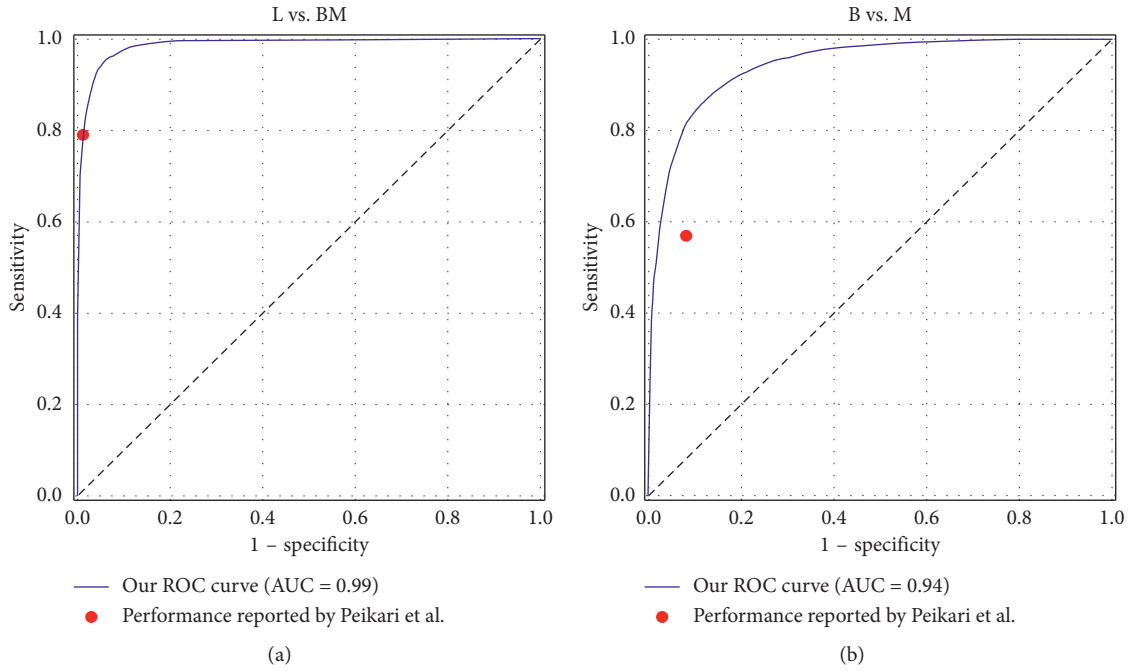

Figure 10: The ROC curves of our nuclei classification based on 5-fold cross-validation on the training set.

Table 3: Performance of the SVMs on the independent test set for nuclei classification.

Class	Accuracy (%)	Sensitivity (%)	Specificity (%)
Lymphocyte (L)	95	84	98
Benign epithelial (B)	88	60	93
Malignant epithelial (M)	89	93	82

It is reasonable to suppose that our methods are ready to be adapted to other histological applications as long as the corresponding data and cellularity labels are provided. The cellularity distribution over breast cancer tumor bed can be easily mapped, and the post-NAT tumor burden can be assessed in a semiautomatic workflow. Future research should focus on the automated segmentation of tumor bed and the estimation of other parameters of RCB.

5. Conclusion

In this work, novel methods for direct cellularity estimation combining deep feature representation, tree boosting, and SVM are proposed. The agreements between the estimations by our methods and those by human pathologists are validated using 3 metrics. Furthermore, the training of our models requires only lightly labeled data instead of annotations on individual nuclei, thus is more generalizable.

Data Availability

The histopathology image data used to support the findings of this study have been deposited in http://spiechallenges.cloudapp.net/competitions/14.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

The data used in this research were acquired from Sunnybrook Health Sciences Centre with funding from the Canadian Cancer Society and were made available for the BreastPathQ challenge, sponsored by the SPIE, NCI/NIH, AAPM, and Sunnybrook Research Institute. This work was supported by the Natural Science Foundation of Guangdong Province (grant no. 2016A030313577) and the Science and Technology Program of Guangdong (grant no. 2018B030333001).

References

[1] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2017," *CA: A Cancer Journal for Clinicians*, vol. 67, no. 1, pp. 7–30, 2017.

[2] A. M. Thompson and S. L. Moulder-Thompson, "Neoadjuvant treatment of breast cancer," *Annals of Oncology*, vol. 23, no. 10, pp. x231–x236, 2012.

[3] M. Kaufmann, G. N. Hortobagyi, A. Goldhirsch et al., "Recommendations from an international expert panel on the use of neoadjuvant (primary) systemic treatment of operable breast cancer: an update," *Journal of Clinical Oncology*, vol. 24, no. 12, pp. 1940–1949, 2006.

[4] G. Von Minckwitz, M. Untch, J.-U. Blohmer et al., "Definition and impact of pathologic complete response on prognosis after neoadjuvant chemotherapy in various intrinsic breast cancer subtypes," *Journal of Clinical Oncology*, vol. 30, no. 15, pp. 1796–1804, 2012.

[5] W. F. Symmans, F. Peintinger, C. Hatzis et al., "Measurement of residual breast cancer burden to predict survival after neoadjuvant chemotherapy," *Journal of Clinical Oncology*, vol. 25, no. 28, pp. 4414–4422, 2007.

[6] Breast Cancer Residual Cancer Burdern, https://www.mdanderson.org/for-physicians/clinical-tools-resources/clinical-calculators/residual-cancer-burden.html.

[7] A. J. J. Smits, J. A. Kummer, P. C. de Bruin et al., "The estimation of tumor cell percentage for molecular testing by pathologists is not accurate," *Modern Pathology*, vol. 27, no. 2, pp. 168–174, 2014.

[8] A. S. Pakurar and J. W. Bigbee, *Digital Histology: An Interactive CD Atlas with Review Text*, John Wiley & Sons, Hoboken, NJ, USA, 2011.

[9] L. Pantanowitz, "Digital images and the future of digital pathology," *Journal of Pathology Informatics*, vol. 1, no. 1, p. 15, 2010.

[10] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, "Breast cancer histopathological image classification using convolutional neural networks," in *Proceedings of the 2016 International Joint Conference on Neural Networks (IJCNN)*, pp. 2560–2567, IEEE, Vancouver, Canada, July 2016.

[11] T. Araújo, G. Aresta, E. Castro et al., "Classification of breast cancer histology images using convolutional neural networks," *PLoS One*, vol. 12, no. 6, Article ID e0177544, 2017.

[12] A. Cruz-Roa, H. Gilmore, A. Basavanhally et al., "Accurate and reproducible invasive breast cancer detection in whole-slide images: a Deep Learning approach for quantifying tumor extent," *Scientific Reports*, vol. 7, no. 1, article 46450, 2017.

[13] A. Janowczyk and A. Madabhushi, "Deep learning for digital pathology image analysis: a comprehensive tutorial with selected use cases," *Journal of Pathology Informatics*, vol. 7, no. 1, p. 29, 2016.

[14] S. Akbar, L. Jordan, A. M. Thompson, and S. J. McKenna, "Tumor localization in tissue microarrays using rotation invariant superpixel pyramids," in *Proceedings of the IEEE 12th International Symposium on Biomedical Imaging*, pp. 1292–1295, IEEE, Brooklyn, NY, USA, April 2015.

[15] S. Akbar, L. B. Jordan, C. A. Purdie, A. M. Thompson, and S. J. McKenna, "Comparing computer-generated and pathologist-generated tumour segmentations for immunohistochemical scoring of breast tissue microarrays," *British Journal of Cancer*, vol. 113, no. 7, pp. 1075–1080, 2015.

[16] J. Cheng, J. Zhang, Y. Han et al., "Integrative analysis of histopathological images and genomic data predicts clear cell renal cell carcinoma prognosis," *Cancer Research*, vol. 77, no. 21, pp. e91–e100, 2017.

[17] X. Zhu, J. Yao, F. Zhu, and J. Huang, "WSISA: making survival prediction from whole slide histopathological images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7234–7242, IEEE, Honolulu, HI, USA, July 2017.

[18] G. Litjens, C. I. Sánchez, N. Timofeeva et al., "Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis," *Scientific Reports*, vol. 6, no. 1, article 26286, 2016.

[19] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[20] Y. Deng, Z. Ren, Y. Kong, F. Bao, and Q. Dai, "A hierarchical fused fuzzy deep neural network for data classification," *IEEE Transactions on Fuzzy Systems*, vol. 25, no. 4, pp. 1006–1012, 2017.

[21] Y. Deng, F. Bao, Y. Kong, Z. Ren, and Q. Dai, "Deep direct reinforcement learning for financial signal representation and trading," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 3, pp. 653–664, 2017.

[22] M. Peikari, S. Salama, S. Nofech-Mozes, and A. L. Martel, "Automatic cellularity assessment from post-treated breast surgical specimens," *Cytometry Part A*, vol. 91, no. 11, pp. 1078–1087, 2017.

[23] S. Akbar, M. Peikari, S. Salama, A. Y. Panah, S. Nofech-Momes, and A. L. Martel, "Automated and manual quantification of tumour cellularity in digital slides for tumour burden assessment," *bioRxiv*, vol. 2019, article 571190, 2019.

[24] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[25] P.-S. Liao, T.-S. Chen, and P.-C. Chung, "A fast algorithm for multilevel thresholding," *Journal of Information Science and Engineering*, vol. 17, no. 5, pp. 713–727, 2001.

[26] M. Peikari and A. L. Martel, "Automatic cell detection and segmentation from H and E stained pathology slides using colorspace decorrelation stretching," in *Medical Imaging 2016: Digital Pathology*, International Society for Optics and Photonics, Bellingham, WA, USA, 2016.

[27] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision,"

in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2818–2826, IEEE, Las Vegas, NV, USA, June 2016.

[28] A. Rakhlin, A. Shvets, V. Iglovikov, and A. A. Kalinin, "Deep convolutional neural networks for breast cancer histology image analysis," in *Proceedings of the International Conference Image Analysis and Recognition*, pp. 737–744, Springer, Póvoa de Varzim, Portugal, June 2018.

[29] F. A. Spanhol, L. S. Oliveira, P. R. Cavalin, C. Petitjean, and L. Heutte, "Deep features for breast cancer histopathological image classification," in *Proceedings of the 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 1868–1873, IEEE, Miyazaki, Japan, October 2017.

[30] N. Weiss, H. Kost, and A. Homeyer, "Towards interactive breast tumor classification using transfer learning," in *Proceedings of the International Conference Image Analysis and Recognition*, pp. 727–736, Springer, Póvoa de Varzim, Portugal, June 2018.

[31] C. B. Do and A. Y. Ng, "Transfer learning for text classification," in *Proceedings of the Advances in Neural Information Processing Systems*, pp. 299–306, Barcelona, Spain, December 2006.

[32] R. Raina, A. Y. Ng, and D. Koller, "Constructing informative priors using transfer learning," in *Proceedings of the 23rd International Conference on Machine Learning*, pp. 713–720, ACM, Pittsburgh, PA, USA, June 2006.

[33] F. Chollet, "Xception: deep learning with depthwise separable convolutions," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1800–1807, IEEE, Honolulu, HI, USA, July 2017.

[34] SPIE-AAPM-NCI BreastPathQ, "Cancer cellularity challenge," https://breastpathq.grand-challenge.org/.

[35] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu et al., "Convolutional neural networks for medical image analysis: full training or fine tuning?," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1299–1312, 2016.

[36] S. Hoo-Chang, H. R. Roth, M. Gao et al., "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1285–1298, 2016.

[37] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[38] S. Roy, A. kumar Jain, S. Lal, and J. Kini, "A study about color normalization methods for histopathology images," *Micron*, vol. 114, pp. 42–61, 2018.

[39] M. Macenko, "A method for normalizing histology slides for quantitative analysis," in *Proceedings of the IEEE International Symposium on Biomedical Imaging: From Nano to Macro (ISBI'09)*, pp. 1107–1110, IEEE, Boston, MA, USA, July 2009.

[40] A. C. Ruifrok and D. A. Johnston, "Quantification of histochemical staining by color deconvolution," *Analytical and Quantitative Cytology and Histology*, vol. 23, no. 4, pp. 291–299, 2001.

[41] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: a large-scale hierarchical image database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255, Miami, FL, USA, June 2009.

[42] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proceedings of the International Conference on Learning Representations*, Lille, France, July 2015.

[43] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, Las Vegas, NV, USA, June 2016.

[44] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?," in *Proceedings of the Advances in Neural Information Processing Systems*, pp. 3320–3328, Montreal, Canada, December 2014.

[45] Y. L. Boureau, J. Ponce, and Y. LeCun, "A theoretical analysis of feature pooling in visual recognition," in *Proceedings of the 27th International Conference on Machine Learning*, pp. 111–118, Haifa, Israel, June 2010.

[46] Y. Xu, Z. Jia, L.-B. Wang et al., "Large scale tissue histopathology image classification, segmentation, and visualization via deep convolutional activation features," *BMC Bioinformatics*, vol. 18, no. 1, p. 281, 2017.

[47] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[48] T. Chen and C. Guestrin, "XGBoost: a scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, ACM, San Francisco, CA, USA, August 2016.

[49] G. Ke, Q. Meng, T. Finley et al., "LightGBM: a highly efficient gradient boosting decision tree," in *Proceedings of the Advances in Neural Information Processing Systems*, pp. 3146–3154, Long Beach, CA, USA, December 2017.

[50] C. Burges, "Learning to rank using gradient descent," in *Proceedings of the 22nd International Conference on Machine Learning*, pp. 89–96, ACM, New York, NY, USA, July 2005.

[51] V. N. Vapnik, *The Nature of Statistical Learning Theory*, Springer, Berlin, Germany, 1995.

[52] H. Drucker, C. J. C. Burges, L. Kaufman, A. J. Smola, and V. Vapnik, "Support vector regression machines," in *Advances in Neural Information Processing Systems*, pp. 155–161, MIT Press, Cambridge, MA, USA, 1997.

[53] T. Joachims, "Optimizing search engines using clickthrough data," in *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 133–142, ACM, Edmonton, AB, Canada, July 2002.

[54] M. G. Kendall, "The treatment of ties in ranking problems," *Biometrika*, vol. 33, no. 3, pp. 239–251, 1945.

[55] M. G. Kendall, "A new measure of rank correlation," *Biometrika*, vol. 30, no. 1-2, pp. 81–93, 1938.

[56] W. D. Smith, R. C. Dutton, and N. T. Smith, "A measure of association for assessing prediction accuracy that is a generalization of non-parametric roc area," *Statistics in Medicine*, vol. 15, no. 11, pp. 1199–1215, 1996.

[57] K. O. McGraw and S. P. Wong, "Forming inferences about some intraclass correlation coefficients," *Psychological Methods*, vol. 1, no. 1, pp. 30–46, 1996.

[58] T. K. Koo and M. Y. Li, "A guideline of selecting and reporting intraclass correlation coefficients for reliability research," *Journal of Chiropractic Medicine*, vol. 15, no. 2, pp. 155–163, 2016.

[59] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, p. 27, 2011.

[60] F. Pedregosa, "Scikit-learn: machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.