Renal Cell Carcinoma Whole-Slide Image Classification and Search Using Deep Learning

Amir Safarpoor
Sobhan Shafiei
Ricardo Gonzalez
https://orcid.org/0000-0003-3853-2582
Anil Parwani
Hamid Tizhoosh (tizhoosh@uwaterloo.ca)
University of Waterloo https://orcid.org/0000-0001-5488-601X

Article

Keywords: Representation Learning, Digital Pathology, Image Search

Posted Date: October 22nd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-971708/v1

Renal Cell Carcinoma Whole-Slide Image Classification and Search Using Deep Learning

Amir Safarpoor[a], Sobhan Shafiei[a], Ricardo Gonzalez[a], Anil V. Parwani[b], H.R. Tizhoosh[a,*]

[a]*Kimia Lab, University of Waterloo, Ontario, Canada*
[b]*Department of Pathology, The Ohio State University, Columbus, OH, USA*

Abstract

Background: The size of whole slide images (WSIs) in digital pathology can vary from millions to billions of pixels. Accordingly, training state-of-the-art deep learning models with WSIs may not be feasible due to existing memory and computational constraints. In addition, not all WSI pixels may contain useful information about the scanned biopsy sample, e.g., background, debris, and artifact pixels. Furthermore, distilling expressive features from WSIs is a challenging task as there is generally no region-of-interest annotation in real-world archives. Hence, many methods focus on patch processing which can result in improper representation.

Methods: Unlike the patching approach, in this work, we propose a novel framework for learning and the creation of unique, meaningful, and compact features that are critical for indexing kidney cancer WSIs. In this method, a deep convolutional neural network is trained on low magnification patches. Then, slide-level features are extracted from the feature maps at low magnification using the same deep model by considering tissue location and the corresponding filter responses. This procedure enables us to represent a large image using a small set of features.

Results: We used data from the publicly available TCGA dataset to train our model, and it was assessed by both the TCGA and an additional external test cohort of 141 patients from the Ohio State University. We achieved state-of-the-art performance for WSI image search and classification in Renal Cell Carcinoma (RCC) subtypes on both datasets.

Conclusions: Our study depicted that deep neural networks can be used to learn morphological patterns required for accurately representing large whole-slide images. These features can be applied to build WSI search engines to help reducing inter- and intra-observer variability.

Keywords: Representation Learning, Digital Pathology, Image Search

1. Introduction

Digital pathology is the practice of converting a glass slide into a digital image for analysis and review by a pathologist or subject the scanned biopsy samples to image analysis or deep learning networks. Therefore, digitization of the tissue slides allows us to examine the histopathology images by applying the vast range of computer vision algorithms currently available for other related tasks. More importantly, computers can support the diagnostic process in order to reduce the inter- and intra-observer variability among pathologists. Taking into account the advantages of going digital compared to conventional microscopy [1,2], an automatic approach for representing a WSI histopathology image is a fundamental requirement for many tasks including similarity search, detection and classification.

The pathology whole-slide images (WSIs) usually are made up of many different regions that may or may not contain relevant information for any task of interest. Because of their huge scale, these images may incorporate a wide range of patterns ranging from local textures to global structures. Due to present hardware limits, state-of-the-art techniques such as deep convolutional neural networks (CNNs) cannot be easily straightforwardly adapted to WSIs. Extracting the information from these images has several aspects and is understandably quite challenging [1,2].

Given the challenges and complexities of gigapixel image analysis, researchers frequently attempt to tackle simpler versions of the problem by making assumptions. The most popular substitute for processing WSIs is to aggregate the result of processed smaller patches/tiles extracted from the whole image. Faust *et al.* [3] fine-tuned a deep CNN to classify 13 different tissue and lesion classes common in surgical specimens of the central nervous system. Coudray *et al.* [4] trained a deep CNN to differentiate

*Corresponding author
Email address: tizhoosh@uwaterloo.ca (H.R. Tizhoosh)

lung cancer subtypes at 5× and 20× magnification. Riasatian et al.[5] trained a CNN using 1000 × 1000 pixel high cellularity tiles at 20× from 32 cancer subtypes available on TCGA to construct a visual extractor, particularly for general histopathology image representation. Hou *et al.*[6] suggested a patch selection scheme to reduce sample redundancy while working at high magnifications. Bejnordi *et al.*[7] proposed a framework for training a stacked CNNs on both high and low magnification patches. Kalra *et al.*[8] devised an approach, called Yottixel, for representing WSIs using a handful of high magnification tiles sampled from distinctive color regions. Later they applied the Yottixel search engine[8] on the TCGA WSIs and presented their findings in[9].

On the other hand, there are few works on WSI analysis at the gigapixel level, namely representing the WSI in its entirety. Graham *et al.*[10] classified WSIs using features that are extracted from the probability maps of a trained CNN. Kong *et al.*[11] suggested a method based on 2D long short-term memory (LSTM) networks to embed the spatial context into the representation. Lin *et al.*[12] proposed, *ScanNet*, a new neural network architecture that can be trained using larger regions. Wang *et al.*[13] presented a weakly-supervised approach to WSI representation based on *ScanNet*[12]. Tellez *et al.*[14] embedded all patches of a WSI and then learn a representation based on the latter tensor of the WSI patches. Shaban *et al.*[15] proposed a framework, that applied multitask learning, namely classification, and segmentation, on a tensor of spatially ordered patch embeddings, while utilizing attention.

According to the most recent global cancer statistics report, in 2020, there were an estimated 431,288 new cases of kidney cancer and 179,368 deaths globally[16]. The Renal Cell Carcinoma (RCC) is the most common kidney cancer that is responsible for 85% malignant cases[17]. From a single malignant phenotype to a heterogeneous group of tumors, our knowledge about RCC has evolved over time[17]. Among all RCC histologic subtypes, "Clear Cell", "Papillary", and "Chromophobe" make almost 75%, 16%, and 7% of the whole RCC cases, respectively[17]. RCC subtypes differ in their histology, molecular characteristics, clinical outcomes, and therapeutic responsiveness as a result of this heterogeneity. For instance, because the 5-year survival rate differs across different subtypes, proper subtype diagnosis is critical[18].

In this study, we demonstrated a novel approach based on deep learning that can be used for classification and search of RCC subtypes WSIs. We proposed a framework that is capable of learning WSI patterns using CNNs at a low cost at 2.5× magnification. The application of low magnification WSIs has been reported by other researchers, previously. For instance, WSIs at 2.5×[19], 4×[13], and 5×[4,8] are used by other researchers for wide range of machine learning and computer vision applications in pathology. Finally, a fixed-length feature vector is extracted from a WSI using a trained deep model. As a result, not only are we capable of learning task-specific features efficiently but also we can represent a large image with a fixed-length compact representation that is useful for a fast similarity search in image databases or other applications. Unlike classification, search helps us understand how our model reaches a consensus by providing the most similar and relevant cases.

The rest of the paper is organized as follow. In the Section 2, we introduced our dataset and presented our suggested approach for indexing WSIs. We explain our experiments and discuss the corresponding results in Section 3. Finally, we concluded our achievements and presented future research directions in Section 4.

2. Materials and Methods

In this section, first, we introduced the datasets that we used in this study. Then, we presented the fundamental ideas and steps that we used in this work to encode WSIs into a compact fixed-length feature vector.

2.1. Datasets

2.1.1. Ethics Statement

This study was approved by the Ohio State University institutional research board. Informed consent was obtained from all individual patients included in the study. All the data was de-identified using a honest broker system.

2.1.2. TCGA Kidney Cancer Whole Slide Image Dataset

We used kidney WSIs available on The Cancer Genome Atlas (TCGA) repository through the Genomic Data Commons Data Portal to demonstrate the performance of our methodology. In this work, we only used formalin-fixed paraffin-embedded (FFPE) samples. Total number of 883 WSIs from 849 distinct cases were retrieved from TCGA repository. This subset contains 488 (484), 118 (106), and 277 (259) WSIs (cases) from clear cell carcinoma, ICD-O 8310/3, (ccRCC), chromophobe type - renal cell carcinoma, ICD-O 8317/3, (crRCC), and papillary carcinoma, ICD-O 8260/3, (pRCC), respectively. We randomly partition our dataset into a training set (80% of cases), a validation set (10% of cases), and a test set (10% of cases). Because no patients were shared by any of the sets, they were mutually exclusive.

2.1.3. Ohio State University Kidney Cancer Whole Slide Image Dataset

For the *external validation* we used an internal dataset from the Ohio State University. The surgical pathology files of the pathology department were searched for consecutive cases of renal cell carcinoma with the classification of clear cell carcinoma (ccRCC), chromophobe renal cell carcinoma (crRCC) and papillary renal cell carcinoma (pRCC). At the end of the search the dataset was generated which included a total of 141 cases of renal cell carcinomas. These comprised of 48, 44, and 49 WSIs from ccRCC, crRCC, and pRCC, respectively. One representative cancer slide was selected from each case, reviewed by a board certified pathologist (AP) and then scanned at 20× using a aperio XT scanscope (Leica biosystems, CA). The WSI images were reviewed and the classifications were confirmed a second time by a board-certified pathologist (AP) to ensure the quality of the image and accuracy of the diagnosis.

2.2. Method

2.2.1. Stain Normalization

Color differences in sample preparation, raw material use, staining methods, and slide scanners can all affect visual examination and computer-aided image analysis. As a result, several approaches for normalizing image stain have been presented in recent years[20]. In the realm of digital pathology, the influence of stain normalization on image processing methods and CNNs has been investigated[21]. In this study, we deployed a stain normalization technique proposed by Macenko *et al.*[22]. We normalized the color of WSIs at 2.5× magnification to reduce stain variation.

2.2.2. Tissue Localization

The next step in our algorithm is to localize the tissue region. The benefits for localizing tissue is twofold. First, we can extract patches from the informative part of the WSI and avoid background pixels. Second, we use the same masks to confine the deep feature maps extracted from the CNN model to important information. We extract tissue masks from WSIs by applying the following steps:

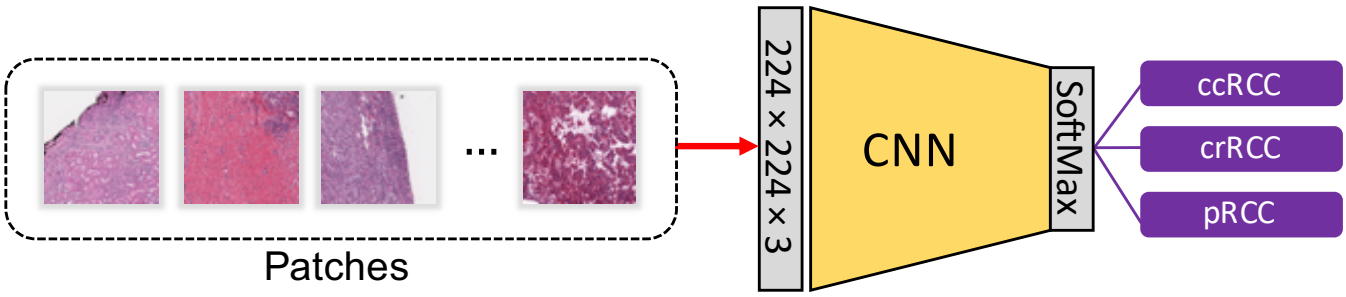
1. First, red, green, and blue pen markers (manual markings on glass slides) were filtered out using a set of fuzzy rules[1].
2. The standard deviation (SD) was calculated within the three color channels. All pixels with SD < 5 were considered as background region.
3. The resulting primary mask is then closed using morphological operation with a 3×3 structuring element. All pixels associated with the background region are set to white (i.e., [255, 255, 255] for a pixel in 24-bit RGB image).
4. By using RGB to HSD (hue, saturation, and density) color space transformation[23], the density channel of each image is evaluated. By applying a global threshold, 0.05, on the mean filtered density channel, the rest of the background pixels are transformed to white pixels.
5. All white pixels are assigned to the background pixels, while the rest of the pixels are considered as tissue pixels.

2.2.3. Representation Learning

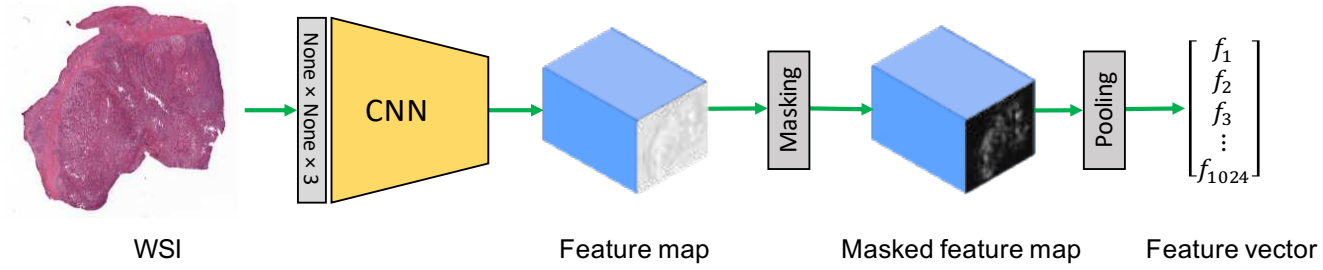
A visual feature extractor is a cornerstone of any image recognition task. It extracts key patterns from an image and quantifies it as a feature vector. With the introduction of deep learning, CNNs outperformed handcrafted features in image recognition tasks by a considerable margin[24]. Therefore, we used the DenseNet-121 model[25], as the visual feature extractor in this work. The DenseNet-121 is frequently utilized in image recognition applications[26,8,27]. DenseNet-121 is made up of many convolutional, pooling, and fully-connected layers, similar to most CNNs. Additionally, this design employs skip connections in such a way that the input for each layer is comprised of feature maps from all previous levels, and its feature maps are utilized as inputs for all subsequent layers[25].

We randomly selected 25%, 50%, 75%, and 100% of the cases from our training set to create four different subsets. For each subset, we trained our model using patches of size 224×224 extracted at 2.5× magnification level. The validation set was used for error estimation throughout the training. We set the parameters of our network to the best parameter set attained in the ImageNet competition[28]. We used the kidney cancer patches from ccRCC, crRCC, and pRCC subtypes in the training set to optimize all weights. To update all network weights, we utilized the backpropagation algorithm guided by the cross-entropy loss function and the Adam optimization technique.

[1] Available at `https://github.com/deroneriksson/python-wsi-preprocessing`



(a)



(b)

Figure 1: The outline of the proposed algorithm for encoding WSIs. (a) shows the training step in our algorithm. In this step, we train a DenseNet-121 [25] using 224×224 patches extracted from 2.5× WSIs in the training subsets to classify different RCC subtypes. Our DenseNet-121 [25] is initialized with ImageNet weights. (b) depicts how we encode a complete WSI at 2.5× using the DenseNet-121 [25] model that we train earlier. First, we pass our 2.5× WSI into the model and calculate the feature maps. Then, we mask the feature maps using the associated tissue mask. As the WSI goes through a series of convolution operations with a stride larger than one, the ultimate size of the feature map is smaller than the actual WSI. As a result, the tissue mask is resized first. Finally, we take an average of the positive values that are inside the tissue region within each feature map to calculate the ultimate feature vector that represents the WSI.

2.2.4. Representing Whole Slide Images

After training a DenseNet-121 [25] model using patches associated with the kidney cancer subtypes, the model learns how to represent salient morphology patterns at 2.5× magnification. Consequently, we can apply our model to extract feature maps from an image of any size. Although the proposed method can facilitate feature extraction from the entire WSI, the feature maps may still contain redundant information (like background pixels in filter responses). To address this issue, we mask the feature maps using tissue masks that we extracted earlier from each WSI. To begin, we downsampled each tissue mask to match the height and width of the associated feature map. Then, we maintain positive values within the tissue region and set the remainder of the values to zero for each feature map. Each feature map now depicts how each area of the WSI has a certain pattern. Finally, to produce a fixed-length feature vector related to the WSI, each feature map is replaced with its average value. In this study, each WSI has a feature length of 1024 since DenseNet-121 has a last convolutional layer with 1024 feature values. The outline of our proposed algorithm is shown in Figure 1.

3. Results

We assessed the performance of our proposed technique in two scenarios, namely classification, and search. In WSI classification and search, we compared our method against state-of-the-art methods [29,8]. To provide a fair comparison, all methods were evaluated on the same test set and their performance was measured using established criteria. We used *weighted F1 score* to compare our technique to [29] in the classification experiment. For each label, the F1 scores are determined, and the average is then weighted by the number of true cases for each label. The *precision@K* is used as a metric for the WSI search experiment. Precision@K is the number of relevant results among K retrieved cases [30].

3.1. Experiments on the TCGA Dataset

To begin, we used models trained with various training sets to compute percision@K on the TCGA test set for the WSI search. First, following Section 2.2.4, all test WSIs are embedded using our model. Next, we computed the distance matrix for each query WSI and returned the k nearest WSIs as our search results to identify the top-k instances for search. The distances between two feature vectors were calculated using the Pearson correlation. To avoid obtaining samples from the same patient in the results, we searched on a *leave-one-patient-out* basis. The outcomes for masking feature maps are also included. Table 1 summarizes these findings. As it is depicted in Table 1 overall precision@K value increased proportional to the size of the

training set. Also, masking feature maps enhanced and stabilized the precision@K outcomes, when we retrieved more cases (from precision@3 to precision@10).

Table 1: The impact of different training dataset sizes and masking feature maps on precision@K using TCGA test set.

Training Set	Methods	Precision									
		@1	@2	@3	@4	@5	@6	@7	@8	@9	@10
25%	NM	0.89	0.84	0.80	0.78	0.77	0.75	0.73	0.73	0.72	0.70
	M	0.86	0.85	0.84	0.83	0.83	0.82	0.81	0.80	0.80	0.80
50%	NM	0.85	0.85	0.83	0.82	0.80	0.78	0.77	0.76	0.76	0.74
	M	0.88	0.88	0.88	0.87	0.87	0.85	0.83	0.82	0.81	0.80
75%	NM	0.86	0.89	0.87	0.85	0.83	0.83	0.81	0.80	0.79	0.79
	M	0.85	0.87	0.89	0.90	0.89	0.89	0.88	0.88	0.88	0.87
100%	NM	0.94	0.94	0.90	0.89	0.87	0.86	0.85	0.84	0.83	0.81
	M	0.90	0.93	0.93	0.91	0.91	0.91	0.90	0.90	0.89	0.88

NM and M are corresponding to "not masked" and "masked" approaches, respectively.

For the second experiment, we explored the performance of our method against[8]. To construct a larger search database for the WSI search, we first added the samples that were left out of the training set to the test set. We made sure that the left-out set and the training set do not share patients. As a result, while training, the model did not use data from any of the slides in the search database. Then, our method was used to embed all samples in the search database. Figure 2 shows the two-dimensional t-SNE embeddings of the samples in the search databases for various models. Our method could represent various RCC subtypes in distinct clusters, as illustrated in t-SNE diagrams in Figure 2. We compared our method to the Yottixel search engine[8] in terms of WSI retrieval over the search datasets. The precision@K diagrams for three RCC subtypes for both methods are shown in Figure 2. The standard deviations are shown in the diagrams by horizontal lines, indicating that Yottixel[8] used a random patch selection strategy in the WSI representation process. The Yottixel results are the outcome of ten independent runs, with the mean and standard deviations of these runs used to create the graphs. In the RCC subtype WSI search, we showed that our technique performs superior. However, we must acknowledge that Yottixel is designed as a universal search engine and has been tested for all TCGA anatomical sites and their subtypes of TCGA data. Finally, we included top three searches for the TCGA test dataset in our qualitative evaluation of our WSI search framework in Figure 3.

For the third experiment, we investigated our framework's performance against CLAM[29] for classification. We reported the weighted F1 score of different approaches on the TCGA test set in Table 2. The first two approaches, namely patch level aggregation (PLA) and patch probabilities aggregation (PPA), are patch-based methods for predicting WSI labels based on patch-level labels. We used our model to estimate probabilities for patches of size 224×224 in the test dataset for patch-based models. The bulk of the patch labels in a WSI are used to compute the WSI label in PLA. In other words, the predicted label for WSI was chosen from the class with the most patches. However, in PPA, we consider the average of all patch probabilities in a WSI and infer the WSI label based on the class with the highest average. More significantly, because our method converts a WSI into a fixed-length feature vector, we could apply various classifiers directly on the WSI features. Thus, we employed different strategies to do this, namely logistic regression (LR) and Gaussian processes (GP). The embedded WSIs in the training set are used to train all classifiers. Also, we provided the F1 score on the test set using the approach described in[29]. In terms of weighted F1 score, all approaches based on our trained model outperform CLAM[29], as shown in Table 2.

Table 2: The weighted F1 score of WSI classification using different approaches considering different training subsets.

Methods	% of training set used (number of slides)			
	25 (176)	50 (354)	75 (531)	100 (710)
PLA	0.84	0.91	0.93	0.95
PPA	0.85	0.90	0.92	0.95
LR	0.84	0.90	0.93	0.94
GP	0.86	0.90	0.94	0.94
CLAM[29]	0.63	0.50	0.69	0.66

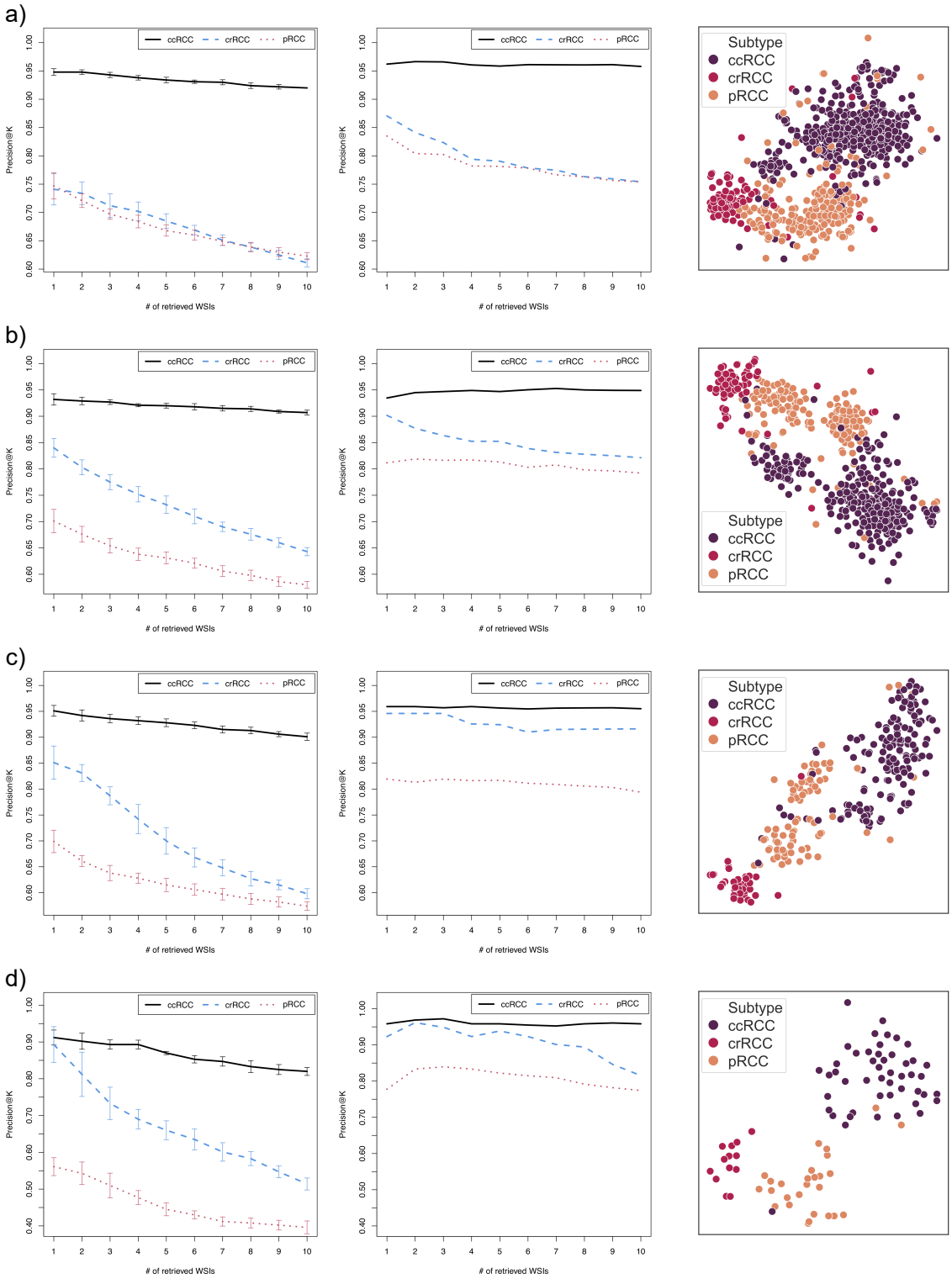


Figure 2: TCGA WSI search results. The precision@K for the Yottixel[8] approach on the left. The precision@K for our approach in the middle. The two-dimensional t-SNE embedding for TCGA WSIs using our method on the right. From top to bottom diagrams are related to the test WSIs alongside the WSIs excluded from the 25% to 100% training subsets.

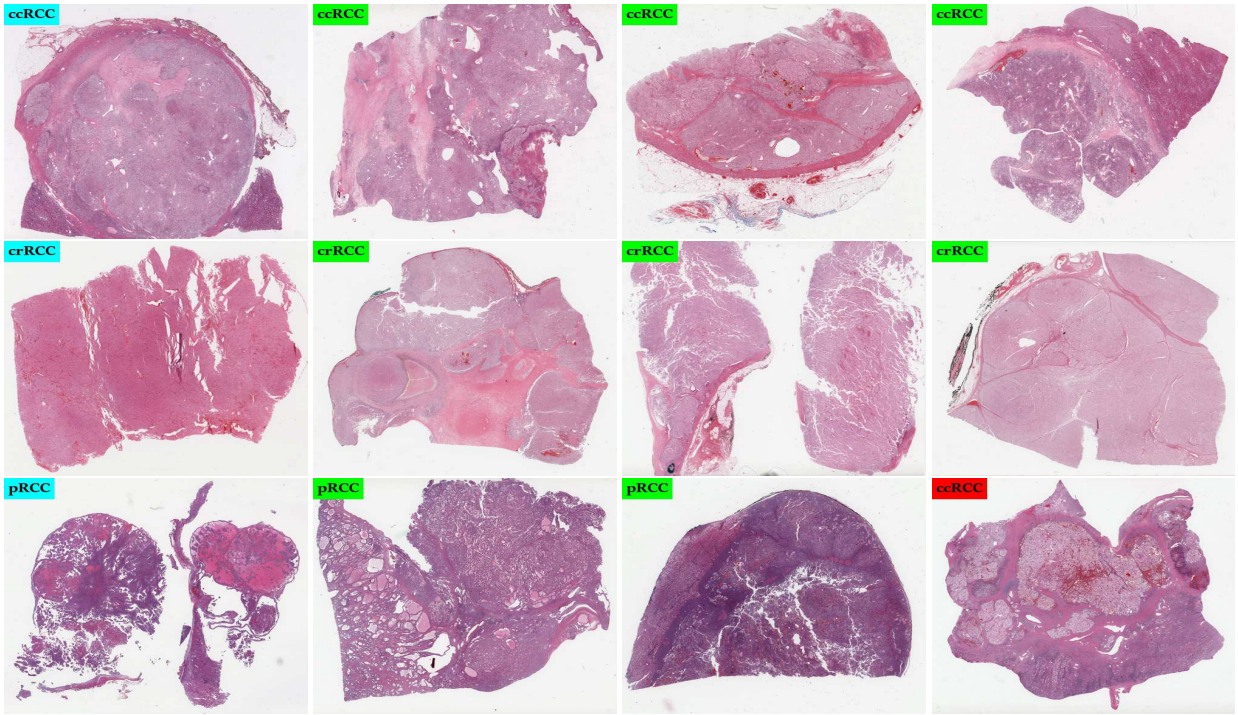


Figure 3: The top three search results for queries related to different RCC subtypes from the TCGA search dataset. The images with a light blue tag are the query WSIs, while the correct and the wrong retrievals are shown in green and red tags, respectively. There are a variety of similar cases that shows the robustness of our method in terms of different colors, shapes, size, and the number of tissue segments. For visualization purposes, all images are resized to a fixed-sized square. So, the size of the WSIs may vary.

We drew the receiver operating characteristic (ROC) curves and provided the micro area under the curve (AUC) alongside their associated confidence intervals for different classification strategies, as shown in Figure 4. We could infer that the GP classifier trained with WSI-level embeddings had the best overall performance, as reported in Table 2 and depicted in Figure 4. The GP classifier's confusion matrices are shown alongside the ROC curves. We can see, according to the confusion matrices, that increasing the amount of training samples enhanced the performance of our algorithm.

3.2. External Validation

External Validation for Search – We used a private dataset from the Ohio State University to better assess the generalizability of our method beyond the TCGA dataset. We analyzed the WSI search performance on 141 WSIs from the Ohio State University dataset. In Table 3, we provide the precision@K for our framework alongside Yottixel[8]. To demonstrate the benefits of the steps we incorporated in our approach, we presented the search results for the DenseNet-121[25] with *ImageNet* weights as well. Our approach exceeds alternative methods by a substantial margin, as demonstrated in Table 3. We included top three searches for the external dataset from the Ohio State University in our qualitative evaluation of our WSI search pipeline in Figure 5.

Table 3: The precision@K for WSI search in Ohio State University dataset.

	Precision									
Methods	@1	@2	@3	@4	@5	@6	@7	@8	@9	@10
Yottixel[8]	0.78± 0.02	0.76± 0.01	0.74± 0.01	0.72± 0.01	0.71± 0.01	0.70±0.01	0.69± 0.01	0.68± 0.01	0.67± 0.01	0.66± 0.01
DenseNet-121	0.76	0.71	0.68	0.65	0.64	0.64	0.62	0.61	0.59	0.59
Proposed method	0.85	0.80	0.78	0.78	0.76	0.75	0.75	0.74	0.74	0.73

External Validation for Classification – For the last experiment, we used the external dataset to test the performance of the classification algorithms mentioned in the preceding sections. The GP classifier (our approach) generated the best AUC among all other strategies, according to the ROC curves in Figure 7. The confusion matrix for the GP classifier, as well as the two-dimensional t-SNE representations of the WSIs, are displayed in Figure 7.

3.3. Interpretability and Convolution Activation Map (CAM) Visualization

By comparing pathologist annotations with patterns that are relevant to our model, we demonstrate the interpretability of our model. First, we selected the most important features by computing the mutual information between the training set WSI

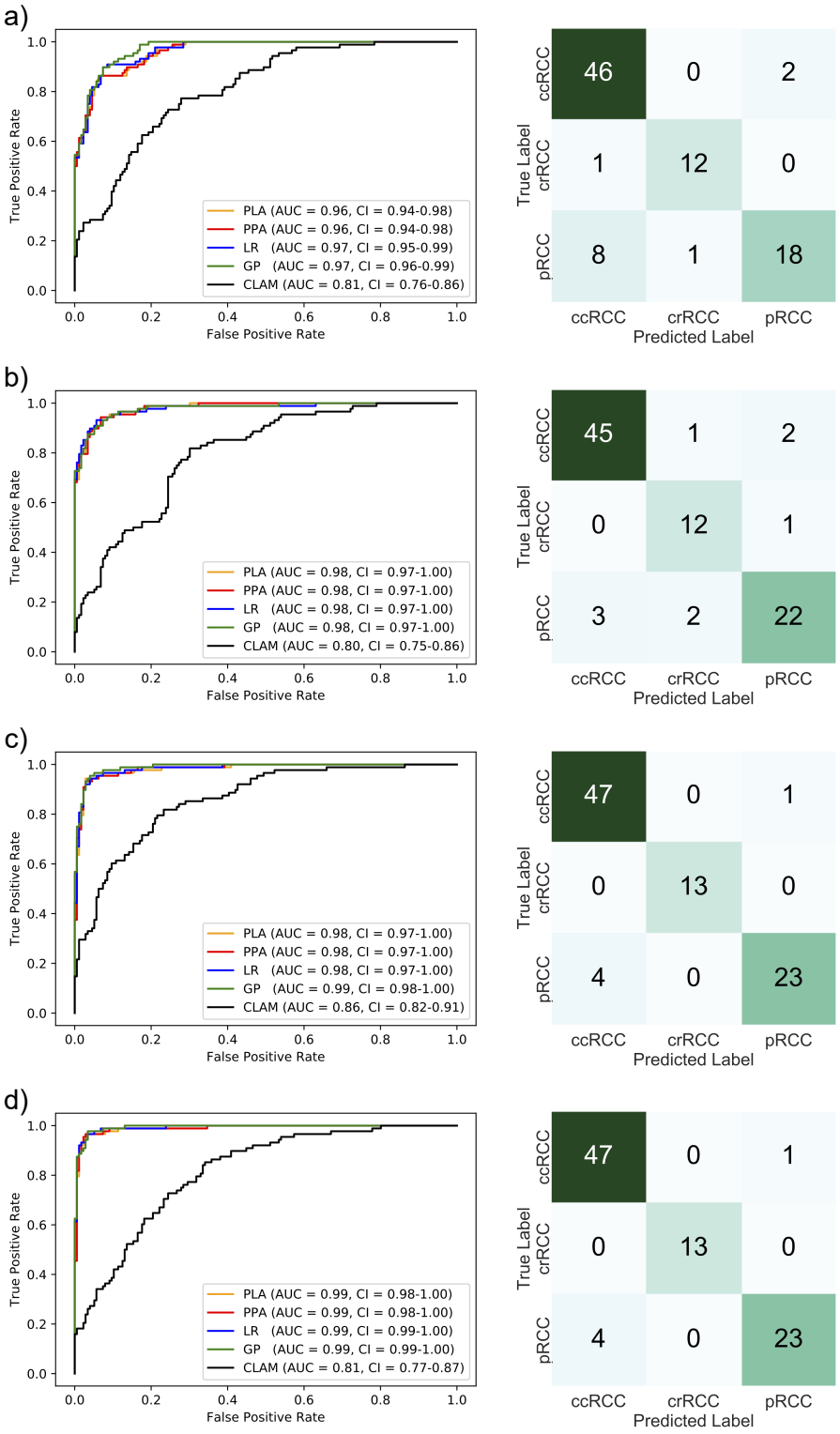


Figure 4: TCGA test classification results. The ROC curves of PLA, PPA, LR, GP, and CLAM [29] on the left. The AUC and the confidence interval values are shown in the legend of the ROC diagrams. The confusion matrices for the GP classifier on the right. From top to bottom models trained on 25%, 50%, 75%, and 100% TCGA training sets.

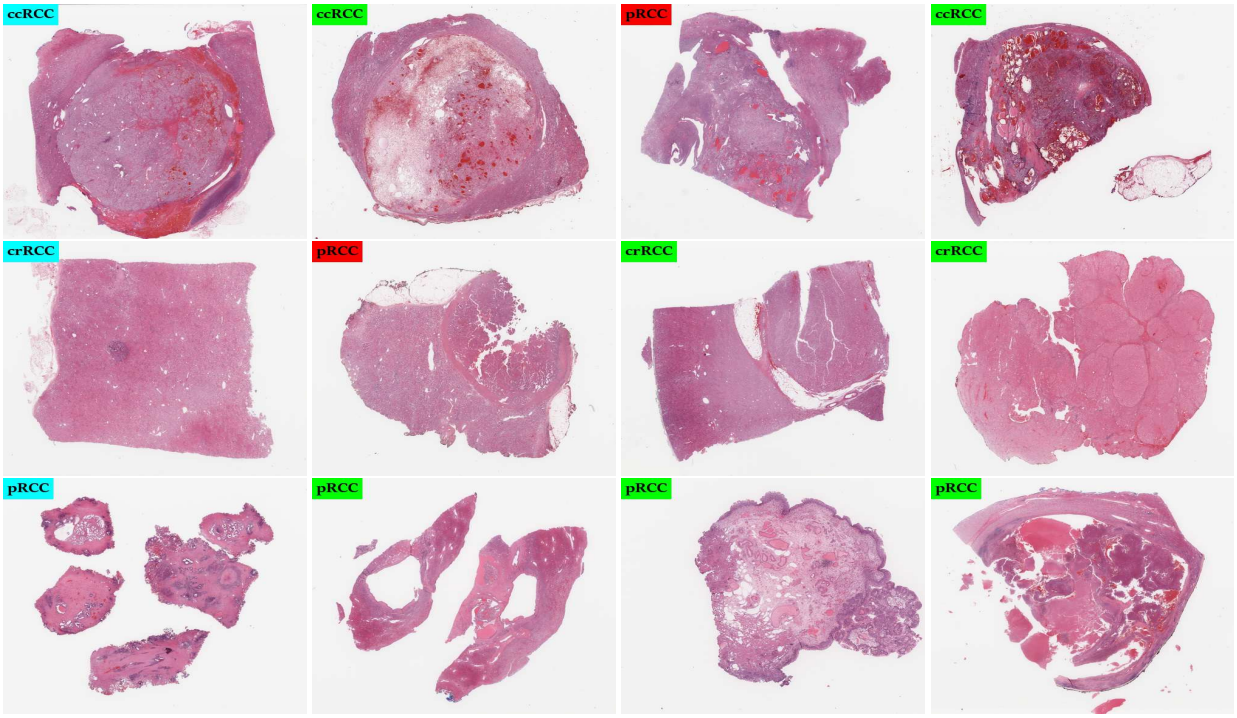
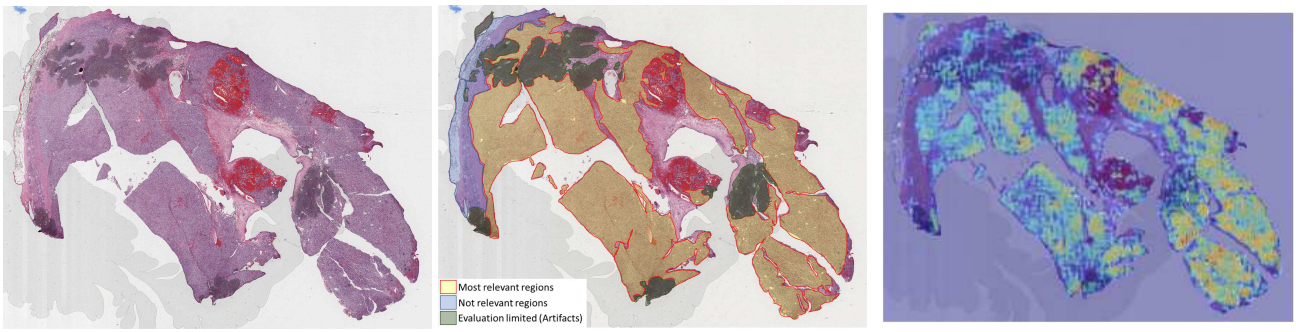


Figure 5: The top three search results for queries related to different RCC subtypes from the Ohio State University dataset. The images with a light blue tag are the query WSIs, while the correct and the wrong retrievals are shown in green and red tags, respectively. There are a variety of similar cases that shows the robustness of our method in terms of different colors, shapes, size, and the number of tissue segments. For visualization purposes, all images are resized to a fixed-sized square. So, the size of the WSIs may vary.



(a) A ccRCC WSI from the TCGA test set.

(b) The annotated WSI by the pathologist.

(c) The CAM for a deep feature.

Figure 6: Interpretability and visualization.

features and class labels. The mutual information values ranged from 0.68 to 0. Then, we selected the deep features that had mutual information greater than 0.5. Out of 1024 features, we were left with a subset of 82 features. Because the number of features was reduced, it was easy to explore each one visually one at a time. Figure 6 depicts the convolution activation map (CAM) corresponding to the deep feature in our model that complies with the pathologist's annotation. There is a high degree of consistency between the fine annotation area and the heatmap. According to Figure 6, the model does in fact focus on significant patterns associated with the diagnosis while excluding irrelevant data and artifacts.

4. Discussion

We present a pipeline for representation learning for kidney WSIs. We demonstrate that by utilizing only slide-level labels and no extra annotations, our method achieves better performance in WSI multi-class classification and search in comparison with state-of-the-art algorithms. We show that our method enhances data efficiency and generalizability over an independent test cohort. Our method learns relevant textural information for RCC subtyping at 2.5× magnification. Choosing a magnification of 2.5× helps to balance the computational cost while still obtaining essential information at both the cell and structural levels. At test time, our technique processes the WSI in a single step, whereas competing approaches require tens of thousands of patches to cover the WSI at 20×[29]. In addition, unlike methods that rely on patch selection for WSI representation, we create a fixed-length vector to characterize each instance. The latter benefit allows us to significantly minimize the number of comparisons at retrieval

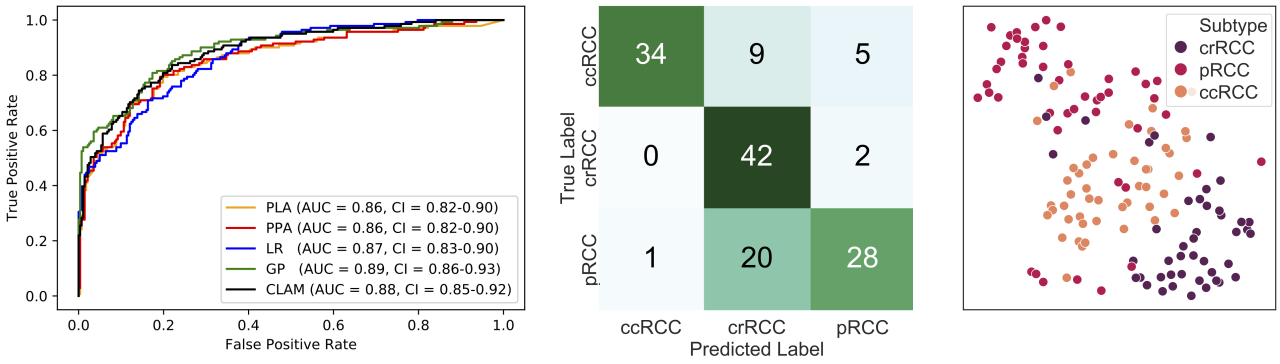


Figure 7: Ohio State University dataset classification results. The ROC curve for PLA, PPA, LR, GP, and CLAM[29] on the left. The area under the curve and the confidence interval values are shown in the legend of the ROC diagram. The confusion matrix for the GP classifier (our approach) in the middle. The two-dimensional t-SNE embedding for the external test WSIs using our approach on the right.

time. It also assists in the removal of variance in the outcomes by dealing with the unpredictability caused by sample selection. Overall, our pipeline provides a more efficient and accurate approach for encoding gigapixel kidney WSIs, which can be utilized to design and train classifiers and search engines. It will be of interest for future works to extend current method to include multi-magnification information, attention mechanism, cross-modal retrieval, and unsupervised representation learning.

Author contributions

A.S. and S.S. designed and performed the research, analyzed and interpreted the results, and wrote the paper. HR.T. and AV.P. have conceived and oversaw the study.

Competing interests

All authors have no conflict of interests and nothing to declare.

Data availability

The NCI Genomic Data Commons Portal (https://portal.gdc.cancer.gov/) has all of the TCGA digital slides available to the public.

Code availability

Upon publishing, our source codes will be made publicly available on our lab's website: "kimia.uwaterloo.ca".

References

[1] Niazi, M. K. K., Parwani, A. V. & Gurcan, M. N. Digital pathology and artificial intelligence. *The Lancet Oncology* **20**, 253–261 (2019).

[2] Tizhoosh, H. R. & Pantanowitz, L. Artificial intelligence and digital pathology: challenges and opportunities. *Journal of Pathology Informatics* **9**, 38 (2018).

[3] Faust, K. *et al.* Visualizing histopathologic deep learning classification and anomaly detection using nonlinear feature space dimensionality reduction. *BMC Bioinformatics* **19**, 173 (2018).

[4] Coudray, N. *et al.* Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nature Medicine* **24**, 1559–1567 (2018).

[5] Riasatian, A. *et al.* Fine-tuning and training of densenet for histopathology image representation using tcga diagnostic slides. *Medical Image Analysis* **70**, 102032 (2021).

[6] Hou, L. *et al.* Patch-based convolutional neural network for whole slide tissue image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2424–2433 (2016).

[7] Bejnordi, B. E. *et al.* Context-aware stacked convolutional neural networks for classification of breast carcinomas in whole-slide histopathology images. *Journal of Medical Imaging* **4**, 044504 (2017).

[8] Kalra, S. *et al.* Yottixel–an image search engine for large archives of histopathology whole slide images. *Medical Image Analysis* **65**, 101757 (2020).

[9] Kalra, S. *et al.* Pan-cancer diagnostic consensus through searching archival histopathology images using artificial intelligence. *npj Digital Medicine* **3**, 31 (2020).

[10] Graham, S. *et al.* Classification of lung cancer histology images using patch-level summary statistics. In *Medical Imaging 2018: Digital Pathology*, vol. 10581, 1058119 (International Society for Optics and Photonics, 2018).

[11] Kong, B., Wang, X., Li, Z., Song, Q. & Zhang, S. Cancer metastasis detection via spatially structured deep network. In *International Conference on Information Processing in Medical Imaging*, 236–248 (Springer, 2017).

[12] Lin, H. *et al.* Scannet: A fast and dense scanning framework for metastastic breast cancer detection from whole-slide image. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 539–546 (IEEE, 2018).

[13] Wang, X. *et al.* Weakly supervised deep learning for whole slide lung cancer image analysis. *IEEE Transactions on Cybernetics* **50(9)**, 3950–3962 (2020).

[14] Tellez, D., Litjens, G., van der Laak, J. & Ciompi, F. Neural image compression for gigapixel histopathology image analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **43**, 567–578 (2021).

[15] Shaban, M. *et al.* Context-aware convolutional neural network for grading of colorectal cancer histology images. *IEEE Transactions on Medical Imaging* **39**, 2395–2405 (2020).

[16] Sung, H. *et al.* Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians* **71**, 209–249 (2021).

[17] Shuch, B. *et al.* Understanding pathologic variants of renal cell carcinoma: distilling therapeutic opportunities from biologic complexity. *European Urology* **67**, 85–97 (2015).

[18] Tabibu, S., Vinod, P. & Jawahar, C. Pan-renal cell carcinoma classification and survival prediction from histopathology images using deep learning. *Scientific Reports* **9**, 10509 (2019).

[19] Qaiser, T. & Rajpoot, N. M. Learning where to see: A novel attention model for automated immunohistochemical scoring. *IEEE Transactions on Medical Imaging* **38**, 2620–2631 (2019).

[20] Tosta, T. A. A., de Faria, P. R., Neves, L. A. & do Nascimento, M. Z. Computational normalization of h&e-stained histological images: Progress, challenges and future potential. *Artificial Intelligence in Medicine* **95**, 118–132 (2019).

[21] Swiderska-Chadaj, Z. *et al.* Impact of rescanning and normalization on convolutional neural network performance in multi-center, whole-slide classification of prostate cancer. *Scientific Reports* **10**, 14398 (2020).

[22] Macenko, M. *et al.* A method for normalizing histology slides for quantitative analysis. In *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, 1107–1110 (IEEE, 2009).

[23] Van der Laak, J. A., Pahlplatz, M. M., Hanselaar, A. G. & de Wilde, P. C. Hue-saturation-density (hsd) model for stain recognition in digital images from transmitted light microscopy. *Cytometry: The Journal of the International Society for Analytical Cytology* **39**, 275–284 (2000).

[24] Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems* **25**, 1097–1105 (2012).

[25] Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4700–4708 (2017).

[26] Rajpurkar, P. *et al.* Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225* (2017).

[27] He, B. *et al.* Integrating spatial gene expression and breast tumour morphology via deep learning. *Nature Biomedical Engineering* **4**, 827–834 (2020).

[28] Deng, J. *et al.* Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255 (IEEE, 2009).

[29] Lu, M. Y. *et al.* Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering* **5**, 555–570 (2021).

[30] Baeza-Yates, R. & Ribeiro-Neto, B. *Modern Information Retrieval: The Concepts and Technology Behind Search* (Addison-Wesley Professional, 2011), 2nd edn.