Decoding Expertise from Pathologists’ Diagnostic Processes on Whole Slide Images

Abstract Based on the expertise of pathologists, the pixelwise manual annotation has provided substantial support for training deep learning models of whole slide images (WSI)-assisted diagnostic. However, the collection of pixelwise annotation demands massive annotation time from pathologists, leading to waste medical manpower resources, hindering to construct larger datasets and more precise diagnostic models. To obtain pathologists’ expertise with minimal pathologist workloads then achieve precise diagnostic, we collected the image review patterns of pathologists by eye-tracking devices. Simultaneously, we designed a deep learning system: Pathology Expertise Acquisition Network (PEAN), based on the collected visual patterns, which can decode pathologists’ expertise then diagnoses WSIs. In terms of time cost, using eye tracking annotation reduces the time required to 4%, with an average of only 36.5 seconds per annotation. We evaluated PEAN on a multi-center dataset comprising 5,881 WSIs and 5 categories of skin lesions, achieving a significantly higher AUC of 0.984 and an accuracy of 93.1% compared to other models based on fully supervised or weakly supervised learning. This study fills the gap in existing models’ inability to learn from the diagnostic processes of pathologists. Its efficient data annotation and precise diagnostics provide assistance in both large-scale data collection and clinical care.

1 Introduction

The pathology diagnosis forms the basis of clinical and pharmaceutical research and is fundamental in determining patient treatment modalities [1, 2]. The quantitative analysis of digital pathology images (whole slide image, WSI) and the development of computer-aided diagnostic systems provide crucial support to pathologists [3]. This not only saves a significant amount of medical manpower resources but also enables faster and more accurate patient care.

The development of deep learning (DL)-assisted diagnostic systems [4-6] in the field of WSI classification has garnered widespread attention [7]. Traditionally, such methods have relied on the manual extraction of pathologists’ professional knowledge, achieved through pixelwise annotation on ultra-large WSI images [8-10]. (A single WSI typically contains billions of pixels, and therefore needs to be divided into 224×224-pixel patches for processing; manual pixelwise annotations are usually provided as labels at the patch level [10].) Through fine guidance based on pathologists’ professional knowledge, DL models have achieved precise diagnostics [11]. For instance, in the Camelyon16 dataset [12] for breast cancer metastasis detection (comprising 400 WSIs with pixel-wise annotations), DL models have exhibited performance surpassing that of pathologists. However, due to significant variations in clinical samples, successful results in small datasets are not yet sufficient to confirm practicality in clinical practice [11]. While the substantial workload associated with large-scale images and the high demand for pathological expertise exacerbate the scarcity of large annotated datasets in the field of computational pathology.

Although weakly supervised learning methods [13-15], represented by multiple instance learning, only require reported diagnostic outcomes as labels for training, thus alleviating the issue of high annotation costs [11, 16-20]. These methods, however, exhibit lower performance and robustness due to the lack of guided experience from pathologists [21-24]. The performance degradation due to cross-dataset issues is particularly evident, a result that our findings also confirm. Weakly supervised learning methods struggle to meet clinical requirements, even when trained on large datasets, and the expansion of datasets further exacerbates the strain on medical resources and potential privacy concerns. Additionally, due to the inability to form a direct association with lesion regions in the images, weakly supervised learning lacks interpretability, posing potential safety concerns for clinical use.

The use of eye-tracking devices to capture human expertise from gaze behavior is a hot topic in computer vision, often applied to robot control or autonomous driving [25-31]. However, current research on the development of WSI diagnostic systems is largely focused on obtaining guidance from traditional manual annotations or reported diagnostic reports. There is insufficient research on extracting professional knowledge from pathologists’ image review processes or collecting visual annotations to replace traditional manual annotations [32, 33]. This situation highlights the significant cost of data annotation, as well as the poor interpretability of the models: the diagnostic process of the model is detached from the pathologists’ diagnostic process. Actually, both time-efficient diagnostic reports and time-consuming manual annotations stem from the visual image review process of pathologists. In other words, collecting visual data from pathologists incurs almost no additional time cost. The absence of this data from existing available datasets represents a waste of medical resources. We hypothesize that the visual data obtained using eye-tracking devices from pathologists' image review processes can reflect their areas of interest, thus forming an alternative to traditional "pixel-wise annotation". The core issue of this study is to extract pathologists' professional knowledge from their visual behavior and effectively apply it to DL, surpassing the performance brought about by traditional manual annotations while reducing data annotation costs. Furthermore, this study aims to fill the gap by guiding DL models through the learning of pathologists' diagnostic processes.

This study aims to decode the expertise of pathologists from their visual behavior and utilize it in the development of the first DL system that learns from the diagnostic processes of pathologists. The objective is to achieve more precise and interpretable diagnostic assistance at a lower data annotation cost, ultimately saving medical manpower resources in the construction of diagnostic systems and providing better care for patients. First, we acquired WSIs and pathologists' slide-reviewing data using custom-developed software and an eye-tracking device and reported the details of their reading behavior, which included the pathologists' eye movements, zooming or panning the WSIs, and the final diagnoses. 5,881 WSIs covering 5 categories of skin lesions were collected from two medical research centers. Slide-reviewing data and manual pixelwise annotations were collected for approximately 25% of the WSIs and used as a training set, while the remaining WSIs were divided into two testing sets. (The manual pixelwise annotation was only used for training comparative algorithms and was not involved in the development of our model.)

Second, a DL system called Pathology Expertise Acquisition Network (PEAN) was designed to extract the pathologists' expertise from their slide-reviewing data (as shown in Figure 1.a). We defined the value of this expertise as the "pathologist's attention level", with each patch corresponding to an "expertise value". PEAN computes the "expertise values" for all patches in the WSI, simulating the pathologist's regions of interest (ROIs) for diagnostic assistance. To validate the correlation between the expertise extracted by PEAN and the actual diagnostic evidence attended to by the pathologist (ground truth), i.e., the ground truth region having higher expertise value, we compared the pathologist's manual pixelwise annotated map, the pathologist's visual attention map, the expertise value heatmap, and the suspicious region map selected by PEAN to imitate pathologists. We found the overlap among the four types of regions, thus validating the effectiveness of the expertise value. Driven by this expertise, models for WSI classification and imitating pathologists' visual diagnostic process were developed. PEAN achieved an accuracy of 93.1% and an AUC of 0.984 on the internal test set, and an accuracy of 85.5% and an AUC of 0.950 on the external test set. Its classification performance and robustness significantly surpassed existing fully supervised and weakly supervised learning models. Furthermore, learning from multiple pathologists' experiences concurrently can enhance the classification ability. The existing DL models in this field lack consideration for the diagnostic process of pathologists and do not possess the ability to learn from human expertise and imitate it. PEAN can achieve "human-like" pathological diagnosis by imitating the diagnostic process of pathologists, i.e., conducting a step-by-step search on the WSI to form a "pseudo-visual trajectory". We observed a high degree of overlap between the regions of interest identified by PEAN and those of the pathologists. Additionally, through hypothesis testing, we found that the regions of interest identified by PEAN effectively improve the accuracy of the classification model (p=0.005). This validates the effectiveness and interpretability of the "imitator", filling the gap in human-like diagnosis. This study represents the first DL model to decode human expertise from visual behavior and apply it to assist in the WSI diagnosis. The integration of pathologists' diagnostic processes with DL has enhanced model performance and data collection efficiency. Unlike existing fully supervised and weakly supervised learning approaches, this study offers a novel approach to computational pathology.

2 Results

1. Specifics of the Dataset. The unique retrospective dataset constructed in this study comprised two types of data: hematoxylin and eosin (H&E)-stained pathologic images produced by a whole slide scanner, (the WSIs), and slide-reviewing data generated by eye-tracking devices. A total of 5,881 WSIs representing different skin conditions (benign moles [nevus] and four skin diseases [basal cell carcinoma (BCC), melanoma, squamous cell carcinoma (SCC), and seborrheic keratosis (SK)]) were collected (Figure 1.b and 1.c). of these, 3,899 and 1,982 WSIs were collected from the First Affiliated Hospital of China Medical University (Hospital F) and the General Hospital of Shenyang Military Region (Hospital G), respectively. All image data were paired with slide-level labels generated from previously recorded diagnostic reports. WSIs collected from Hospital F were divided into the training (1,468) and internal testing (2,431) datasets, while those collected from Hospital G were used as an external testing dataset. A total of 92 WSIs in the internal testing set and all 1,473 WSIs in the training set were reviewed by a group of five dermatopathologists, yielding a total of 1,565 reviewed WSIs; of these, 552 were reviewed by all five pathologists (Figure 1.d).

The slide-reviewing data consist of the visual attention patterns of pathologists collected via "EasyPathology", a self-developed

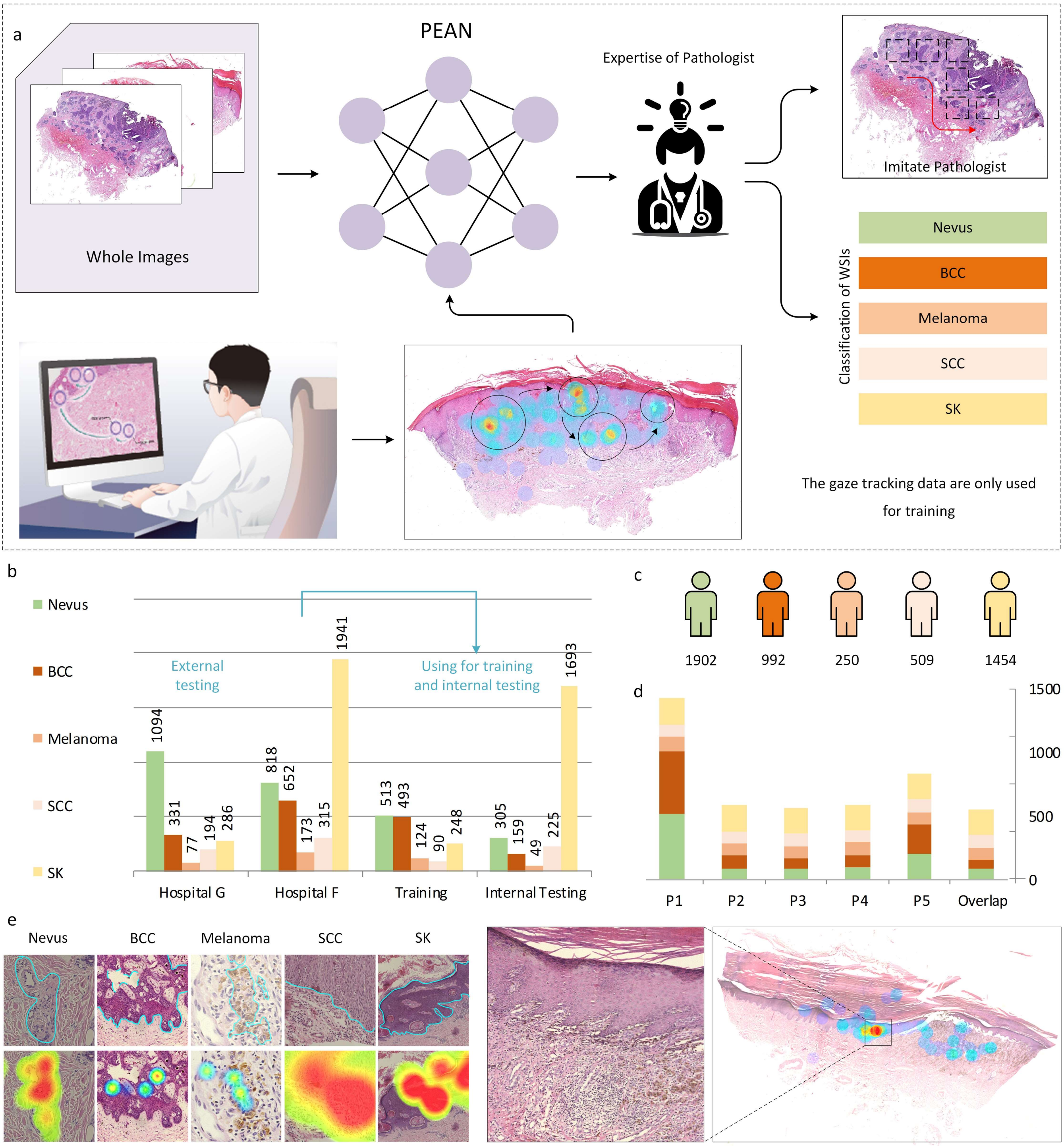


Figure 1. PEAN model and dataset. a. PEAN Model: After training on pathologists' slide-reviewing data, the model is capable to both perform a multiclassification task and imitate the pathologists' slide-reviewing behaviors. **b.** Data distribution of the training dataset, internal testing dataset, and external testing dataset. The color legend representing various diseases is utilized in c and d. **c.** Total number of patients with different skin conditions in the dataset. **d.** Numbers of slide-reviewing operations performed by the different pathologists. The "Overlap" column includes the images listed for each pathologist. **e.** Images at high magnification showing the regions of interest (heatmaps, second row) in which the pathologist's gaze highly overlaps with the actual tumor tissue (marked in blue in the first row). At lower magnifications, the distribution of the pathologist's gaze areas approximately corresponds with the actual tumor tissue; more examples are illustrated in Figure 2.b. We also observed that areas on which the pathologists focused more attention typically contained chaotic tumor boundaries. Even at high magnification, manual annotation of scattered tumor cells within these areas is challenging, underscoring one of the advantages of using eye tracking for "visual annotation".

eye-tracking system (detailed in Appendix 1). The data encompass the eye movements of the pathologists during their review of the WSIs, the two-dimensional mappings of the corresponding gaze points (example gaze heatmaps are shown in Figure 1.e), the magnification levels employed when viewing WSIs, and the diagnostic results. The external environment (such as light and room

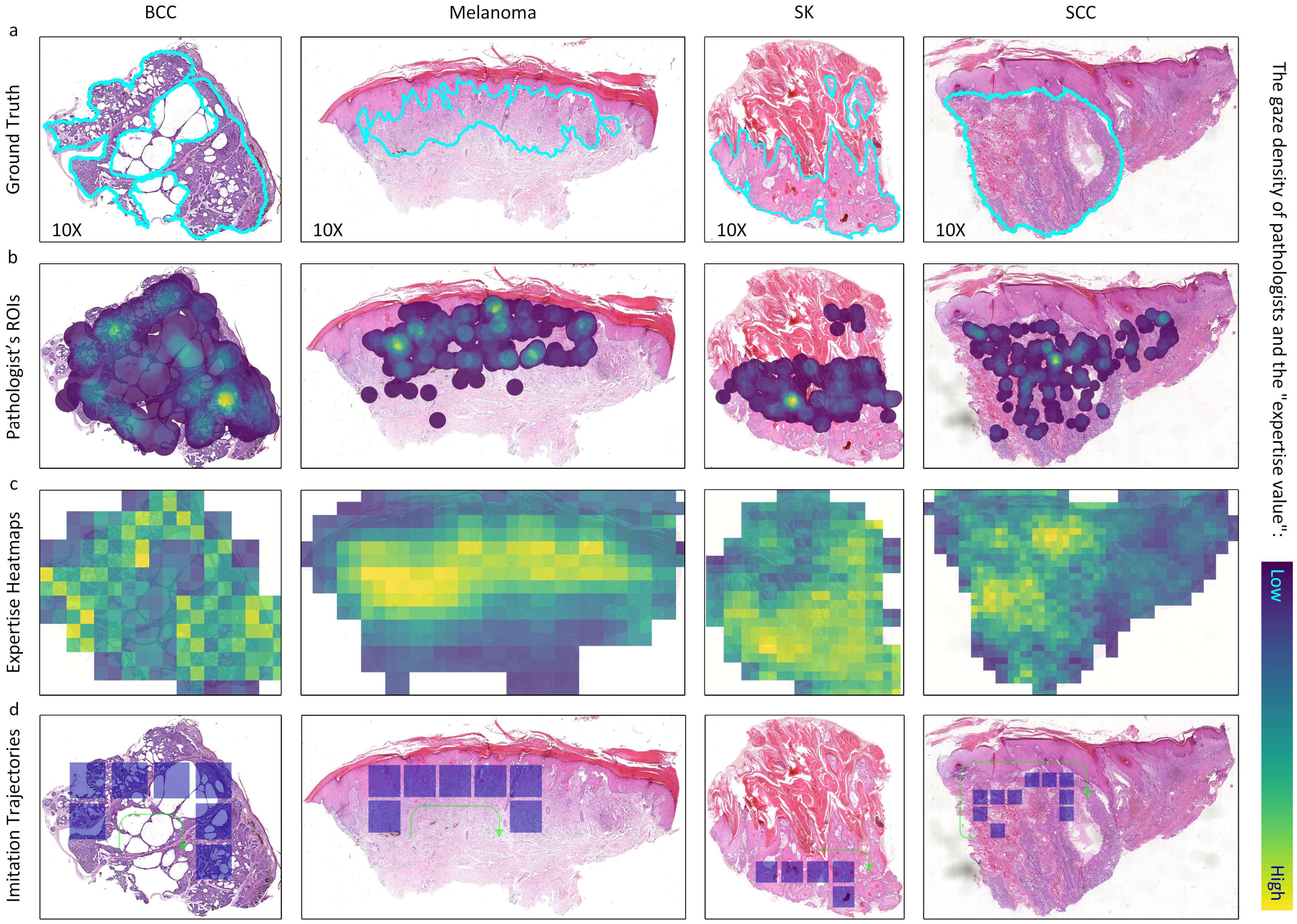


Figure 2. Map Comparison between the pathologists' pixel-level manual annotations for identifying tumor regions **(a)**, heatmaps representing the pathologists' selected ROIs **(b)**, heatmaps generated by PEAN as the pathologists' "expert knowledge" **(c)**, and heatmaps depicting the attention trajectories generated by PEAN imitating the pathologists' slide-review behavior **(d)**.

temperature) for slide-reviewing data collection was standardized as much as possible to minimize the influence of external disturbances on the pathologists' review. After conducting fatigue tests on the pathologists, we determined that data could be collected continuously for 50 minutes at a time (detailed in Appendix 2). During this process, the actual labels of the WSIs were concealed, and the pathologists were asked to relabel the images. Prior to commencement of the official data collection, the pathologists underwent thorough training to acclimate to the procedure; the data collected from their first five WSIs, considered a set of pretraining images, were excluded from the final collected dataset.

In addition, all WSIs reviewed by the pathologists using the eye-tracking device, were annotated manually to compare the effort involved in traditional fully supervised learning. The manual annotations were completed by the five pathologists involved in the slide-reviewing data collection process. Collecting manual annotations was shown to be extremely labor intensive; the monitoring records of our dataset showed that pathologists spent an average of 14.2 minutes manually annotating a single WSI, while the average time for collecting slide-reviewing data per WSI is 36.5 seconds. Thus, the pathologists' workload in viewing one WSI was substantially reduced to less than 5% of that required for manual annotation. This indicates that within the same time frame, a pathologist can "visually annotate" a significantly larger number of WSIs, potentially increasing the power in training more accurate and robust DL models.

For DL model training, the WSIs had their backgrounds removed [20], then segmented into 224×224-pixel patches at 10× magnification (the monitoring records showed it is the magnification at which the pathologists most frequently exhibited their gaze behaviors), resulting in approximately 2.8 million patches in our dataset. All these patches are encoded into feature vectors by a Resnet50 [34] pretrained by ImageNet [35]; the encoder does not participate in model training, a common practice in computational pathology research [16-20].

Table 1. Comparison of PEAN-C with baseline models in classifying WSI

Methods		Internal Testing Set		External Testing Set	
		ACC (%)	AUC	ACC (%)	AUC
Fully	DLCCP	89.3	0.969	74.2	0.905
Supervised	SLC	90.3	0.976	75.5	0.893
Learning	HSL	91.2	0.978	76.7	0.908
Weakly	CLAM-SB	86.1	0.961	70.1	0.864
Supervised	ABMIL	82.6	0.933	62.4	0.804
Learning	TransMIL	88.8	0.965	73.7	0.899
Ours	PEAN-C	**93.1**	**0.984**	**85.5**	**0.950**

2. Overlap of the pathologist's manual annotations, the ROIs representing the pathologist's visual behaviors, and the subregions with high expertise value as recognized by PEAN. We attempted to demonstrate that the pathological expertise decoded by PEAN accurately reflects the pathologists' own knowledge, including their manual annotations and visual behavior. Figure 2 shows this comparison for four WSIs (corresponding to the four malignant diseases investigated in this study); their precise lesion area contours (Figure 2.a) are shown alongside the pathologists' ROI heatmaps (Figure 2.b) and the heatmap of expertise values output by PEAN (Figure 2.c, where higher values indicate higher diagnostic value perceived by PEAN; the corresponding computational methods are described in detail in Section 5.2). The maps show visually visible overlap among the pathologist's manually annotated tumor boundaries, the visual ROIs and the areas with high expertise values output by PEAN. This indicates that the PEAN-generated areas of focus tend to match the parts of the WSIs identified by the pathologists as the ground truth. Notably, the regions shown in Figure 2.a were manually outlined by the pathologists, and so they, too, are manifestations of the pathologists' expertise. The overlap among the three types of region suggests that as a result from learning from the pathologists' slide-reviewing data, the PEAN-output pathology image features can be considered to represent human expertise well, that is, to effectively capture their pathology knowledge. In the inference process, this "expertise" is manifested by PEAN as higher expertise values. This intuitive map comparison provides evidence for the validity of PEAN and the expertise values it decodes.

3. In classification tasks, compared to other models, PEAN has demonstrated superior performance. PEAN was compared with six other models in the classification task to demonstrate its excellence in the field of pathological diagnostic assistance. The baselines included Fully supervised learning: DLCCP [7], SLC [8], HSL [6]; Weakly supervised learning: CLAM-SB [20], ABMIL [16], TransMIL [17]. We utilized the official published code where available. The classification module of PEAN (hereafter referred to as "PEAN-C") and that of the other 6 models were trained and tested using the same training and testing datasets. Each model was used to classify the images into one of 5 classes: no malignancy or one of 4 types of skin disease (malignancy). The performance of each model was evaluated using two indices, accuracy (ACC) and area under the receiver operating characteristic (ROC) curve (AUC), the latter of which was calculated through macro averaging across all 5 classes in this multiclassification task. Additionally, confusion matrices were computed to assess the models' ability to identify the different diseases. The classification results are presented in Table 1 (ACC and AUC) and Appendix Figure 2 (confusion matrices). The results indicate that the performance of the fully supervised learning models exceeded that of the weakly supervised learning models and that PEAN-C outperformed the other 6 models. Specifically, in the internal test dataset, PEAN-C achieved an ACC of 93.1% and an AUC of 0.984, compared to the fully supervised learning algorithm HSL (which was ranked second on both ACC and AUC), PEAN-C exhibited a 1.85% greater ACC and a 0.62 greater AUC, indicative of a slight improvement. However, PEAN-C surpassed the baseline models by substantial margins in the external test set, achieving an ACC of 85.5% and an AUC of 0.950; these values were 8.8% and 0.042 larger, respectively, than those of the second-best model. The results indicate that PEAN-C demonstrates outstanding performance and strong generalization in WSI classification tasks. Combined with its lower data collection costs, it can be efficiently applied in clinical settings, alleviating the pressure on medical resources.

4. PEAN-C also demonstrates outstanding performance on a small training dataset. In addition to testing on the external dataset, we further evaluated the robustness of PEAN by reducing the training data volume. The models were trained on 5 random samples of 30 WSIs per class from the training set and tested them with the complete testing dataset. Even under these conditions,

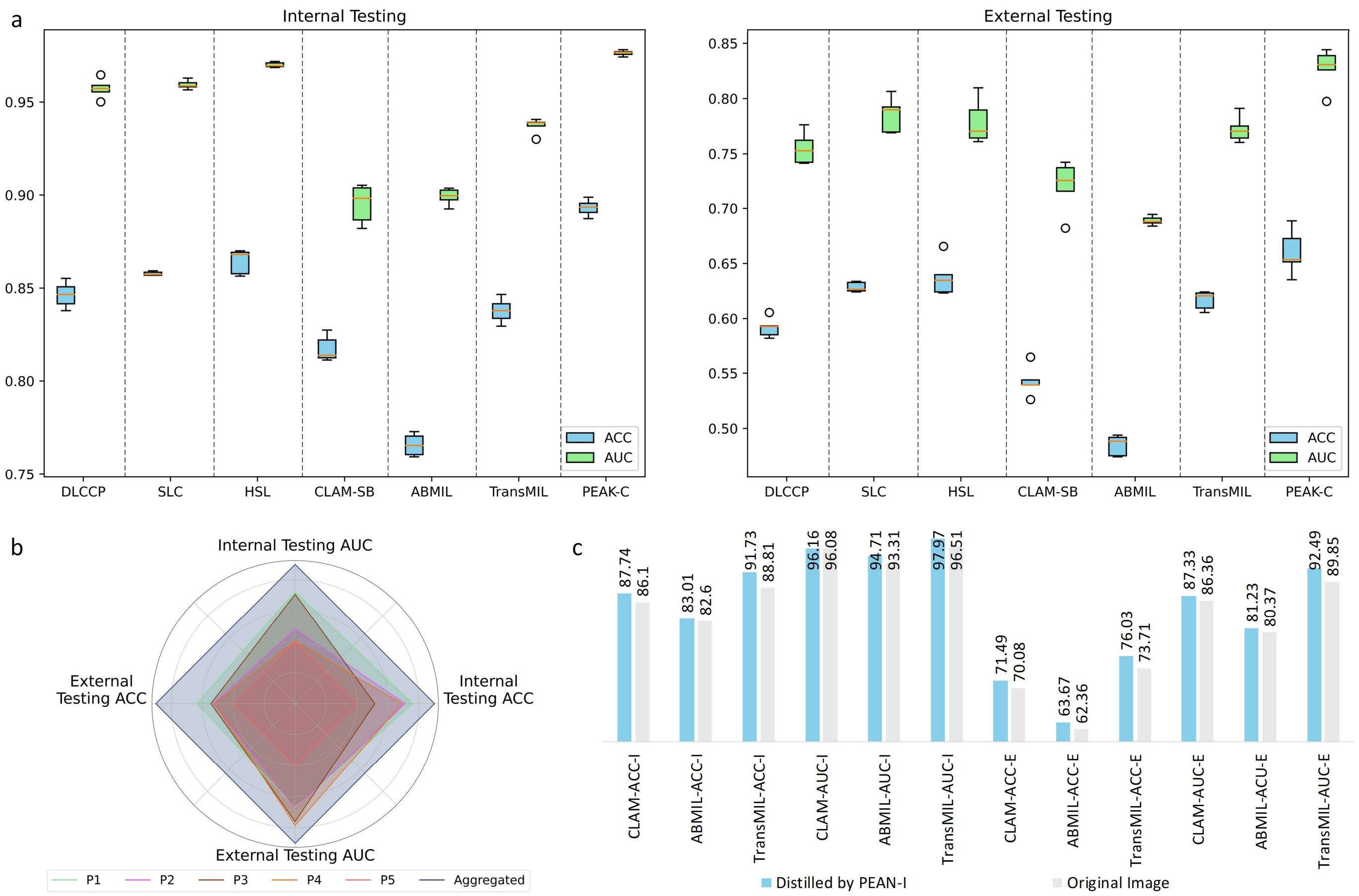


Figure 3. a. Model performance with a lower amount of training data (30 WSIs per class). PEAN-C achieved the best results for this training sample size. **b.** Comparison of the performance of models trained using the slide-reviewing data from 5 pathologists. The axes of the radar chart have been normalized. Polygons P1-P5 represent the models trained on the data from Pathologists 1-5, respectively; the size of the polygon represents the performance of the model. **c.** Comparison of results for testing different regions (I: internal testing dataset, E: external testing dataset).

PEAN-C still exhibited the best performance among all the models, as shown in Figure 3.a. In the internal testing set, PEAN-C achieved an average ACC of 89.3% and average AUC of 0.976, surpassing those of the second-ranked HSL by 2.88% and 0.63%, respectively. In the external testing set, PEAN-C obtained an average ACC of 66.0% and an average AUC of 0.827; these values were 2.29% and 4.84% greater, respectively, than those of the second-ranked HSL. In addition to the ACC on the external testing set, the performance gap between PEAN-C and the second-best model, HSL, is larger when trained on a smaller dataset compared to when trained on the original dataset (Table 1). The results indicate that PEAN-C is a superior choice when dealing with small datasets because of its ability to learn from the diagnostic processes of pathologists. Due to its greater consistency and robustness, this approach has the potential to be widely adopted by studies relying on small-sized datasets, thereby avoiding the additional costs and privacy risks associated with collecting large amounts of image data.

5. PEAN integrates the visual behaviors of different pathologists to increase classification performance. The discrepancies in manual annotations due to variations in reviewers' cognition are a well-known issue. Previous studies have attempted to address this problem by introducing additional and more experienced pathologists as "arbiters" of the annotators' delineations [7]. However, we found that PEAN can leverage this cognitive difference, integrating the diverse experiences of multiple pathologists, and thereby improve diagnostic performance. Six PEAN-Cs were constructed, five trained separately using the individual review data from each of the pathologists and one trained collectively using the mixed data from five pathologists. Due to variations in the volume of data reviewed by each pathologist, all models were trained using the "overlap" WSIs reviewed by all five pathologists (a total of 552 WSIs, as shown in Figure 1.c) to control for confounding variables. The ACC and AUC in the two testing sets were compared, as shown in Figure 3.b. The performance of the individual pathologist models varied; however, the model trained using the data from all the pathologists achieved the best performance, with ACCs of 91.98% and 74.7% and AUCs of 0.984 and 0.903 in the internal and external testing datasets, respectively; with respect to those of the top-performing individual DL model (trained using data from pathologist 1 (P1)), the ACCs were 1.48% and 2.44% greater, and the AUCs were 0.004 and 0.016 greater, respectively.

Unlike manual pixelwise annotations, which serve as the primary ground truth diagnostic standard (hard label) for patches, slide-reviewing data is not used in this manner but rather as a "soft label". When 2 different pathologists review the same WSI, although their diagnoses for contentious subregions may differ, as long as these regions have been observed by the 2 pathologists, PEAN interprets these regions as having a higher "attention level". Therefore, when training with slide review data from multiple pathologists, PEAN does not experience a performance decline due to label confusion, a common issue that affects traditional supervised learning methods. By learning from a more diverse set of visual behaviors, PEAN can further refine pathologists' expertise, thereby improving classification performance.

6. PEAN can imitate the visual behavior of pathologists and maps out review trajectories on WSIs. Reinforcement learning (RL) [36-39] was used to develop the imitation module of PEAN (PEAN-I), which is capable of imitating pathologists' visual behaviors for selecting regions on WSIs (details are discussed in Section 5.4). PEAN-I is an agent capable of autonomously selecting a series of regions on WSIs by scanning the WSI in a manner similar to the gaze patterns used by the pathologists but with a fixed step size and movement direction each time (up, upper-right, right, and so on, eight directions in total), as shown in Figure 2.d. The regions selected by PEAN-I also exhibited high degree of overlap with the ROIs manually annotated by the pathologists and the ground-truth tumor region. This indicates that in addition to reflecting pathological knowledge learned from the pathologists' "expertise", PEAN can imitate the pathologists' slide-reviewing behavior, truly learning human priors.

Furthermore, PEAN-I can be effectively integrated with existing weakly supervised learning models. The regions selected can be used to select ROIs from the original WSIs, followed by further training of weakly supervised learning models, leading to improved classification performance. As shown in Figure 3.c, when CLAM, ABMIL, and TransMIL were trained with pathology images generated by PEAN-I, both the ACC and AUC were increased in the two testing datasets. This improvement in performance was statistically significant, with p values of 0.0053 and 0.0161, respectively, as determined by paired t tests. This effective enhancement of DL model performance demonstrates the efficacy of imitating pathologist behavior, reflecting its ability to "learn" pathologists' expertise while providing strong evidence for the validity of this expertise.

3 Conclusion

In this study, we collected a large dataset of skin pathology images, as well as manual pixelwise annotations and visual slide-reviewing data from five pathologists. Subsequently, we developed a novel DL system, PEAN, which decodes pathologists' professional knowledge using their eye-tracking data, and designed a new paradigm for the auxiliary diagnosis of WSIs. In this way, PEAN fills the gap in the integration of DL models with the diagnostic processes of pathologists. Not only does it significantly reduce the workload of pathologists in manually annotating a large number of WSIs while achieving more precise diagnostics, but it also effectively simulates pathologists' slide review behavior, automatically identifying regions of high diagnostic value. Importantly, the pathologist expertise decoded by this method can be used not only for the two tasks in this study, but also has high scalability, which provides strong support for the efficient application of future large-scale studies.

4 Data and Code Availability

The data involved in this study, including 1,565 skin WSIs along with their corresponding slide-reviewing data and manual annotations, will be made available for access. The dataset does not contain any personally identifiable information, and its release has been approved by the Medical Science Research Ethics Committee of the First Affiliated Hospital of China Medical University, with ethical code AF-SOP-07-1.1-01.

All the code involved in this study, including the open-source version of the "EasyPathology" system and the PEAN model code, will be made publicly available on open-source platforms. The code can be obtained through the following link: https://github.com/MasyerN/PEAN. The link also includes the data acquisition channels for the dataset.

The demonstration of PEAN imitating pathologists' slide-reviewing behavior has been uploaded to the following link: https://www.tankaai.com/eponline/. We have also recorded the process of collecting pathologists' slide-reviewing data and the operation of PEAN, which is presented as a demonstration video in the supplementary files "Demo-xxxx.mp4".

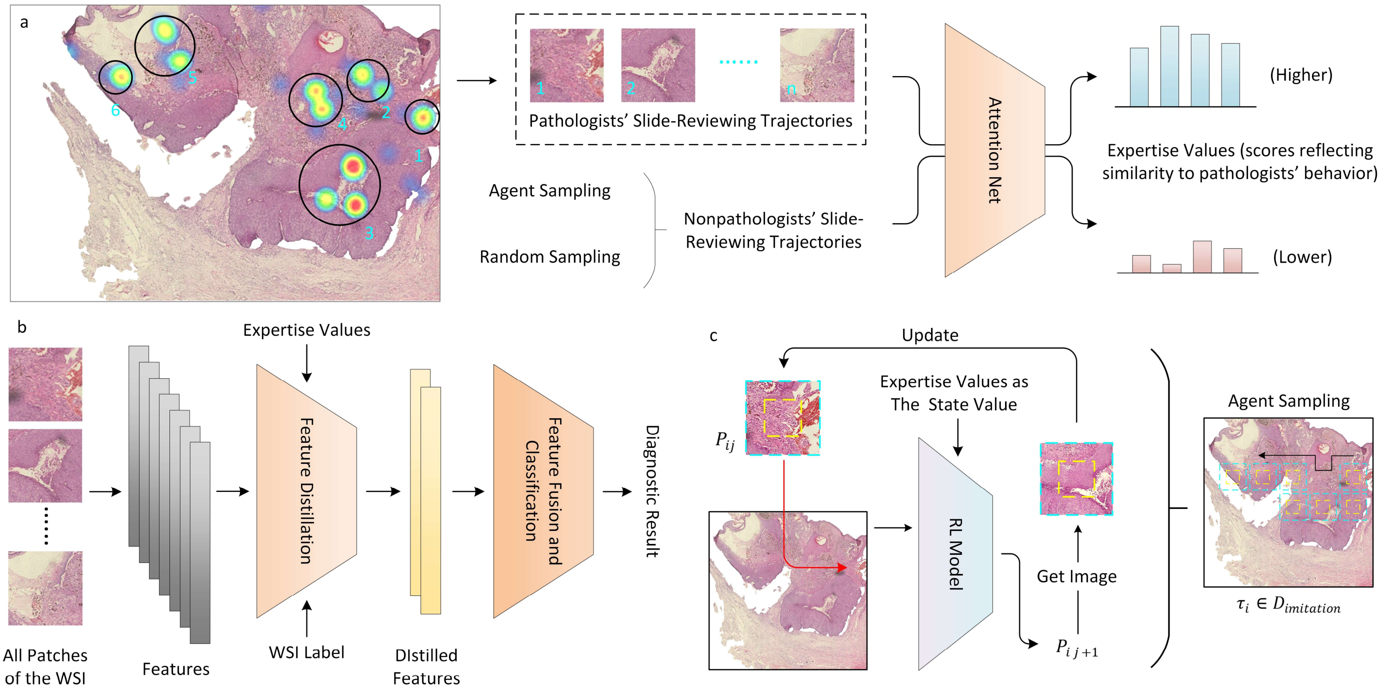

a
1
2
3
4
5
6
1
2
......
n
Pathologists' Slide-Reviewing Trajectories
Agent Sampling
Random Sampling
Nonpathologists' Slide-Reviewing Trajectories
Attention Net
(Higher)
Expertise Values (scores reflecting similarity to pathologists' behavior)
(Lower)
b
Expertise Values
All Patches of the WSI
Features
Feature Distillation
WSI Label
DIstilled Features
Feature Fusion and Classification
Diagnostic Result
c
Update
P_{ij}
Expertise Values as The State Value
RL Model
Get Image
$P_{i\,j+1}$
Agent Sampling
$\tau_i \in D_{imitation}$

Figure 4. Model details of the three-part PEAN. a. Model for extracting pathologists' expertise; **b.** Feature distillation classification model using the extracted pathologists' expertise; and **c.** Model imitating pathologists' slide-reviewing behavior using reinforcement learning.

5 Methods

1. Data acquisition and preprocessing. The image dataset used in this study consisted of 5,881 H&E-stained, previously diagnosed pathological WSIs, including images of one benign skin condition (nevus) and four diseases: BCC, melanoma, SCC, and SK. These WSIs were collected from 5,107 patients between 2016 and 2022 from the First Affiliated Hospital of China Medical University and the Shenyang Military Region General Hospital, respectively. To ensure that the WSIs were of sufficient quality for analysis, two pathologists inspected the data, including checking the data labels and assessing data quality issues such as image blurring, low contrast, uneven staining, background stains, duplicated samples, and insufficient tissue proportions. Only WSIs deemed satisfactory by both pathologists during this process were included in the dataset, resulting in the exclusion of one nevus WSI and one BCC WSI. The process of data inclusion is described in more detail in Appendix 3

To collect the eye-tracking-based slide-reviewing data from the pathologists, we used "EasyPathology", pathology slide-reviewing software that allows an individual to review WSIs while simultaneously capturing his or her eye movements using an eye tracker, which are then mapped onto the WSIs in 2D, as shown in Figure 1.f. Five pathologists participated in the collection of slide-reviewing data. The pathologists' reviewing behavior was recorded in a controlled experimental environment (at the same time and with minimal external environmental disturbances such as external auditory or visual distractions) using "EasyPathology" software in conjunction with the "Tobii Pro Spectrum" eye-tracking system [40] at a sampling frequency of 60 Hz. This allows the fully automated and unobtrusive collection of slide-reviewing data.

Before the formal data collection began, the pathologists underwent rigorous training; the first 5% of the slide-reviewing data were used for practice but were not included in the final dataset. During the slide-reviewing process, various aspects of the pathologists' visual behavior are captured, including the movement of the of the WSI within the software window, magnification changes, gaze positions, angular velocity of the eye, head posture, and visual behaviors such as saccades, fixations, and blinks. Pathologists can also annotate and diagnose WSIs using the software platform (with the true labels hidden from the pathologists). Before reviewing each WSI, the eye-tracking system was recalibrated to minimize offsets. The duration of each data collection session was set to 50 minutes based on our quantitative analysis of the pathologists' slide-reviewing behavior. Specifically, we compared the values of two feature variables with the highest correlation with the pathologists' diagnostic accuracy, "the first fixation scale" and "searching number", over the slide-reading time during the pathologists' review process. Both indicators showed significant changes around the 50-minute mark, as illustrated in Appendix Figure 1. The collected slide-reviewing data was subject to a review process to check for completeness, stability of gaze points, absence of missing data, and any occurrence of offsets. This review process resulted in the removal of the slide-reviewing data for 101 WSIs. In the end, a total of 3,978 sets of slide-reviewing data from five pathologists were collected; of these, 542 WSIs were reviewed by all pathologists (Figure 1.c and d). For a given WSI W, the slide-reviewing data collected by the eye tracker can be regarded as a point set $\{Points\} \in R^{points\ number \times (\mathbb{1},\mathbb{1})}$ located above W. The tissue area of W is segmented to obtain the maximum tissue frame b_W as an irregular virtual frame surrounding the tissue image. Additionally, to align with the range displayed on the computer screen during the pathologist's reading process, W is divided into windows of size $[d \times d]$ at a magnification M, corresponding to the screen frame of the pathologist’s eye movement data; this is recorded as $\{w_i^M\}_i^K \in R^{K \times [d \times d]}$, where K is the total number of windows contained in W, and d is set to be the same dimension as the screen used by the pathologist. The set of $\{Points\}$ is then preprocessed as follows:

1) During the data collection process, points where the pathologists’ visual field extended beyond the window of the computer screen or beyond the boundaries of the pathological tissue are considered indicative of nonmeaningful visual behaviors and are not included.
2) When the pathologists’ eye movement angular velocity is greater than q_1°/s, the system records this as an eyelid twitch, and the corresponding point is deleted.

The preprocessed point set is then denoted as $\{Points'\}$. We categorize the ocular movements of the experts into two types based on the degree of fixation at the same time point: fixation and searching. We postulate that the areas with greater fixation, known as fixation zones, are likely to be pathological regions showing a relatively high degree of pathological suspicion. The pathologists' fixation regions were extracted using the DBSCAN [43] clustering algorithm. Prior studies have confirmed that the mere extraction of image features from the fixation zone is inadequate for definitively localizing diseased regions, a finding that was corroborated by our experiments. We hypothesize that (1) shifts in focus, particularly in the fixation zone, encompass underlying logical relationships and that (2) areas of focus continuously scrutinized by pathologists in the corresponding images, in addition to containing key features helpful for diagnosis, also contribute to latent logical relationships. Thus, by constructing these sequences and analyzing them as a whole, we can infer the logic and expertise that pathologists employ to interpret the images and apply this information to assist in diagnosing the images. We refer to these sequential data as expert trajectory samples $\tau \in D_{demo}$.

The sequence $< P_{i0}^M, P_{i1}^M, P_{i2}^M \ldots\ldots >$ formed by gaze areas $P_{ij}^M \in w_i^M$ constitutes the expert trajectory $\tau_i \in D_{demo}$ within window w_i^M. In addition, the gaze duration coefficient E_{time} and gaze point density coefficient $E_{density}$ are introduced as indicators for evaluating the differences in importance between different gaze regions:

$$E_{time}(i,j) = \beta_1 ** \frac{mean(\{Pointsnum\}_w)}{Pointsnum(i,j)} \tag{1}$$

$$E_{density}(i,j) = \beta_2 ** \frac{Regiondistance(i,j)}{mean(\{Regiondistance\}_W)} \tag{2}$$

where β_1 and β_2 are weight coefficients, both of which are positive numbers not greater than 1; $Pointsnum(i,j)$ indicates the total number of $Points \in \{Points'\}$ contained in P_{ij}^M; $mean(\{Pointsnum\}_W)$ indicates the average value of $Points$ contained in each gaze area in the full slice W; $Regiondistance(i,j)$ is the average distance from P_{ij}^M to other gaze regions $P_{ik}^M (i \neq k)$ under W_i^M; and $mean(\{Regiondistance\}_W$ represents the average value of the $Regiondistance()$ function over all gaze regions in the full slice W. $E_{time}(i,j)$ is positively correlated with the gaze duration of region P_{ij}^M, while $E_{density}(i,j)$ is positively correlated with the degree of aggregation of P_{ij}^M in W.

When the image is magnified to magnification M, each w_i^M is partitioned into patches of dimensions $[l \times l]$. These patches are centered on point P_{ij}^M and recorded as $\{x_{ij}^M\}_j^N \in R^{N \times [l \times l]}$. Simultaneously, x_{ij}^{2M} is also segmented with the same center point. x_{ij}^{2M} is the patch sampled at the same center point as x_{ij} under a magnification of $M \times 2$, with a size of $[l \times l]$ (thus, the image covered by this patch has a size of $[l/2 \times l/2]$ under magnification M). Images $\{w_i^M, < x_{i0}^M, x_{i1}^M, x_{i2}^M \ldots >, < x_{i0}^{2M}, x_{i1}^{2M}, x_{i2}^{2M} \ldots >\}$ are incorporated into the developed model as the pathologists’ visual inputs for the area of focus with the goal of learning from expert pathologists’ expertise and aiding in the diagnostic process.

2. Extraction of pathologists’ expertise. The expertise of the pathologists can be described as an attention score based on the pathologists’ manual sampling of WSIs (Figure 4.a), which can also be considered the degree of similarity between any sequence$< P_{i0}^M, P_{i1}^M, P_{i2}^M \ldots\ldots >$ in the WSI and the pathologists' manual sampling at the level of the aggregated image features. We construct an optimal control framework based on the principle of maximum entropy, which essentially posits that the sampled experts' behavior results from random, nearly optimal responses based on an unknown cost function. Specifically, under the expertise extraction model $f_{experience}$, it is assumed that the expert samples the demonstration trajectory τ from a distribution:

$$p(\tau) = \frac{1}{Z} exp\big(-C_\theta(\tau)\big) \tag{3}$$

$\tau_i = \{S_0, S_1 \ldots S_T\}$ can be viewed as the trajectory of the pathologist's ROIs under w_i^M. $C_\theta(\tau) = \sum_0^T S_t$ is an unknown value function parameterized by θ. S_t represents the set of images $\{x_{i\,t}^M, x_{i\,t}^{2M}, w_i^M\}$ at current time t, and $Z = \int exp\big(-C_\theta(\tau)\big) d\tau$ is a partition function. Under this specification, trajectories with higher values have a greater probability of being selected, and while the expert pathologist may select optimal actions, suboptimal actions may also occur.

The randomly sampled trajectory and the subsequent trajectory generated by the imitation model $f_{mimicry}$ are introduced as nonexpert pathologist trajectories $\tau \in D_{samp}$ into this part of the model, which then undergoes adversarial learning with expert trajectories $\tau \in D_{demo}$. In this way, the behavioral trajectories generated by $f_{mimicry}$ are "guided" toward a distribution closer to that of the expert behaviors, and the expertise extraction model $f_{experience}$ acquires the ability to distinguish between the two types of trajectories. $f_{experience}$ takes as input τ_i. The input images are passed through a pretrained encoder to obtain the feature vectors $\{u_i^M, \sum_t^T v_{ij}^M\}$. Here, u_i^M is considered to represent the global information at the window level, and $\sum_t^T v_{ij}^M$ is considered to carry information contained in the transitions among the pathologist's fixation points. v_{it}^M is concatenated with a learnable "pose embedding" vector and passed through a transformer layer, which outputs the second-layer feature vector r_{it}^M at the current time. The r_{it}^M of each moment is concatenated with the global feature u_i^M, and the predicted attention scores $C'_\theta(t)$ are obtained through a multilayer perceptron (MLP) using the mean square error loss:

$$loss_1 = \frac{1}{T-t} \sum_t^T \big(c'_\theta(t) - c_\theta(t)\big)^2 \tag{4}$$

$$\begin{cases} c_\theta(t) = \big(\lambda_1 * e_{time} + \lambda_2 * E_{density}\big) * \beta^{T-t}, \\ \qquad\qquad\qquad\qquad \tau_t \in D_{demo} \\ \quad c_\theta(t) = 0, \qquad \tau_t \in D_{samp} \end{cases} \tag{5}$$

where coefficients λ_1 and λ_2 are used to balance the importance of gaze duration and gaze point density in the region, satisfying $\lambda_1 + \lambda_2 \equiv 1$. β is a number no greater than 1.

We combine $f_{experience}$ based on sampling with $f_{mimicry}$, which is essentially an RL model. The core idea is to optimize the trajectory distribution for the current cost $C_\theta(\tau)$ through $f_{mimicry}$ and to assign higher values to trajectories that are closer to expert behavior. This method allows us to make reverse optimal choices in an infinite state space, even without a known system model.

3. Feature distillation and classification with weakly supervised models. Although basic MIL has been widely applied to WSI classification tasks, numerous studies have shown that attention-based MIL can yield better results. However, although such methods integrate complete pathological image information, they also introduce more interference and do not fully utilize the known information. To address this issue, we designed a feature distillation MIL classification module that utilizes the pathologists' slide-reviewing expertise. This module is orthogonal to various existing MIL methods and can endow them with consistent performance improvements. The network architecture of the classification module is shown in Figure 4.b. For a given WSI W, a series of patches can be obtained through tissue-region and instance-level segmentation. A pretrained image encoder is utilized to extract features from these patches. However, this portion of the model is not involved in the training process; the model designed in this study focuses solely on learning from the extracted features. The patches yield instance features $X = \{x_1, x_2, \ldots, x_K\}$, where K represents the total number of patches contained in W. Each individual patch x_i possesses a latent label y_i that indicates the disease type to which the tissue in x_i belongs and that are unknown to the model. The task of feature distillation is to distill the areas on which pathologists are most likely to focus and that best represent a certain disease, i.e., patches with high c_θ values representing a high probability of belonging to a specific disease. Specifically, the top-k distilled features $\{\hat{o}_1, \ldots \hat{o}_k\}$ satisfy the following relationship:

$$\{\hat{o}_1, \ldots \hat{o}_k\} = argmax(\{c_1 + \hat{y}_1, \ldots c_K + \hat{y}_K\}, k) \tag{6}$$

Here, c_i represents the cost value corresponding to x_i, and $\hat{y}_i$ is the probability of x_i being predicted as belonging to its disease type. The optimization method for $\hat{y}_i$ involves guiding the maximum value of the corresponding disease type in $\{\hat{y}_0, \ldots \hat{y}_K\}$

with the WSI-level label Y, enabling the model to autonomously learn the "patch most likely to belong to a certain disease" during the optimization process:

$$loss_2 = Y * log\big(argmax(\{\hat{y}_0, \dots \hat{y}_K\})\big) + (1 - Y) * log\big(1 - argmax(\{\hat{y}_0, \dots \hat{y}_K\})\big) \qquad (7)$$

However, the true disease type to which x_i belongs is unknown. Therefore, during the feature distillation stage, representative patches of all disease types are packaged together and used for WSI-level classification:

$$\{\hat{o}_1, \dots \hat{o}_k\} = \sum_{Disease\ Types} argmax(\{c_1 + \hat{y}_1, \dots c_K + \hat{y}_K\}, \frac{k}{5})_{Disease\ Type} \qquad (8)$$

$$Disease\ Types = \{Nevus, BCC, Melanoma, SCC, SK\} \qquad (9)$$

Additionally, the k features obtained are transformed into WSI-level features through feature fusion $f_{fusion}$and used for WSI classification. There are numerous suitable feature fusion methods, and so this function can be interchanged with other bag-based MIL approaches. Common methods include feature score weighting [18] or self-attention mechanisms [19].

$$\hat{Y} = f_{mlp}\big[f_{fusion}(\hat{o}_1, \dots \hat{o}_k)\big] \qquad (10)$$

$$loss_3 = Y * log\big(\hat{Y}\big) + (1 - Y) * log\big(1 - \hat{Y}\big) \qquad (11)$$

4. Construction of an RL model to imitate the slide-reviewing behavior of the pathologists. For a given WSI W, an RL [38-40] task can emulate the visual behavior of expert pathologists to conduct a rapid search on a two-dimensional plane and locate areas potentially harboring lesions [41]. As shown in Figure 4.c, the objective is to generate a "human behavior -like" search trajectory.

The trajectory under the window is set to imitate the pathologist's behavior and is regarded as a Markov decision process (MDP). At time t, the agent acquires patch $x_{(it)}$ corresponding to a certain position P_t in w_i (more specifically, the physical location corresponding to a pixel point in the pathological image), and, along with w_i, constitutes the state S_t at time t. Given the irregular characteristics of pathological images, there can be significant variability between WSIs derived from specimens of the same type of tissue. Therefore, the reinforcement learning framework in the context of pathological images can be considered to possess an infinite state space. The action a_t is described as a change in position within w_i, namely, starting from P_t, movement occurs in one of the eight preestablished directions (upper-left, up, upper-right, right, etc.) with a fixed step length l, resulting in a new position P_{t+1}. Based on the expert pathologists' actions discussed in Section 2.1, which are both sequential and continuous, we choose to generate a state-action sequence $\{S_t, S_{t+1} \dots, S_T, a_t, a_{t+1} \dots, a_{T-1}\}$ after repeating this pattern multiple times before assigning a reward sequence $\{R_t, R_{t+1} \dots, R_{T-1}\}$. Note that this is not calculated as a reward during the single $S_t + a_t \Rightarrow S_{t+1}$ process. This approach draws inspiration from the classic RL model DRQN [24] and reflects the complete observational information in the process of pathologists reviewing slices, rather than the simplistic model based on a single patch as the basis for action selection. This RL model integrates global observational information with expertise accumulated prior to the current moment t.

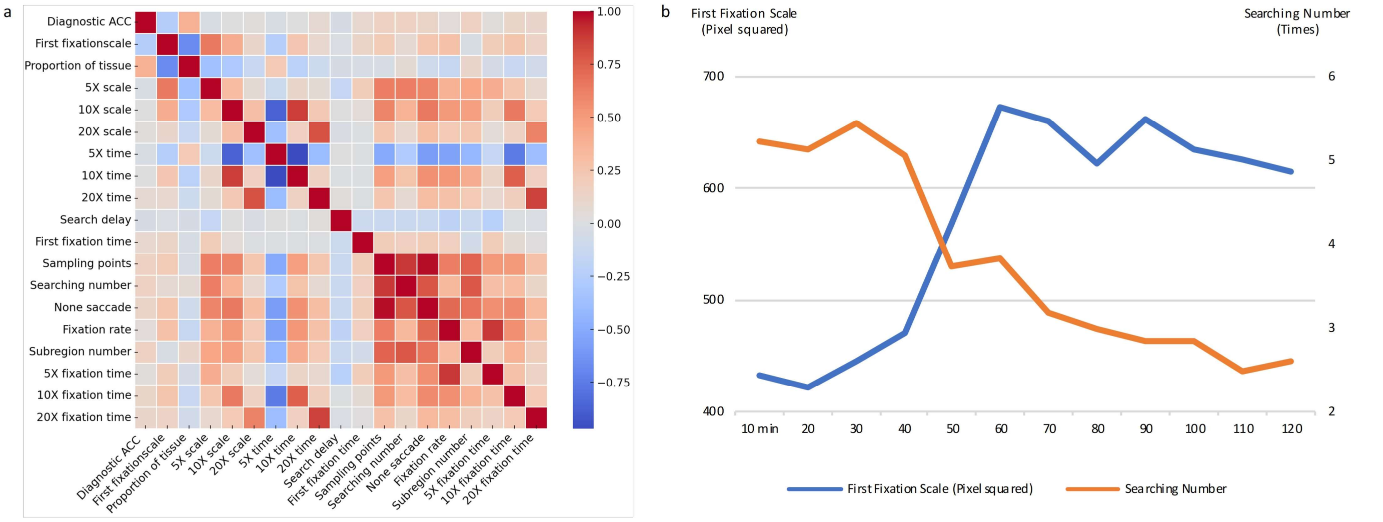
The $c_\theta(t+1)$ obtained from $f_{experience}$ will act as the reward R_t for the process $\{S_t, a_t\} \Rightarrow S_{t+1}$, the purpose of which is to determine the reward size for executing action a_t under the current state S_t based on the value $c_\theta(t+1)$ possessed by the next moment's state S_{t+1}. The RL model uses structurally identical but differently parameterized Q_{eval} and Q_{target} networks. The Q_{eval} network obtains the estimated rewards $\{\widehat{r_{tn}}\}_n^8$ for all actions corresponding to S_t and chooses the action $a_t = argmax(\widehat{r_t})$ with the highest reward. The parameters θ_Q of Q_{eval} are updated in real time, while the parameters $\theta_{\bar{Q}}$ of Q_{target} are assigned by θ_Q after a certain step length of training. a_t is combined with P_t to calculate P_{t+1} and obtain the state S_{t+1} at moment $t+1$, which is then used by f_{IRL} to obtain R_t and thus guide the learning of the RL model. Additionally, during the execution process, the RL model saves the sequence $\{S_t, a_t, R_t, \phi_t\}$ (where ϕ_t indicates whether $t = T$, i.e., whether it is the last item of the continuous state sequence) to the expertise replay pool D_{RL}. The RL model is then randomly sampled from D_{RL} for learning. The Q-network is optimized as follows:

$$r_t = R_{tk} + \gamma * max\left(Q_{target}(S_{t+1})\right) \qquad (12)$$

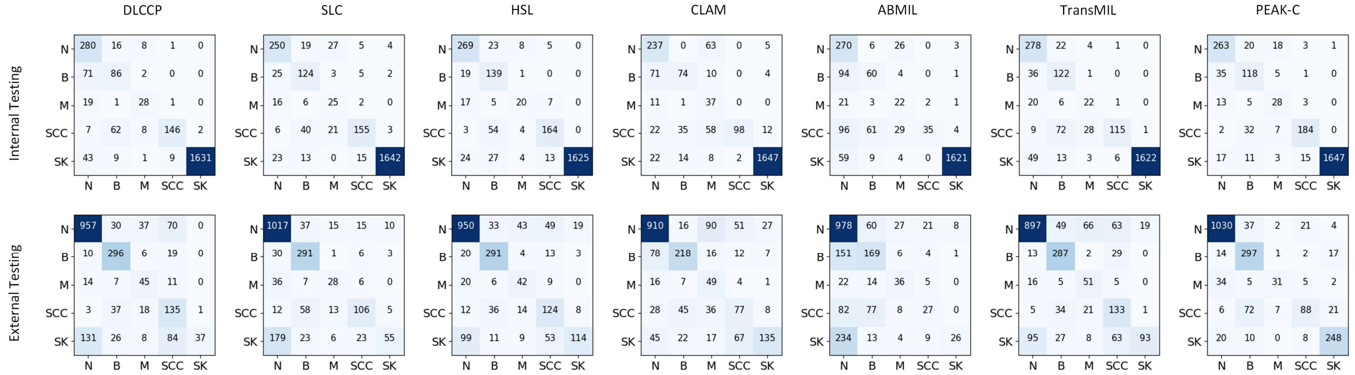
$$n = argmax\big(q_{eval}(s_t)\big) \qquad (13)$$

$$r'_t = max\big(Q_{eval}(S_t)\big) \qquad (14)$$

$$loss_4\big(\theta_Q\big) = \sum_t^T \frac{1}{T-t}(r'_t - t_t)^2 \qquad (15)$$



Appendix Figure 1. a. Correlation analysis of pathologists' slide-reviewing behaviors. **b.** The variation curves of Plots of two key pathologist slide-reviewing features of pathologist over continuous working time.



Internal Testing

DLCCP	N	B	M	SCC	SK
N	280	16	8	1	0
B	71	86	2	0	0
M	19	1	28	1	0
SCC	7	62	8	146	2
SK	43	9	1	9	1631

SLC	N	B	M	SCC	SK
N	250	19	27	5	4
B	25	124	3	5	2
M	16	6	25	2	0
SCC	6	40	21	155	3
SK	23	13	0	15	1642

HSL	N	B	M	SCC	SK
N	269	23	8	5	0
B	19	139	1	0	0
M	17	5	20	7	0
SCC	3	54	4	164	0
SK	24	27	4	13	1625

CLAM	N	B	M	SCC	SK
N	237	0	63	0	5
B	71	74	10	0	4
M	11	1	37	0	0
SCC	22	35	58	98	12
SK	22	14	8	2	1647

ABMIL	N	B	M	SCC	SK
N	270	6	26	0	3
B	94	60	4	0	1
M	21	3	22	2	1
SCC	96	61	29	35	4
SK	59	9	4	0	1621

TransMIL	N	B	M	SCC	SK
N	278	22	4	1	0
B	36	122	1	0	0
M	20	6	22	1	0
SCC	9	72	28	115	1
SK	49	13	3	6	1622

PEAK-C	N	B	M	SCC	SK
N	263	20	18	3	1
B	35	118	5	1	0
M	13	5	28	3	0
SCC	2	32	7	184	0
SK	17	11	3	15	1647

External Testing

DLCCP	N	B	M	SCC	SK
N	957	30	37	70	0
B	10	296	6	19	0
M	14	7	45	11	0
SCC	3	37	18	135	1
SK	131	26	8	84	37

SLC	N	B	M	SCC	SK
N	1017	37	15	15	10
B	30	291	1	6	3
M	36	7	28	6	0
SCC	12	58	13	106	5
SK	179	23	6	23	55

HSL	N	B	M	SCC	SK
N	950	33	43	49	19
B	20	291	4	13	3
M	20	6	42	9	0
SCC	12	36	14	124	8
SK	99	11	9	53	114

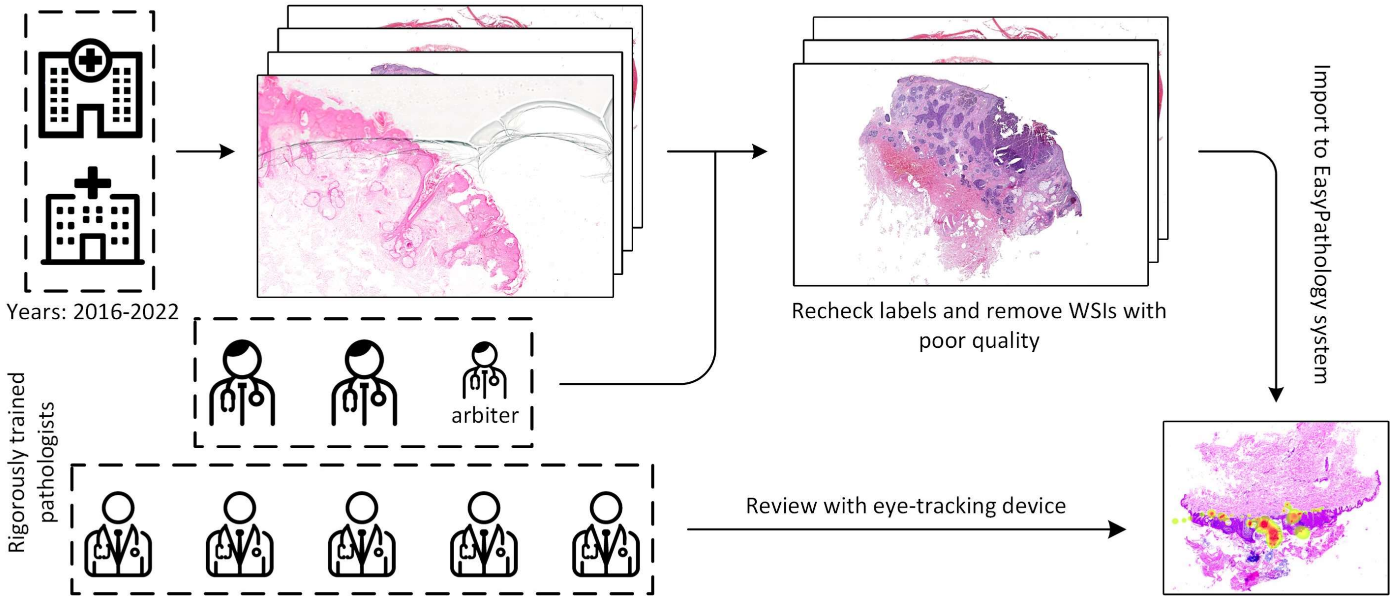
CLAM	N	B	M	SCC	SK
N	910	16	90	51	27
B	78	218	16	12	7
M	16	7	49	4	1
SCC	28	45	36	77	8
SK	45	22	17	67	135

ABMIL	N	B	M	SCC	SK
N	978	60	27	21	8
B	151	169	6	4	1
M	22	14	36	5	0
SCC	82	77	8	27	0
SK	234	13	4	9	26

TransMIL	N	B	M	SCC	SK
N	897	49	66	63	19
B	13	287	2	29	0
M	16	5	51	5	0
SCC	5	34	21	133	1
SK	95	27	8	63	93

PEAK-C	N	B	M	SCC	SK
N	1030	37	2	21	4
B	14	297	1	2	17
M	34	5	31	5	2
SCC	6	72	7	88	21
SK	20	10	0	8	248

Appendix Figure 2. Confusion matrices for the classification models.


Years: 2016-2022
Recheck labels and remove WSIs with poor quality
Import to EasyPathology system
arbiter
Rigorously trained pathologists
Review with eye-tracking device

Appendix Figure 3. The process for dataset collection.

6 Appendices

Appendix 1: EasyPathology

EasyPathology is a proprietary system developed for collecting pathologists' slide-reviewing data. The technical architecture of the EasyPathology project can be categorized into the following layers:

1) Physical Layer:

   Developed in-house to collect multidimensional eye-tracking data, this module is primarily responsible for the real-time capture and recording of pathologists' eye movements during a slide review. It utilizes high-performance eye-tracking technology [42] combined with high-definition video and audio recording devices. Through a unified data collection and processing platform, it enables functionalities such as gaze tracking, video and audio recording, slide-reviewing operation capture, and multidimensional data information recording from pathologists during the slide review process.

2) Data Processing and Analysis Layer:

   By utilizing the brightness and pupil segmentation algorithm of the eye-tracking system, this layer processes the gaze data to track the pathologists' screen gazes. Density-based clustering algorithms are employed for nonrelevant viewpoint removal, optimizing storage efficiency by discarding irrelevant data and improving recognition computational efficiency.

Appendix 2:

Correlation analysis of pathologists' slide-reviewing behaviors and fatigue testing. The various behavioral characteristics were assessed with regard to their correlations with and impact on pathologists' diagnostic accuracy, as depicted in Appendix Figure 1.a. Diagnostic accuracy was defined as the accuracy of the diagnoses given by pathologists in each consecutive slide review session relative to the true diagnoses, to which they were blinded. Among the numerous factors affecting pathologists' diagnostic accuracy, the one with the greatest impact was the proportion of tissue in the WSI (correlation coefficient 0.38). Among the slide-reviewing behaviors of the pathologists, the one with the strongest correlation with accuracy was the initial fixation scale (-0.25), followed by the number of searches (0.15). These results suggest that experienced pathologists tend to focus on the specific areas with a smaller scope during the initial fixation process and engage in more searching behavior to gain deeper insights into the WSI. Additionally, there were significant correlations among most slide-reviewing behavior characteristics; however, most of these indicators had relatively weak correlations with pathologists' diagnostic accuracy.

Based on our correlation analysis of the pathologists' slide-review behavior, we compared the two feature variables, "Initial Fixation Scale" and "Number of Searches", over the continuous slide-reading time during the pathologists' review process. As shown in Appendix Figure 1.b, both of these indicators exhibited substantial changes around the 50-minute mark. Therefore, the pathologists were considered in a "fatigue state" after working continuously for 50 minutes and were encouraged to take breaks to ensure that the collected data were suitably robust.

Appendix 3:

Inclusion Criteria for the Dataset. The process for dataset collection is illustrated in Appendix Figure 3. In this study, data were collected from 2016 to 2022, yielding 3,899 WSIs from Hospital F and 1,982 WSIs from Hospital G. All WSIs were accompanied by final diagnostic results provided by clinical physicians and pathologists at the time of collection. Subsequently, three pathologists conducted a review and screening process. The review involved a reexamination of the labels assigned to all WSIs by two pathologists; WSIs with unanimous labels were directly accepted, while in cases of disagreement, a third pathologist with more extensive experience acted as an "arbiter". The screening process involved excluding WSIs deemed by the pathologists to be of inferior quality, including images containing 1. tissue samples that were too small or had diagnostic areas too limited to adequately represent the entirety of the disease; 2. folded sections; 3. overstaining; 4. high levels of tissue fragmentation; and 5. a significant presence of bubbles.

Following rigorous training, five pathologists reviewed the organized WSIs using the EasyPathology system and eye-tracking device. All five pathologists, identified as P1-P5, possessed relevant qualifications, and had 24, 16, 14, 14, and 10 years, respectively, of experience reviewing WSIs. During the review process, environmental conditions within the laboratory, such as temperature and lighting, were strictly controlled for consistency across the five pathologists. The first five WSIs initially reviewed by each pathologist were considered pretests and were not included in the dataset. Pathologists took a break after 50 minutes of continuous work (as determined by the analysis described in Appendix 1), at which point the eye tracker was recalibrated before the next review session to prevent data distortion. The slide-reviewing data were collected over a total of three months.

References

[1] Toby C Cornish, Ryan E Swapp, and Keith J Kaplan. Whole slide imaging: routine pathologic diagnosis. Advances in Anatomic Pathology, 19(3):152–159, 2012. 1 Litjens, Geert, et al. "Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis." Scientific reports 6.1 (2016): 26286. DOI: 10.1097/PAP.0b013e318253459e

[2] Madabhushi, Anant. "Digital pathology image analysis: opportunities and challenges." Imaging in medicine 1.1 (2009): 7. DOI: 10.2217/IIM.09.9

[3] Pantanowitz, Liron, et al. "Review of the current state of whole slide imaging in pathology." Journal of pathology informatics 2.1 (2011): 36. DOI: https://doi.org/10.4103/2153-3539.83746

[4] LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." nature 521.7553 (2015): DOI: 436-444. https://doi.org/10.1038/nature14539

[5] Guo, Yanming, et al. "Deep learning for visual understanding: A review." Neurocomputing 187 (2016): 27-48. DOI: https://doi.org/10.1016/j.neucom.2015.09.116.

[6] Pinckaers, Hans, Bram Van Ginneken, and Geert Litjens. "Streaming convolutional neural networks for end-to-end learning with multi-megapixel images." IEEE transactions on pattern analysis and machine intelligence 44.3 (2020): 1581-1590. DOI: 10.1109/TPAMI.2020.3019563. 10.1109/TPAMI. 2020.3019563

[7] Tellez, David, et al. "Neural image compression for gigapixel histopathology image analysis." IEEE transactions on pattern analysis and machine intelligence 43.2 (2019): 567-578. DOI: 10.1109/TPAMI.2019.2936841

[8] Li, Jiahui, et al. "Hybrid supervision learning for pathology whole slide image classification." International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer International Publishing, 2021. DOI: https://doi.org/10.1007/978-3-030-87237-3_30

[9] Korbar, Bruno, et al. "Deep learning for classification of colorectal polyps on whole-slide images." Journal of pathology informatics 8.1 (2017): 30. DOI: https://doi.org/10.4103 /jpi.jpi_34_17

[10] Ghosh, Arna, Satyarth Singh, and Debdoot Sheet. "Simultaneous localization and classification of acute lymphoblastic leukemic cells in peripheral blood smears using a deep convolutional network with average pooling layer." 2017 IEEE international conference on industrial and information systems (ICIIS). IEEE, 2017. DOI: 10.1109/ICIINFS.2017.8300425

[11] Campanella, Gabriele, et al. "Clinical-grade computational pathology using weakly supervised deep learning on whole slide images." Nature medicine 25.8 (2019): 1301-1309. DOI: https://doi.org/10.1038/s41591-019-0508-1

[12] B. E. Bejnordi et al., "Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer," Jama, vol. 318, no. 22, pp. 2199-2210, 2017. doi:10.1001/jama.2017.14585

[13] Chen, Zenghai, et al. "Multi-instance multi-label image classification: A neural approach." Neurocomputing 99 (2013): 298-306. DOI: https://doi.org/10.1016/j.neucom.2012.08.001

[14] Chen, Richard J., et al. "Multimodal co-attention transformer for survival prediction in gigapixel whole slide images." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021. DOI:10.1109/iccv48922.2021.00398

[15] Yao, Jiawen, et al. "Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks." Medical Image Analysis 65 (2020): 101789. DOI: https://doi.org/10.1016/j.media.2020. 101789

[16] Ilse, Maximilian, Jakub Tomczak, and Max Welling. "Attention-based deep multiple instance learning." International conference on machine learning. PMLR, 2018. DOI: 10.48550/arXiv.1802.04712

[17] Shao, Zhuchen, et al. "Transmil: Transformer based correlated multiple instance learning for whole slide image classification." Advances in neural information processing systems 34 (2021): 2136-2147. DOI: https://doi.org/10.48550/arXiv.2106.00908

[18] Zhang, Hongrun, et al. "DTFD-MIL: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022. DOI: https://doi.org/10.48550/arXiv.2203.12081

[19] Lin, Tiancheng, et al. "Interventional bag multi-instance learning on whole-slide pathological images." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023. DOI: https://doi.org/10.48550/ arXiv.2303.06873

[20] Lu, Ming Y., et al. "Data-efficient and weakly supervised computational pathology on whole-slide images." Nature biomedical engineering 5.6 (2021): 555-570. DOI: https://doi.org/10.1038/s41551-020-00682-w

[21] Li, Shaohua, et al. "Multi-instance multi-scale CNN for medical image classification." Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part IV 22. Springer International Publishing, 2019. DOI: https://doi.org/10.1007/978-3-030-32251-9_58

[22] Hou, Le, et al. "Patch-based convolutional neural network for whole slide tissue image classification." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016. DOI: https://doi.org/10.48550/arXiv.1504.07947

[23] Kanavati, Fahdi, et al. "Weakly-supervised learning for lung carcinoma classification using deep learning." Scientific reports 10.1 (2020): 9297. DOI: https://doi.org/10.1038/s41598-020-66333-x

[24] Lerousseau, Marvin, et al. "Weakly supervised multiple instance learning histopathological tumor segmentation." Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part V 23. Springer International Publishing, 2020. DOI: https://doi.org/10.1007/978-3-030-59722-1_45

[25] P ̈aivi Majaranta and Andreas Bulling. "Eye tracking and eye-based human–computer interaction". In: Advances in physiological computing. Springer, 2014, pp. 39–65. DOI: https://doi.org/10.1007/978-1-4471-6392-3_3

[26] Xucong Zhang et al. "MPIIGaze: Real-World Dataset and Deep Appearance-Based Gaze Estimation". In: IEEE Transactions on Pat- tern Analysis and Machine Intelligence 41.1 (2019), pp. 162–175. DOI: 10.1109/TPAMI.2017. 2778103

[27] Yusuke Sugano, Xucong Zhang, and Andreas Bulling. "Aggregaze: Collective estimation of audience attention on public displays". In: Proceedings of the 29th Annual Symposium on User Interface Software and Technology. 2016, pp. 821–831. DOI: https://doi.org/10.1145/2984511.2984536

[28] Hosnieh Sattar et al. "Prediction of search targets from fixations in open-world settings". In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015, pp. 981–990. DOI: https://doi.org/10.48550/arXiv.1502.05137

[29] Matthias K ̈ummerer, Lucas Theis, and Matthias Bethge. "Deep Gaze I: Boosting Saliency Prediction with Feature Maps Trained on ImageNet". In: CoRR abs/1411.1045 (2014). DOI: https://doi.org/10.48550/arXiv.1411.1045

[30] Matthias K ̈ummerer, Thomas S. A. Wallis, and Matthias Bethge. "DeepGaze II: Reading fixations from deep features trained on object recognition". DOI: https://doi.org/10.48550/arXiv.1610.01563

[31] Matthias K ̈ummerer, Thomas S. A. Wallis, and Matthias Bethge. "DeepGaze III: Using Deep Learning to Probe Interactions Between Scene Content and Scanpath History in Fixation Selection". In: 2019 Conference on Cognitive Computational Neuroscience (2019)

[32] Komal Mariam et al. "On Smart Gaze Based Annotation of Histopathology Images for Training of Deep Convolutional Neural Networks". In: IEEE Journal of Biomedical and Health Informatics 26 (2022), pp. 3025–3036. DOI: 10.1109/JBHI.2022.3148944

[33] Tad T. Bruny ́e et al. "Eye tracking reveals expertise-related differences in the time-course of medical image inspection and diagnosis". In: Journal of Medical Imaging 7 (2020), pp. 051203–051203. DOI: https://doi.org/10.1117/1. JMI.7.5.051203

[34] Theckedath, Dhananjay, and R. R. Sedamkar. "Detecting affect states using VGG16, ResNet50 and SE-ResNet50 networks." SN Computer Science 1 (2020): 1-7. DOI: https://doi.org/10.1007/s42979-020-0114-9

[35] Deng, Jia, et al. "Imagenet: A large-scale hierarchical image database." 2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009. DOI: 10.1109/CVPR.2009.5206848

[36] Kaelbling, Leslie Pack, Michael L. Littman, and Andrew W. Moore. "Reinforcement learning: A survey." Journal of artificial intelligence research 4 (1996): 237-285. DOI: https://doi.org/10.1613/jair.301

[37] Arulkumaran, Kai, et al. "Deep reinforcement learning: A brief survey." IEEE Signal Processing Magazine 34.6 (2017): 26-38. DOI: 10.1109/MSP.2017.2743240

[38] Barata, C., Rotemberg, V., Codella, N.C.F. et al. A reinforcement learning model for AI-based decision support in skin cancer. Nat Med 29, 1941–1946 (2023). DOI: https://doi.org/10.1038/s41591-023-02475-5

[39] Zhao, Boxuan, et al. "RLogist: fast observation strategy on whole-slide images with deep reinforcement learning." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 37. No. 3. 2023. DOI: https://doi.org/10.1609/ aaai.v37i3.25467

[40] " Tobii Pro Spectrum." URL: https://www.tobii.com/ products/eye-trackers/screen-based/tobii-pro-spectrum

Authors

Tianhang Nan[1], Song Zheng[2, 3], Siyuan Qiao[4], Hao Quan[1], Jun Niu[5], Bin Zheng[1], Chunfang Guo[6], Yue Zhang[7], Xiaoqin Wang[7], Liping Zhao[8], Ze Wu[5], Yaoxing Guo[2, 3], Xingyu Li[1], Mingchen Zou[1], Shuangdi Ning[1], Yue Zhao[1], Wei Qian[1], Hongduo Chen[2, 3], Ruiqun Qi*[2, 3], Xinghua Gao*[2, 3], Xiaoyu Cui*[1]

1. College of Medicine and Biological Information Engineering, Northeastern University, Shenyang, China
2. Department of Dermatology, The First Hospital of China Medical University, Shenyang, China.
3. Key Laboratory of Immunodermatology, Ministry of Education, and National Health Commission; National Joint Engineering Research Center for Theranostics of Immunological Skin Diseases, Shenyang, China
4. College of Computer Science and Technology, Fudan University, Shanghai, China
5. Department of Dermatology, General Hospital of Northern Theater Command, Shenyang, China
6. Department of Dermatology, Shenyang Seventh People's Hospital, Shenyang, China
7. Department of Dermatology, Shengjing hospital of China Medical University, Shenyang, China
8. Department of Dermatology, Zhongyi Northeast International Hospital, Shenyang, China

The authors Tianhang Nan and Song Zheng have the same contribution.
The authors marked with * (Ruiqun Qi, Xinghua Gao, Xiaoyu Cui) are corresponding authors.

Contributors: A. Conceived and designed the experiments; B. Performed the experiments; C. Analyzed the data; D. Contributed materials/analysis tools; E. Wrote the paper

1. Tianhang Nan, contributed A, B, C, D and E.
   tianhang_nan@foxmail.com
   College of Medicine and Biological Information Engineering, Northeastern University, Shenyang, China
2. Song Zheng, contributed A, B, C, D and E.
   nyaadzs@163.com
   Department of Dermatology, The First Hospital of China Medical University, Shenyang, China.
   Key Laboratory of Immunodermatology, Ministry of Education, and National Health Commission; National Joint Engineering Research Center for Theranostics of Immunological Skin Diseases, Shenyang, China
3. Siyuan Qiao, contributed A, C and D.
   23210240038@m.fudan.edu.cn
   College of Computer Science and Technology, Fudan University, Shanghai, China
4. Hao Quan, contributed A, C and D.
   quan@siat.ac.cn
   College of Medicine and Biological Information Engineering, Northeastern University, Shenyang, China
5. Jun Niu, contributed A, C and D.
   niujun06@126.com
   Department of Dermatology, General Hospital of Northern Theater Command, No. 83 Wenhua Road, Shenhe District, 110016 Shenyang, China
6. Bin Zheng, contributed A, C and E.
   bin.zheng@myyahoo.com
   College of Medicine and Biological Information Engineering, Northeastern University, Shenyang, China

7. Chunfang Guo, contributed A and D.
   1211541310@qq.com
   Shenyang Seventh People's Hospital, Shenyang, China
8. Yue Zhang, contributed A and D.
   zhangyue-80@qq.com
   Department of Dermatology, Shengjing hospital of China Medical University, Shanghai, China
9. Xiaoqin Wang, contributed A and D.
   343442083@qq.com
   Department of Dermatology, Shengjing hospital of China Medical University, Shanghai, China
10. Liping Zhao, contributed A and D.
    2608917317@qq.com
    Zhongyi Northeast International Hospital, Shenyang, China
11. Ze Wu, contributed A, C and D.
    wuze15841391120@163.com
    Department of Dermatology, General Hospital of Northern Theater Command, No. 83 Wenhua Road, Shenhe District, 110016 Shenyang, China
12. Yaoxing Guo, contributed A, C and D.
    yxguo@cmu.edu.cn
    Department of Dermatology, The First Hospital of China Medical University, Shenyang, China.
    Key Laboratory of Immunodermatology, Ministry of Education, and National Health Commission; National Joint Engineering Research Center for Theranostics of Immunological Skin Diseases, Shenyang, China
13. Xingyu Li, contributed A, B and D.
    whiteli1925@outlook.com
    College of Medicine and Biological Information Engineering, Northeastern University, Shenyang, China
14. Mingchen Zou, contributed A, B and D.
    1936891494@qq.com
    College of Medicine and Biological Information Engineering, Northeastern University, Shenyang, China
15. Shuangdi Ning, contributed A and D.
    shundyning@outlook.com
    College of Medicine and Biological Information Engineering, Northeastern University, Shenyang, China
16. Yue Zhao, contributed A, C and D.
    zhaoyue@bmie.neu.edu.cn
    College of Medicine and Biological Information Engineering, Northeastern University, Shenyang, China
17. Wei Qian, contributed A, C and E.
    wqian@bmie.neu.edu.cn
    College of Medicine and Biological Information Engineering, Northeastern University, Shenyang, China
18. Hongduo Chen, contributed A, C and E.
    hongduochen@hotmail.com
    Department of Dermatology, The First Hospital of China Medical University, Shenyang, China.
    Key Laboratory of Immunodermatology, Ministry of Education, and National Health Commission; National Joint Engineering Research Center for Theranostics of Immunological Skin Diseases, Shenyang, China

19. *Ruiqun Qi, contributed A, B, C, D and E.
xiaoqiliumin@163.com
Department of Dermatology, The First Hospital of China Medical University, Shenyang, China.
Key Laboratory of Immunodermatology, Ministry of Education, and National Health Commission; National Joint Engineering Research Center for Theranostics of Immunological Skin Diseases, Shenyang, China

20. *Xinghua Gao, contributed A, B, C, D and E.
gaobarry@hotmail.com
Department of Dermatology, The First Hospital of China Medical University, Shenyang, China.
Key Laboratory of Immunodermatology, Ministry of Education, and National Health Commission; National Joint Engineering Research Center for Theranostics of Immunological Skin Diseases, Shenyang, China

21. *Xiaoyu Cui, contributed A, B, C, D and E.
cuixy@bmie.neu.edu.cn
College of Medicine and Biological Information Engineering, Northeastern University, Shenyang, China