Published in final edited form as:

Comput Methods Programs Biomed. 2020 October ; 195: 105630. doi:10.1016/j.cmpb.2020.105630.

Interactive Thyroid Whole Slide Image Diagnostic System using Deep Representation

Pingjun Chen^{a,*}, Xiaoshuang Shi^a, Yun Liang^a, Yuan Li^b, Lin Yang^a, Paul D. Gader^c

^aJ. Crayton Pruitt Family Department of Biomedical Engineering, University of Florida

^bDepartment of Pathology, Peking Union Medical College Hospital

^cDepartment of Computer and Information Science and Engineering, University of Florida

Abstract

Background and objectives—The vast size of the histopathology whole slide image poses formidable challenges to its automatic diagnosis. With the goal of computer-aided diagnosis and the insights that suspicious regions are generally easy to identify in thyroid whole slide images (WSIs), we develop an interactive whole slide diagnostic system for thyroid frozen sections based on the suspicious regions preselected by pathologists.

Methods—We propose to generate feature representations for the suspicious regions via extracting and fusing patch features using deep neural networks. We then evaluate region classification and retrieval on four classifiers and three supervised hashing methods based on the feature representations. The code is released at <https://github.com/PingjunChen/ThyroidInteractive>.

Results—We evaluate the proposed system on 345 thyroid frozen sections and achieve 96.1% cross-validated classification accuracy, and retrieval mean average precision (MAP) of 0.972.

Conclusions—With the participation of pathologists, the system possesses the following four notable advantages compared to directly handling whole slide images: 1) Reduced interference of irrelevant regions; 2) Alleviated computation and memory cost. 3) Fine-grained and precise suspicious region retrieval. 4) Cooperative relationship between pathologists and the diagnostic

*Corresponding author: pingjunchen@ufl.edu (Pingjun Chen).

The authors declare no conflicts of interest.

Conflict of Interest and Authorship Conformation Form

Please check the following as appropriate:

- ☐ All authors have participated in (a) conception and design, or analysis and interpretation of the data; (b) drafting the article or revising it critically for important intellectual content; and (c) approval of the final version.
- ☐ This manuscript has not been submitted to, nor is under review at, another journal or other publishing venue.
- ☐ The authors have no affiliation with any organization with a direct or indirect financial interest in the subject matter discussed in the manuscript
- ☐ The following authors have affiliations with organizations with direct or indirect financial interest in the subject matter discussed in the manuscript:

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

system. Additionally, experimental results demonstrate the potential of the proposed system on the practical thyroid frozen section diagnosis.

Keywords

Thyroid frozen section; Whole slide image; Suspicious region; Deep representation; Region retrieval

1. Introduction

Histopathology is taken as the gold-standard for the diagnosis of many cancer diseases [1, 2, 3]. In current clinical practice, most pathologists are still directly examining thyroid glass slides using the microscope to make the diagnosis. However, there are several shortcomings in using glass slides, including costly storage, transportation inconvenience, and limitation on training novices. The evolution of hardware and software for whole slide imaging, also known as digital pathology, can overcome these shortcomings with just an extra slide digitization process on glass slides [4]. A large-scale multi-center evaluation of the whole slide image (WSI) versus microscopy for primary diagnosis in surgical pathology concludes that WSI is non-inferior to microscopy. Moreover, pathologists can inspect the slide on a computer screen, where the image can be spatially navigated, zoomed in, and zoomed out the same manner with standard microscopy [5]. Most importantly, computer-aided diagnosis (CAD) can be introduced into the histopathology diagnosis [6, 7].

Due to the limited computation and storage capability, studies on digital pathology mainly focus on cropped patches from WSI [8, 9], including many cell-level tasks, such as cell detection, segmentation, and retrieval [10, 11, 12, 13, 14, 15, 16, 17, 18, 19]. There are also some studies on patch-based classification [20, 21, 22]. However, from the viewpoint of practical assistance to pathologists, analysis on the WSI would be more beneficial, because their routine work is to inspect WSIs for the diagnosis.

Studies on the WSI become the mainstream in digital pathology recently. Early in 2012, Xu *et al.* propose multiple clustered instance learning (MCIL) to classify, segment, and cluster cancer cells in colon histopathology images [23]. Mousavi *et al.* combine cell segmentation, cell detection, and decision tree to discriminate gliomas of lower and higher grade [24]. Besides the cell-based method, WSI analysis also employs the patch-based method. Vu *et al.* presents a discriminative feature-oriented dictionary learning to learn class-specific dictionaries for patches and diagnose the test sample based on the fusion of patch classification [25]. Hou *et al.* propose a patch-based convolutional neural network (CNN) for WSI classification of glioma and non-small-cell lung carcinoma, in which all patches are predicted using a CNN and then a decision fusion model is to aggregate patch-level predictions [26]. Patch-based methods gradually become the standard framework for WSI classification, which mainly contains two steps: 1) splitting WSI to patches and extract patch's features, and 2) fusion of all patches' feature for the diagnosis. Most recent studies on WSI analysis make use of this framework. Wang *et al.* propose a patch-based CNN to find discriminative regions and a context-aware feature selection scheme for patch feature

aggregation [27]. The attention mechanism is also proposed by Ilse *et al.* for patch feature fusion, in which higher weights are given to those discriminative patches [28].

These studies on WSI analysis achieve some promising results. Nevertheless, most of these studies formulate the diagnosis to be a binary or multiple classification task, which is a drastic simplification of the real scenery. Because of the complex nature, pervasive variability of histopathology, the practical pathology diagnosis usually takes advantage of cognition, clinical context, and practical experience. Sometimes descriptive terminology is also necessary for severe and rare cases [29]. In summary, pathologists should dominate in the diagnosis process, but the artificial intelligence (AI) system can provide supporting knowledge to assist them. Moreover, the AI system needs to consider the time cost of diagnosing a WSI. The dimension of a WSI larger than 50,000 by 50,000 pixels is quite common. When directly handling the WSI, even the loading and pre-processing takes more than 10 secs, and current existing diagnosis systems can hardly finish a diagnosis within one minute. Additionally, WSIs are created through a series of procedures, thereby potentially generating low-quality WSIs. Tissues in WSI can be bent, wrinkled; dust may also contaminate the slides. Other factors, such as blurring and severe color variation, may also affect the AI diagnosis [30]. The AI system cannot easily recognize these situations. Based on the practical diagnosis experience from expert pathologists, the suspicious regions in most thyroid WSI are not hard to identify even under the low-power field in microscopy, and the number of required suspicious regions for careful examination is usually less than five. Fig. 1 shows four thyroid slide examples, in which suspicious regions are located using a contour. However, it is difficult to accurately differentiate these suspicious regions to be benign or malignant in many cases.

Based on these insights, we propose an interactive thyroid WSI diagnostic system to support the diagnosis of pathologists. In the whole slide image viewer, the pathologists first inspect and draw a suspicious region, which can be either rectangle or contour. Based on the drawn region, the AI system starts to analyze and then provide diagnosis information for pathologists to refer to. Meanwhile, the pathologists can make diagnosis irrespective of the AI system. The pathologists give the final diagnosis based on their expertise, as well as the AI provided information.

As for the diagnosis of the drawn region of interest (ROI) in a WSI, the most important challenge is to obtain a discriminative representation for further analysis. Before the thriving of deep learning, traditional feature extraction methods, such as local binary pattern (LBP) [31], scale-invariant feature transform (SIFT) [32], and GIST [33] are widely used to extract image features. The main advantage of these feature extractors lies in that they are unsupervised, and thus there is no need for annotations. With the great success of convolutional neural networks (CNNs) in the ImageNet recognition competition [34, 35], deep learning gradually becomes the mainstream for representation learning because of its superior performance [36, 37, 38]. With the rapid development of CNNs, new classification networks, such as VGG [39], Inception [40], and ResNet [35] are proposed successively and achieve state-of-the-art performance on ImageNet. In this study, we propose to employ the CNN to extract patch features for the representation of the suspicious region in thyroid WSIs.

In addition to classification, searching for similar cases from the database as references can be of tremendous assistance for the diagnosis as well as for training novice pathologists. Content-based image retrieval (CBIR) provides a new approach to mine similar samples from the database to support clinical diagnosis [41, 42]. Recently, Zheng *et al.* introduce an automatic CBIR method by using super-pixels to locate suspicious regions and retrieve similar regions [43]. Two potential issues with their system are real suspicious regions that can be missed by the super-pixel segmentation method, and the super-pixel segmentation process tends to be slow when applying to WSI. Researchers in Google recently present a SMILY system to search similar images for histopathology and demonstrate its ability to retrieve results with similar histologic features, organ site, and prostate cancer Gleason grade [44]. However, SMILY is developed to retrieve similar image patches, which is different from most clinical diagnosis requirements. A multi-magnification correlation hashing (MMCH) framework is proposed to learn discriminative binary codes [45]. They build a patch-link graph with patches from low-mag images and labeled high-mag images, and propagate the semantic similarity from high-mag images to low-mag patches. Both the complexity and time cost of the framework can be an issue. Zheng *et al.* firstly introduce a content-based image retrieval (CBIR) framework for a database consisting of WSIs and size-scalable query ROIs. They only retrieve region proposals with a similar size to the query ROI, and this framework requires a fine-annotated WSI database [46]. The requirement for annotation is too strict, and the size of the retrieval ROI is limited. Afterwards, they present a content-based histopathological image retrieval (CBHIR) framework for regions retrieval based on hierarchical graph neural networks (GNNs) [47]. They allow the pathologist to define the query region using free curves, which can avoid the interference of irrelevant regions. However, their retrieval targets are diagnostically relevant regions, which are not predefined, making the task unnecessarily challenging. Jimenez-del-Toro *et al.* propose a retrieval system for whole-slide histopathology images utilizing both the image and its diagnostic report information [48]. Fusing multi-modality information can boost the discrimination ability of the fused representation. Nevertheless, the report shall not necessarily always exist.

In line with [43, 47], we deem free-curved region retrieval to be the most appropriate manner for the reference of WSI diagnosis. Directly retrieving slides has several disadvantages. First, there would be a vast region of other tissues affecting the critical suspicious region's representation. Second, even a similar slide is retrieved; the user may still do not know which part of the slide to refer to. Third, the computation and memory cost for the generation of slide-level representation is burdensome. Based on the interactive design of our proposed system, pathologists would first preselect the suspicious region from the thyroid WSI, and the retrieval in our system takes advantage of the selected region in WSIs, which would be more fine-grained and precise.

In this paper, we propose an interactive thyroid WSI diagnostic system to support the pathologists in their diagnosis. Based on pre-selected suspicious regions, we construct deep representations based on the CNN feature extractor. With the ROI representation, the system provides classification and retrieved similar regions to pathologists for references. Fig. 2 shows an overview of the proposed interactive thyroid WSI diagnostic system.

2. Material and Methods

In this section, we first describe the thyroid frozen section dataset that is used for the evaluation of the proposed system. Then we introduce how to train the patch-based classification model for patch feature extraction as well as the ROI representation generation. Next, we present the ROI-based diagnosis and retrieval techniques.

2.1. Slides collection and annotation

We collect frozen sections of thyroid from the Peking Union Medical College Hospital (Beijing, China) with a cohort of patients having suspected thyroid nodules admitted between January 2017 and October 2017. The hospital's ethics review board approves the study. Firstly, the specimens dissected from the diseased region are rapidly frozen using cryostat for three minutes. Subsequently, they are sliced as thin as three micrometers with a microtome, and next picked up on a glass slide and stained with hematoxylin and eosin (H&E). All collected slides are de-identified before the analysis. Thyroid frozen sections are categorized as benign, uncertain, or malignant, based on the requirements for developing surgery plans. Table 1 presents the subtypes and surgery plan for each diagnosis. One certified pathologist first annotates those interested regions. Another certified pathologist checks the label for the annotated region, and disagreements are resolved through discussion. Those slides would be excluded for model training and evaluation if two pathologists still have different opinions after the discussion. Pathologists annotate not only malignant regions, but also annotate benign and uncertain regions for the evaluation of ROI-based classification and retrieval. Fig. 3 shows examples of ROIs belonging to these three different categories.

2.2. Patch-based feature exaction

As the large size of the thyroid WSI, the annotated ROI might be larger than $10,000 \times 10,000$ pixels. A large number of diagnostic details would be lost if directly resizing ROI as input for CNN. Besides, the size of the ROI can vary in a vast range. The width and height of a small ROI could be less than 1,000 pixels. Resizing all ROIs to the same size would cause inconsistent image resolutions. Also, the total number of annotated ROI is less than 700 in the current study. These factors can affect classification performance if taking ROI as a whole. Following the widely used analysis manner of the WSI, we adopt the patch-based scheme to generate a representation for the ROI, such that a patch classifier needs to be firstly trained for patch feature extraction.

2.2.1. Patch cropping—After pathologists' annotation on WSIs, we have annotated ROIs with category labels. As the appearance inside the ROI is comparatively consistent, which can be seen in Fig. 3, we randomly crop patches from ROI based on its minimum enclosing rectangle and set the label of the cropped patch to be the same as the ROI. We crop the patch with a size of 256×256 loading from level 2 of WSI, corresponding to a patch image of 1024×1024 in level 0 of WSI. Since we digitize the physical slide using a $40\times$ lens with a resolution of 0.25 $\mu\text{m}/\text{pixel}$, and the downsampling factor of each level is 2.0. The spatial resolution of each pixel cropped from level 2 is 1.0 $\mu\text{m}/\text{pixel}$. Besides, we only keep

those patches with at least 75% of its pixels inside the ROI, and we set the number of cropped patches in an ROI to be proportional to its area.

2.2.2. Patch classifier—A large number of patches of these three different categories are produced after the patch cropping from ROIs. Because of CNN's state-of-the-art performance in image recognition, we employ the CNN model for the patch classification. In order to find the optimal CNN network and to have a comparison on current mainstream CNN networks, we explore vgg16bn [39], inceptionv3 [40], and resnet50 [35] for the thyroid patch classification. Fig. 4 shows the sketch of thyroid patch classification based on ROI annotations.

2.3. ROI representation generation

The CNN model combines representation learning with a fully connected layer to perform the classification task. With the labels as supervision, parameters in the CNN are tuned through multiple epoch training. We take the output of the penultimate layer, namely the layer before the fully connected layer, for thyroid patch representation.

With the trained patch feature extraction model, the ROI representation generation can be divided into three steps: 1) splitting ROI bounding rectangle into multiple patches; 2) extracting features for all valid patches. 3) fusing patches' features to generate a fixed-length descriptor for the ROI. Fig. 5 describes the ROI representation generation pipeline.

In the patch sampling process, we take a self-overlapping patch cropping manner. In this manner, all neighboring patches overlap with nearly the same length with each other, and the overlap length is enforced to be as small as possible, and meanwhile to ensure the sufficient usage of all pixels from top-left to bottom-right of the ROI bounding rectangle. Some patches near the border of the ROI may have a large ratio of pixels outside the ROI. We only keep those patches with at least 75% pixels inside the ROI.

After putting the patches from ROIs into the CNN model, we obtain multiple patch feature vectors. As the size of ROIs is different, the number of feature vectors differs among ROIs. Here in this study, as the ROI is preselected, thus the appearance inside ROI is roughly consistent. There will not have a significant difference among patches from the same ROI. Therefore we propose all patches to have the same weight and use power mean with exponent p of all patch features to generate the representation for an ROI, which is formulated as:

$$R_{rep}^p = \left(\frac{1}{n} \sum_{i=1}^n f_i^p \right)^{\frac{1}{p}}, p \in \mathbb{Z}^+, \quad (1)$$

where R_{rep}^p stands for ROI's representation, n is the number of valid patches in an ROI, f_i is the CNN model extracting features of the i th patch in the ROI. p is the fusion power exponent, which is a positive integer. f_i^p is defined as $[f_{i1}^p, f_{i2}^p, \dots, f_{ik}^p]^T$, where k is the number of elements in f_i . We design this feature fusion formula borrowing the idea of the

norm definition and thus exploring more feature fusion manners with the hyperparameter p holding different values.

2.4. ROI classification

With the patch fused deep representation for the ROI, we explore and compare four widely used classifiers for ROI diagnosis, which include decision tree (DT) [49], support vector machine (SVM)[50], random forest (RF)[51], and multilayer perceptron (MLP). The decision tree constructs a tree-like graph for classification, which is simple to understand and interpret. SVM tries to find a hyperplane that maximizes the margin to different categories. It can efficiently perform a non-linear classification with the kernel trick. Random forest is an ensemble of a multitude of decision trees, aiming to be more robust and generalized. MLP is a vanilla neural network with one or more hidden layers followed by non-linear activation function to separate points in different classes.

2.5. ROI retrieval

Based on the fixed-length representation and label for each ROI, we propose to employ supervised hashing methods to retrieve ROIs which have the same labels and close content with the query ROI [52]. We experiment with three supervised hashing methods, which are supervised hashing with kernels (KSH) [53], supervised discrete hashing (SDH) [54], and column sampling based discrete supervised hashing (COSDISH) [55], to evaluate the region-based WSI retrieval performance.

In supervised hashing with kernels (KSH), the ROI representation is firstly mapped to high dimensional space with kernel functions and then projected into binary codes. The pairwise labels between different ROI representations are adopted to improve the discrimination of the binary codes. Supervised discrete hashing (SDH) formulates hashing objective function by introducing an auxiliary variable and thus can be effectively solved by a regularization algorithm. The optimal binary codes then can be generated by the high-quality discrete regularization algorithm. Column sampling based discrete supervised hashing (COSDISH) directly learns the discrete hashing code from semantic label information. Specifically, it iteratively samples several columns from the semantic similarity matrix and decomposes the hashing code into two parts to be alternately optimized discretely.

3. Results

In this section, we first present the details of the dataset and the implementation settings. Then we introduce the evaluation metrics for classification and retrieval. Next, we present the experimental results and analysis.

3.1. Dataset

In total, we collect and annotate 345 frozen section thyroid slides, including 114 benign, 50 uncertain, and 181 malignant samples. The pathologists annotate at least one region in each slide. They would also annotate benign regions in “Uncertain” and “Malignant” slides. After the region annotation, we obtain 671 ROI regions, which contain 371 benign, 56 uncertain, and 244 malignant ROIs.

To validate the effectiveness of the proposed system, we use k-fold cross-validation to split the collected thyroid slides, in which k is set to 5. To note that, as the number of ROIs in a slide can be varied; thus, the total number of ROIs in training and testing sets of different splits can vary with each other. For each splitting, we further divide training data into Train/Val sets with a ratio of 4:1 for patch classifier training and evaluation.

3.2. Implementation details

As stated in the ‘‘Patch cropping’’, each patch is cropped from level 2 of the slide with a size of 256×256 pixels. We explore vgg16bn, inceptionv3, and resnet50 for patch classification. Among them, the input size of inceptionv3 is 299×299 , and the input size of both vgg16bn and resnet50 is 224×224 . The patches are resized to the corresponding input size of the network and are normalized based on the mean and standard deviation of all training patches. We adopt ImageNet based transfer learning to fine-tune the classification model. Image augmentation techniques, including rotation, flipping, and color jittering, are used to improve the model’s generalization ability. The SGD is used for the optimization of model parameters with a Nesterov momentum of 0.9 and a weight decay of 0.0005. The initial learning rate is set to 0.002, and it decays every two epochs with a ratio of 0.3. We train the network with a batch size of 32 for ten epochs and evaluate the trained model after each epoch on validation patches. We select the model with the best accuracy on validation patches for the following ROI representation generation. We also extract patch features using GIST [33] to compare with the proposed deep feature extractors.

Four classifiers for ROI classification are fitted and evaluated using the scikit-learn package [56]. We set the maximum depth of decision tree (DT) as 2, and choose ‘‘rbf’’ as the kernel for SVM, and the remaining settings for all four classifiers adopt scikit-learn defaults.

3.3. Evaluation metrics

In the evaluation, we assume there are C categories in the histopathology diagnosis, and the overall number of testing ROIs is T_{roi} .

3.3.1. Classification evaluation metrics—For each category $c \in [1, C]$, TP_c denotes the number of ROIs predicted as the c -th category, with the ground truth label c . FP_c denotes the number of ROIs predicted as the c -th category while they do not belong to this category, and FN_c denotes the number of ROIs that their ground truth labels are c while their predictions are wrong. Based on these notations, we use the following four metrics to evaluate ROI classification:

$$\begin{aligned}
 Precision_c &= \frac{TP_c}{TP_c + FP_c}, c \in [1, C] \\
 Recall_c &= \frac{TP_c}{TP_c + FN_c}, c \in [1, C] \\
 F_{1c} &= 2 * \frac{Precision_c * Recall_c}{Precision_c + Recall_c}, c \in [1, C] \\
 Acc &= \frac{1}{T_{roi}} \sum_{i=1}^C TP_c
 \end{aligned} \tag{2}$$

3.3.2. Retrieval evaluation metrics—For each query ROI $q \in [1, T_{roi}]$, assume we retrieve K ROIs from the database. We denote TP_q as the number of retrieved ROIs with the same label as the query ROI q , and FN_q as the number of ROIs in the database with the same label with query ROI q but not in the retrieved results. Based on these notations, we use the mean average precision (MAP), precision, recall, and accuracy for the evaluation of ROI retrieval. MAP is the mean of AP, which is defined as:

$$AP@K = \frac{\sum_{k=1}^K P(k)\delta(k)}{\sum_{k=1}^K \delta(k)} \quad (3)$$

where $P(k)$ is the precision at cut-off k in the list, $\delta(k) = 1$ if the ROI ranked at k -th position is relevant, otherwise, $\delta(k) = 0$. We measure the precision and recall following the usage of the information retrieval domain.

$$\begin{aligned} Precision_K &= \frac{1}{T_{roi}} \sum_{q=1}^{T_{roi}} \frac{TP_q}{K} \\ Recall_K &= \frac{1}{T_{roi}} \sum_{q=1}^{T_{roi}} \frac{TP_q}{TP_q + FN_q}, \end{aligned} \quad (4)$$

We follow the usage in [14] to define the retrieval accuracy, namely the majority vote of the top images retrieved by the system. We select the top nine retrieved images to evaluate the retrieval accuracy.

3.4. ROI classification evaluation

We evaluate the performance of three deep feature extractors (vgg16bn, inceptionv3, and resnet50) as well as the GIST feature extractor on five splits with feature fusion power exponent p in Eq. 1 ranging from 1 to 3. We average the evaluation results of five splits for comparison and analysis. Fig. 6 shows the comparison of overall classification accuracy of using different fusion power exponent. The accuracy of using different fusion power exponent is very close when adopting the same feature extraction method and classifier. The main reason lies in that the annotated ROI has a very consistent appearance. Thus the cropped patches from ROI would have similar feature descriptors. Therefore, the fused ROI representation is less sensitive to the value of fused power exponent. We set the fused power exponent $p = 1$ in the remaining analysis.

Table 2 shows the results of four classifiers using four different feature extraction methods with ROI feature fusion power exponent set to 1. First, comparing the three deep feature extractors with GIST, the classification accuracy of three deep feature extractors on all four classifiers are higher than 0.90. However, the GIST-based ROI feature representation only obtains accuracy around 0.8 when using all these classifiers, thereby validating the superior performance using CNN-based deep representation. Second, comparing the three different deep neural networks, their performance is very close to each other, while inceptionv3 and resnet50 show tiny better performance than vgg16bn. Overall, resnet50 shows the most consistent superiority. Third, when it comes to the ROI classifiers, random forest and MLP achieve higher accuracy than decision tree and SVM, with MLP having the best performance

on all four feature extraction methods. Based on the comparison of feature extraction methods as well as ROI classifiers, resnet50+MLP tends to be the best choice for thyroid frozen section ROI classification. The best classification accuracy (display as bold in Table 2) higher than 0.96 demonstrates the excellence of the ROI-based WSI diagnosis system.

When inspecting the category-wise classification performance, the precision and recall of the “Uncertain” category are lower than “Benign” and “Malignant”. These results could be rooted in the less number of “Uncertain” samples in training as well as “Uncertain” is an intermediate state between “Benign” and “Malignant” and prone to be misclassified. With the combination of resnet50 and MLP, the precision and recall of both “Benign” and “Malignant” are higher than 0.96.

Moreover, we embed the extracted representations of all ROIs using four feature extraction methods into 2D using t-Distributed Stochastic Neighbor Embedding (t-SNE) [57] and show the embeddings in Fig. 7. On all three deep representation methods, the ROIs of different thyroid categories are well-separated, except a few points mixed in other categories, mainly the “Uncertain” category. While GIST cannot isolate different ROI categories, the embedded points are highly mixed. These embeddings validate that three deep feature extractions can achieve better performance, and the “Uncertain” category would have lower performance than the other two categories in the thyroid diagnosis.

3.5. ROI retrieval evaluation

Same with classification evaluation, we also calculate five splits' average value on all evaluation metrics to analyze the retrieval performance. We first compare the performance of KSH on different numbers of binary codes, including 8 bits, 16 bits, and 32 bits, as can be seen in Fig. 8, there is no notable difference in using these three different lengths of bits. Probably because 1) The number of diagnosis category of the thyroid frozen section is three, which can be even represented using only 2 bits; 2) The number of ROIs in the retrieval database is between 500 to 600, 8 bits binary can well separate them. In the following analysis for different retrieval methods, we fix the number of binary codes to 8.

Fig. 9 and Table 3 show the retrieval comparison using three different supervised hashing strategies on four feature extraction methods based on the evaluation metric of MAP. In contrast with GIST, three deep methods have significant advantages. For three feature extraction methods, KSH and SDH obtain close MAP value, COSDISH on resnet50 features achieves lower MAP compared with the rest. As a whole, three different retrieval methods attain a very consistent MAP with retrieved samples ranging from 50 to 500. The results corroborate the outstanding performance of the proposed ROI-based representation generation. Among KSH, SDH, and COSDISH, KSH shows tiny superiority over the other two, and obtain the best MAP, e.g., 0.972.

In Fig. 10, we show precision-recall curves as well as precision and recall curve of KSH based on different numbers of retrieval samples using 8-bit binary codes. The performance on three deep feature extraction methods shows substantial agreement. The precision is higher than 0.9 when the number of retrieved samples is less than 200, which indicates the top retrieved samples having the same label with the query ROI. The performance

dramatically declines when the retrieved number of samples is more substantial than 200, as the number of samples of each category is close to 200. It can also be validated by observing the recall curve in Fig. 10, in which the recall gradually approximates to 1.0 when the number of retrieved samples is larger than 200.

We present the classification accuracy of the retrieval system via a majority vote of the top nine retrieved samples in Table 4. Among these three compared hashing algorithms, KSH obtains superior performance on all the four feature extraction methods over SDH and COSDISH, even using the feature extraction method GIST. As for the number of binary codes, the retrieval method with a higher number of binary codes achieves better accuracy in an overall trend. For those three deep learning-based feature extraction, the performance on these three retrieval methods is very close. The best accuracy of 0.988 is achieved by the retrieval method of KSH with 32 binary codes on the feature extraction method resnet50. The best retrieval accuracy is superior to the best accuracy of 0.961 achieved by all the classification methods, including DT, SVM, RF, and MLP.

In addition to these quantitative results, we also show the top-5 retrieved samples for each category of three different thyroid diagnosis categories in Fig. 11. Comparing with the pixels inside the annotated ROI, the top retrieved ROIs have very similar color and texture with the query ROI, which indicates the retrieval method has found ROIs with similar content from the ROI database.

3.6. Time analysis

The primary time cost in the WSI analysis lies in the representation generation. Assuming that the dimension of a testing WSI is $80,000 \times 80,000$ pixels, the level down-sampling factor is 2 in the WSI pyramid storage structure, and the size of annotated ROI to be one-fourth in both height and width of the slide, the dimension of the ROI region would be $5,000 \times 5,000$ pixels in the level 2 slide image, corresponding to $20,000 \times 20,000$ in level 0 of WSI. The size of extracted patches for the CNN model is 256×256 , the number of patches we obtain for the ROI would be less than $20 \times 20 = 400$ with the self-overlapping patch cropping manner. We run experiments to extract patch features using three deep neural networks, and the time cost is less than 0.04 sec with a batch size of 32 using one Nvidia GeForce GTX 1080 Ti on all of them. Thus the ROI representation generation takes 0.5 sec for a normal-size WSI. With the ROI representation, the time cost of ROI-classification can almost be ignored. As for the ROI retrieval, all annotated ROIs are represented by binary codes in the database, and the retrieval time cost can also be ignored compared with ROI representation generation. Overall, the time cost for analyzing an ROI is within 1 sec, which can well support pathologists' diagnosis without delay. The overall time cost can be further reduced by using multiple GPUs and paralleling computing techniques.

4. Discussion

In this study, we propose an interactive diagnostic system for the thyroid frozen section WSI diagnosis, with the pathologists first drawing ROIs, and then the proposed system providing classification and retrieval information to pathologists for decision making. The main novelties of our approach are: 1) A ROI-based histopathology diagnostic system is proposed

and deemed to fit well to assist pathologists in making the diagnosis. 2) We propose to use deep feature extraction and fusion to generate a discriminative representation for suspicious regions. By performing extensive experimental analysis on the ROI-based classification and retrieval on the thyroid frozen section dataset, the proposed system achieves prominent results, demonstrating its potential to apply in a practical setting to assist pathologists in the future.

The advantages of drawing ROI

In the WSI analysis, the main challenges are dimension obstacle, appearance variability, and bad slide quality. With the ROI-based solution, the dimension of the region to be analyzed can have a certain degree of control. Based on the time analysis, an ROI with $20,000 \times 20,000$ pixels in level 0 of WSI only spends less than 1 sec with GPU for patch feature extraction. As for the appearance variance, the annotated ROI can discard the background, benign, as well as tissue tears/folds to focus on the suspicious region. Thus the representation for the ROI would be more discriminative. Besides, content-based image retrieval is proposed in the diagnostic system, ROIs with close appearance can be retrieved to assist the recognition, which can be very helpful, particularly when diagnosing rare samples. When it comes to the low-quality slides, such as bent, wrinkled, blurring, color variation, *etc.*, pathologists can decide whether to select the slide for analysis or request a new one for diagnosis.

Assistant to pathologists

In this framework, pathologists can draw the suspicious region in the thyroid slide and exploit the computed knowledge provided by our system. Although we can design detection or segmentation methods to locate suspicious regions as Zheng *et al.*'s study does [43]. While current methods need at least three minutes for a slide with typical dimensions, moreover, malignant regions can be missed, or false-positive regions can weaken the discriminative ability of slide representation. Therefore, it would be better to introduce pathologists into the system to select suspicious regions and make the final diagnosis based on their expertise in pathology, computer-aided information, as well as extra patient information.

Instead of competing with pathologists, they are vital to computer analysis as well as decision-making. The system introduces ROI classification and retrieval functionality into the diagnosis pipeline to assist them, thereby being a win-win solution for the AI system and pathologists. With the support of the AI system, pathologists can increase the robustness and reliability of their diagnosis. As the pathologists make the final decision, the responsibility of AI misdiagnosis would be largely reduced, thus would propel the deployment of AI systems in practical applications. The pathologists can even make their own decision without drawing ROI and referring to AI diagnosis when they are highly confident about their diagnosis. They can use AI assistance only when they have doubts about the slide.

Boosting slide digitization

In traditional glass slide diagnosis, pathologists receive the slide, make the diagnosis, write the report, and then the slide would be archived. The slides and the diagnosis can hardly be

reused and referred. With our proposed diagnostic system, the annotated contours and their diagnosis can be collected in the diagnosis process and used for further training, retrieval, *etc.*, which demonstrates the advantages of digital pathology. With pathologists' annotations and diagnosis, the system can be consistently upgraded along with its ongoing usage, and thus it is a knowledge accumulation process. With a large number of annotations, the system would become more robust, accurate, and trustworthy. These advantages can promote the popularization of digital pathology.

Further extensions

With the generated representation for the ROI, other functionalities, such as survival analysis [58] and report generation [37], can be added to this system if corresponding survival and report information are available. The current analysis is based on thyroid frozen sections. However, the system can be easily extended to other types of histopathology tissues without modification. Additionally, we adopt pathologists' selected ROIs. This is mainly caused by the slow speed of automated suspicious region detection and the unsatisfying performance. Fast and accurate methods for suspicious region detection and segmentation need further exploration.

Acknowledgment

This study is supported by the US National Institutes of Health (R01 AR065479–02). We thank Dingrong Zhong, Yong Jiang, Shuhao Wu, and Zhiyuan Li for thyroid slide collection, annotation, and management.

References

- [1]. Sardanelli F, Giuseppetti GM, Panizza P, Bazzocchi M, Fausto A, Simonetti G, Lattanzio V, Del Maschio A, Sensitivity of mri versus mammography for detecting foci of multifocal, multicentric breast cancer in fatty and dense breasts using the whole-breast pathologic examination as a gold standard, *American Journal of Roentgenology* 183 (4) (2004) 1149–1157. [PubMed: 15385322]
- [2]. Jeelani S, Ahmed QM, Lanker AM, Hassan I, Jeelani N, Fazili T, Histopathological examination of nail clippings using pas staining (hpepas): gold standard in diagnosis of onychomycosis, *Mycoses* 58 (1) (2015) 27–32. [PubMed: 25346218]
- [3]. Macedo AA, Pessotti HC, Almansa LF, Felipe JC, Kimura ET, Morphometric information to reduce the semantic gap in the characterization of microscopic images of thyroid nodules, *Computer methods and programs in biomedicine* 130 (2016) 162–174. [PubMed: 27208531]
- [4]. Williams BJ, Bottoms D, Treanor D, Future-proofing pathology: the case for clinical adoption of digital pathology, *Journal of clinical pathology* 70 (12) (2017) 1010–1018. [PubMed: 28780514]
- [5]. Mukhopadhyay S, Feldman MD, Abels E, Ashfaq R, Beltaifa S, Cacciabeve NG, Cathro HP, Cheng L, Cooper K, Dickey GE, et al., Whole slide imaging versus microscopy for primary diagnosis in surgical pathology: A multicenter blinded randomized noninferiority study of 1992 cases (pivotal study), *The American journal of surgical pathology* 42 (1) (2018) 39. [PubMed: 28961557]
- [6]. Madabhushi A, Digital pathology image analysis: opportunities and challenges, *Imaging in medicine* 1 (1) (2009) 7. [PubMed: 30147749]
- [7]. Hipp J, Flotte T, Monaco J, Cheng J, Madabhushi A, Yagi Y, Rodriguez-Canales J, Emmert-Buck M, Dugan MC, Hewitt S, et al., Computer aided diagnostic tools aim to empower rather than replace pathologists: Lessons learned from computational chess, *Journal of pathology informatics* 2.
- [8]. Veta M, Pluim JP, Van Diest PJ, Viergever MA, Breast cancer histopathology image analysis: A review, *IEEE Transactions on Biomedical Engineering* 61 (5) (2014) 1400–1411. [PubMed: 24759275]

- [9]. Xing F, Yang L, Robust nucleus/cell detection and segmentation in digital pathology and microscopy images: a comprehensive review, *IEEE reviews in biomedical engineering* 9 (2016) 234–263. [PubMed: 26742143]
- [10]. Gabriel E, Venkatesan V, Shah S, Towards high performance cell segmentation in multispectral fine needle aspiration cytology of thyroid lesions, *Computer methods and programs in biomedicine* 98 (3) (2010) 231–240. [PubMed: 19720425]
- [11]. Cire DC, san, Giusti A, Gambardella LM, Schmidhuber J, Mitosis detection in breast cancer histology images with deep neural networks, in: *International Conference on Medical Image Computing and Computer-assisted Intervention*, 2013, pp. 411–418.
- [12]. Meijering E, Cell segmentation: 50 years down the road [life sciences], *IEEE Signal Processing Magazine* 29 (5) (2012) 140–145.
- [13]. Qi X, Xing F, Foran DJ, Yang L, Robust segmentation of overlapping cells in histopathology specimens using parallel seed detection and repulsive level set, *IEEE Transactions on Biomedical Engineering* 59 (3) (2011) 754–765. [PubMed: 22167559]
- [14]. Zhang X, Liu W, Dundar M, Badve S, Zhang S, Towards large-scale histopathological image analysis: Hashing-based image retrieval, *IEEE Transactions on Medical Imaging* 34 (2) (2014) 496–506. [PubMed: 25314696]
- [15]. Xie Y, Xing F, Kong X, Su H, Yang L, Beyond classification: structured regression for robust cell detection using convolutional neural network, in: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2015, pp. 358–365.
- [16]. Shi X, Xing F, Xu K, Xie Y, Su H, Yang L, Supervised graph hashing for histopathology image retrieval and classification, *Medical image analysis* 42 (2017) 117–128. [PubMed: 28783503]
- [17]. Shi X, Sapkota M, Xing F, Liu F, Cui L, Yang L, Pairwise based deep ranking hashing for histopathology image classification and retrieval, *Pattern Recognition* 81 (2018) 14–22.
- [18]. Xie Y, Xing F, Shi X, Kong X, Su H, Yang L, Efficient and robust cell detection: A structured regression approach, *Medical image analysis* 44 (2018) 245–254. [PubMed: 28797548]
- [19]. Xing F, Xie Y, Shi X, Chen P, Zhang Z, Yang L, Towards pixel-to-pixel deep nucleus detection in microscopy images, *BMC bioinformatics* 20 (1) (2019) 1–16. [PubMed: 30606105]
- [20]. Nayak N, Chang H, Borowsky A, Spellman P, Parvin B, Classification of tumor histopathology via sparse feature learning, in: *2013 IEEE 10th international symposium on biomedical imaging*, 2013, pp. 410–413.
- [21]. Irshad H, Veillard A, Roux L, Racoceanu D, Methods for nuclei detection, segmentation, and classification in digital histopathology: a reviewcurrent status and future potential, *IEEE reviews in biomedical engineering* 7 (2013) 97–114.
- [22]. Spanhol FA, Oliveira LS, Petitjean C, Heutte L, A dataset for breast cancer histopathological image classification, *IEEE Transactions on Biomedical Engineering* 63 (7) (2015) 1455–1462. [PubMed: 26540668]
- [23]. Xu Y, Zhu J-Y, Chang E, Tu Z, Multiple clustered instance learning for histopathology cancer image classification, segmentation and clustering, in: *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 964–971.
- [24]. Mousavi HS, Monga V, Rao G, Rao AU, Automated discrimination of lower and higher grade gliomas based on histopathological image analysis, *Journal of pathology informatics* 6.
- [25]. Vu TH, Mousavi HS, Monga V, Rao G, Rao UA, Histopathological image classification using discriminative feature-oriented dictionary learning, *IEEE transactions on medical imaging* 35 (3) (2015) 738–751. [PubMed: 26513781]
- [26]. Hou L, Samaras D, Kurc TM, Gao Y, Davis JE, Saltz JH, Patch-based convolutional neural network for whole slide tissue image classification, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2424–2433.
- [27]. Wang X, Chen H, Gan C, Lin H, Dou Q, Huang Q, Cai M, Heng P-A, Weakly supervised learning for whole slide lung cancer image classification, *Medical Imaging with Deep Learning*.
- [28]. Ilse M, Tomczak JM, Welling M, Attention-based deep multiple instance learning, *arXiv preprint arXiv:1802.04712*.
- [29]. Tizhoosh HR, Pantanowitz L, Artificial intelligence and digital pathology: Challenges and opportunities, *Journal of pathology informatics* 9.

- [30]. Komura D, Ishikawa S, Machine learning methods for histopathological image analysis, Computational and structural biotechnology journal 16 (2018) 34–42. [PubMed: 30275936]
- [31]. Ahonen T, Hadid A, Pietikäinen M, Face recognition with local binary patterns, in: European conference on computer vision, Springer, 2004, pp. 469–481.
- [32]. Lowe DG, Distinctive image features from scale-invariant keypoints, International journal of computer vision 60 (2) (2004) 91–110.
- [33]. Oliva A, Torralba A, Modeling the shape of the scene: A holistic representation of the spatial envelope, International journal of computer vision 42 (3) (2001) 145–175.
- [34]. Krizhevsky A, Sutskever I, Hinton GE, Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, 2012, pp. 1097–1105.
- [35]. He K, Zhang X, Ren S, Sun J, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
- [36]. LeCun Y, Bengio Y, Hinton G, Deep learning, nature 521 (7553) (2015) 436. [PubMed: 26017442]
- [37]. Zhang Z, Chen P, McGough M, Xing F, Wang C, Bui M, Xie Y, Sapkota M, Cui L, Dhillon J, et al., Pathologist-level interpretable whole-slide cancer diagnosis with deep learning, Nature Machine Intelligence 1 (5) (2019) 236.
- [38]. Zhang Z, Chen P, Shi X, Yang L, Text-guided neural network training for image recognition in natural scenes and medicine, IEEE Transactions on Pattern Analysis and Machine Intelligence.
- [39]. Simonyan K, Zisserman A, Very deep convolutional networks for largescale image recognition, arXiv preprint arXiv:1409.1556.
- [40]. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z, Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2818–2826.
- [41]. Zhang S, Metaxas D, Large-scale medical image analytics: Recent methodologies, applications and future directions (2016).
- [42]. Sapkota M, Shi X, Xing F, Yang L, Deep convolutional hashing for low-dimensional binary embedding of histopathological images, IEEE journal of biomedical and health informatics 23 (2) (2018) 805–816. [PubMed: 29993648]
- [43]. Zheng Y, Jiang Z, Zhang H, Xie F, Ma Y, Shi H, Zhao Y, Histopathological whole slide image analysis using context-based cbir, IEEE transactions on medical imaging 37 (7) (2018) 1641–1652. [PubMed: 29969415]
- [44]. Hegde N, Hipp JD, Liu Y, Emmert-Buck M, Reif E, Smilkov D, Terry M, Cai CJ, Amin MB, Mermel CH, et al., Similar image search for histopathology: Smily, npj Digital Medicine 2 (1) (2019) 56. [PubMed: 31304402]
- [45]. Gu Y, Yang J, Multi-level magnification correlation hashing for scalable histopathological image retrieval, Neurocomputing 351 (2019) 134–145.
- [46]. Zheng Y, Jiang Z, Zhang H, Xie F, Ma Y, Shi H, Zhao Y, Size-scalable content-based histopathological image retrieval from database that consists of wsis, IEEE journal of biomedical and health informatics 22 (4) (2017) 1278–1287. [PubMed: 28692995]
- [47]. Zheng Y, Jiang B, Shi J, Zhang H, Xie F, Encoding histopathological wsis using gnn for scalable diagnostically relevant regions retrieval, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2019, pp. 550–558.
- [48]. Jimenez-del Toro O, Otálora S, Atzori M, Müller H, Deep multimodal case-based retrieval for large histopathology datasets, in: International Workshop on Patch-based Techniques in Medical Imaging, Springer, 2017, pp. 149–157.
- [49]. Safavian SR, Landgrebe D, A survey of decision tree classifier methodology, IEEE transactions on systems, man, and cybernetics 21 (3) (1991) 660–674.
- [50]. Cortes C, Vapnik V, Support-vector networks, Machine learning 20 (3) (1995) 273–297.
- [51]. Ho TK, Random decision forests, in: Proceedings of 3rd international conference on document analysis and recognition, Vol. 1, 1995, pp. 278–282.
- [52]. Xia R, Pan Y, Lai H, Liu C, Yan S, Supervised hashing for image retrieval via image representation learning, in: Twenty-eighth AAAI conference on artificial intelligence, 2014.

- [53]. Liu W, Wang J, Ji R, Jiang Y-G, Chang S-F, Supervised hashing with kernels, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 2074–2081.
- [54]. Shen F, Shen C, Liu W, Tao Shen H, Supervised discrete hashing, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 37–45.
- [55]. Kang W-C, Li W-J, Zhou Z-H, Column sampling based discrete supervised hashing, in: Thirtieth AAAI conference on artificial intelligence, 2016.
- [56]. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al., Scikit-learn: Machine learning in python, Journal of machine learning research 12 (Oct) (2011) 2825–2830.
- [57]. Maaten L. v. d., Hinton G, Visualizing data using t-sne, Journal of machine learning research 9 (Nov) (2008) 2579–2605.
- [58]. Mobadersany P, Yousefi S, Amgad M, Gutman DA, Barnholtz-Sloan JS, Vega JEV, Brat DJ, Cooper LA, Predicting cancer outcomes from histology and genomics using convolutional networks, Proceedings of the National Academy of Sciences 115 (13) (2018) E2970–E2979.

Highlights

- We develop an interactive thyroid whole slide image classification and retrieval system using deep representation.
- With the participation of pathologists, the proposed system can focus on the suspicious regions in thyroid frozen sections, thus speeding up the analysis and improving the diagnostic performance.
- We propose retrieval of the whole slide image to base on the region of interest (ROI), which is more fine-grained, precise, and reliable.
- We introduce a win-win scheme for pathologists and the AI system, enable the diagnostic system to be more easily accepted by both pathologists and patients.

Figure 1:

Annotated suspicious regions in thyroid frozen section whole slide images. The regions inside the green contours are suspicious, which have pronounced appearance differences with tissues around.

Figure 2:

Interactive thyroid whole slide image (WSI) diagnostic system. The pathologist first draws a region of interest (ROI) in the thyroid slide viewer. Based on the drawn contour, the diagnostic system can provide classification and retrieval information. The pathologist gives the final diagnosis based on his/her expertise with the system supplied knowledge.

Figure 3:

Examples of three different types of regions in thyroid slides. From left to right are benign, uncertain, and malignant, respectively.

Figure 4:

An overview of thyroid patch cropping and classifier training. Three different patch categories are randomly cropped from corresponding annotated ROIs. The CNN model is adopted for patch classification training and feature extraction.

Figure 5:

The pipeline of ROI representation generation. Patches are firstly sampled from the preselected ROI. The CNN model is applied to extract features for patches, and all patch features are combined to generate a fixed-length representation.

Figure 6:
ROI classification performance comparison using different fusion power exponent ranging from 1 to 3 on different ROI representation generation methods and ROI classifiers.

Figure 7:

Thyroid WSI ROI representation embedding using t-SNE on four different patch feature extraction methods.

Figure 8:
Comparison of KSH retrieval method using 8, 16, and 32 binary hash codes.

Figure 9:
Comparison of three retrieval methods, KSH, SDH, and COSDISH, on four feature extraction methods using 8 bits binary codes.

Figure 10:

The retrieval precision and recall of KSH using 8 bits binary codes.

Figure 11:

Retrieval examples for three different types of regions using KSH with binary codes of 8 bits. Based on the appearance of these retrieved samples, they are very close to the query sample, thus can provide a reference for the diagnosis of the query ROI.

Summary of thyroid frozen sections’ diagnostic categories, the subtypes included, and their corresponding surgery plan for each category.

Table 1:

Diagnosis	Subtypes included
Benign	Normal thyroids, nodular hyperplasia, multinodular goiter, lymphocytic thyroiditis, Hashimoto
Malignant	PTC, medullary carcinoma, poorly differentiated thyroid carcinoma, thyroid lymphoma
Uncertain	Follicular adenoma, follicular carcinoma, follicular PTC, adenomatous nodules

Note: PTC-papillary thyroid carcinoma

ROI classification performance using four different feature extraction methods on four classifiers with ROI feature fusion power exponent set as 1. Note: the highest classification accuracy is bold.

power exponent p=1		Decision Tree				SVM				Random Forest				MLP			
Methods	Class	Precision	Recall	F ₁	Acc	Precision	Recall	F ₁	Acc	Precision	Recall	F ₁	Acc	Precision	Recall	F ₁	Acc
vgg16bn	Benign	0.930	0.979	0.954	0.925	0.936	0.998	0.966	0.945	0.942	0.997	0.969	0.952	0.965	0.982	0.973	0.956
	Uncertain	0.763	0.621	0.685		0.983	0.464	0.630		0.936	0.691	0.795		0.863	0.724	0.787	
	Malignant	0.944	0.915	0.929		0.956	0.974	0.965		0.973	0.946	0.959		0.961	0.970	0.965	
inceptionv3	Benign	0.949	0.975	0.962	0.937	0.963	0.987	0.975	0.947	0.945	0.988	0.966	0.945	0.975	0.984	0.979	0.957
	Uncertain	0.801	0.582	0.674		0.907	0.508	0.651		0.872	0.600	0.711		0.861	0.647	0.739	
	Malignant	0.941	0.965	0.953		0.924	0.989	0.955		0.956	0.966	0.961		0.948	0.988	0.968	
resnet50	Benign	0.956	0.982	0.969	0.944	0.945	0.990	0.967	0.943	0.946	0.998	0.971	0.957	0.962	0.984	0.973	0.961
	Uncertain	0.640	0.655	0.647		0.933	0.470	0.625		0.967	0.649	0.777		0.857	0.745	0.797	
	Malignant	0.952	0.906	0.928		0.932	0.981	0.956		0.975	0.968	0.971		0.981	0.979	0.980	
GlST	Benign	0.748	0.933	0.830	0.752	0.555	1.000	0.714	0.555	0.808	0.919	0.860	0.777	0.833	0.947	0.886	0.817
	Uncertain	0.480	0.182	0.264		0.000	0.000	0.000		0.312	0.160	0.212		0.470	0.241	0.319	
	Malignant	0.805	0.605	0.691		0.000	0.000	0.000		0.789	0.706	0.745		0.810	0.748	0.778	

Retrieval MAP of four feature extraction methods using three supervised retrieval manners with the number of retrieved samples ranging from 50 to 500. The maximum MAP values for different number of retrieved samples are bold.

Table 3:

Retrieval Sample Num.		50	100	150	200	250	300	350	400	450	500
Ffeature	Retrieval	MAP	MAP	MAP	MAP	MAP	MAP	MAP	MAP	MAP	MAP
	KSH	0.963	0.963	0.966	0.968	0.970	0.971	0.971	0.971	0.972	0.972
vgg16bn	SDH	0.950	0.951	0.952	0.952	0.953	0.954	0.955	0.957	0.958	0.959
	COSDISH	0.951	0.953	0.953	0.954	0.954	0.955	0.956	0.956	0.957	0.958
	KSH	0.953	0.955	0.958	0.961	0.963	0.964	0.965	0.966	0.966	0.966
inceptionv3	SDH	0.938	0.9-10	0.941	0.942	0.942	0.943	0.945	0.946	0.947	0.948
	COSDISH	0.937	0.945	0.949	0.951	0.952	0.953	0.954	0.955	0.956	0.957
	KSH	0.954	0.960	0.962	0.963	0.965	0.965	0.966	0.966	0.966	0.965
resnet50	SDH	0.053	0.955	0.956	0.956	0.957	0.958	0.959	0.960	0.961	0.962
	COSDISH	0.873	0.904	0.918	0.926	0.930	0.931	0.932	0.933	0.934	0.934
	KSH	0.761	0.755	0.750	0.746	0.742	0.740	0.741	0.742	0.744	0.745
GIST	SDH	0.773	0.778	0.781	0.783	0.788	0.792	0.800	0.808	0.814	0.820
	COSDISH	0.769	0.779	0.784	0.784	0.785	0.789	0.796	0.806	0.814	0.820

Table 4:

Classification accuracy of four feature extraction methods using three supervised retrieval manners with different length of binary codes. We adopt near neighbors with $k = 9$ for classification. The maximum accuracy value is bold.

Feature	vgg16bn			inceptionv3			resnet50			GIST		
	8	16	32	8	16	32	8	16	32	8	16	32
Bits												
KSH	0.984	0.985	0.987	0.985	0.984	0.984	0.982	0.982	0.988	0.980	0.978	0.978
SDH	0.953	0.952	0.956	0.944	0.947	0.957	0.959	0.959	0.965	0.790	0.794	0.802
COSDISH	0.960	0.964	0.958	0.965	0.956	0.965	0.960	0.957	0.968	0.803	0.804	0.818