


Optimization of deep learning models for the prediction of gene mutations using unsupervised clustering

Zihan Chen^{1†}, Xingyu Li^{2†}, Miaomiao Yang³, Hong Zhang^{2*} and Xu Steven Xu^{4*} 

¹School of Data Science, University of Science and Technology of China, Hefei, PR China

²Department of Statistics and Finance, School of Management, University of Science and Technology of China, Hefei, PR China

³Clinical Pathology Center, The Fourth Affiliated Hospital of Anhui Medical University, Hefei, PR China

⁴Clinical Pharmacology and Quantitative Science, Genmab Inc., Princeton, NJ, USA

*Correspondence to: Hong Zhang, Department of Statistics and Finance, School of Management, University of Science and Technology of China, Hefei, Anhui 230026, PR China. E-mail: zhangh@ustc.edu.cn and Xu Steven Xu, Clinical Pharmacology and Quantitative Science, Genmab Inc., Princeton, NJ, USA. E-mail: sxu@genmab.com

†These authors contributed equally to this work.

Abstract

Deep learning models are increasingly being used to interpret whole-slide images (WSIs) in digital pathology and to predict genetic mutations. Currently, it is commonly assumed that tumor regions have most of the predictive power. However, it is reasonable to assume that other tissues from the tumor microenvironment may also provide important predictive information. In this paper, we propose an unsupervised clustering-based multiple-instance deep learning model for the prediction of genetic mutations using WSIs of three cancer types obtained from The Cancer Genome Atlas. Our proposed model facilitates the identification of spatial regions related to specific gene mutations and exclusion of patches that lack predictive information through the use of unsupervised clustering. This results in a more accurate prediction of gene mutations when compared with models using all image patches on WSIs and two recently published algorithms for all three different cancer types evaluated in this study. In addition, our study validates the hypothesis that the prediction of gene mutations solely based on tumor regions on WSI slides may not always provide the best performance. Other tissue types in the tumor microenvironment could provide a better prediction ability than tumor tissues alone. These results highlight the heterogeneity in the tumor microenvironment and the importance of identification of predictive image patches in digital pathology prediction tasks.

Keywords: deep learning; whole-slide images; H&E image; gene mutation; unsupervised clustering

Received 9 June 2022; Revised 3 October 2022; Accepted 27 October 2022

Conflict of interest statement: XSX is an employee of Genmab Inc. Genmab did not provide any funding for this study.

Introduction

The diagnosis of cancer is typically based on a histopathological assessment of tissue sections, supplemented by genetic and other molecular tests [1–6]. The identification of molecular biomarkers and gene mutations is becoming increasingly important for the development of novel treatment options. For example, *KRAS* mutations, present in about 30–50% of colorectal cancers (CRCs), are associated with poor prognosis and advanced disease [7–13]. In lung adenocarcinoma (LUAD), *EGFR* has been reported to be mutated in about 20% of patients. As a result, multiple *EGFR* therapies aimed at targeting these mutations have been developed and approved by the Food and Drug

Administration [14,15]. However, due to the long turn-around time, tissue usage, and costs in the current oncology workflows for genetic mutations from tissue samples [16], there is a growing need for the development of cheap, scalable fast alternatives to predict genetic mutations.

Deep learning-based algorithms have been developed to predict gene mutations using pathology images [17–23]. Coudray *et al* [24] proposed a deep convolutional neural network (DeepPATH) to predict gene mutations in LUAD based on whole-slide images (WSIs). Kather *et al* [23] proposed, optimized, and extensively validated a one-stop-shop workflow based on the lightweight neural network, ShuffleNet. They showed that a wide range of genetic mutations,

molecular tumor subtypes, gene expression signatures, and standard pathology biomarkers could be inferred from WSIs.

A large number of image patches (ranging from hundreds to thousands) are available for each WSI, and not all areas within the WSIs are relevant to gene mutations. Therefore, use of all image patches of a WSI to construct a prediction model may lead to suboptimal prediction performance for certain gene mutations as, intuitively, pooling patches with little predictive value with relevant, predictive patches may dilute the predictive ability of relevant patches and reduce prediction performance. It has therefore often been postulated that certain image regions or patches within the WSI (e.g. tumor regions) could carry more predictive value. Commonly, pathologist-annotated tumor regions relevant to the diagnostic task have been used to train predictive models [23,25–31]. Scientists also trained tissue classifiers (tumor and nontumor) to automatically select tumor-like tiles to predict mutated genes in different cancers [18,24,32,33].

In the field of digital pathology, unsupervised clustering has been widely used to reduce the dimensionality of patches to facilitate multiple instance learning (e.g. patches from WSIs can be fitted on a graphics processing unit (GPU) at once) [34]. This method was also used to derive additional cluster-based features, and to identify rare events. Dooley *et al* [35] and Zhu *et al* [36] clustered patches and used the frequency of patches in each cluster as a new feature to predict heart transplant rejection. Similarly, Abbet *et al* [34] proposed a self-supervised learning method that jointly learns from a representation of tissue regions as well as a clustering metric to identify spatial tissue features such as cluster probabilities and cluster transition probabilities.

In addition, unsupervised clustering has been commonly used in image-based deep learning survival analysis. Yao *et al* [37] clustered the patches in each WSI individually into different phenotype clusters. One patch from each cluster was then sampled and was used to predict survival in CRC patients. Sharma *et al* [38] deployed a local cluster-based (clustering patches from a single WSI) sampling approach for identifying children with celiac disease. Zhu *et al* and Yue *et al* used global clustering of patches from all patients to train a survival model based on the information derived for each cluster. The features from the most predictive clusters were then aggregated across the patches from each cluster to predict the outcome. Muhammad *et al* [39] used patch features grouped by global centroids to calculate the local slide-level centroid and concatenated the nearest patches to local

centroids to represent each slide. The model was then trained with survival data. Their approach performed better than other approaches used in the modeling of intrahepatic cholangiocarcinoma.

Although various applications have been developed for unsupervised clustering in digital pathology, very few studies evaluated the use of unsupervised clustering for the identification of image patches linked with genetic mutations. Therefore in this paper, we proposed an unsupervised clustering-based multiple-instance learning method to develop a deep learning model for optimization of the prediction of genetic mutations using the WSIs of three common cancer types obtained from The Cancer Genome Atlas (TCGA).

Materials and methods

Datasets, image preprocessing, and feature extraction

Datasets of WSIs for three tumor types, including CRC, head and neck squamous cell carcinoma (HNSCC), and LUAD, were retrieved from TCGA available on <https://portal.gdc.cancer.gov>. The corresponding TCGA gene mutation data and subtype data were downloaded from the <https://xenabrowser.net/datapages/> website.

Clinically relevant genes for each cancer type reported in [40,41] were selected for analysis (Table 1). We assigned each patch with label 1 or 0, depending on the presence or absence of the mutation in that patient.

The background region with no tissue from each H&E stained WSI was excluded using an adapted Otsu method [42]. This technique involves separating the pixels in each image on the grayscale space into foreground and background. The background was then removed leaving only the tissue. The tissue areas of the image were then tiled into small nonoverlapping patches, each with a dimension of 224×224 pixels. Macenko's method [43] was then used to normalize the color patches synchronously according to a reference pathology image patch (supplementary material, Figure S1). For the LUAD, HNSCC, and CRC, the number of patches extracted from the WSIs ranged from about 100 to 50,000 (average = 12,664), 100 to 30,000 (average = 12,772), and a few hundreds to 30,000 (average = 7,888), respectively.

We employed a fine-tuned Xception model to extract features from the image patches [44], which was pre-trained on the ImageNet datasets and fine-tuned on a CRC dataset [45]. It has been shown that

Table 1. Mutant and wild type numbers for each gene in three cancers

Gene	Total	Mutant	Wild type	Mutation frequency
LUAD				
<i>TP53</i>	434	224	210	0.52
<i>STK11</i>	434	63	371	0.15
<i>KEAP1</i>	434	77	357	0.18
<i>EGFR</i>	434	55	379	0.13
<i>ALK</i>	434	24	410	0.06
<i>KRAS</i>	434	132	302	0.30
HNSCC				
<i>TP53</i>	431	320	111	0.74
<i>CASP8</i>	431	47	384	0.11
<i>NSD1</i>	431	51	380	0.12
<i>HRAS</i>	431	27	404	0.06
<i>PTEN</i>	431	10	421	0.02
<i>DNAH5</i>	431	58	373	0.13
CRC				
<i>TP53</i>	414	260	154	0.63
<i>PIK3CA</i>	414	158	256	0.38
<i>ATM</i>	414	120	294	0.29
<i>MET</i>	414	38	376	0.09
<i>BRAF</i>	414	91	323	0.22
<i>RET</i>	414	27	387	0.07
<i>ERBB2</i>	414	30	384	0.63

the fine-tuned Xception model using pathology images from CRC samples improved the feature extraction and the predictive performance across different cancer types [44]. A feature vector with a dimension of 256 was extracted from each patch. For each patient, an $n \times 256$ feature matrix was obtained whereby n represents the number of patches in the patient of interest.

Best-cluster optimized multiple-instance learning using unsupervised clustering

We randomly selected 100 patients from each of the three tumor types (LUAD, HNSCC, and CRC) in the TCGA dataset. After pooling all patches from the 100 patients of each cancer type, we used K-means clustering to cluster these patches into four groups. A k-NN algorithm was then used to assign cluster labels to the rest of the patches of that cancer type, which were not included in the process of building the clustering model.

To our knowledge, this is the first study that has made use of unsupervised clustering to optimize the prediction of genetic mutations on WSIs. However, we leveraged studies currently available in other areas to select the number of clusters [37,38]. It was found that 4–6 clusters usually provide optimal or close to optimal predictive performance for survival prediction [37] and breast cancer classification [38]. We

performed preliminary analysis using six genes from LUAD and confirmed that a cluster number of 4 may provide satisfactory predictive performance for prediction of gene mutations as well (data not shown). Also, since global clustering (i.e. clustering was done for patches from all patients) was used in this study, implementing large cluster numbers may result in missing data (e.g. missing certain clusters in some patients due to lack of certain type of tissues for those patients). Therefore, a cluster number of 4 was selected to not only ensure close to optimal predictive performance, but also to mitigate the potential missing data issue and to ensure fair comparisons across different clusters.

Semiautomatic annotation was used to classify the patches in each cluster for all three cancers (LUAD, HNSCC, and CRC), while patches from CRC tumor were also annotated using an automatic, supervised tissue classifier (Supplementary methods).

To study the effect of clustering on each cluster, we trained a patch-level multilayer perceptron classifier that used the features of the patches from each cluster as the input to estimate the mutation probability of each patch. The Adam algorithm was used to optimize the cross-entropy. After averaging the predicted probability, we obtained a classifier for each slide level. The algorithm was then tested on WSIs obtained from the TCGA dataset. Figure 1 shows the pipeline method used to develop our model.

Model comparisons

The predictive performance of the best-cluster optimized method was compared with (1) a WSI-based approach without unsupervised clustering, (2) a tumor region based method (CRC only), and (3) other published algorithms that utilized unsupervised clustering as follows.

WSI-based approach

As a benchmark comparison, we also trained the multiple instance learning (MIL) classifier using all patches from patients as input without clustering the patches. All the patches obtained from the WSIs were used to train a patch-level network. The average predicted probability of patches was used to predict the slide-level mutation [24].

Tumor-region-based approach

Patches from tumor areas have often been used to train prediction models for gene mutations [18,24,32,33]. For CRC, we selected tumor tiles using a fine-tuned

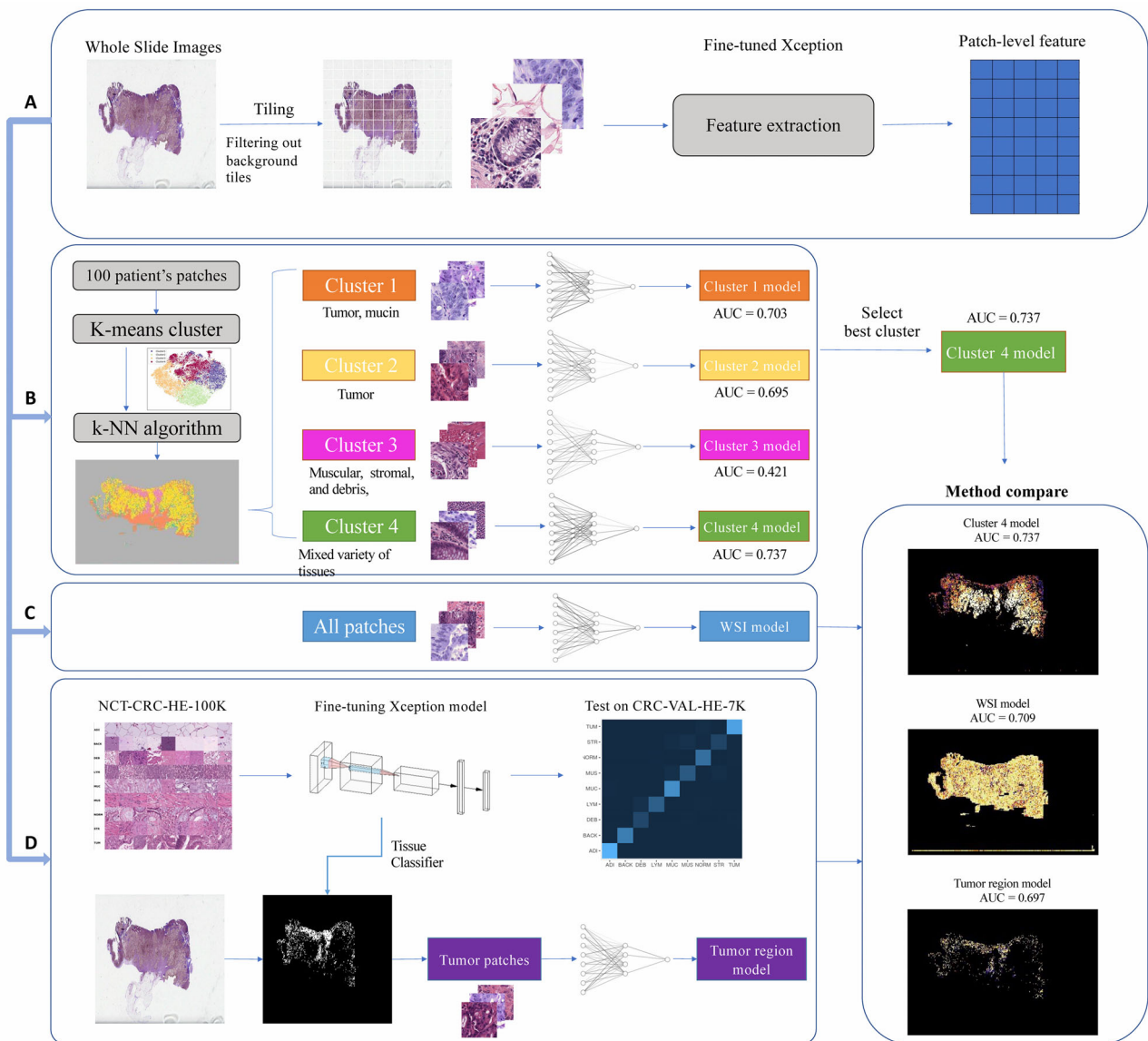


Figure 1. Framework of unsupervised clustering-based deep learning modeling for prediction of gene mutations. (A) Each whole-slide H&E image was preprocessed to (i) remove the background areas using the Otsu method, (ii) split into nonoverlapping tiles with a size of 224×224 pixels, and (iii) color normalized. A fine-tuned Xception model-based feature extractor was used to generate patch representations. (B) For each cancer type, K-means clustering was used to group patches into four clusters. The cluster labels of the patch were then assigned by k-NN algorithm. A neural network was trained on each cluster data and the model with best predictive performance among the four clusters was selected based on five-fold cross-validation (the average AUC values on unseen test fold was reported). (C) For the WSI model, all patches extracted from WSIs were used to train the model. (D) For the tumor region model, we used the NCT-CRC-HE-100K dataset to train a CRC tissue classifier and tested it on the CRC-VAL-HE-7K dataset. For each WSI, the tumor patches were selected by the classifier and were used to train the mutation prediction model. Finally, we compared the best-cluster optimized model, the WSI-based model, and the tumor-region-based model using the average AUC of five-fold cross validation.

Xception-based tissue-type classifier [44]. We then trained the mutation prediction model on the tumor patches and compared the performance of the model with the best-cluster-based model. Since patch-level

labels for tissue types were only available for the CRC dataset, we compared the performance of the best-cluster optimized approach to the tumor-region based approach only for the CRC dataset.

Published baseline algorithms

In addition, we compared our method with two recently published baseline methods by Dooley *et al* [35] and Zhu *et al* [46] that also utilized unsupervised clustering to improve the accuracy of their prediction models. For details, please refer to Supplementary methods.

Experiment setup

In all experiments, we used stratified five-fold cross-validation according to the mutation status for each gene whereby the dataset obtained from the TCGA for each cancer type was divided into five folds to avoid the data imbalance problem. In each fold, 80% of the data were used for model training and 20% of the data were used to test the performance of the model. During training, we used the Adam optimization method with an initial learning rate of 0.00005 and a cosine annealing schedule with a maximum number of 20 iterations. Training and validation were done over 1,000 iterations (supplementary material, Figure S2). The performance of the model was evaluated by calculating the area under the curve (AUC) of a receiver operating characteristic curve. When calculating the cross-entropy loss, we assigned more weight to classes with a small number of training images so that the network was punished more if it falsely predicted the labels of these classes.

Results

Composition of the tissue clusters within the LUAD, HNSCC, and CRC datasets

Supplementary material, Figure S3 shows that K-means clustered the tiles into four distinct clusters for the three TCGA datasets. Supplementary material, Figures S4, S5, and S6A show the image annotations based on the semiautomated annotation approach where two sequential unsupervised clustering procedures were performed to group the image patches into 16 clusters/subclusters and the centers of the 16 subclusters were manually annotated by a pathologist (Supplementary methods). The image tiles in supplementary material, Figures S4, S5, and S6A represent the most common tissue type among the four neighborhood patches near the center of each subcluster (four subclusters for each cluster) for the LUAD, HNSCC, and CRC, respectively.

For the LUAD, based on the pathologist's manual annotation, Cluster 2 mainly consisted of tumor tissues, while Cluster 4 primarily included stromal cells. Clusters 1 and 3 consisted of a mix of red blood cells,

stromal cells, pulmonary alveolus, tumor, lymphocytes, proliferating fibroblasts, and other nontumoral cells (supplementary material, Figure S4).

For the HNSCC cohort, according to the pathologist's annotation, Clusters 3 and 4 consisted of the nontumor and tumor compartments, respectively. Cluster 1 from the HNSCC cohort was a mix of lymphocytes and tumor cells, while Cluster 2 comprised mostly nontumor cells with some tumor cells (supplementary material, Figure S5).

For CRC, supplementary material, Figure S6A shows that based on the pathologist's annotation, Cluster 1 consisted mainly of tumor and mucin cells, while almost all patches in Cluster 2 were tumor cells. Cluster 3 of the CRC primarily included muscular and stromal cells with some debris and tumor cells as well, whereas lymphocytes, adipose, tumor as well as nontumor tissues were identified in Cluster 4.

We also examined the tissue types of each cluster for CRC tumors using the supervised, automatic tissue type classifier [47] (supplementary material, Figure S6B). Similar to the results according to the semiautomated approach and the pathologist's annotation (supplementary material, Figure S6A), the supervised, automatic tissue type classifier (supplementary material, Figures S6B and S7) also predicted that the patches from Cluster 1 were primarily tumor and mucin while the patches from Cluster 2 were dominantly tumor tissue. Similar annotations between the manual- and auto-annotations were also observed for the four patches selected from Cluster 3 and Cluster 4 (supplementary material, Figure S6) except for that the nontumor patches in both clusters (Subcluster 2 in Cluster 3 and Subcluster 4 in Cluster 4) from the semiautomated approach (supplementary material, Figure S6A) were further identified as stromal and normal by the supervised tissue classifier, prospectively (supplementary material, Figure S6B). Furthermore, supplementary material, Figure S7 demonstrated that Cluster 3 of the CRC primarily included muscular and stromal cells with some debris and tumor cells as well, and various tissue types were present in Cluster 4, which included all eight tissue types (plus a small number of background patches). These results indicate that the annotations based on the tissue classifier were generally consistent with the manual annotations provided by the pathologist in the semiautomated approach (supplementary material, Figure S6).

Prediction of gene mutations by tissue clusters and by WSIs

Tables 2–4 illustrate the average AUC values obtained from the five-fold cross-validation using the three

TCGA datasets (LUAD, HNSCC, and CRC, respectively) for the four prediction models based on the image tiles from the four individual clusters. The results demonstrated that the image tiles from the different clusters had different predictive abilities. In addition, the cross-validation also showed that, comparing to the method using all patches from WSIs, the cluster with the best predictive performance consistently provided an improvement in the prediction of genetic mutations for all the three cancer types (i.e. LUAD, HNSCC, and CRC) (Figure 2 for genes that can be robustly predicted [AUC > 0.6] and supplementary material, Figure S8 for all genes regardless of AUC).

LUAD

For the LUAD (Table 2), the tumor cells in Cluster 2 provided the best prediction for the *TP53* and *STK11* mutations, suggesting that the mutant-like image features for *TP53* and *STK11* are mainly found within the tumor region (refer to the heatmap in Figure 3). This finding is consistent with the results obtained by Coudray *et al* [24]. The tumor patches also predicted the *EGFR* mutations well (Table 2). The stromal cells in Cluster 4 provided the highest AUC for the prediction of *ALK* gene mutation (Table 2) and the image tile with the highest likelihood of *ALK* mutation demonstrated stromal features

Table 2. Average AUC (standard deviation) from five-fold cross validation for different clusters in TCGA LUAD. Bold numbers represent the highest AUC values for each gene. Cluster 2 of LUAD mainly consisted of tumor tissues.

Gene	Cluster				Whole image (N = 434)
	Cluster 1 (N = 433)	Cluster 2 (N = 428)	Cluster 3 (N = 434)	Cluster 4 (N = 434)	
<i>TP53</i>	0.655 ± 0.077	0.692 ± 0.082	0.609 ± 0.070	0.584 ± 0.075	0.679 ± 0.084
<i>STK11</i>	0.608 ± 0.095	0.647 ± 0.100	0.553 ± 0.100	0.563 ± 0.157	0.586 ± 0.122
<i>EGFR</i>	0.649 ± 0.126	0.643 ± 0.118	0.584 ± 0.107	0.595 ± 0.123	0.624 ± 0.130
<i>ALK</i>	0.549 ± 0.192	0.609 ± 0.151	0.544 ± 0.123	0.655 ± 0.233	0.604 ± 0.209
<i>KRAS</i>	0.517 ± 0.053	0.536 ± 0.054	0.608 ± 0.068	0.564 ± 0.070	0.562 ± 0.059
<i>KEAP1</i>	0.630 ± 0.137	0.594 ± 0.152	0.619 ± 0.084	0.611 ± 0.081	0.629 ± 0.170

Table 3. Average AUC (standard deviation) from five-fold cross validation for different clusters in TCGA HNSCC. Bold numbers represent the highest AUC values for each gene. Cluster 4 of HNSCC mainly consisted of tumor tissues.

Gene	Cluster				Whole image (N = 431)
	Cluster 1 (N = 431)	Cluster 2 (N = 431)	Cluster 3 (N = 430)	Cluster 4 (N = 430)	
<i>TP53</i>	0.690 ± 0.073	0.611 ± 0.128	0.596 ± 0.068	0.719 ± 0.061	0.700 ± 0.093
<i>DNAH5</i>	0.462 ± 0.090	0.604 ± 0.064	0.505 ± 0.067	0.479 ± 0.088	0.521 ± 0.068
<i>HRAS</i>	0.590 ± 0.152	0.665 ± 0.140	0.454 ± 0.103	0.658 ± 0.178	0.598 ± 0.182
<i>CASP8</i>	0.666 ± 0.124	0.564 ± 0.105	0.638 ± 0.061	0.665 ± 0.072	0.664 ± 0.082
<i>PTEN</i>	0.540 ± 0.286	0.625 ± 0.230	0.577 ± 0.204	0.552 ± 0.225	0.568 ± 0.234
<i>NSD1</i>	0.630 ± 0.100	0.632 ± 0.121	0.657 ± 0.089	0.639 ± 0.070	0.629 ± 0.102

Table 4. Average AUC (standard deviation) from five-fold cross validation for different clusters in TCGA CRC. Bold numbers represent the highest AUC values for each gene. Cluster 2 of CRC mainly consisted of tumor tissues.

Gene	Cluster				Tumor patches*	Whole image (N = 414)
	Cluster 1 (N = 413)	Cluster 2 (N = 411)	Cluster 3 (N = 414)	Cluster 4 (N = 414)		
<i>TP53</i>	0.657 ± 0.034	0.642 ± 0.059	0.575 ± 0.061	0.653 ± 0.028	0.665 ± 0.043	0.660 ± 0.021
<i>PIK3CA</i>	0.759 ± 0.071	0.706 ± 0.083	0.721 ± 0.081	0.766 ± 0.048	0.737 ± 0.085	0.757 ± 0.085
<i>BRAF</i>	0.729 ± 0.043	0.666 ± 0.066	0.625 ± 0.072	0.703 ± 0.078	0.687 ± 0.054	0.695 ± 0.067
<i>ERBB2</i>	0.554 ± 0.145	0.551 ± 0.135	0.546 ± 0.049	0.633 ± 0.167	0.598 ± 0.137	0.567 ± 0.122
<i>ATM</i>	0.738 ± 0.036	0.733 ± 0.056	0.720 ± 0.030	0.743 ± 0.026	0.734 ± 0.054	0.747 ± 0.021
<i>MET</i>	0.703 ± 0.097	0.695 ± 0.062	0.696 ± 0.096	0.737 ± 0.112	0.697 ± 0.052	0.709 ± 0.121
<i>RET</i>	0.685 ± 0.094	0.529 ± 0.111	0.510 ± 0.123	0.677 ± 0.076	0.623 ± 0.048	0.631 ± 0.081

*Model trained on tumor patches identified by a tissue classifier for CRC.



Figure 2. Comparison of model performance (average AUC values) of the proposed best-cluster optimized algorithm with the WSI-based model using all patches from WSIs without patch selection. Red points represent the best-cluster results; green points represent models using WSIs. The bar charts show the difference in average AUC between the best-cluster optimized model and the WSI-based model. The genes that can be robustly predicted (AUC > 0.6) are displayed.

(Figure 3). Models based on image tiles from Clusters 1 and 3 which consisted of a mix of red blood cells, stromal cells, interalveolar septum cells, and other nontumoral cells provided the best prediction for *KEAP1* and *KRAS* mutations, respectively.

Compared to the method using all patches from WSIs, a substantial improvement in the AUC was

observed for the *STK11* (0.061, $p = 0.018$), *ALK* (0.051, $p = 0.016$), and *KRAS* (0.046, $p = 0.055$) mutations. In addition, standard deviations of AUC values for the *TP53*, *STK11*, *EGFR*, and *KEAP1* mutations from models based on the cluster providing the best prediction appear smaller than those based on all the patches on a WSI (Table 2), indicating that the

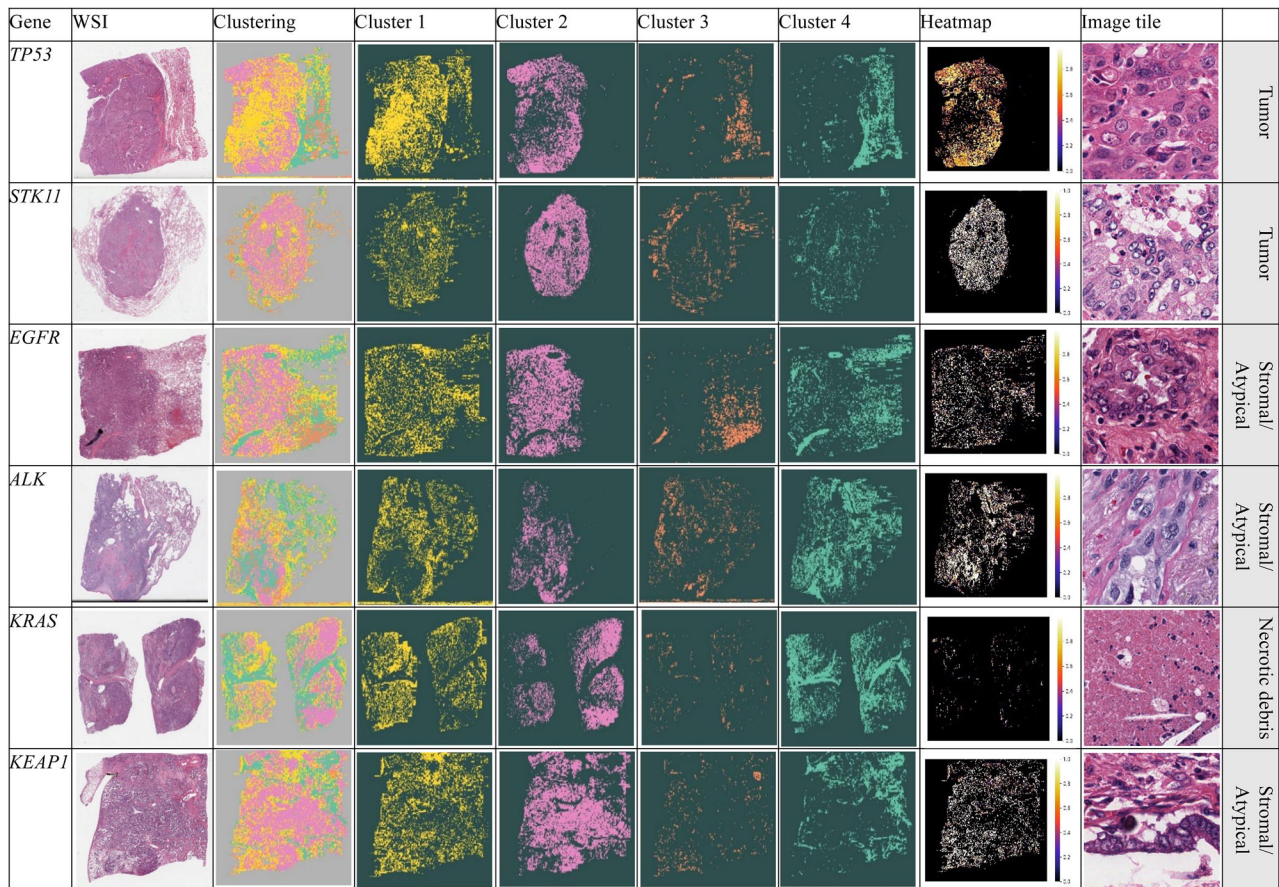


Figure 3. Visualization of the proposed algorithm for different genes in LUAD. The deep learning-based unsupervised clustering and mutation predictions are visualized to understand the spatial locations of each cluster, to identify the spatial regions related to mutation of a specific gene via the resolved probability scores, and to highlight the heterogeneity of a predicted genotype in the tumor microenvironment. The heatmap shows the probability scores of the gene mutations in the identified best cluster. The tile with the highest probability of mutations for each gene is displayed and the corresponding tissue type is provided.

model based on the best cluster may provide a more robust prediction for these genes than the models based on WSIs.

HNSCC

Similarly, in the HNSCC cohort, marked differences in the AUC were noted for the prediction of gene mutations for the different clusters (Table 3). The predictive performance of the tumor cells in Cluster 4 was very high for the *TP53*, *HRAS*, *CASP8*, and *NSD1* mutations. The heatmap in Figure 4 shows that the *TP53* mutant-like features are highly present in the tumor compartment of HNSCC (Cluster 4). Conversely, nontumor cells in Cluster 3 provided the best prediction performance for the *NSD1* mutations, while the mix of lymphocytes and tumor cells in Cluster 1 was better at predicting the *CASP8* mutation. The

image tiles with the highest likelihood of predicting the *NSD1* and *CASP8* mutations were nontumoral (stromal) and tumor cells, respectively (Figure 4). The models based on the nontumor cells (including blood vessels, debris, etc) in Cluster 2 outperformed the other three clusters in terms of the prediction for the *DNAH5*, *HRAS*, and *PTEN* mutations. Consistently, Figure 4 shows that the image tiles with the highest likelihood of *DNAH5*, *HRAS*, and *PTEN* consisted of red blood cells, stroma/red blood cells, and tumor/blood cells, respectively (Figure 4).

When compared to the method using all patches from WSIs (Figure 2), a substantial improvement in the AUC based on the best clusters was noted for the *DNAH5* (0.083, $p = 0.015$), *HRAS* (0.067, $p = 0.06$), and *PTEN* (0.057, $p = 0.031$) mutations, whereas the improvement for *TP53*, *CASP8*, and *NSD1* was

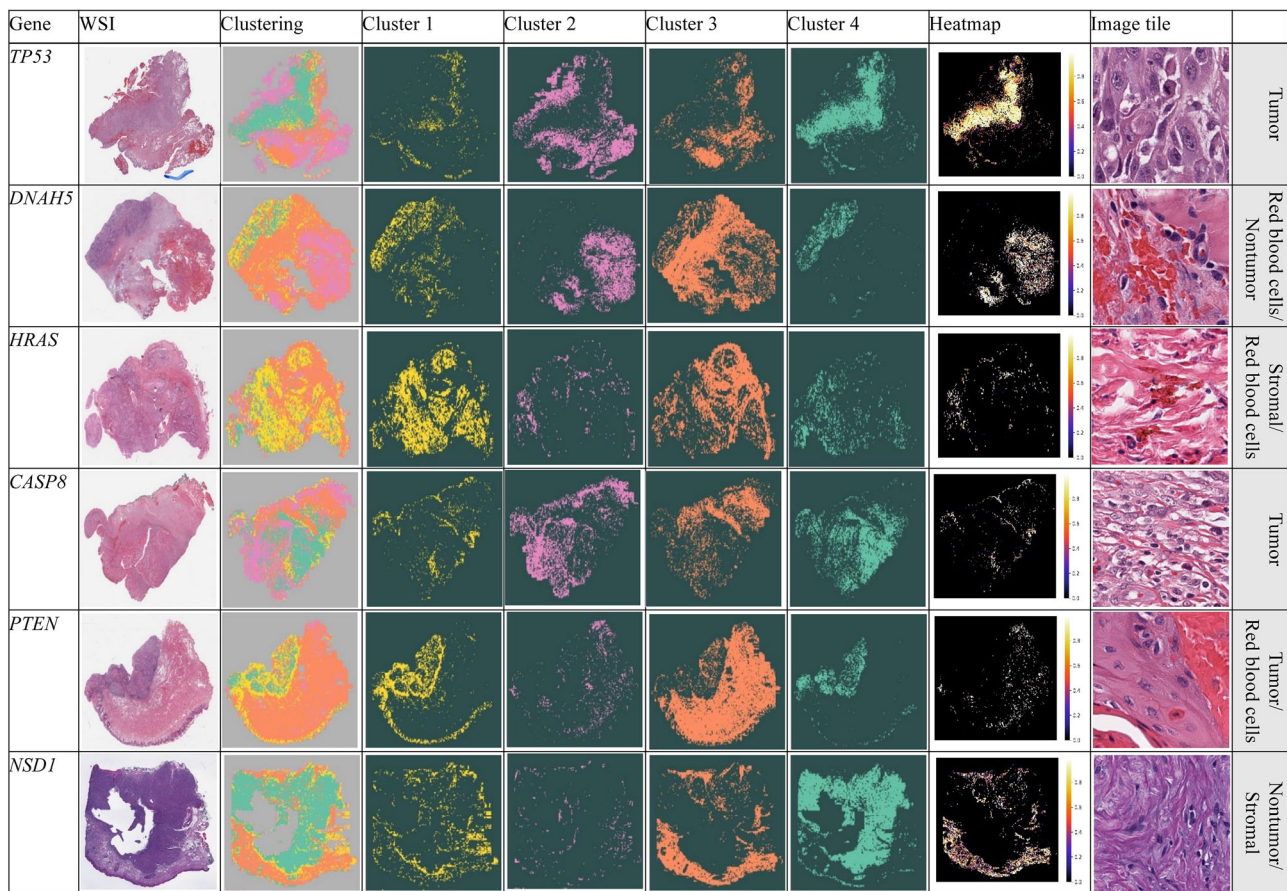


Figure 4. Visualization of the proposed algorithm for different genes in HNSCC. The deep learning-based unsupervised clustering and mutation predictions are visualized to understand the spatial locations of each cluster, to identify the spatial regions related to mutation of a specific gene via the resolved probability scores, and to highlight the heterogeneity of a predicted genotype in the tumor microenvironment. The heatmap shows the probability scores of the gene mutation in the identified best cluster. The tile with the highest probability of mutation for each gene is displayed and the corresponding tissue type is provided.

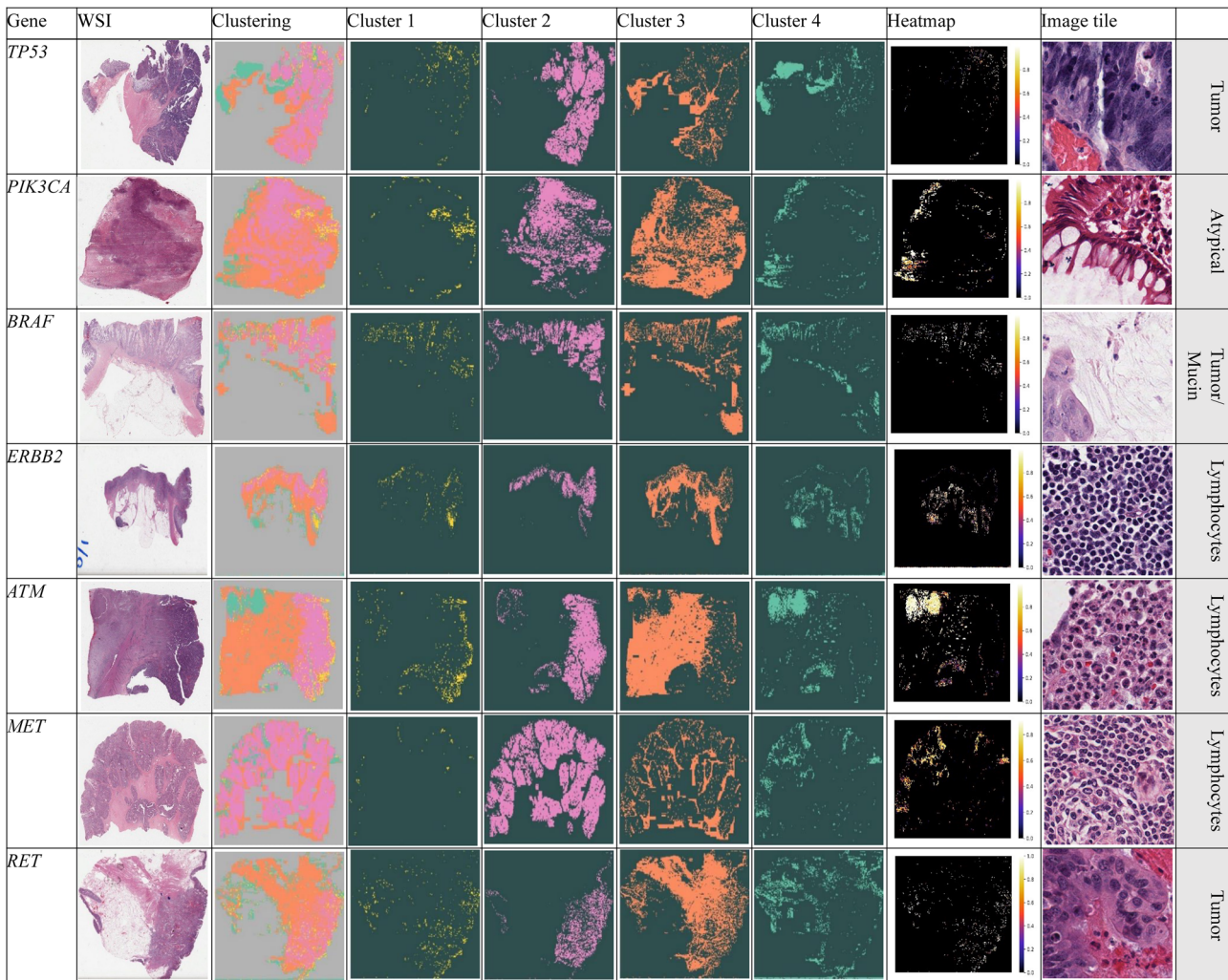
relatively smaller. Again, the standard deviations of the AUC values based on the WSIs were greater than those based on the best clusters (Table 3).

CRC

For CRC (Table 4), features from both Cluster 1 (primarily tumor and mucin tissues) and Cluster 4 (a mixture of all types of tissues) had the best predictive performance for the vast majority of the genes, although the tumor compartment (Cluster 2) also had a relatively good prediction performance. These results suggest that the mutant-like image features for these clinically relevant genes in CRC are not exclusively confined to the tumor regions (Figure 5). Other tissues particularly mucin also have a great predictive value for predicting genetic mutations on CRC WSIs. These findings are consistent with the work of Nguyen *et al*, whereby image patches for tumor and mucus regions

tended to better predict the microsatellite instability (MSI) status and other biomarkers for CRC patients [48]. In addition, Figure 5 demonstrates that the image tiles in Cluster 4 consisting of lymphocytes had the highest predictive ability for the *ERBB2*, *ATM*, and *MET* mutations, while an image tile with normal-tissue-like features had the highest likelihood of predicting *PIK3CA* mutation. For *TP53*, *BRAF*, and *RET* mutations, the highest scoring tiles were tumor or tumor/mucin tiles (Figure 5).

Compared to the WSI approach (Figure 2 and Table 4), the prediction based on the best clusters was remarkably improved for the *ERBB2* (0.066, $p = 0.05$) and *RET* (0.054, $p = 0.021$) mutations in CRC. Numerical improvement was also observed for *MET* and *BRAF* mutations while similar predictive performance was observed for *PIK3CA*, *ATM*, and *TP53* between these two approaches. In addition, for the



Figures 5. Visualization of the proposed algorithm for different genes in CRC. The deep learning-based unsupervised clustering and mutation predictions are visualized to understand the spatial locations of each cluster, to identify the spatial regions related to mutation of a specific gene via the resolved probability scores, and to highlight the heterogeneity of a predicted genotype in the tumor microenvironment. The heatmap shows the probability scores of the gene mutation in the identified best cluster. The tile with the highest probability of mutation for each gene is displayed and the corresponding tissue type is provided.

PIK3CA, *BRAF*, *MET*, and *RET* mutations, more robust prediction (i.e. smaller standard deviation in AUC values from cross-validation) was observed based on the best-cluster approach than the WSI approach. Particularly, for *PIK3CA* mutation prediction, the standard deviation for the best-cluster approach was 0.048 compared to 0.085 for the WSI approach.

These results from LUAD, HNSCC, and CRC suggest that the best-cluster optimized method can provide not only more accurate, but also more precise prediction. For most gene predictions, using all image patches for the WSI approach may result in

reduced predictive performance as patches with low predictive ability will reduce the accuracy of prediction when aggregated with mutation-related patches.

Mutation prediction comparison between the best-cluster optimized model and the tumor area model. Tumor regions are commonly selected in computational pathology pipelines for prediction tasks [18,23,24,28,32]. In the previous experiment, it is apparent that the clusters primarily containing tumor tissues may not provide the best prediction for certain gene mutations. For instance,

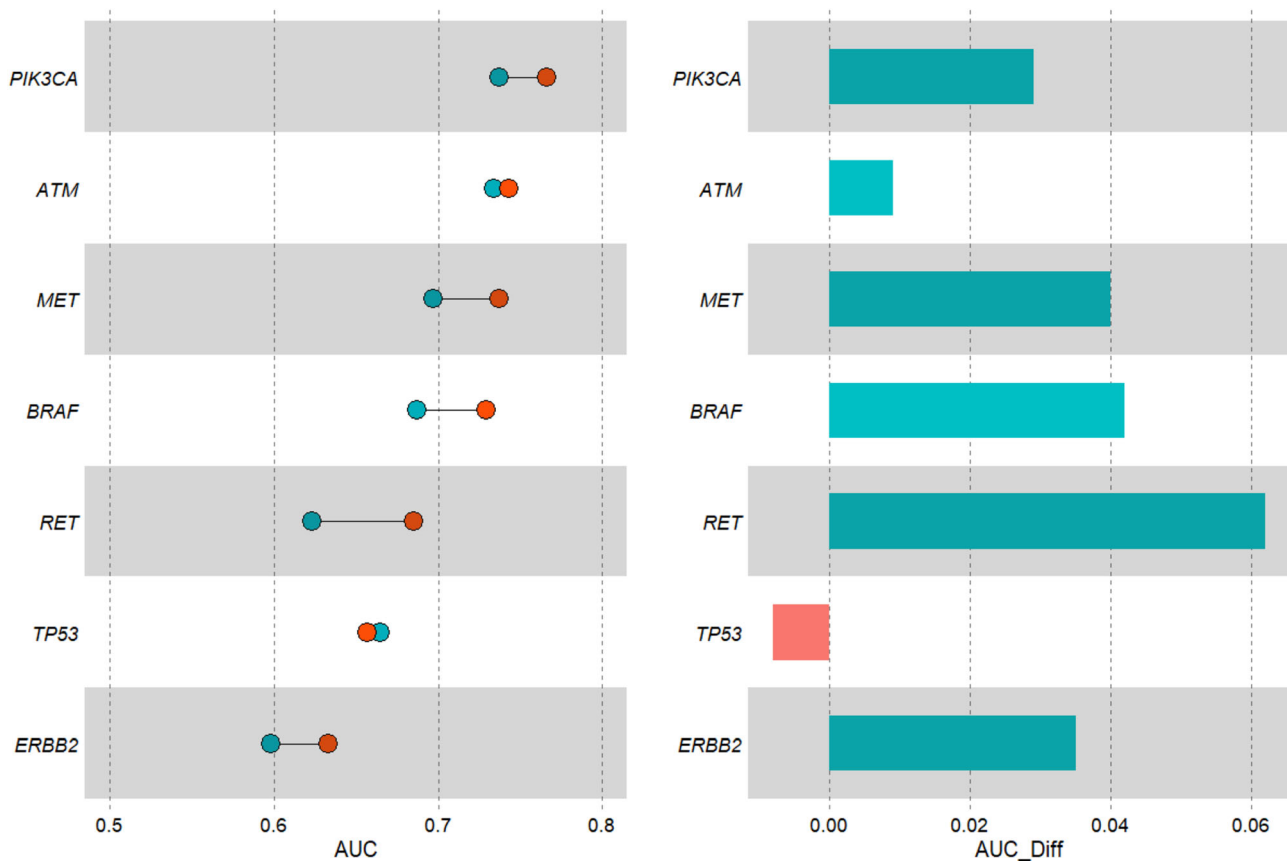


Figure 6. Comparison of the proposed best-cluster optimized model with tumor-region-based for CRC. Red points represent the best-cluster results; green points represent models trained on tumor patches. The bar charts show the difference in average AUC between clustering model and all-patch model.

based on the unsupervised clustering, the cluster with mixed types of tissues (e.g. Cluster 4 [a mix of all tissue types]) better predicted gene mutations for CRC than the cluster primarily with tumor tissue (Cluster 2) (Table 4). In this experiment, we further evaluated the concept and compared the best-cluster optimized approach to tumor-region-based approach using CRC tumors, for which a supervised tissue classifier was developed to select the tumor patches. Figure 6 shows the comparison of average AUC values between the model trained on the best clusters and the model trained using tumor patches selected from the supervised tissue classifier. The best tissue cluster model generally outperformed models trained only on tumor regions, and improvement was observed for *RET*, *BRAF*, *MET*, *PIK3CA*, and *ATM* mutations. Particularly, the predictions of *RET*, *BRAF*, and *MET* mutations were significantly improved by 0.062 ($p = 0.026$), 0.042 ($p = 0.051$), and 0.04 ($p = 0.043$), respectively. For *TP53* mutation, the supervised model-identified tumor patches

provided similar or slightly higher predictive performance compared to the best-cluster approach.

Mutation prediction comparison between the best-cluster optimized model and two baseline algorithms

In this experiment, we compared the best-cluster optimized method with two recent published machine learning methods that utilized unsupervised clustering techniques in different ways (Supplementary methods). The cluster distribution algorithm proposed by Dooley *et al* [35] uses frequency of patches of individual clusters as a new feature for prediction, while Zhu *et al* [46] selected all clusters with predictive values (i.e. AUC > 0.5). Supplementary material, Figure S9 shows that our proposed best-cluster method outperformed both baseline algorithms proposed by Dooley *et al* [35] and Zhu *et al* [46]. The superiority over the Zhu *et al* model suggests that the combining of all clusters with an AUC higher than 0.5 reduces the predictive ability of the

model. Intuitively, involving data from less predictive clusters may reduce the predictive performance. The cluster distribution algorithm of Dooley *et al* achieved an AUC mostly around or lower than 0.5, indicating that the model did not perform well on the gene mutation data, probably because summary statistics of distribution of clusters such as tile frequency of a cluster may be an oversimplified approach for complex prediction tasks for gene mutations, possibly due to substantial loss of information of pathology images like texture features and cell morphology in patches.

Overall, these results suggest that unsupervised clustering can facilitate the identification of patches with better predictive values and exclude patches that lack predictive information. Furthermore, as expected, the introduction of less predictive clusters reduced the performance of the model. As a result, our proposed best-cluster approach outperformed Zhu's method when all the clusters with an AUC > 0.5 were combined and used to construct the prediction model.

Discussion

WSIs are widely used in digital pathology to predict gene mutations, molecular subtypes, and clinical outcomes. Since WSIs are too large (Giga pixels) to fit on a GPU at once, they are usually split into small image patches for training neural networks and prediction models. However, since patch-level labels are usually not available, we cannot directly perform classification on each patch. Therefore, multiple instance learning is often implemented to develop prediction models for patients. It is commonly assumed that tumor regions carry the most predictive information. Therefore, the development of deep learning models for the prediction of genetic mutations on WSIs is usually based solely on tumor tiles. In this paper, we propose an unsupervised clustering method to segment WSIs according to the different morphologic features. Additionally, we also aim to identify the best tissue tiles for the training of deep learning models for the prediction of gene mutations in three different types of cancers.

We demonstrate that the different clusters possessed had different predictive abilities. In addition, the clustering of image patches facilitated the identification of predictive patches and therefore improved the prediction of gene mutations for all three cancer types (LUAD, HNSCC, and CRC from TCGA) when compared with a model trained on all patches obtained from WSIs. These results suggest that unsupervised clustering can facilitate the identification of patches with better predictive values and exclude patches that lacked predictive information. Furthermore, our proposed algorithm outperformed two recently

published baseline algorithms based on leveraging unsupervised clustering. Finally, the unsupervised clustering-based deep learning mutation prediction models made use of resolved probability scoring to facilitate the identification of spatial locations from each cluster that are most likely to be related to specific genetic mutations. This method further highlighted the importance of evaluating the heterogeneity of the tumor microenvironment to predict gene mutations.

Image tiles from tumor regions of a WSI are usually selected for constructing deep learning digital pathology models based on the assumption that tumor cells possess most of the predictive information. Our findings have shown that while this hypothesis may be true for HNSCC in Cluster 4 (where tumor patches best predicted the mutations for HNSCC), for the LUAD cohort, tumor-like image tiles seem to be less predictive of the ALK, KRAS, and KEAP1 mutations (Table 3). Similarly, for the CRC cohort, neither the tumor tiles (Cluster 2) identified by the unsupervised clustering nor the tumor patches identified by the supervised classifier (Table 4) provided a superior prediction performance for the gene mutation status. This suggests that the selection of tumor regions on WSIs is not always the best way to identify patches for the prediction of gene mutations, and other tissue types in the tumor microenvironment may provide a better prediction ability for certain phenotypes than tumor tissues. Previous studies have also shown that the mucin-to-tumor area ratio is highly correlated with the consensus molecular subtypes, MSI status, and the expression of mucin-producing genes [48].

Finally, we also demonstrate that unsupervised clustering could help reduce the workload for pathologist-based manual annotation. We assumed that a limited number of tissue types are present in WSIs, and the repeated clustering of the tiles could separate individual tissue types based on their morphologic appearance. Additionally, through further clustering of each cluster, we selected a small number of tiles near the center of each subcluster (e.g. four tiles). Therefore, the pathologist only had to annotate the selected clusters. We showed that this semi-automatic annotation approach could identify similar tissue types on CRC WSIs to those identified by an automatic tissue classifier for CRC. This technique could be used to improve the interpretability of the unsupervised clustering-based deep learning model.

Acknowledgements

The research of Zihan Chen, Xingyu Li, and Hong Zhang was partially supported by National Natural

Science Foundation of China (No. 12171451 and No. 72091212) and Anhui Center for Applied Mathematics.

Author contributions statement

XSX, XL, ZC and HZ contributed to design of the research. XL, ZC and XSX contributed to data acquisition. XL, ZC, MY and XSX contributed to data analysis. ZC, XL, XSX, MY and HZ contributed to data interpretation. All authors were involved in writing the paper and had final approval of the submitted and published versions.

Data availability statement

The TCGA dataset is publicly available at the TCGA portal (<https://portal.gdc.cancer.gov>). Xception model weights are available at https://github.com/fchollet/deep-learning-models/releases/download/v0.4/xception_weights_tf_dim_ordering_tf_kernels_notop.h5.

References

- Abeshouse A, Ahn J, Akbani R, *et al.* The molecular taxonomy of primary prostate cancer. *Cell* 2015; **163**: 1011–1025.
- Bailey P, Chang DK, Nones K, *et al.* Genomic analyses identify molecular subtypes of pancreatic cancer. *Nature* 2016; **531**: 47–52.
- Dienstmann R, Vermeulen L, Guinney J, *et al.* Consensus molecular subtypes and the evolution of precision medicine in colorectal cancer. *Nat Rev Cancer* 2017; **17**: 79–92.
- Lindeman NI, Cagle PT, Beasley MB, *et al.* Molecular testing guideline for selection of lung cancer patients for EGFR and ALK tyrosine kinase inhibitors: guideline from the College of American Pathologists, International Association for the Study of Lung Cancer, and Association for Molecular Pathology. *J Thorac Oncol* 2013; **8**: 823–859.
- Russnes HG, Lingjærde OC, Børresen-Dale A-L, *et al.* Breast cancer molecular stratification: from intrinsic subtypes to integrative clusters. *Am J Pathol* 2017; **187**: 2152–2162.
- Woodman SE, Lazar AJ, Aldape KD, *et al.* New strategies in melanoma: molecular testing in advanced disease. *Clin Cancer Res* 2012; **18**: 1195–1200.
- Bazan V, Migliavacca M, Zanna I, *et al.* Specific codon 13 K-ras mutations are predictive of clinical outcome in colorectal cancer patients, whereas codon 12 K-ras mutations are associated with mucinous histotype. *Ann Oncol* 2002; **13**: 1438–1446.
- Castagnola P, Giaretti W. Mutant KRAS, chromosomal instability and prognosis in colorectal cancer. *Biochim Biophys Acta* 2005; **1756**: 115–125.
- Liu X, Jakubowski M, Hunt JL. KRAS gene mutation in colorectal cancer is correlated with increased proliferation and spontaneous apoptosis. *Am J Clin Pathol* 2011; **135**: 245–252.
- Nosho K, Irahara N, Shima K, *et al.* Comprehensive biostatistical analysis of CpG island methylator phenotype in colorectal cancer using a large population-based sample. *PLoS One* 2008; **3**: e3698.
- Poehlmann A, Kuester D, Meyer F, *et al.* K-ras mutation detection in colorectal cancer using the pyrosequencing technique. *Pathol Res Pract* 2007; **203**: 489–497.
- Russo A, Bazan V, Agnese V, *et al.* Prognostic and predictive factors in colorectal cancer: Kirsten Ras in CRC (RASCAL) and TP53CRC collaborative studies. *Ann Oncol* 2005; **16**: iv44–iv49.
- Suehiro Y, Wong CW, Chirieac LR, *et al.* Epigenetic–genetic interactions in the APC/WNT, RAS/RAF, and P53 pathways in colorectal carcinoma. *Clin Cancer Res* 2008; **14**: 2560–2569.
- Blumenthal GM, Kluetz PG, Schneider J, *et al.* Oncology drug approvals: evaluating endpoints and evidence in an era of breakthrough therapies. *Oncologist* 2017; **22**: 762–767.
- Pérez-Soler R, Chachoua A, Hammond LA, *et al.* Determinants of tumor response and survival with erlotinib in patients with non-small-cell lung cancer. *J Clin Oncol* 2004; **22**: 3238–3247.
- Rusch M, Nakitandwe J, Shurtleff S, *et al.* Clinical cancer genomic profiling by three-platform sequencing of whole genome, whole exome and transcriptome. *Nat Commun* 2018; **9**: 1–13.
- Srinidhi CL, Ciga O, Martel AL. Deep neural network models for computational histopathology: a survey. *Med Image Anal* 2021; **67**: 101813.
- Qu H, Zhou M, Yan Z, *et al.* Genetic mutation and biological pathway prediction based on whole slide images in breast carcinoma using deep learning. *NPJ Precis Oncol* 2021; **5**: 1–11.
- Ding K, Liu Q, Lee E, *et al.* Feature-enhanced graph networks for genetic mutational prediction using histopathological images in colon cancer. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer: Lima, Peru, 2020; 294–304.
- Liao H, Long Y, Han R, *et al.* Deep learning-based classification and mutation prediction from histopathological images of hepatocellular carcinoma. *Clin Transl Med* 2020; **10**: e102.
- Chen M, Zhang B, Topatana W, *et al.* Classification and mutation prediction based on histopathology H&E images in liver cancer using deep learning. *NPJ Precis Oncol* 2020; **4**: 1–7.
- Wang X, Zou C, Zhang Y, *et al.* Prediction of BRCA gene mutation in breast cancer based on deep learning and histopathology images. *Front Genet* 2021; **12**: 661109.
- Kather JN, Heij LR, Grabsch HI, *et al.* Pan-cancer image-based detection of clinically actionable genetic alterations. *Nat Cancer* 2020; **1**: 789–799.
- Coudray N, Ocampo PS, Sakellaropoulos T, *et al.* Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nat Med* 2018; **24**: 1559–1567.
- Gurcan MN, Boucheron LE, Can A, *et al.* Histopathological image analysis: a review. *IEEE Rev Biomed Eng* 2009; **2**: 147–171.
- Yuan Y, Failmezger H, Rueda OM, *et al.* Quantitative image analysis of cellular heterogeneity in breast tumors complements genomic profiling. *Sci Transl Med* 2012; **4**: 157ra143.
- Zhu X, Yao J, Huang J. Deep convolutional neural network for survival analysis with pathological images. In: *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE: Shenzhen, China, 2016: 544–547.

28. Zhu X, Yao J, Luo X, et al. Lung cancer survival prediction from pathological images and genetic data—an integration study. In: *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*. IEEE: Prague, Czech Republic, 2016; 1173–1176.
29. Cheng J, Mo X, Wang X, et al. Identification of topological features in renal tumor microenvironment associated with patient survival. *Bioinformatics* 2018; **34**: 1024–1030.
30. Sirinukunwattana K, Domingo E, Richman SD, et al. Image-based consensus molecular subtype (imCMS) classification of colorectal cancer using deep learning. *Gut* 2021; **70**: 544–554.
31. Levy J, Haudenschild C, Barwick C, et al. Topological feature extraction and visualization of whole slide images using graph neural networks. In: *BIOCOMPUTING 2021: Proceedings of the Pacific Symposium*. World Scientific: Kohala Coast, Hawaii, 2020; 285–296.
32. Bilal M, Raza SEA, Azam A, et al. Development and validation of a weakly supervised deep learning framework to predict the status of molecular pathways and key mutations in colorectal cancer from routine histology images: a retrospective study. *Lancet Dig Health* 2021; **3**: e763–e772.
33. Jang H-J, Lee A, Kang J, et al. Prediction of genetic alterations from gastric cancer histopathology images using a fully automated deep learning approach. *World J Gastroenterol* 2021; **27**: 7687–7704.
34. Abbet C, Zlobec I, Bozorgtabar B, et al. Divide-and-rule: self-supervised learning for survival analysis in colorectal cancer. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer: Lima, Peru, 2020; 480–489.
35. Dooley AE, Tong L, Deshpande SR, et al. Prediction of heart transplant rejection using histopathological whole-slide imaging. In: *2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*. IEEE: Las Vegas, NV, 2018; 251–254.
36. Zhu Y, Tong L, Deshpande SR, et al. Improved prediction on heart transplant rejection using convolutional autoencoder and multiple instance learning on whole-slide imaging. In: *2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*. IEEE: Chicago, IL, 2019; 1–4.
37. Yao J, Zhu X, Jonnagaddala J, et al. Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks. *Med Image Anal* 2020; **65**: 101789.
38. Sharma Y, Shrivastava A, Ehsan L, et al. Cluster-to-conquer: a framework for end-to-end multi-instance learning for whole slide image classification. *arXiv preprint, arXiv:210310626* 2021 [Not peer reviewed].
39. Muhammad H, Xie C, Sigel CS, et al. EPIC-survival: end-to-end part inferred clustering for survival analysis, featuring prognostic stratification boosting. *arXiv preprint, arXiv:210111085* 2021 [Not peer reviewed].
40. Mosele F, Remon J, Mateo J, et al. Recommendations for the use of next-generation sequencing (NGS) for patients with metastatic cancers: a report from the ESMO Precision Medicine Working Group. *Ann Oncol* 2020; **31**: 1491–1505.
41. The Cancer Genome Atlas Network. Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature* 2015; **517**: 576–582.
42. Otsu N. A threshold selection method from gray level histograms. *IEEE Trans Syst Man Cybern* 1979; **9**: 62–66.
43. Macenko M, Niethammer M, Marron JS, et al. A method for normalizing histology slides for quantitative analysis. In: *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. IEEE: Boston, MA, 2009; 1107–1110.
44. Li X, Cen M, Xu J, et al. Improving feature extraction from histopathological images through a fine-tuning ImageNet model. *J Pathol Inform* 2022; **13**: 100115.
45. Kather JN, Krisam J, Charoentong P, et al. Predicting survival from colorectal cancer histology slides using deep learning: a retrospective multicenter study. *PLoS Med* 2019; **16**: e1002730.
46. Zhu X, Yao J, Zhu F, et al. WSISA: making survival prediction from whole slide histopathological images. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE: Honolulu, HI, 2017; 7234–7242.
47. Li X, Jonnagaddala J, Yang S, et al. A retrospective analysis using deep-learning models for prediction of survival outcome and benefit of adjuvant chemotherapy in stage II/III colorectal cancer. *arXiv preprint, arXiv:211103532* 2021 [Not peer reviewed].
48. Nguyen H-G, Lundström O, Blank A, et al. Image-based assessment of extracellular mucin-to-tumor area predicts consensus molecular subtypes (CMS) in colorectal cancer. *Mod Pathol* 2022; **35**: 240–248.

SUPPLEMENTARY MATERIAL ONLINE

Supplementary methods

Figure S1. Target reference patch of for Macenko's color normalization

Figure S2. Training details

Figure S3. t-SNE visualization of clustering results

Figure S4. Representative image tiles for each cluster/subcluster for LUAD

Figure S5. Representative image tiles for each cluster/subcluster for HNSCC

Figure S6. Representative image tiles for each cluster/subcluster for CRC

Figure S7. Sankey diagram of CRC clustering results

Figure S8. Comparison of model performance (average AUC score) of the proposed best-cluster-based algorithm with the WSI-based models using all patches from whole-slide images without patch selection

Figure S9. Comparison of model performance (average AUC score) of the proposed clustering-based algorithm with the two baseline methods

Figure S10. Framework of the two baseline methods (referred to in Supplementary methods)

Figure S11. Framework of semi-automatic annotation of clusters (referred to in Supplementary methods)

Figure S12. Predictive performance of the supervised CRC tissue-type classifier (referred to in Supplementary methods)