Using transfer learning on whole slide images to predict tumor mutational burden in bladder cancer patients

Hongming Xu[1], Sunho Park[1], Sung Hak Lee[2], and Tae Hyun Hwang[1⋆]

[1] Department of Quantitative Health Sciences, Cleveland Clinic, Cleveland, OH 44195, USA

[2] Department of Hospital Pathology, Seoul St.Mary's Hospital, College of Medicine, The Catholic University of Korea, Seoul 06591, Korea.

Abstract. The tumor mutational burden (TMB) is a genomic biomarker, which can help in identifying patients most likely to benefit from immunotherapy across a wide range of tumor types including bladder cancer. DNA sequencing, such as whole exome sequencing (WES) is typically used to determine the number of acquired mutations in the tumor. However, WES is expensive, time consuming and not applicable to all patients, and hence it is difficult to be incorporated into clinical practice. This study investigates the feasibility to predict bladder cancer patients TMB by using histological image features.

We design an automated whole slide image analysis pipeline that predicts bladder cancer patient TMB via histological features extracted by using transfer learning on deep convolutional networks. The designed pipeline is evaluated to publicly available large histopathology image dataset for a cohort of 253 patients with bladder cancer obtained from The Cancer Genome Atlas (TCGA) project. Experimental results show that our technique provides over 73% classification accuracy, and an area under the receiver operating characteristic curve of 0.75 in distinguishing low and high TMB patients. In addition, it is found that the predicted low and high TMB patients have statistically different survivals, with the p value of 0.047. Our results suggest that bladder cancer patient TMB is predictable by using histological image features derived from digitized H&E slides. Our method is extensible to histopathology images of other organs for predicting patient clinical outcomes.

Keywords: bladder cancer, tumor mutational burden, whole slide image, transfer learning

1 Introduction

Tumor mutational burden (TMB) is a quantitative genomic biomarker that measures the number of mutations within a tumor genome. TMB has been shown to be associated with improved responses to checkpoint inhibitor immunotherapies

⋆ Corresponding author email: hwangt@ccf.org

in lung cancer, melanoma and bladder cancer. Higher TMB levels are correlated with higher levels of neoantigens which help immune system to recognize tumors [2]. It has been found that patients with high TMB had better clinical outcomes such as improved immunotherapy response rate and progression free survival after immunotherapy [3]. DNA sequencing (e.g., WES) is typically used to assess the number of acquired mutations in the tumor. However, one key challenge is that not all patients would have adequate tumor tissue excised for genomic testing to measure TMB. Although the blood-based TMB measurement (e.g., liquid biopsies) recently becomes available, this approach also posies many technical challenges to measure TMB accurately. Due to these challenges and high expense in genome sequencing, TMB measurement is difficult to be incorporated into current clinical practice [1]. Therefore, to explore novel methods that are effective and economical for measuring TMB, such as by making use of widely available histological image information, is promising.

Routine histopathological examination is the gold standard for diagnosis and grading of various cancer types such as skin cancer, bladder cancer, and prostate cancer. By evaluating cellular morphologies and its changes in pathology slides, tumor information such as histologic grade, mitotic rate, lymph node status and tumor stage is provided in pathology reports. With the development to acquire digitized whole slide images (WSI), there have been plenty of studies addressing computer-aided pathology image analysis, which try to overcome human subjectivity and reduce clinical workload. Barker *et al.* [4] proposed a WSI analysis framework that computes cytological and textural features from local representative tiles for brain tumor subtype classification. Yu *et al.* [5] designed a method for lung adenocarcinoma and squamous cell carcinoma distinction, which distilled hand-crafted histological features from selected densest tiles of the WSI. Cheng *et al.* [7] proposed to extract topological features from manually selected tumor tiles of the WSI and associate them with renal tumor patient survival. Mobadersany *et al.* [6] trained a convolutional neural network model to extract and correlate histological features with glioma patient survival. More recent studies on computer-aided diagnosis of pathology slides have been reported on different cancer types including such as colon cancer [10], breast cancer [8][9], skin cancer [11] and prostate cancer [14].

Genetic change (e.g., gene mutation) is an alteration in the DNA sequence that makes up a gene. Since genetic change is not visible on pathology slides, it is very difficult or even impossible for human evaluators to predict gene-level clinical information based purely on H&E stained microscopic slides. Recently, there have been two studies that explored and suggested to predict certain somatic gene mutations by using computerized histological image features. Schaumberg *et al.* [15] proposed a quantitative model to predict SPOP mutation state in prostate cancer using H&E stained whole slide images. Their technique first determines a cohort of dominant tumor tiles based on tumor nuclei densities in WSI. Ensembles of residual networks [16] are then trained and integrated to predict SPOP mutation state. Coudray *et al.* [17] trained an inceptionV3 deep learning model [18] on lung adenocarcinoma (LUAD) whole slide images to pre-

dict ten most commonly mutated genes in LUAD. They reported that six of those ten commonly mutated genes including such as STK11, EGFR and FAT1 are predictable from pathology images by using deep learning models. These two studies are the pioneer works that try to discover whether gene mutations would affect tumor patterns on prostate and lung cancer pathology slides, respectively. Although they have reported promising results about the correlation between genetic changes and tumor morphologies within WSI, more studies are desired to verify if these correlations are generalizable on other cancer types and other clinic outcomes.

In this study, we hypothesize that abnormal changes in terms of tumor cell nuclei morphologies and tumor micro-environment reflected in WSI are functions of the underlying genetic drivers, and hence histological image features could reversely predict genetic status such as TMB. Using a TCGA cohort of bladder cancer WSI, we explore the feasibility to predict TMB of bladder cancer patients by using histological image features. The major contributions for this study are: (1) we design a general pipeline for WSI analysis, which is, to our best knowledge, the first study to predict bladder cancer patient TMB; (2) several state-of-the-art deep learning models are explored to extract image features by using transfer learning; (3) Our experiments thoroughly evaluated on a TCGA cohort suggest that histological image features are informative to patient TMB state. The organization of this paper is as follows. Section 2 describes the dataset and our designed method, followed by evaluations in Section 3. Section 4 provides a short discussion and conclusion.

2 Materials and Methods

2.1 Dataset

A cohort of 386 bladder cancer patients (and corresponding clinical information) with 457 diagnostic H&E stained WSI was downloaded from TCGA data portal (project TCGA-BLCA). Based on the percentile of total number single nucleotide variants [3], 386 TCGA bladder cancer patients were categorized into 3 groups: 128 low, 128 intermediate and 130 high TMB patients. Since prediction of intermediate TMB patients is not the clinical interest, this study only focuses on prediction of high vs low TMB patients. Due to either lack of patient survival information or poor quality slides with severe image artifacts [12], 5 patients belonging to low or high TMB categories are excluded. Overall, there are 124 low and 129 high TMB patients that are adopted for TMB prediction in this study. The remaining 128 intermediate TMB patient slides were sent to a pathologist (Dr. Sung Hak Lee), who randomly selected 60 representative patient slides and manually annotated visible tumor regions in these slides (using freely available viewing software ImageScope from Aperio). The annotation of tumor regions by the pathologist is used to train a classifier for tumor detection, which is a key step in our TMB prediction pipeline. Although there are a small number of patients with more than 1 diagnostic slides, for simplicity we only select the first diagnostic slide (i.e., with DX1 suffix) from every patient. Each digitized

slide had been scanned at multiple resolutions ranging from 2.5× to 40×, with all images containing a maximum scanning resolution scan at least 20× ($0.5\mu m$ per pixel). Fig. 1 illustrates an example of digitized whole pathology slide with multiple resolutions. Our study will make use of different image resolutions to perform histological image analysis.

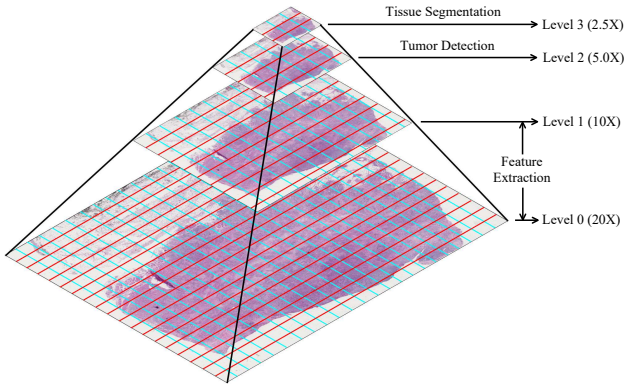


Fig. 1: Illustration of digitized whole pathology slide with multiple resolutions.

2.2 Method

Fig. 2 shows the pipeline of the proposed technique. It is observed from Fig. 2 that the proposed technique has four main modules. First, bladder cancer regions are detected from the whole slide image. Affinity propagation clustering is then applied to select a number of representative tumor tiles. After that, histological features are extracted from color-normalized tumor tiles based on transfer learning on a pre-trained deep learning model. Finally, the integrated features are input into a SVM classifier, which predicts patient slide into different categories: high TMB or low TMB. Details of these four modules are presented in the following.

Tumor detection Since TCGA bladder cancer pathology slides include both tumor and non-tumor (e.g., surrounding normal tissues) regions (see Fig. 3(a)), it is necessary to first detect tumor regions such that the subsequent histological image analysis is focused on tumor regions. In this module, we train and apply a tumor classifier to distinguish tumor and non-tumor regions in the WSI.

1) Establishing classifier: To establish a tumor classifier, our cooperated pathologist was asked to annotate tumor and non-tumor regions from 60 different representative patient slides. Since the pathologist's annotation was mainly performed at 5.0× magnification, we first extract those annotated ground truth regions at 5.0× magnification using the Openslide library [19]. The extracted

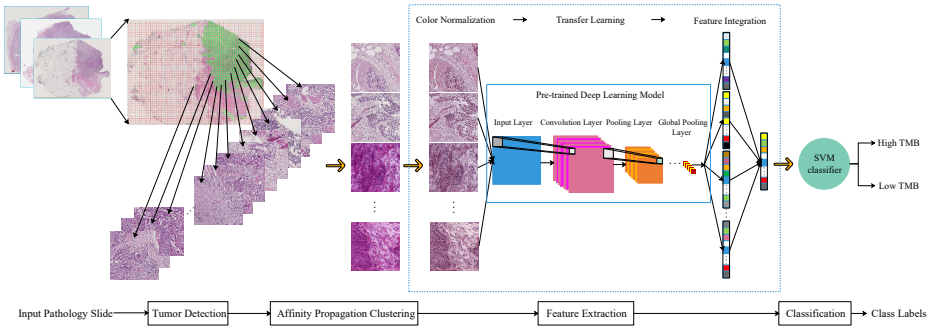


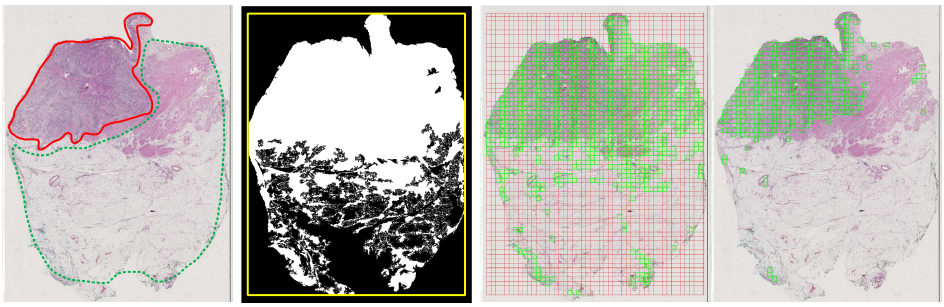
Fig. 2: Pipeline of the proposed technique. The technique includes four main modules: tumor detection, affinity propagation clustering, feature extraction and classification.

regions are then divided into a set of non-overlap tiles, where each tile has a size of 256×256 pixels (for examples see supplementary Fig.1). Overall, there are 13,635 tumor tiles and 14,191 non-tumor tiles obtained from 60 annotated patient slides. After that, the multi-scale local binary patterns (LBP) operator [20] is applied on every annotated image tile to compute a feature representation. The LBP operator characterizes the spatial structure of a local image neighborhood by encoding the differences between the pixel value of the central point and those of its neighbors, i.e.,

$$LBP_{p,r}\left(g_c\right) = \sum_{n=0}^{p-1} s\left(g_n - g_c\right) 2^n, \quad s\left(x\right) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases} \tag{1}$$

where $\{g_n\}_{n=0}^{p-1}$ are the p evenly distributed surrounding pixels on a radius r centered at pixel g_c (see supplementary Fig.2). After LBP encoding by Eq. (1), the rotation invariant uniform $riu2$ scheme is applied to compress encoded pixel values from the range $0 \sim 2^p - 1$ to $0 \sim p+1$. The histogram derived from the LBP encoded feature map is used as the texture feature representation. In this study, the (p, r) values were set as (8,2), (8,4), (8,6) and (8,8), respectively, for multi-scale LBP operations, and hence each image tile corresponds to a 40 dimensional feature vector. Finally, a SVM classifier (with a Gaussian kernel) is trained using tumor and non-tumor tiles represented by the LBP-based feature vectors. The trained SVM classifier is used to detect tumor regions in bladder WSI. For experimental results about tumor detection classifier training and testing, please see Results Section 3.1.

2) Detecting tumor: Using the trained tumor classifier, we employ a multi-resolution analysis strategy to detect tumor regions. First, we extract the WSI with 2.5× magnification (see Fig. 1), and binarize gray-scale WSI by an empirically selected threshold τ ($\tau = 210$), which generates a binary mask I_b with bladder tissue regions (with relatively dark color) as the foreground. Morphological operations [21] are then applied to remove noisy regions and fill small



(a) (b) (c) (d)

Fig. 3: Illustration of tumor detection. (a) A H&E stained bladder WSI, where tumor and non-tumor regions are highlighted by red and (dashed) green contours, respectively. (b) Bladder tissue mask I_b, where white regions are segmented bladder tissue regions, and superimposed yellow rectangle is the bounding box of tissue regions. (c) Image tiling, where green squares highlight selected image tiles containing a relatively high proportion of bladder tissue pixels. (d) Tumor detection result, where green squares highlight all predicted tumor tiles.

holes within the foreground. In Fig. 3(b) foreground white regions correspond to segmented bladder tissue regions. After bladder tissue segmentation, the bounding box surrounding foreground regions is computed (see superimposed yellow rectangle in Fig. 3(b)). The image within the bounding box is divided into many non-overlapping tiles, from which every tile containing tissue regions over than half of its size is selected. In Fig. 3(c), the selected image tiles are highlighted by green squares. Finally, the selected image tiles with 5.0× magnification are extracted and represented by multi-scale LBP-based feature vectors. The trained SVM classifier predicts each image tile as tumor or non-tumor tile. Let us denote the detected tumor tiles as $\{T_i\}_{1\leq i\leq t}$, and corresponding LBP feature vectors as $\{F_i\}_{1\leq i\leq t}$, where t is the number of tumor tiles and each F_i is a 40 dimensional feature vector. In Fig. 3(d), the predicted tumor tiles (e.g., $t = 490$ for this slide) are highlighted by green squares. As observed from Figs. 3(a)(d), almost all tumor regions are correctly detected by the trained SVM classifier.

Affinity propagation clustering After obtaining tumor tiles $\{T_i\}_{1\leq i\leq t}$, we apply image clustering to select a subset of representative tumor tiles for subsequent feature extraction. Image clustering is performed for two main reasons. First, since the whole slide image typically has a huge size (usually several gigabytes) and includes hundreds to thousands of detected tumor tiles, it is very time-consuming to compute high-level features from all tumor tiles at high magnification. By contrast, computation of high-level features only from representative tumor tiles is much faster than that from all tumor tiles. Second, since there often exist noisy tumor tiles (e.g., poor quality tiles with image artifacts)

detected from TCGA WSI, feature extraction only from representative tumor tiles is likely to suppress contributions from those noisy tiles and improve classification performance.

In this study, affinity propagation (AP) clustering [22] is used to select representative tumor tiles. The selection of AP clustering rather than other popular clustering techniques such as k-mean algorithm is based on three considerations. (1) AP clustering does not require to pre-specify the number of clusters, instead it adaptively produces different number of clusters based on input feature vectors. (2) The solution of AP clustering is unique and does not change with different runs of the algorithm. (3) It has been shown that AP clustering is usually more accurate and efficient than the k-means algorithm. Considering that geometrically close tumor tiles usually have similar texture patterns and should belong to the same cluster, we represent each tumor tile by a 42 dimensional feature vector $\{F_i, (r_i, c_i)\}_{1 \leq i \leq t}$, where F_i is a 40 dimensional LBP feature vector and (r_i, c_i) represents the geometric location of a tile (i.e., the center point of tile T_i) in the image. The affinity propagation algorithm then treats each feature vector as a node in a network and recursively transmits real-valued messages along edges of the network until it finds a good set of exemplars and corresponding clusters. Following suggestions in the reference [22], we define the similarities between feature vectors of tumor tiles as the negative square Euclidean distance between them. The "preferences" that influence the number of finally generated clusters are set as the median value of similarities between feature vectors. Let us assume that AP clustering adaptively group tumor tiles into r clusters. Each cluster includes λ_j tumor tiles and a corresponding representative tumor tile R_j, where $1 \leq j \leq r$. Fig. 4 illustrates AP clustering of tumor tiles on a WSI, where tumor tiles belonging to different clusters are indicated by different color of blocks in the image. Note that there are 56 (i.e., $r = 56$) representative tiles selected among 490 tumor tiles for the patient slide shown in Fig. 4. In other words, the number of tumor tiles used for subsequent feature extraction reduces about 9 times after AP clustering.

Feature extraction In this module, we analyze representative tumor tiles obtained by the AP clustering and computes a high-level feature representation for the whole pathology slide. Since some of TCGA bladder cancer pathology slides only have the highest magnification at 20× (not 40×), in this study all representative tumor tiles $\{R_j\}_{1 \leq j \leq r}$ are extracted at 20× for high-level feature extraction. This module includes three steps, which are described as follows.

1) Color normalization: Because TCGA pathology images were collected from many different institutions, there exist severe color variations due to different staining procedures and storages. To suppress the influence of color variations on feature extraction, a color deconvolution based method [23] is utilized to normalize tumor tile into a standard color appearance. First, using the binary mask I_b shown in Fig. 3(b), we extract bladder tissue pixels at 2.5× magnification from the WSI. The singular decomposition method [23] is then applied to adaptively determine H&E stain vectors based on the extracted bladder tissue pixels

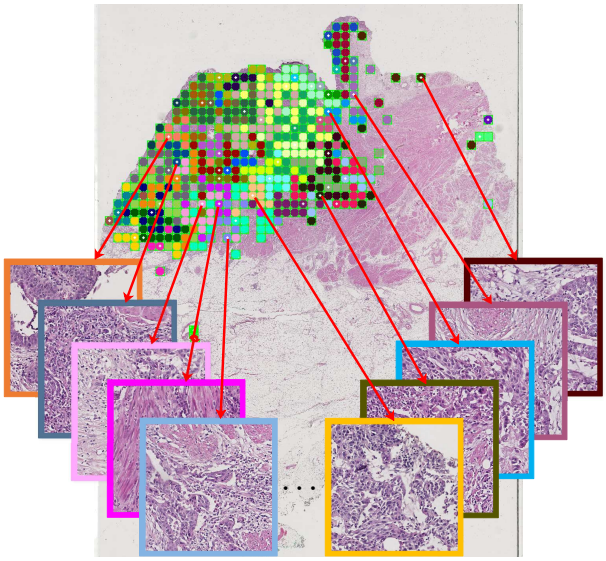


Fig. 4: Illustration of AP clustering. Note that tumor tiles belonging to different clusters are indicated by different color of blocks in the image. Several representative tumor tiles indicated by arrows are zoomed-in for better viewing.

from the WSI. Finally, the color deconvolution method [24] with the determined H&E stain vectors is separately applied to all representative tumor tiles, which are normalized to have similar color appearances (for examples see supplementary Fig.3). Note that we adaptively determine H&E stain vectors from the WSI rather than individual tumor tile, as some tumor tiles may include unexpected image artifacts (e.g., pen markers) which would impair determined H&E stain vectors for image color normalization.

2) Transfer learning: The application of deep learning models pre-trained on large annotated dataset such as ImageNet to solve another classification task with images of different modalities is referred to as transfer learning. Due to the challenge of acquiring sufficient training data in medical imaging domain, transfer learning on pre-trained convolutional neural networks have been applied for different classification problems, such as skin cancer [25], retinal diseases [26] and breast cancer [27]. Unlike these transfer learning studies where all images have explicit labels, our study explores to predict TMB from gigabyte whole pathology slides. Tumor representative tiles are determined from WSI for analysis, but they may not contain information relevant to the class assigned to the whole slide image. In other words, there is no guarantee that a patient slide has the same label with tumor representative tiles extracted from it. Therefore, instead of fine-tuning pre-trained models directly for TMB prediction, we make use of pre-trained models as the feature extractor. Motivated by the superior performance on ImageNet classification, this study utilizes a Xception CNN architecture [31]

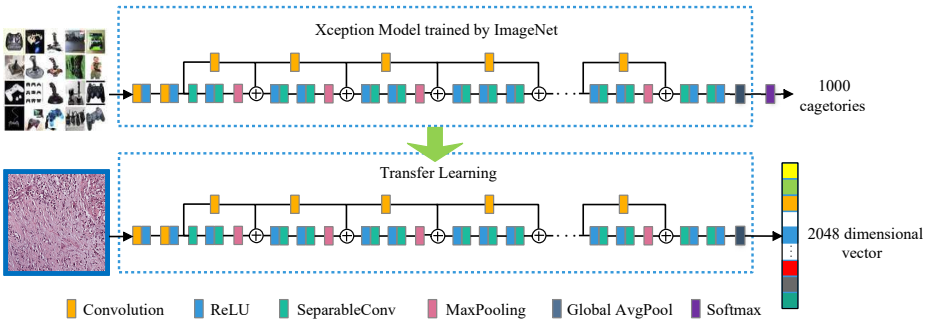


Fig. 5: Illustration of transfer learning on Xception model trained by ImageNet. Note that we remove the output layer of the pre-trained Xception model and use it as a feature extractor which outputs a 2048 dimensional vector for each input image.

that was pre-trained on approximately 1.28 million images (1,000 object categories) from the 2014 ImageNet Large Scale Visual Recognition Challenge [28]. The Xception module is extended from the Google Inception module [18] and residual learning [16], which replaces regular convolutions with depthwise separable convolutions and has shown an improved classification performance on several public image dataset. Fig. 5 illustrates transfer learning on the pre-trained Xception model. As illustrated in Fig. 5, we remove the output layer of the pre-trained Xception model and re-use all other weights trained by ImageNet. Given an input tumor tile R_j at 20× magnification (with a size of 1024×1024 pixels), the transfer learning model finally outputs a high-level feature representation V_j which is a 2048 dimensional vector.

3) Feature integration: After extracting histological features from all representative tumor tiles $\{R_j\}_{1\leq j\leq r}$, the feature vector $\overline{V}$ representing the whole patient slide is obtained by integrating features of individual tumor tiles together, i.e.,

$$\overline{V} = \sum_{j=1}^{r} \frac{\lambda_j}{\sum_{j=1}^{r} \lambda_j} V_j \tag{2}$$

Note that λ_j represents the number of tumor tiles belonging to the jth cluster. The feature vector $\overline{V}$ is the weighted mean of features extracted from representative tumor tiles. The representative tumor tile belonging to the cluster with more images is assigned a higher weight, such that the WSI could be effectively described by representative tumor tiles.

Classification In this module, we build the classifier to predict patient TMB using extracted histological image features. The classifier is trained on randomly selected training samples, which is performed in following three steps:

1) Feature dimension reduction: Since the feature vector $\overline{V}$ of the WSI has a high dimension (i.e., 2048) that is much larger than the number of patient

samples, feature dimension is reduced to prevent over-fitting to the training dataset. To ensure efficiency and simplicity, principal component analysis (PCA) is utilized, which selects a small number of principal feature components (e.g., 50 or 100) to train the classifier.

2) Feature standardization: To make every feature component contribute classification more equally, feature standardization is performed on each feature component, which makes its values have zero mean and unit variance.

3) Classifier construction: Using the normalized image features of training samples, we train a SVM classifier (with a Gaussian kernel) to predict patient TMB state.

After building up the TMB prediction classifier, the testing is ready to be performed on testing samples. Likewise the training sample, each testing sample corresponds to a high dimension feature vector with values in different scales. Therefore, feature dimension reduction and standardization are first performed based on the PCA transformation matrix and scaling factors computed from training samples. The trained SVM classifier is then applied to predict the patient slide into low TMB or high TMB category.

3 Results

In this section, we present evaluations of the proposed technique and provide comparisons with other baseline methods.

3.1 Tumor detection performance

For tumor detection, a pathologist was asked to annotate tumor and non-tumor regions in 60 different representative patient slides. To build the tumor detection classifier, 48 annotated patient slides were randomly selected as the training set, and the remaining 12 annotated patient slides were used for testing. Due to the large amount of time and effort needed to produce expert ground truth annotations, the pathologist only annotated obvious tumor and non-tumor regions in selected patient slides at about 5× magnification. In other words, there may exist some uncertain regions in annotated slides. Therefore, instead of evaluating tumor detection slide by slide, the evaluation is quantitatively performed by measuring classification performance of small annotated tiles. The widely-used classification evaluation criteria including accuracy, specificity, sensitivity and area under the receiver operating characteristic curve (AUC) [11][30] are used in this study.

Table 1 lists evaluation results of tumor detection on 12 testing patient samples (including 2722 tumor tiles and 2844 non-tumor tiles). As observed in Table 1, LBP texture features with SVM classifier provides a good performance in tumor detection, with over 94% accuracy and 0.98 of AUC value (please see plotted ROC curve in supplementary Fig.4). The trained tumor detection classifier is then qualitatively evaluated on 253 patient slides that are used for TMB prediction. The pathologist visually examined all tumor detection results and

confirmed that most of visible tumor regions had been correctly detected by the tumor detection classifier. For visual examples of tumor detections, please see supplementary Fig.5.

Table 1: Performance of tumor detection on testing samples

Technique	Accuracy (%)	Specificity (%)	Sensitivity (%)	AUC
LBP+SVM	94.18	95.14	93.19	0.98

3.2 TMB prediction performance

In this subsection, we first evaluate the proposed technique for TMB prediction, and then provide comparative results with two baseline methods.

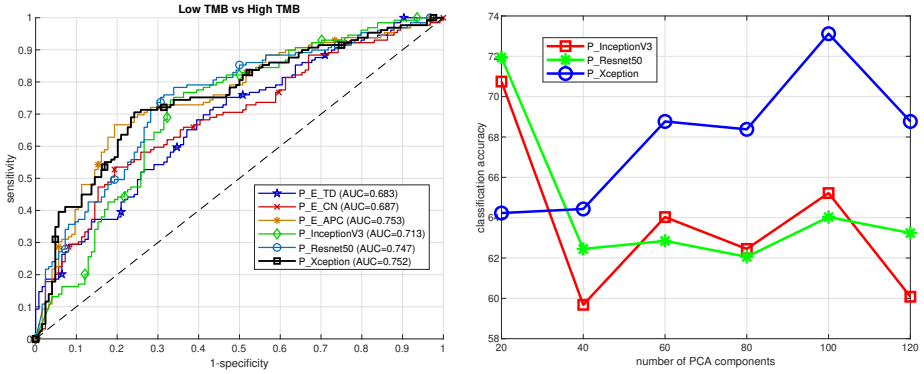
Evaluation of the proposed technique The proposed technique includes four modules: tumor detection, affinity propagation clustering, feature extraction and classification. To verify necessities of these modules, we evaluated the proposed technique by excluding a certain of modules. Specifically, the proposed technique excluding tumor detection (abbreviated as P-E-TD), the proposed technique excluding affinity propagation clustering (abbreviated as P-E-APC), and the proposed technique excluding color normalization (abbreviated as P-E-CN), are quantitatively evaluated. Besides using the Xception model (abbreviated as P-Xception) for transfer learning, we explored two more well-known deep learning models including InceptionV3 [18] and Resnet50 [16] as feature extractors in our pipeline, where are separately abbreviated as P-InceptionV3 and P-Resnet50 in subsequent sections. To quantitatively evaluate these variants of the proposed technique, we make use of leave-one-out validation method on 253 patient slides for TMB prediction.

Table 2: Performance evaluation of the proposed technique

Techniques	Accuracy (%)	Specificity (%)	Sensitivity (%)	AUC
P-E-TD	64.03	62.90	65.12	0.683
P-E-CN	64.43	66.13	62.79	0.687
P-E-APC	71.94	71.77	72.09	**0.753**
P-InceptionV3	70.75	66.94	74.42	0.713
P-Resnet50	71.94	68.55	**75.19**	0.747
P-Xception	**73.12**	**75.81**	70.54	0.752

Table 2 lists accuracy, specificity, sensitivity, and AUC values of the proposed technique with different settings. As observed in Table 2, the P-E-TD technique provides the overall poorest performance, with an AUC value of 0.683. This indicates that tumor detection module is crucial for TMB prediction, as non-tumor regions tend to be uninformative about tumor genetic changes and thus should be excluded for computerized histology image analysis. The P-E-CN technique provides a similar performance with the P-E-TD technique, which provides an AUC value of 0.687. The poor performance of the P-E-CN technique is because TCGA bladder cancer pathology slides were collected from many different institutions and there exist severe color variations among digitized slides. Therefore, color normalization before feature extraction plays a significant role in boosting TMB prediction performance. The P-E-APC technique provides a much better performance than the P-E-TD and P-E-CN techniques, which provides over 71% of accuracy, specificity, and sensitivity values, and the highest AUC value of 0.753. However, due to excluding the affinity propagation clustering module, the P-E-APC technique extracts image features from all tumor tiles and thus has the highest computational complexity. In this study, the number of tumor tiles utilized for feature extraction is reduced from 125,358 to 11,654 after AP clustering, which indicates that the P-E-APC technique has about 10 times computational complexity in comparison with other techniques with AP clustering. Compared with other techniques, the P-InceptionV3 provides an intermediate performance, with an AUC value of 0.713. The P-Resnet50 provides a better performance than the P-InceptionV3, which provides the highest sensitivity value of 75.19% and an AUC value of 0.747. Overall, the P-Xception provides the best performance among all variants of the proposed technique, which yields the highest accuracy (73.12%) and specificity (75.81%) values and an AUC value of 0.752. The transfer learning on Xception model yielding a better performance than that of either InceptionV3 or Resnet50 model could be due to two reasons. First, both InceptionV3 and Resnet50 models were developed with a focus on ImageNet and may thus be relatively less applicable to other datasets than the Xception model [31]. Second, the Xception model inherits advantages from both residual network and inception modules together, which tends to more powerful to learn image features. Note that although the P-E-CN provides the highest AUC value, we are inclined to refer our proposed technique as P-Xception due to its highest classification accuracy and relatively low computational complexity. Fig. 6(a) shows ROC curves for different variants of the proposed technique.

The PCA is utilized in the proposed pipeline for feature dimension reduction. However, it is usually not intuitive to determine the number of feature components that should be selected by the PCA transformation. To select the best number of feature components, we did a series of evaluations by adjusting the number of feature components from 20 to 120 with a step of 20. Fig. 6(b) compares classification accuracy changes of three transfer learning models with different number of selected PCA components. As shown in Fig. 6(b), both P-InceptionV3 and P-Resnet50 techniques provide the highest classification accuracy (as provided in Table 2) when 20 feature components are selected. However,



(a) (b)

Fig. 6: Evaluation of the proposed technique. (a) ROC curves for variants of the proposed technique. (b) Evaluation of three transfer learning models with different number of selected PCA components.

when the number of feature components ranges from 40 to 120, their accuracies are greatly reduced and fluctuates between 60% and 66%. In comparison, the P-Xception technique always provides the classification accuracy more than 64%, and achieves the highest classification accuracy about 73% when 100 feature components are selected. It could be concluded that the extracted features by transfer learning on Xception model are more robust than those of InceptionV3 and Resnet50 models.

Comparisons To our best knowledge, there are no other existing studies that tried to predict TMB by using histological image features. For comparisons, we implemented a traditional machine learning method that uses LBP textural features. Specifically, instead of extracting features by using transfer learning on pre-trained deep learning models, we extracted multi-scale LBP texture features in high resolution representative tumor tiles (i.e., 20× magnification), and then performed TMB prediction by training a SVM classifier. The (p, r) values (see Eq. (1)) for LBP texture computation were set as (16,2),(16,4),(16,6) and (16,8), respectively, and hence each patient slide finally corresponds to a 72 dimensional feature vector. To select the best number of PCA components for LBP features, we adjusted the number from 20 to 60 by a step of 10. It was found that 30 PCA components provide the best performance on our evaluated dataset. The leave-one-out validation method was used to evaluate bladder cancer patient TMB prediction performance using LBP features.

Table 3 lists comparisons of bladder cancer patient TMB prediction between the proposed technique and the traditional baseline method. As observed in Table 3, the proposed technique achieves a great improvement over the baseline method, where all gains in terms of accuracy, specificity, sensitivity and AUC

Table 3: Comparison with the baseline method

Techniques	Accuracy (%)	Specificity (%)	Sensitivity (%)	AUC	p-value
LBP	60.87	64.52	57.36	0.623	0.93
Proposed	73.12	75.81	70.54	0.752	0.047

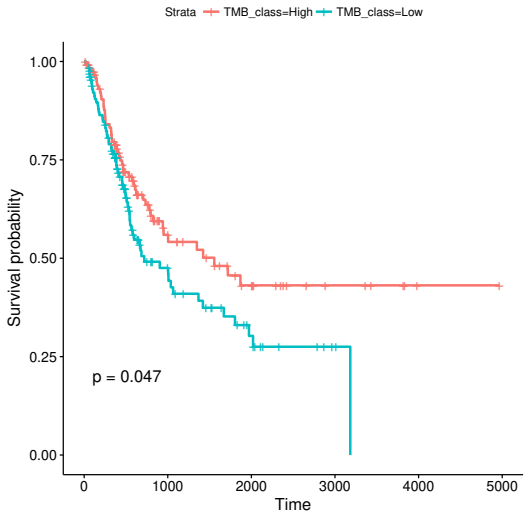


Fig. 7: Kaplan-Meier curve of bladder cancer patients stratified by TMB prediction using the proposed technique.

values are about 12% to 13%. In addition, we generates the Kaplan-Meier curve of bladder cancer patients stratified by TMB prediction using the automatic techniques. The proposed technique provides a p-value of 0.047, which is more significant than that (i.e., 0.93) of the baseline method. Fig. 7 shows the plotted Kaplan-Meier curve based on TMB predictions using the proposed technique.

4 Discussion

Visually examination of H&E stained histopathology slides by pathologists is a standard clinical practice to diagnose and grade different type of cancers. Many previous studies have been focusing on detection/segmentation of tissue objects [33][34], or extracting histological image features to assist pathologists in computer-aided tumor diagnosis [5][11]. Until recently, there are two existing studies [15][17] which explored and found that some gene mutations (e.g., SPOP, KRAS mutations) are predictable by using histological image features in prostate and lung cancer pathology slides. Their findings suggest that tumor

morphology or pattern changes in pathology slides are reflecting genetic aberrations driving cancer. Therefore, histological image features capturing tumor morphology changes could be informative to predict patient genetic status, and other clinical outcomes.

Tumor mutational burden (TMB) is an important genome biomarker that has been shown to be associated with treatment response to immunotherapy in many cancer types including bladder cancer. However, due to economical and technical issues, not all cancer patients are able to undergo a tumor biopsy to generate sequencing based genomic testing to measure TMB. Considering that histopathology slides used by pathologists are widely-available for most cancer patients, in this study, we explore and present the first computerized pipeline to explore if bladder cancer patient TMB is predictable by using histological image features. Our study make use of state-of-the-art deep learning models to extract histological image features in tumor tiles, and integrate them together to predict patient TMB status. By experimenting on TCGA bladder cancer patients, our results suggest that histological image features extracted from whole slide images can predict patient TMB status ($\sim$0.73 accuracy and $\sim$0.75 AUC).

There are still limitations in our study. First, the patients with intermediate TMB status were only used to train a tumor detection classifier, but they were excluded for TMB prediction evaluation. This is because intermediate TMB status is generally not informative for immunotherapy treatment and hence is usually not the clinical interest. In other words, even a patient is correctly predicted with an intermediate TMB status, it is still hard to determine whether immunotherapy would be beneficial for the patient. But still it would be more promising if intermediate TMB patients could be correctly predicted by computerized algorithms. Second, although we have evaluated our technique on a large cohort of TCGA bladder cancer patients via cross-validation, it is still eager to fully validate it on more patient slides or other cancer type of patient slides in order to verify its generalizability. Future works (also an ongoing effort) are to collect more patient slides and corresponding patient genetic information, which are to be used for further improving and validating our discoveries in this study.

In summary, this study explores the feasibility to predict bladder cancer patient TMB status by using histological image features extracted by the pretrained deep learning models. Our results suggest that histological image features revealing tumor morphology changes have prognostic values in predicting bladder cancer patient genetic changes, i.e., TMB status. These results pave the way for future studies to further refine the presented technique and validate it in prospective patient cohorts with bladder cancer or other cancer types.

References

1. Goodman, A. M. *et al.* (2017) Tumor mutational burden as an independent predictor of response to immunotherapy in diverse cancers. *Molecular cancer therapeutics*, **molcanther**, 0386.

2. Brown, S. D. *et al.* (2014) Neo-antigens predicted by tumor genome meta-analysis correlate with increased patient survival. *Genome research*, doi:10.1101/gr.165985.113
3. Robertson, A. G. *et al.* (2017). Comprehensive molecular characterization of muscle-invasive bladder cancer. *Cell*, **171(3)**, 540-556.
4. Barker, J. *et al.* (2016). Automated classification of brain tumor type in whole-slide digital pathology images using local representative tiles. *Medical image analysis*, **30**, 60-71.
5. Yu, K. H. *et al.* (2016). Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. *Nature communications*, **7**, 12474.
6. Mobadersany, P. *et al.* (2018). Predicting cancer outcomes from histology and genomics using convolutional networks. *Proceedings of the National Academy of Sciences*, 201717139.
7. Cheng, J. *et al.* (2017). Identification of topological features in renal tumor microenvironment associated with patient survival. *Bioinformatics*, **34(6)**, 1024-1030.
8. Liu, Y. *et al* (2017). Detecting cancer metastases on gigapixel pathology images. *arXiv preprint* arXiv:1703.02442.
9. Bejnordi, B. E. *et al*(2018). Using deep convolutional neural networks to identify and classify tumor-associated stroma in diagnostic breast biopsies. *Modern Pathology*, 1.
10. Bychkov, D. *et al.* (2018). Deep learning based tissue analysis predicts outcome in colorectal cancer. *Scientific reports*, **8(1)**, 3395.
11. Xu, H. *et al.* (2018). Automated analysis and classification of melanocytic tumor on skin whole slide images. *Computerized medical imaging and graphics*, **66**, 124-134.
12. Kothari, S. *et al.* (2013). Pathology imaging informatics for quantitative analysis of whole-slide images. *Journal of the American Medical Informatics Association*, **20(6)**, 1099-1108.
13. Wang, D. *et al.* (2016). Deep learning for identifying metastatic breast cancer. *arXiv preprint*, arXiv:1606.05718
14. Arvaniti, E. *et al.* (2018). Automated Gleason grading of prostate cancer tissue microarrays via deep learning. *bioRxiv*, 280024.
15. Schaumberg, A. J. *et al* (2018). H&E-stained whole slide image deep learning predicts SPOP mutation state in prostate cancer. *BioRxiv*, 064279.
16. He, K. *et al.* (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770-778.
17. Coudray, N. *et al.*, (2018). Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nature medicine*, **24(10)**, 1559.
18. Szegedy, C. *et al* (2016). Rethinking the inception architecture for computer vision. In Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2818-2826.
19. Goode, A. *et al* (2013). OpenSlide: A vendor-neutral software foundation for digital pathology. *Journal of pathology informatics*, 4.
20. Ojala, T. *et al* (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on pattern analysis and machine intelligence*, **24(7)**, 971-987.
21. Xu, H. *et al* (2015). Epidermis segmentation in skin histopathological images based on thickness measurement and k-means algorithm. *EURASIP Journal on Image and Video Processing*, **2015(1)**, 18.
22. Frey, B. J., & Dueck, D. (2007). Clustering by passing messages between data points. *science*, **315(5814)**, 972-976.

23. Macenko, M. *et al.*, (2009). A method for normalizing histology slides for quantitative analysis. In IEEE International Symposium on Biomedical Imaging: From Nano to Macro (ISBI). pp. 1107-1110.
24. Ruifrok, A. C., & Johnston, D. A. (2001). Quantification of histochemical staining by color deconvolution. *Analytical and quantitative cytology and histology*, **23(4)**, 291-299.
25. Esteva, A. *et al.*, (2017). Dermatologist-level classification of skin cancer with deep neural networks. **Nature**, **542(7639)**, 115.
26. Kermany, D. S. *et al.*, (2018). Identifying medical diagnoses and treatable diseases by image-based deep learning. **Cell**, **172(5)**, 1122-1131.
27. Araújo, T. *et al.*, (2017). Classification of breast cancer histology images using convolutional neural networks. **PloS one**, **12(6)**, e0177544.
28. Russakovsky, O. *et al.*, (2015). Imagenet large scale visual recognition challenge. **International Journal of Computer Vision**, **115(3)**, 211-252.
29. Lin, M. *et al.*, (2013). Network in network. arXiv preprint arXiv:1312.4400.
30. Lu, C. *et al.*, (2017). An oral cavity squamous cell carcinoma quantitative histomorphometric-based image classifier of nuclear morphology can risk stratify patients for disease-specific survival. **Modern Pathology**, 30(12), 1655.
31. Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. arXiv preprint, 1610-02357.
32. Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556
33. Xu, H. *et al.*, (2017). Automatic Nuclear Segmentation Using Multiscale Radial Line Scanning With Dynamic Programming. IEEE Transactions on Biomedical Engineering, 64(10), 2475-2485.
34. Hou, L. *et al.*, (2019). Sparse autoencoder for unsupervised nucleus detection and representation in histopathology images. Pattern recognition, 86, 188-200.