nature portfolio

Peer Review File

Decoding Expertise from Pathologists' Diagnostic Processes on Whole Slide Images

Corresponding Author: Professor Xiaoyu Cui

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
Title: Decoding Expertise from Pathologists' Diagnostic Processes on Whole Slide Images
Dear Authors and the Editor,

I have critically and carefully analyzed the manuscript titled "Decoding Expertise from Pathologists' Diagnostic Processes on Whole Slide Images. The manuscript presents a novel approach to enhancing deep learning models for pathology by leveraging eye-tracking data to capture pathologists' diagnostic processes. The proposed Pathology Expertise Acquisition Network (PEAN) aims to use this data to improve the accuracy and efficiency of diagnostic models for whole slide images (WSIs). The study claims superior performance over traditional fully supervised and weakly supervised learning methods.

Detailed Review:

I have fundamental concerns about the hybrid approach proposed in this paper. The reliance on pathologists' eye-tracking data to train diagnostic models presents significant challenges in terms of scalability, reliability, and practicality. The inherent variability and subjectivity in human visual inspection raise questions about the consistency and accuracy of the resulting models. Moreover, the resource-intensive nature of continuous eye-tracking data collection makes this approach impractical for large-scale implementation. Compared to weakly supervised learning methods, which offer greater scalability, flexibility, and robustness, the proposed method appears limited and potentially less effective in diverse clinical settings. These fundamental issues cast doubt on the viability and long-term value of the proposed approach.

Details on my concerns:

1. Conceptual Flaws:

1.1 Fundamental Dependence on Pathologist Gaze:

The central hypothesis of this study is that pathologists' gaze data, collected through eye-tracking, can be utilized to train deep learning models, thereby encapsulating their diagnostic expertise. However, this approach is intrinsically flawed due to the inherent variability and subjectivity involved in human visual inspection. Pathologists can and do make errors, and their gaze patterns may not consistently target the most diagnostically relevant areas. This reliance on human gaze data introduces a significant risk: the model may learn and perpetuate these human errors. Consequently, this could compromise the overall reliability and safety of the diagnostic system, as the model may inherit and propagate inaccuracies present in the pathologists' visual inspection process. In contrast, artificial intelligence (AI) models, particularly those employing fully supervised or weakly supervised learning, do not mimic human behavior but rather analyze data SYSTEMATICALLY and COMPREHENSIVELY. Unlike pathologists who might focus on specific regions of a slide based on their expertise and experience, AI models can evaluate the entire slide without any inherent biases. This comprehensive analysis ensures that all potential diagnostic features are considered, thereby reducing the risk of missing critical diagnostic information. Moreover, AI approaches leverage vast amounts of data to learn patterns and make predictions, improving their accuracy over time. They are not limited by human constraints such as fatigue, variability in skill level, or individual biases. AI models can be trained on diverse datasets, which enhances their ability to generalize across different types of pathology and patient populations. This systematic and unbiased approach makes AI inherently more reliable and consistent compared to

methods that depend on human gaze patterns. Additionally, AI models are capable of identifying subtle patterns and correlations that may not be apparent to human observers. This capability is particularly valuable in medical diagnostics, where early detection of diseases can significantly improve patient outcomes. By systematically analyzing the entire slide, AI models ensure a thorough examination, which is critical for accurate and timely diagnosis. In summary, while the use of pathologist gaze data introduces significant risks due to potential human errors and biases, AI approaches provide a more reliable and comprehensive method for diagnostic analysis. AI models systematically analyze the entire slide, ensuring that no critical information is overlooked, thereby enhancing the accuracy, reliability, and safety of diagnostic systems.

1.2 Scalability and Practicality:

The requirement to continuously collect eye-tracking data from pathologists is both impractical and resource-intensive. This approach is not scalable in real-world clinical settings where the deployment of AI systems needs to be efficient, cost-effective, and minimally disruptive. Continuously tracking and recording the gaze patterns of pathologists demands specialized equipment, ongoing data collection, and significant manual effort to annotate and process the data. These requirements introduce substantial logistical and financial burdens. In contrast, AI models that leverage fully supervised or weakly supervised learning approaches are inherently more scalable and practical. These models rely on large, pre-existing datasets that can be annotated with high-level labels, such as diagnostic outcomes, without the need for detailed manual input. This makes the data collection process much more efficient and less costly. As medical data continues to grow exponentially, AI models can continuously improve by incorporating new data, thus enhancing their performance over time without requiring additional specialized data collection efforts. Moreover, AI systems designed for scalability can be easily integrated into existing clinical workflows. They do not require the ongoing participation of pathologists for data collection, allowing medical professionals to focus on patient care rather than being burdened by additional tasks. This integration is crucial for widespread adoption, as it minimizes disruptions and maximizes the utility of AI systems in clinical practice.

2. Methodological Concerns:

2.1 Limited Generalization:

The study focuses on specific skin conditions, and it is unclear how well the proposed method would generalize to other types of pathology or medical imaging domains. The heavy dependence on data from a limited number of pathologists further questions the robustness and adaptability of the model to broader applications.

2.2 Marginal Performance Gains:

The reported performance improvements (AUC of 0.984 and accuracy of 93.1% on the internal test set) over fully supervised and weakly supervised learning methods are not sufficiently substantial to justify the added complexity and cost of incorporating eye-tracking data. Given the rapid advancements in weakly supervised learning, which can leverage large-scale, automatically collected datasets consisting of data coming from millions of patients, the marginal gains presented here may not hold long-term competitive value.

2.3 Interpretability Issues:

While the method aims to enhance interpretability by mimicking pathologists' gaze patterns, the decision-making process of the model remains complex and opaque. The study does not provide sufficient mechanisms to ensure that the learned gaze patterns are indeed the most diagnostically relevant, thereby limiting the transparency and clinical trustworthiness of the model. This opacity is problematic, as clinicians need to understand and trust AI recommendations to integrate them effectively into their diagnostic workflow. In comparison, weakly supervised learning methods often provide greater explainability. These methods can leverage high-level labels, such as diagnostic outcomes, which are more straightforward and interpretable. By using broader annotations, weakly supervised models can identify and learn relevant patterns and features directly associated with diagnostic outcomes. This direct association can enhance the transparency of the model, making it easier for clinicians to understand how the AI arrived at its conclusions. Furthermore, weakly supervised methods can incorporate various techniques to improve explainability, such as attention mechanisms and feature importance mapping. These techniques allow the model to highlight the regions of the image that contributed most to its decision, providing a visual and interpretable explanation of its diagnostic process. This level of transparency is critical for gaining clinical trust and facilitating the adoption of AI tools in healthcare. Additionally, the flexibility and scalability of weakly supervised methods mean they can be applied across a wide range of medical imaging tasks without the need for specialized and continuous data collection efforts. They can adapt to different datasets and clinical contexts, maintaining their robustness and reliability while also providing interpretable results.

In summary, while the proposed method attempts to improve interpretability by mimicking pathologists' gaze patterns, it falls short in ensuring transparency and clinical trustworthiness. Weakly supervised learning methods offer a more practical, scalable, and explainable alternative, making them better suited for real-world clinical applications. These methods can leverage broader, more interpretable data and incorporate techniques to enhance transparency, ultimately providing more reliable and trustworthy diagnostic support.

3. Practical Considerations:

3.1 Resource Intensiveness:

The process of collecting and annotating eye-tracking data is labor-intensive and time-consuming. The manuscript underestimates the practical challenges and costs associated with this approach, which include the need for specialized equipment, ongoing data collection, and extensive manual annotation.

3.2 Risk of Propagating Errors:

By training the model to mimic pathologists' gaze patterns, there is a significant risk of the model learning and replicating the diagnostic errors made by the pathologists. This raises serious concerns about the safety and reliability of the system in clinical practice, where errors can have critical consequences.

More Detailed Comparison with Weakly Supervised Methods:

4.1 Scalability:
Weakly supervised learning methods are inherently more scalable compared to the proposed approach of using eye-tracking data from pathologists. These methods can leverage large amounts of readily available data with minimal annotation, such as diagnostic outcomes or general labels, which significantly reduces the need for detailed and labor-intensive manual annotations. This minimal annotation requirement allows for the continuous and automated expansion of datasets, as new patient data can be incorporated into the training process with minimal additional effort.

This scalability is crucial for the widespread adoption of AI in clinical practice. As the volume of medical data grows, weakly supervised learning methods can efficiently handle the influx of new data, ensuring that the AI models remain up-to-date and capable of learning from the latest clinical information. This ability to scale effortlessly with increasing data volumes makes weakly supervised methods particularly well-suited for large healthcare institutions and research centers that generate vast amounts of data daily.

Moreover, weakly supervised methods enable the development of robust models that can generalize well across different patient populations and medical conditions. By training on diverse datasets, these models can learn to recognize patterns and anomalies that are consistent across various contexts, thereby improving their diagnostic accuracy and reliability. This generalizability is essential for creating AI systems that can be deployed in a wide range of clinical settings, from large hospitals to smaller clinics, without requiring extensive customization or retraining.

In contrast, the proposed approach of using eye-tracking data from pathologists faces significant scalability challenges. The continuous collection of gaze data is resource-intensive, requiring specialized equipment and significant manual effort to process and annotate the data. This approach is not feasible for large-scale implementation, as it imposes substantial logistical and financial burdens. Additionally, the reliance on a small number of pathologists for data collection limits the diversity of the training data, potentially reducing the model's ability to generalize across different clinical settings.

In summary, weakly supervised learning methods offer a highly scalable and practical solution for the development of AI diagnostic models. They leverage large, minimally annotated datasets to continuously improve and generalize well across diverse clinical environments. This scalability makes them far more suitable for widespread adoption in healthcare compared to the resource-intensive and limited approach of using pathologists' eye-tracking data.

4.2 Cost-Effectiveness:

Weakly supervised methods significantly reduce the cost associated with data annotation. By using high-level labels (e.g., diagnostic outcomes) instead of detailed pixel-wise annotations, these methods minimize the need for extensive manual input, thereby conserving valuable medical resources and reducing operational costs.

4.3 Robustness and Generalization:

Models trained with weakly supervised learning are often more robust and generalizable due to their exposure to a broader range of data variations. Unlike approaches that rely on detailed manual annotations or specific data collection methods like eye-tracking, weakly supervised learning methods can utilize large and diverse datasets that represent a wide array of clinical scenarios and patient demographics. This extensive exposure enables the models to learn more generalized and representative features, which enhances their ability to perform accurately across different datasets and clinical settings.

The inherent variability in medical data, including differences in patient populations, imaging techniques, and disease presentations, is better captured by weakly supervised learning methods. By training on such diverse datasets, these models develop a more comprehensive understanding of the underlying patterns and nuances associated with various medical conditions. This comprehensive learning process improves the model's robustness, making it less sensitive to specific biases or anomalies present in any single dataset.

Furthermore, weakly supervised models can leverage high-level labels, such as diagnostic outcomes, which are easier to obtain and less prone to subjective interpretation errors compared to detailed annotations. This reliance on broader labels reduces the risk of overfitting to specific features that may not be generalizable across different clinical environments. As a result, these models can maintain high performance even when applied to new and unseen data, ensuring their utility in real-world medical practice.

In contrast, the proposed approach of using pathologists' eye-tracking data is limited by the variability and subjectivity

inherent in human visual inspection. Pathologists may focus on different regions of interest based on their individual experiences and biases, leading to inconsistencies in the training data. These inconsistencies can negatively impact the model's ability to generalize, as it may learn to rely on specific patterns that do not hold true across different clinical settings.

Additionally, the limited scope of data collected from a small group of pathologists further restricts the model's exposure to diverse diagnostic scenarios. This narrow focus can result in a model that performs well in the specific context in which it was trained but fails to adapt to broader applications. The lack of generalization is a significant drawback, as it undermines the model's reliability and effectiveness when deployed in varied clinical environments.

In summary, weakly supervised learning methods offer superior robustness and generalization due to their ability to train on large, diverse datasets and leverage high-level labels. This comprehensive exposure to a wide range of data variations enables these models to perform consistently across different clinical settings, making them a more reliable and practical choice for AI-assisted diagnostics compared to the limited and subjective approach of using eye-tracking data from pathologists.

4.4 Flexibility:

Weakly supervised learning methods offer a high degree of flexibility, making them easily adaptable to various types of medical imaging and pathology without the need for specialized data collection procedures, such as eye-tracking. This inherent flexibility significantly enhances their versatility and applicability across a wide range of diagnostic tasks.

One of the key advantages of weakly supervised methods is their ability to utilize broadly available high-level labels, such as diagnostic outcomes, which are common across many medical disciplines. These labels do not require intricate and time-consuming manual annotations, allowing the models to be trained on existing datasets with minimal additional effort. This adaptability means that weakly supervised models can be quickly repurposed and applied to different types of medical imaging, including radiology, histopathology, and ophthalmology, among others.

The lack of dependence on specialized data collection processes like eye-tracking further enhances the practicality of weakly supervised methods. Eye-tracking requires specific equipment and setup, as well as ongoing participation from pathologists to gather data. This not only adds complexity and cost but also limits the approach's applicability to settings where such resources are available. In contrast, weakly supervised methods can be deployed using standard medical imaging data, which is readily accessible in most clinical environments.

Moreover, weakly supervised models can easily integrate new types of medical data as they become available. This capability is particularly valuable in a rapidly evolving field like medical diagnostics, where new imaging techniques and modalities are continually being developed. Weakly supervised methods can be updated with these new data types without extensive reengineering, ensuring that the models remain relevant and effective over time.

The flexibility of weakly supervised learning also extends to its potential for hybrid approaches. These methods can be combined with other learning paradigms, such as semi-supervised or self-supervised learning, to further enhance their performance and adaptability. This combinatory approach allows for the leveraging of both labeled and unlabeled data, maximizing the utility of all available information and improving the robustness of the models.

In summary, the flexibility of weakly supervised learning methods makes them exceptionally versatile and applicable to a wide array of diagnostic tasks. Their ability to utilize readily available data without the need for specialized collection procedures allows for quick adaptation to different types of medical imaging and pathology. This versatility, combined with the potential for hybrid approaches, positions weakly supervised methods as a more practical and broadly applicable solution for AI-assisted diagnostics compared to the constrained and resource-intensive approach of using eye-tracking data from pathologists.

4.5 Future-Proofing:

As the availability of medical data continues to grow, weakly supervised methods are well-positioned to take advantage of this trend. They can continuously improve as more data becomes available, ensuring that their performance keeps pace with advancements in data collection and annotation technologies.

Conclusion and Recommendation:

In conclusion, while the manuscript presents an innovative approach to integrating human expertise into AI models, the fundamental flaws in the core concept, scalability issues, limited generalization, and practical challenges outweigh the potential benefits. The marginal performance gains do not justify the added complexity and resource requirements. Additionally, the risk of perpetuating diagnostic errors further undermine the viability of the approach.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)
In this paper the authors developed a deep learning system, named Pathology Expertise Acquisition Network (PEAN), to decode pathologists' expertise from eye movements recorded while they are looking at digital pathology images. This learned pathologists' expertise can then be used for precise diagnostics of whole slide image assisted diagnostics and thus reduce manual annotations effort for pathologists. The results show that there approach reduces the manual annotation time of the digital pathology images to 4%, with an average time of 36.5 seconds per annotation. Therefore, the effort of cumbersome and time consuming manual annotation of digital pathology images can be reduced significantly.

Overall review comments:

Although the paper addresses an important topic (reduction of time-consuming manual annotation times) and the chosen approach of learning from eye movement data is a sensible and promising approach, a major revision of the paper seems sensible for the following reasons:

- Eye Tracking (Chapter 5, 1. Data acquisition and preprocessing):
Although important parameters appear to be calculated, it is not possible to trace exactly what has been done, as:
1.) Some parameters are not explained, e.g. gaze area vs. gaze region, calculation of angular velocity)
2.) how exactly the calculated data is interpreted
3) the use of terms is sometimes confusing (e.g. eye tracking is referred to as recorded points, whereas sample data is probably meant here)
4) The setup, design and implementation of the eye tracking experiment and the eye tracker used, as well as its positioning and calibration, are also not described or shown

The same holds for the section 2 (Extraction of pathologists' expertise) and section 3 (Feature distillation and classification with weakly supervised models:

Here too, the initial situation and objectives should be described before the details are discussed. All designators used should also be explained.

The work should also be discussed in relation to existing approaches, such as :
Deane, O., Toth, E. & Yeo, SH. Deep-SAGA: a deep-learning-based system for automatic gaze annotation from eye-tracking data. Behav Res 55, 1372–1391 (2023). https://doi.org/10.3758/s13428-022-01833-4

There is no critical discussion and Future work addressed in the paper

Review comments in detail:

p- 1, l. 10+11: why AUC and seconds per annotation as time costs - why not difference to human annotators and overall annotation time? Additionally, the authors do define AUC on page 5, line 23

p 1, line 22+23: (A single WSI ... level [10]) -> it is unusual to have a whole sentence in brackets. Same on p. 2., l. 12+13.
p 1, line 24+25: Through fine guidance ... precise diagnostics[11]. -> are there precise numbers that can be reported?
p.2, l. 16: what exactly do the authors mean by patch? - Show an example what this patch is on the WSI image
p.2, l. 19+20: pixelwise annotated map, the pathologist’s visual attention map, the expertise value heatmap, and the suspicious region map selected -> the authors should provide more details how they calculated the different maps. For the heatmaps it is also important to report the parameters used to calculate the maps, also if they considered fixation durations - depending on the way and parameters chosen, the heatmaps look different and they are difficult to compare.
p.2, l. 23: Internal and external test set - difference should be explained at first usage and not in line 42+43 - why is hospital G data considered as external data?

p.2, l. 24: Its classification performance ... -> provide more details in how far the authors' methods outperform existing ones
p.2., l. 27+28: PEAN provides a human-like step-by-step search - is it because it follows the pathologists' scan path for analysis?
p.2., l. 29 + 30: provide more details to the statistics, also report the full results of the statistical tests, not only the p value.

p.3., Figure 1, 1e: the heatmaps for BCC and Melanoma look different to the others. It seems to be that there were quite a few fixations on the tumor - e.g. in BCC seems only 3 fixations.

p. 2, l. 15: Heading of Chapter 2 - Results - This heading does not fit the content of section 1 because it is more on methodological descriptions. The subsection 2-6 are more on results. Better to separate this.

p.2., l. 46: “EasyPathology” -> the used eye tracking system is not sufficiently described - even in Appendix 1. Which eye tracking model was used (e.g. vendor, head mounted, remote, mobile) what was the sampling rate, accuracy - how was the system prepared, set-up- calibrated? What was the distance of the participants to the screen as well as the size of the screen, how were they instructed, did they decide when a new WSI was displayed - or was there a fixed time period for each WSI. Were there re-calibrations during the experimental procedure? The reference [42] mentioned in Appendix 1 is not in the references (p. 14+15).

p.4, l. 8: we determined that data could be collected continiously for 50 minutes a time (detailed in Appendix 2) - Appendix 2 does not really explain how "Initial Fixation Scale" and "Number of searches" were measured and analyzed. Also there is a reference to Figure 1.b (Appendix 2, l. 28) which shows this data as a graphic - the figure can not be found in the paper. Also it is not clear how authors came to the numerous factors affecting pathologists' diagnostic accuracy, nor how they came to the correlation coefficients.

p. 4, l. 8: what is meant by during the process and that the images had to be relabeled - in which situations?

p.4., l. 13+14: do it mean that 5 pathologists did the manual annotations and eye tracking test? Although it may be hard to find pathologists it is a quite small number for reliable results. Also, if they do both conditions, there may be priming effects.

p.5, l. 19+20: Why were these fully and weakly supervised learning tools were selected as a performance comparison test? This should be motivated.
p.6 , Figure 3 c: The legend is not clear - what does destilled by PEAN-I and Original Image mean?
p.7: The paper is missing a critical discussion of the approach and suggestions for future research

p.8 , l. 4: why do the authors present the Methods section as the last chapter of the paper. Would it not be better to explain the methods before the results are reported and discussed?

p.8, l. 15: ... as shown in Figure 1.f - there is no label f in Figure 1 on page 3.

p.8, l. 17: There should be more details on "Easy Pathology" and the used Tobii Eye Tracking System. It would be nice to add a picture showing
how Easy Pathology looks like, the monitor where the software runs on and the position of the eye tracker and how the subjects operate it (e.g. what does p.8, l.22 " ... including the movement of the WSI within the software window" means?. And more details on the subjects preparation and calibration/re-calibration.

p.8 , l.17: there are two times the words "of the"

p. 8, l. 30: The collected ... check for completeness, stability of gaze points, absence of missing data, and any occurrence of offsets"- the authors should stated in more detail what they mean by there terms in the context of their work.

p. 9, l. 3: The slide-reviewing data collected by the eye tracker can be regarded as a point set - Do you mean here sample/event data - depending on which data from the eye tracker you used for your analysis. I find the use of the term Points with regard
to eye tracking data quite confusing - because the eye tracker provides sample data every 16,6 ms (1000 Hertz/60 Hertz sampling trate of used eye tracker). You can use the eye tracker vendor software
to calculate event data out of this sample data file.

p. 9, l. 3: The (1,1) in the formular means here over the whole size of W?

p. 9, l. 4: the maximum tissue frame bw - is this similar to what is depicted in Figure 2a but on basis of the fixation cumulations, i.e. the tissue area including the majority of the fixation. The term "maximum tissue frame" is confusing - better to think about a more exploratory term

p. 9, l. 6: W is divided into windows of size [d x d] at a magnification M- I assume that this step is to find those areas in W with the highest amount of fixations - as depicted in Figure 2d. The authors should state what d and M are? Did the pathologists used different screens
during the eye recording? This refers to my previous comments that more details on the "EasyPathology" should be mentioned in order to make the process (WSI checking and eye gaze recording) more clear to the reader.

p. 9, l. 7: in the formular i is the number of subwindows (0,...,K) W is divided into depending on chosen M and d?

p. 9, l. 8: I assume that you mean with set of points the samples in the subwindow here - see my previous comment above.

p. 9, l. 9: How did you check that pathologists visual field extended beyond the window of the computer screen or the boundary of pathologists tissues in 1)

p.9, l.12: what were the values were selected for q1 and s - and why?

p. 9. l. 15: Greater fixation - mean here higher number of fixations or also longer fixation durations?

p. 9, l. 17: DBSCAN[43] is not in the reference list - is used here to extract fixations from the sample data - right? Or also to cluster fixations?

p. 9, l. 17: Prior studies ... - state the references here - also state an explanation why the fixation zone is inadequate to localize diseased regions

p. 9, l. 20: ... shifts in focus, ..., encompass underlying logical relationships ... - specify examples what this mean in the context of this work, the same with latent logical relationships p.9, line 21

p. 9, l. 21-23: Thus by constructing these sequences and analyzing them as a whole, we can infer the logic and expertise of the pathologists ... - how exactly was this done and which conclusions did the authors get form the data analysis?
Has this also been objectively and scientifically tested - e.g. by testing with experts at different levels of expertise?

p. 9, l. 24: How exactly is a gaze area defined and how is this different from gaze regions (p. 9, l. 26)? Sometimes Pij is stated as gaze area (e.g. p. 24) sometimes as gaze region (e.g. p. 9, l. 32 - 33). This should be consistent.

p.9. l. 29: how are the weight coefficients set?

p. 9, l. 39/40: the authors should explicitly state how the different magnification are incorporated into the developed model , instead of just stating that they are incorporated

p. 10 , l.2 : fexperience should be explained.

p. 10, l. 5: The aim of the partition function Z should be explained

p. 10, l. 12+13: the pretrained encoder should be explained in more detail - how was it pretrained to get which output?. Also it should be stated what the single variables of the feature vector exactly stand for

p. 10, l. 19: Equation (5) it should be ETime

Appendix Figures on page 12: why do the authors put these images on an extra page before the Appendix begins on page 13.
Would it not be better that they are part of the Appendix starting on page 13?

p. 10, l. 27: MIL multiple instance learning)

p. 10, l. 28: state references

p. 10, l. 27+ 28: state the difference between basic MIL and attention-based MIL

p. 11, l. 3: it is not clear what loss2 stands for, additionally loss3 and loss4

p. 11, l. 9: Thus, which feature fusion methods the authors used , those in equation (10) / (11) ?

p. 11, l. 12: Qeval and Qtarget should be explained in more detail and also how they are updated in real time.

p. 11, l. 16 : fIRL should be explained

p. 11, l. 41: it should be Qeval in equation (13)

Appendix Figure 1 b: this figure should be explained in more detail in the caption / text - how were the two dimensions First fixation scale (pixel squared) and searching number be calculated

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)
This study reports on the development and performance of an AI-model (PEAN) for classifying 5 common skin lesion categories from whole slide images (WSI). Novel machine learning methods were employed to overcome the large workload required for manual pixel-wise annotation of WSI. The authors used eye-tracking software, and recorded any zooming or panning the WSI combined with final diagnosis, to label image review patterns for ML. As a result, PEAN generates 'expertise of pathologist' output in providing lesion classification, and to map out an imitation of pathologists' review trajectory of the WSI.
The quality of the manuscript is high, the study objective is novel and exciting, I believe will be of high interest to readers. Furthermore, the manuscript is well written and easy to read, with relevant research cited accordingly.

Commendable points:
Authors provided a comparative dataset where manual pixel-wise annotation were completed. They reported on the overlap of manual pixel-wise annotation with eye-tracking records, and PEAN expertise heatmaps, providing validation of these methods.
The PEAN model has shown high accuracy and AUC for classification of the five most relevant skin lesions that would require pathologists' review.
PEAN outperformed 6 other comparative models trained and tested on the same datasets.

Authors have used an external test set collected from a second hospital (and not used in any training data).
PEAN is capable of integrating eye-tracking data from multiple pathologists, resulting in increased classification performance.
PEAN was tested on small training dataset, and outperformed comparative models.
The last feature of PEAN reported, to map out review trajectories, demonstrates the model's ability to recognise regions of interest, and mimic a pathologists review pattern.
Author have made image data along with slide-reviewing data and manual annotations available for access and published their code.

Questions, points to address
- In the abstract and introduction you have referred to manual pixel-wise annotation as a 'waste' of manpower/medical resources. It would be better to describe it as having a high burden, rather than 'waste', as research to date using these methods, including your own methods for ground truth are still important to the research process, and should not be referred to as 'waste'.
- Can you provide a quantitative result of the overlap between pixel-wise annotation, eye-tracking, and PEAN outputs for the whole dataset? Currently only 4 qualitative examples (one from each lesion category) are provided. This does not allow transparency for overall overlap.
- For the comparative models, there is no explanation on why they were selected (beyond picking 3 fully supervised, and 3 weakly supervised learning methods). Was the selection related to performance, e.g. were they the highest performing models reported? Was it related more to availability of code or other factors? Also, Table 1 should include the references with the codes listed (either in the table or in the caption).
- Figures 3, and supplementary 1b refer to model PEAK-C, I assume this is a typo for PEAN-C?
- Did performance differ between lesion categories? There is a class distribution imbalance in the training dataset that is not discussed, and the results of accuracy per class are not reported.
- A little more information on the data acquisition will allow assessment of potential bias, for example, please clarify if the WSI selected from 5,107 represented all available WSI (with the targeted 5 classification categories)? Or were there any other relevant selection criteria? Can you also report the inter-rater agreement between the two pathologists who initially screened the diagnostic data label? i.e. what portion required arbiter review? This may raise a current 'hot-topic' regarding poor interobserver agreement between pathologists for reviewing histopathology, and should be discussed.
- Section 3 of Methods discusses MIL. Please add a sentence or two at the beginning of this section to describe the relevance and use of MIL, including spelling out the acronym. Not all readers will be familiar with MIL.
- Discuss limitations, especially regarding clinical transferability and what you foresee as the proceeding steps required. For example, you have not included any training methods for recognising out-of-distribution classification. This is crucial for clinical transferability – to ensure the model does not incorrectly interpret OOD classes as a false positive for other malignancy.

(Remarks on code availability)
We do not have the expertise to review the code.

Reviewer #4

(Remarks to the Author)

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)
I do not have the expertise to review and remark on the code.

Version 1:

Reviewer comments:

Reviewer #2

(Remarks to the Author)
All my comments from the first review process were answered satisfactorily and in detail and incorporated into the paper in an appropriate manner.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)
The authors have provided a very detailed and comprehensive response to reviewer comments, addressing all concerns with clarity and precision.

After reviewing the revised manuscript along with the detailed rebuttal letter, I have no further issues to raise. I am satisfied with the current state of the manuscript, and in my opinion, it requires no further revisions for publication.

(Remarks on code availability)
The authors have provided a very detailed and comprehensive response to reviewer comments, addressing all concerns with clarity and precision.

Reviewer #4

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)

Reviewer #5

(Remarks to the Author)
This study developed a deep learning-based auxiliary diagnostic system for pathology images, utilizing eye-tracking devices to capture pathologists' image review processes, thereby providing additional supervisory signals for developing AI models. The experimental results presented in the manuscript demonstrate that this method can effectively quantify, and extract pathologists' knowledge based on gazing, thereby facilitating the training of models for tasks such as image classification and automatic annotation. With an additional set-up of a gazing tracker, collecting the eye trajectories as one new type of annotation in clinical settings can improve performance with minimal incremental labor costs. Compared to the currently prevalent fully supervised and weakly supervised learning approaches, it has its advantages.
Four reviewers have conducted peer reviews of this manuscript:
Reviewer 1 expressed concerns about the method of relying on pathologists' eye-tracking data to train diagnostic models, raising three main issues: 1. Reliability: Pathologists' gaze patterns may not consistently target the most diagnostically relevant areas; 2. Scalability: Resource-intensive long-term collection of eye-tracking datasets; 3. Practicality: Weakly supervised methods are positioned as more practical and widely applicable AI-assisted diagnostic solutions compared to using eye-tracking data.
1. I contend that eye-tracking results need not always align with actual lesion images. The authors defined that "diagnostically relevant areas are more likely, but not absolutely, to attract the attention of pathologists." Eye-tracking data is introduced as "soft labels" to address the issue of inconsistency with the most diagnostically relevant areas, avoiding the accumulation of human annotation errors in the model. The experimental results in the manuscript also demonstrate that introducing eye-tracking data in this unique way indeed benefits AI model performance. Furthermore, the authors designed experiments showing that AI model performance improves under the guidance of multiple pathologists' visual patterns, reflecting that this method is not affected by errors in the eye-tracking data collection process.
2. While collecting eye-tracking data increases economic and time costs for clinical work, the workflow designed by the authors can only reduce cost expenditure, not eliminate it. Thus, the long-term resource consumption in promoting this research objectively exists. However, the authors also mentioned that the data collection process can be integrated into pathologists' daily work, allowing data acquisition with little extra labor time.
3. The experimental results included in the manuscript show that this method outperforms weakly supervised learning methods in performance, making it more suitable for complex clinical environments. The authors also mentioned that AI models trained under the guidance of eye-tracking data can continue training with new images and pathological diagnoses, thus the demand for eye-tracking data is relatively small. This allows annotations collected from previous pathologists to be transferred to new image datasets, reducing expansion costs.
Reviewers 2 and 3 acknowledged the promising prospects of this research and suggested revisions to multiple expressions in the manuscript, as well as supplementing relevant experimental results and discussions.
Reviewer 4 indicated that they co-authored a review with another early-career researcher and did not provide an independent review.
The authors provided detailed point-by-point responses to the comments of all four reviewers, making corresponding revisions and additions to the manuscript. Considering the manuscript and the comments from the four reviewers, I believe this study provides a relatively innovative approach. This method has its characteristics and potential application value in computational pathology, offering certain inspiration to peer researchers. After careful consideration, I recommend accepting this research for publication in Nature Communications.

(Remarks on code availability)

Responses to Reviewer #1 (R1)

Comments# I have fundamental concerns about the hybrid approach proposed in this paper. The reliance on pathologists' eye-tracking data to train diagnostic models presents significant challenges in terms of scalability, reliability, and practicality. The inherent variability and subjectivity in human visual inspection raises questions about the consistency and accuracy of the resulting models. Moreover, the resource-intensive nature of continuous eye-tracking data collection makes this approach impractical for large-scale implementation. Compared to weakly supervised learning methods, which offer greater scalability, flexibility, and robustness, the proposed method appears limited and potentially less effective in diverse clinical settings. These fundamental issues cast doubt on the viability and long-term value of the proposed approach.

Response:

Thank you for raising several important concerns that we did not clearly explain in our original manuscript. We have revised our manuscript to better explain these issues. To avoid the previous misunderstanding, we would like to emphasize or summarize our response into the following three key points.

1) We agree that collecting pathologists' eye-tracking data faces the challenges of potentially large inter-reader variability. However, one unique contribution of our new approach is that we do not directly use the raw eye-tracing data. The collected eye-tracking data is further processed by a new Pathology Expertise Acquisition Network (PEAN). In this study, we investigated how such inter-reader variability impacts scalability and robustness of using eye-tracking data to guide PEAN to build a classification model. Our study found that, although there are variations in eye-tracking patterns among different pathologists, a consistence is that all pathologists are able to capture diagnostically critical information from the images. PEAN has been trained to focus on extracting image features which the pathologists tend to emphasize in common, rather than simply replicating the specific diagnoses they make. As a result, PEAN captures the statistical level information rather than individual-level patterns, which alleviates the variability issue. This allows pathology reports to serve as the sole diagnostic label, while incorporating the attention patterns of the pathologists. Thus, by extracting key image features of the targeted disease and disregarding the potential noise from the raw eye-tracking data, the inter-reader variation in detailed eye-tracking patterns has the minimal impact on the accuracy of the final diagnostic outcomes.

2) It is noteworthy that the manual pixel-level annotations collected in this study are used solely for comparison purposes and are not incorporated into the training process of PEAN. Acquiring eye-tracking data only happens in building or training a deep learning (DL) model (PEAN in this study). By integrating our self-developed software "EasyPathology" with an eye-tracking data collection system, this approach can seamlessly be integrated into the routine workflow of pathologists without requiring additional review or manual labor from them. This makes the increased labor cost within an acceptable range even when compared with weak supervised learning. Compared to the traditionally manual annotation (i.e., to draw tumor boundaries) of large quantities of digital pathology images, which is often required to achieve high performance, using eye-tracking method as presented in this study is much faster or more efficient. Our study shows that on average pathologists only need to spend 4% of their time to provide/record their eye-tracking data as compared to the time required in traditionally manual annotation. Thus, our presented eye-tracking method combining with PEAN can be more practically implemented to annotate WSIs in the large-scale studies of developing new DL models.

3) Weakly supervised learning has gained widespread attention due to its ability to function without the need for large labeled datasets. However, its performance and reliability remain low in medical image field due to the rarity in constructing large WSI datasets to date. In this study, we have conducted a detailed comparison between weakly supervised learning, fully supervised learning with manual annotation, and the method we propose. Our experimental results demonstrate that the classification accuracy of our new model significantly surpasses the models trained using a weakly supervised learning method, such as comparing between our new model and a model (CLAM-SB) using weakly supervised learning, the overall classification accuracies (ACCs) are 96.3% vs 88.8% for the internal testing dataset and 93.0% vs 77.9% for the external testing dataset (please see the study data reported in Section 2.3 of the manuscript). Our method consistently outperforms in various settings of weakly supervised learning, positioning it as a promising superior alternative to weakly supervised learning.

In summary, our experimental results fully support our study hypothesis and demonstrate that using this new eye-tracking method with PEAN has unique advantages to achieve better or higher scalability, reliability and practicality than using the existing or traditional methods of either fully supervised learning (based on manual annotation) or weakly supervised learning methods. We believe that our study results may attract more study interest and effort of other researchers to independently verify our study results and make further bigger progress in this research field. We have clarified or discussed these issues in the revised manuscript. The revisions in manuscript are highlighted in Blue Color in the revised manuscript.

Comment #1.1 Fundamental Dependence on Pathologist Gaze:

The central hypothesis of this study is that pathologists' gaze data, collected through eye-tracking, can be utilized to train DL models, thereby encapsulating their diagnostic expertise. However, this approach is intrinsically flawed due to the inherent variability and subjectivity involved in human visual inspection. Pathologists can and do make errors, and their gaze patterns may not consistently target the most diagnostically relevant areas. This reliance on human gaze data introduces a significant risk: the model may learn and perpetuate these human errors. Consequently, this could compromise the overall reliability and safety of the diagnostic system, as the model may inherit and propagate inaccuracies present in the pathologists' visual inspection process. In contrast, artificial intelligence (AI) models, particularly those employing fully supervised or weakly supervised learning, do not mimic human behavior but rather analyze data SYSTEMATICALLY and COMPREHENSIVELY. Unlike pathologists who might focus on specific regions of a slide based on their expertise and experience, AI models can evaluate the entire slide without any inherent biases. This comprehensive analysis ensures that all potential diagnostic features are considered, thereby reducing the risk of missing critical diagnostic information. Moreover, AI approaches leverage vast amounts of data to learn patterns and make predictions, improving their accuracy over time. They are not limited by human constraints such as fatigue, variability in skill level, or individual biases. AI models can be trained on diverse datasets, which enhances their ability to generalize across different types of pathology and patient populations. This systematic and unbiased approach makes AI inherently more reliable and consistent compared to methods that depend on human gaze patterns. Additionally, AI models are capable of identifying subtle patterns and correlations that may not be apparent to human observers. This capability is particularly valuable in medical diagnostics, where early detection of diseases can significantly improve patient outcomes. By systematically analyzing the entire slide, AI models ensure a thorough examination, which is critical for accurate and timely diagnosis. In summary, while the use of pathologist gaze data introduces significant risks due to potential human errors and biases, AI approaches provide a more reliable and comprehensive method for diagnostic analysis. AI models systematically analyze the entire slide, ensuring that no critical information is overlooked, thereby enhancing the accuracy, reliability, and safety of diagnostic systems.

Response:

We would like to clarify the misunderstanding regarding the specific role of pathologists in this study. We fully agree that in the clinical practice of reading pathological images and diagnosing diseases, pathologists' gaze patterns vary and may not always consistently target the most diagnostically relevant areas. However, the designed PEAN is trained to learn the shared image features that the pathologists tend to focus on, without arbitrarily replicating the regions they have observed or the diagnoses they have made. That is, the raw eye-tracking data only serves as initial learning points (or "soft labels"), PEAN is then trained to filter out data noise and capture the statistical level information rather than individual-level patterns, which alleviates the variability issue. Our efforts to achieve this goal are divided into two parts or tasks: rigorous data cleaning and the design of a unique DL model architecture.

1. **Data Cleaning**

    1) The maximum continuous working time for a pathologist is set at 50 minutes. Pathologists will take a break either upon reaching the specified time or when they subjectively feel fatigued. At the start of the next session, gaze calibration is redone (which takes ~ 30 seconds) to avoid capturing bias.

    2) We use features of the Tobii Eye Tracker, such as automatic response to when the subject moves out of the capture range, tracking loss, and latency compensation, to remove incidents where the pathologist is not focused on reviewing the WSI.

    3) While recording the visual behavior of pathologists at a frequency of 60Hz, abnormal behavior is recorded based on the angular velocity of eye movements under the premise that subjects are allowed to move their heads freely within a certain range. Specifically, when the angular velocity of eye rotation exceeds 30 degrees per second, it is considered a saccade, tremor, or similar behavior. Such eye movements, unrelated to reviewing the WSI, are removed or compensated for with delay.

    4) When mapping the subject's eye movement data to a two-dimensional plane on the screen, if the observation point fell outside the screen, it is deleted.

    Through above processes, we maximize the reduction of noise caused by incidental situations during data collection. We have revised manuscript to more clearly explain this task.

2. **The unique DL model architecture in PEAN to avoid or minimize potential noise in the annotations**

The collected raw eye-tracking data is further processed or analyzed by PEAN. Statistics show that pathologists spend 87.4% of their observation time on ground truth areas. The DL model only needs to filter out a small amount of noise to effectively decode the expertise of pathologists from the eye-tracking data. From the perspective of model principles, PEAN can avoid annotation confusion caused by the raw eye-tracking data of incidentally gazing the areas that are irrelevant to diagnosis (i.e., due to unexpected eye fatigue of the pathologists). The detailed steps of developing and applying PEAN model to process the raw eye-tracking data to generate the final annotation results are presented in Methods section (5.1-5.3) of the manuscript. In brief, the basis of a pathologist's diagnosis stems from the images they have observed, meaning the diagnostically relevant images are inherently included. Therefore, we encode these images using a specialized image encoder and attention mechanism to form a holistic set of image features, which are then compared against images that the pathologist never examined. Through this approach, PEAN is trained to identify the image features that pathologists are genuinely inclined to inspect, while being entirely unaffected by observed but non-diagnostic images—since diagnosis itself is not involved at this stage. Regions exhibiting these features are assigned higher weights, thereby receiving greater (non-exclusive) attention in the subsequent WSI-level classification task, which is guided by pathological diagnoses as "hard labels". This approach directly avoids the accumulation of human errors caused by "observing diagnostically irrelevant regions." In this way, PEAN, having learned human prior knowledge, models the "potential pathologist attention" for subregions within the WSI as the proposed pathology expertise in this study. In the subsequent WSI classifier training, this attention is used as the important score for each subregion, and after fully integrating the image features of all subregions, a diagnostic prediction for the WSI is made. This method likes most DL models in that it comprehensively analyzes all images, reducing the risk of missing critical information during the pathologist's slide review process, while effectively excluding interference from diagnostically irrelevant regions that the pathologist may have observed.

Additionally, comparing weakly supervised learning and fully supervised learning, the superior performance demonstrated by PEAN proves that the above work is effective.

Unlike many other application fields in which AI or DL models are trained or learned using "big data" (i.e., over millions of the collected samples), the number of training data in medical image field is much smaller (i.e., typically smaller than several hundred or few thousands as reported in the literature). The collection of large-scale medical image datasets faces greater security challenges. Based on existing discussions on the principles of artificial intelligence and our experimental results, weakly supervised learning is difficult to reliably apply to small datasets. Consequently, most AI models that have been successfully applied in clinical settings to date have been trained using annotated images, which offer significant advantages and higher performance compared to raw, unannotated images. Under guidance of pathologists through image annotation (either manually or using new eye-tracking technology), AI models can more reliably learn and identify more subtle but clinically relevant image features from the targeted tumor areas and distinguish them from other subtle features from the heterogenous background areas. Our experiment results also show that several DL classification models trained using weakly supervised learning methods have lower performance than the models trained using the annotated images (see the data reported in Table 1). Furthermore, when trained with only 30% of the eye-tracking data, PEAN still achieved 93.1% ACC and 0.981 AUC, outperforming weakly supervised learning models trained on the full dataset. This result demonstrates PEAN's long-term competitiveness in extreme scenarios, addressing the reviewers' concerns about the future applicability of this research.

Comment #1.2 Scalability and Practicality:

The requirement to continuously collect eye-tracking data from pathologists is both impractical and resource-intensive. This approach is not scalable in real-world clinical settings where the deployment of AI systems needs to be efficient, cost-effective, and minimally disruptive. Continuously tracking and recording the gaze patterns of pathologists demands specialized equipment, ongoing data collection, and significant manual effort to annotate and process the data. These requirements introduce substantial logistical and financial burdens. In contrast, AI models that leverage fully supervised or weakly supervised learning approaches are inherently more scalable and practical. These models rely on large, pre-existing datasets that can be annotated with high-level labels, such as diagnostic outcomes, without the need for detailed manual input. This makes the data collection process much more efficient and less costly. As medical data continues to grow exponentially, AI models can continuously improve by incorporating new data, thus enhancing their performance over time without requiring additional specialized data collection efforts. Moreover, AI systems designed for scalability can be easily integrated into existing clinical workflows. They do not require the ongoing participation of pathologists for data collection, allowing medical professionals to focus on patient care rather than being burdened by additional tasks. This integration is crucial for widespread adoption, as it minimizes disruptions and maximizes the utility of AI systems in clinical practice.

Response:

We would like to emphasize that this is a misunderstanding. Manual annotation is used for comparison with our proposed method. Eye-tracking is ***only*** required and applied to generate a training dataset to develop AI (or DL) models. There is ***no need*** to continuously collect eye-tracking data from pathologists when applying the trained or optimized model (classifier) to testing images in future clinical practice. Compared to traditionally manual annotation, using our new method that combined eye-tracking data collection and PEAN data processing is much faster and more efficient. In this study we report that on average pathologists only need to spend 4% of time if they manually annotate pathology images. Like all other AI models, once our DL models are trained and developed, they can directly apply to test new WSIs without the need to acquire eye-tracking data from pathologists. Thus, compared to existing annotation methods, our new method is significantly cost-effective.

Developing a high-performance and powerful AI model requires a large training dataset (i.e., > hundreds of thousands or millions of images or other data). However, this is not applicable to current medical image field. Due to relatively small numbers of cancer patients as comparing to general population and protection of patients' privacy in the medical institutions, most previous AI models only trained and tested using a few hundred or a few thousand images (i.e., WSIs) as reported in literature to date. Therefore, developing high-performance and robust models using relatively small image datasets largely depends on the quality of image annotations. In fact, developing more efficient or semi-automatic image annotation methods is a hot research topic among many academic research groups and start-up companies worldwide. For example, in the rapidly advancing field of foundational model research in computational pathology, researchers utilize precisely annotated datasets to train a unique "annotation model", which is then used to annotate much larger datasets lacking detailed annotations [1]. The primary goal of this study is to explore and propose another novel image annotation method, along with a DL model (PEAN) capable of effectively leveraging such annotations. Comparing this to the costs of weakly supervised learning which has inherent diagnostic performance limitations, seems beyond the scope of this study's focus. Nevertheless, we are open to conducting broader comparisons to highlight the significant contributions of this work, especially since we have further demonstrated that our data collection process remains highly competitive even when compared to weakly supervised learning. To avoid confusion to the readers, we re-emphasize the objective of this study in the revised manuscript.

In addition, the collection of eye-tracking data from pathologists can be seamlessly integrated into their existing workflows with minimal labor cost and disruption. The adoption of digital pathology enables this system to operate without requiring additional manual labor, making it nearly imperceptible to pathologists. We have supplemented this with several videos (supplementary files 1-2) demonstrating the process in practice. The advantages of weakly supervised learning in retrospective data collection are diminished, as the data used in such studies also requires repeated review to ensure the correctness of labels. Overall, the one-time investment costs (including equipment) combined with much more efficient image annotation process can make the widespread implementation of this new technology feasible in developing new DL models of medical images in future studies.

[1] Lu, M.Y., Chen, B., Williamson, D.F.K. et al. A visual-language foundation model for computational pathology. Nat Med 30, 863–874 (2024). https://doi.org/10.1038/s41591-024-02856-4

Comment #2.1 Limited Generalization:

The study focuses on specific skin conditions, and it is unclear how well the proposed method would generalize to other types of pathology or medical imaging domains. The heavy dependence on data from a limited number of pathologists further questions the robustness and adaptability of the model to broader applications.

Response:

Our answers are in three folds. First, although this study focuses on skin diseases, the concept of eye-tracking and PEAN is not limited only to skin diseases, and it can also be easily applied to annotate other pathology images acquired from other types of tumors. For example, we recently conducted a new study in which we extended the model to a publicly available Camelyon16 breast cancer WSI dataset for training and testing. To train the model on this dataset, we used our existing data collection platform to gather gaze-tracking data from pathologists. This portion of the gaze-tracking data has also been made publicly available. Camelyon16 includes 399 breast cancer WSIs, of which 129 (80 normal and 49 metastatic breast cancer) are designated as the test set, a division we have adhered to. We compared the classification performance of PEAN with that of eight baselines (six representative classic algorithms from the initial draft and two newly added recent algorithms. In this new study, PEAN achieved an ACC of 96.12% and an AUC of 98.81%, also surpassing the current state-of-the-art weakly supervised learning method IB-MIL by 3.9% in ACC and 4.1% in AUC. To make this paper more focused or concise, we only explained the results of the breast cancer metastasis detection task in the response. The new result provides new evidence to support the potential of applying this new technology or method for broader application in pathological diagnosis across other diseases. We will report this new study in the next follow-up research paper shortly.

Comparison of PEAN-C with baselines on Camelyon16

Methods	ACC	AUC	Recall	Presion
DLCCP	93.0%	0.941	87.7%	93.5%
SLC	93.8%	0.942	87.7%	95.6%
HSL	94.6%	0.953	85.7%	**100%**
CLAM-SB	88.4%	0.896	71.4%	97.2%
ABMIL	86.8%	0.902	89.8%	78.6%
TransMIL	93.0%	0.917	81.6%	**100%**
DS-MIL	91.5%	0.931	81.6%	95.2%
IB-MIL	92.3%	0.947	87.7%	91.5%
PEAN-C	**96.1%**	**0.988**	**89.8%**	**100%**

Second, due to practical constraints and the focus of this study, we have not yet explored the application of PEAN in radiology. However, existing researches have provided theoretical supportd for the use of gaze-tracking data in radiological diagnosis. For example, four recent studies reported by Shen et al. (2022-2024) [1-4] demonstrated that radiologists' gaze data helped mitigate harmful shortcut learning in DL models, which can occur in high-resolution, redundant images such as knee and chest X-rays. By incorporating gaze attention, they improved DL model performance and interpretability. Although their approach differs technically, it confirms that expert gaze data provides valuable prior knowledge for DL model training. Exploring PEAN's feasibility in radiology is beyond the scope of this study, and optimizing this architecture of DL models for medical imaging remains a significant research interest and challenge.

Third, we fully agree that robustness and adaptability are big challenges faced by current AI or DL models of medical images, which were trained using relatively small image datasets. However, such challenge or risk is not caused by our new image annotation method using eye-tracking and PEAN. It is caused by the size and diversity of training image datasets. In this study, we assembled a relatively large dataset involving 5,881 WSIs and 5 pathologists. We have conducted several experiments and data analyses to study performance of DL models trained using different subsets of training samples. We have also tested the trained models using two independent (internal and external) testing datasets to demonstrate robustness or adaptability of our new models. Although the number of pathologists was limited, their performance still exceeded the diagnostic levels of the baselines. Despite such efforts, whether this image dataset can represent diversity of skin disease images in the broad clinical community has not been investigated.

More testing using new image datasets is needed in future studies. We have discussed this limitation in the revised manuscript.

[1] S Wang, X Ouyang, T Liu, Q Wang, D Shen. "Follow my eye: Using gaze to supervise computer-aided diagnosis." IEEE Transactions on Medical Imaging 41.7 (2022): 1688-1698.
[2] C Ma, L Zhao, Y Chen, S Wang, L Guo, T Zhang, D Shen, X Jiang, T Liu. "Eye-gaze-guided vision transformer for rectifying shortcut learning." IEEE Transactions on Medical Imaging (2023).
[3] Z Zhao, S Wang, Q Wang, D Shen. "Mining gaze for contrastive learning toward computer-assisted diagnosis." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 38. No. 7. 2024.
[4] S Wang, Z Zhao, L Zhang, D Shen, Q Wang. "Crafting Good Views of Medical Images for Contrastive Learning via Expert-level Visual Attention." NeuRIPS 2023 Workshop on Gaze Meets ML. 2023.

Comment #2.2 Marginal Performance Gains:

The reported performance improvements (AUC of 0.984 and accuracy of 93.1% on the internal test set) over fully supervised and weakly supervised learning methods are not sufficiently substantial to justify the added complexity and cost of incorporating eye-tracking data. Given the rapid advancements in weakly supervised learning, which can leverage large-scale, automatically collected datasets consisting of data coming from millions of patients, the marginal gains presented here may not hold long-term competitive value.

Response:

We appreciate the reviewer raising this important issue regarding the reported performance comparison. Due to the high-risk nature of medical tasks, it is worthwhile to increase pre-investment to achieve high clinical safety. For example, compared to weakly supervised learning, fully supervised learning, although requiring expensive manual labeling, is a safer method. Although recent studies have attempted to bridge the performance gap by improving the weakly supervised learning framework, there is a lack of clinical applicability. To address this issue and test whether the new model has only achieved marginal gains, we added multiple comparison experiments in the revised manuscript.

1. On the original basis, we newly added two recent and representative weakly - supervised learning models as baselines. Meanwhile, a new image encoder, CONCH, was incorporated to comprehensively evaluate these WSI classifiers. The results indicate that the PEAN developed by us has achieved the optimal performance under various conditions. It has achieved an ACC of 96.3% and an AUC of 0.992 on the internal testing set; and an ACC of 93.0% and an AUC of 0.984 on the external testing set. Compared with the state-of-the-art weakly supervised learning model, IB-MIL, the ACC of PEAN on the external testing set is exceeded by approximately 10%, demonstrating that this is not a marginal gain.
2. We newly added independent test sets from two different medical institutions. PEAN has also demonstrated the highest performance on these test sets, proving its reliability and adaptability in the face of complex clinical environments.

In addition, we would like to emphasize or clarify again the following two points:

1. Using our new method to annotate pathology images is much faster and more cost-effective than other existing image annotation methods. The eye-tracking data collection, which can be seamlessly integrated into the daily work of pathologists, requires no manual labor and is competitive even in the face of the data collection cost of weakly supervised learning.
2. Although we cannot guess or compare with future models trained using “the advanced weakly supervised learning methods” comparing to the existing weakly supervised learning methods, our new model (PEAN-C) yields the highest performance in this study, which clearly indicates the values of our new method and can be added to this research field to help develop new DL models with higher accuracy and robustness in the future. Additionally, we conducted an experiment that only used 30% of the eye-tracking data to train PEAN. This aims to simulate and test the performance of PEAN when it is difficult to collect a large amount of eye-tracking data within a short period. The results indicate that PEAN still achieved an ACC of 93.1% and an AUC of 0.981, outperforming weakly supervised learning models trained on the complete data set. Maintaining an advantage even with a small amount of training data shows that PEAN has long-term competitiveness.

Comparison of PEAN-C with baselines on Added External Testing 1

Methods	ACC	AUC	Recall of Each Disease				
			Nevus	BCC	Melanoma	SCC	SK
DLCCP	82%	0.944	95%	100%	50%	75%	90%
SLC	85%	0.959	95%	95%	60%	85%	90%
HSL	90%	0.973	95%	95%	95%	70%	95%
CLAM-SB	78%	0.943	95%	95%	90%	15%	95%
ABMIL	75%	0.892	80%	80%	95%	60%	70%
TransMIL	81%	0.949	95%	95%	70%	45%	100%
DS-MIL	81%	0.95	95%	95%	60%	60%	95%
IB-MIL	87%	0.971	95%	95%	65%	8%	100%

PEAN-C	**94%**	**0.987**	100%	100%	95%	75%	100%

Comparison of PEAN-C with baselines on Added External Testing 2

Methods	ACC	AUC	Recall of Each Disease				
			Nevus	BCC	Melanoma	SCC	SK
DLCCP	86%	0.953	80%	100%	60%	0.95%	95%
SLC	89%	0.968	75%	95%	75%	100%	100%
HSL	92%	0.969	80%	100%	80%	100%	100%
CLAM-SB	82%	0.947	40%	95%	80%	100%	95%
ABMIL	56%	0.872	85%	0%	95%	100%	0%
TransMIL	88%	0.947	75%	90%	80%	95%	100%
DS-MIL	79%	0.918	70%	90%	50%	90%	95%
IB-MIL	87%	0.96	45%	100%	95%	100%	95%
PEAN-C	**96%**	**0.988**	85%	100%	95%	100%	100%

Comment #2.3 Interpretability Issues:

While the method aims to enhance interpretability by mimicking pathologists' gaze patterns, the decision-making process of the model remains complex and opaque. The study does not provide sufficient mechanisms to ensure that the learned gaze patterns are indeed the most diagnostically relevant, thereby limiting the transparency and clinical trustworthiness of the model. This opacity is problematic, as clinicians need to understand and trust AI recommendations to integrate them effectively into their diagnostic workflow. In comparison, weakly supervised learning methods often provide greater explainability. These methods can leverage high-level labels, such as diagnostic outcomes, which are more straightforward and interpretable. By using broader annotations, weakly supervised models can identify and learn relevant patterns and features directly associated with diagnostic outcomes. This direct association can enhance the transparency of the model, making it easier for clinicians to understand how the AI arrived at its conclusions. Furthermore, weakly supervised methods can incorporate various techniques to improve explainability, such as attention mechanisms and feature importance mapping. These techniques allow the model to highlight the regions of the image that contributed most to its decision, providing a visual and interpretable explanation of its diagnostic process. This level of transparency is critical for gaining clinical trust and facilitating the adoption of AI tools in healthcare. Additionally, the flexibility and scalability of weakly supervised methods mean they can be applied across a wide range of medical imaging tasks without the need for specialized and continuous data collection efforts. They can adapt to different datasets and clinical contexts, maintaining their robustness and reliability while also providing interpretable results.

In summary, while the proposed method attempts to improve interpretability by mimicking pathologists' gaze patterns, it falls short in ensuring transparency and clinical trustworthiness. Weakly supervised learning methods offer a more practical, scalable, and explainable alternative, making them better suited for real-world clinical applications. These methods can leverage broader, more interpretable data and incorporate techniques to enhance transparency, ultimately providing more reliable and trustworthy diagnostic support.

Response:

We fully agree with the reviewer's perspective on the need for high transparency and interpretability in AI applications within clinical settings. Weakly supervised learning faces challenges in addressing certain critical issues. Ideally, a model's predictions should be based solely on small diagnostically relevant regions within each gigapixel image, which are also the basis for pathologists' diagnoses. Due to the complexity of pathology images, patch-level diagnostic results are often associated with incorrect image features (e.g., noise introduced by different scanning devices), while pathologists primarily focus on tissue and cellular morphology for diagnosis. Although the weakly supervised learning has gained attention for its ability to consider the comprehensive information in pathology images, the introduction of overly redundant information hinders further improvements in model performance. A key method to establish a genuine connection between diagnostically relevant regions and model predictions is through out-of-distribution generalization, which separates features of diagnostic patches from non-diagnostic ones. This is where fully supervised learning holds an advantage. While using pathologists' eye-tracking data, PEAN can identify the image features that pathologists tend to focus on. Through this indirect generalization, PEAN can make accurate diagnoses while also providing distilled subregions of interest.

Currently, most DL models act like "black boxes" and the reasoning process of the model is not explainable to the clinicians, which can reduce clinicians' confidence to use DL models if the model is not robust when applying to the new image data. The same issue applies to weakly supervised learning, as the regions of interest identified by weakly supervised models are derived through backpropagation of the model's predictions. This means that the labeled regions can accumulate errors from the model's predictions, especially given that weakly supervised learning typically exhibits lower classification accuracy. In contrast, PEAN identifies the "pathologists' potential regions of interest" before making a diagnosis, effectively avoiding error propagation and providing reliable support to pathologists. Reports indicate that PEAN achieves a mean overlap score of 0.822 with ground truth images and 0.357 with other non-diagnostic images. We have also included additional graphical examples comparing PEAN's labeled regions with ground truth. Both in principle and in results, the diagnostic support provided by PEAN is trustworthy.

In summary, developing new "explainable" or "transparent" DL models is a hot research topic to date, which depends on new model structures and methods to effectively extract more clinically relevant or explainable features. The progress is slow due to many technical or clinical challenges in developing "explainable" DL models. However, we would like to point out that developing an explainable or transparent DL model does not depend on image annotation methods. In addition to the proposed DL model, another important objective of this study is to develop a new more efficient and cost-effective image annotation method, to facilitate

development of DL or AI models. The annotated data can be used to train both non-explainable (“black box”) and explainable (or transparent) DL models in the future.

Comment#3.1 Resource Intensiveness:

The process of collecting and annotating eye-tracking data is labor-intensive and time-consuming. The manuscript underestimates the practical challenges and costs associated with this approach, which include the need for specialized equipment, ongoing data collection, and extensive manual annotation.

Response:

As mentioned in our response to comment 1.2, the collection of eye-tracking data from pathologists does not rely on manual labor and can be seamlessly integrated into existing digital pathology workflows. Our self-developed slide-reviewing software, "EasyPathology", and the eye-tracking equipment do not impose significant financial burdens, especially given EasyPathology's high compatibility and accessibility. Overall, the one-time investment costs (including equipment and the relatively low learning curve for pathologists to adapt to the workflow), combined with the significantly reduced ongoing labor costs, make the widespread implementation of eye-tracking data collection in developing/training DL models at medical institutions feasible.

Comment #3.2 Risk of Propagating Errors:

By training the model to mimic pathologists' gaze patterns, there is a significant risk of the model learning and replicating the diagnostic errors made by the pathologists. This raises serious concerns about the safety and reliability of the system in clinical practice, where errors can have critical consequences.

Response:

While we acknowledge the potential risk of errors in pathologists' gaze patterns, as detailed in our response to comment 1.1, we have taken steps to address these issues through the design of the DL model (PEAN) and rigorous data cleaning efforts. Therefore, the inter-reader variability in pathologists' gaze patterns has a minimal impact on the final diagnostic outcomes, as individual differences are largely mitigated by PEAN. This is also noted in Section 2.5 of the original manuscript. Our study results have demonstrated that the classification model (PEAN-C) achieves the highest classification accuracy than other 8 models trained using the fully supervised learning methods (manual annotation) and weakly supervised learning methods. We believe that these efforts and the experimental results show that the impact of human errors on the AI model can be minimized, thereby enabling the development of an assistive diagnostic model that leverages pathologists' prior knowledge and can be applied in future complex clinical environments. In the revised manuscript, we have added additional study results and more detailed discussions to better address these concerns.

Comment #4.1 Scalability:

Weakly supervised learning methods are inherently more scalable compared to the proposed approach of using eye-tracking data from pathologists. These methods can leverage large amounts of readily available data with minimal annotation, such as diagnostic outcomes or general labels, which significantly reduces the need for detailed and labor-intensive manual annotations. This minimal annotation requirement allows for the continuous and automated expansion of datasets, as new patient data can be incorporated into the training process with minimal additional effort.

This scalability is crucial for the widespread adoption of AI in clinical practice. As the volume of medical data grows, weakly supervised learning methods can efficiently handle the influx of new data, ensuring that the AI models remain up-to-date and capable of learning from the latest clinical information. This ability to scale effortlessly with increasing data volumes makes weakly supervised methods particularly well-suited for large healthcare institutions and research centers that generate vast amounts of data daily.

Moreover, weakly supervised methods enable the development of robust models that can generalize well across different patient populations and medical conditions. By training on diverse datasets, these models can learn to recognize patterns and anomalies that are consistent across various contexts, thereby improving their diagnostic accuracy and reliability. This generalizability is essential for creating AI systems that can be deployed in a wide range of clinical settings, from large hospitals to smaller clinics, without requiring extensive customization or retraining.

In contrast, the proposed approach of using eye-tracking data from pathologists faces significant scalability challenges. The continuous collection of gaze data is resource-intensive, requiring specialized equipment and significant manual effort to process and annotate the data. This approach is not feasible for large-scale implementation, as it imposes substantial logistical and financial burdens. Additionally, the reliance on a small number of pathologists for data collection limits the diversity of the training data, potentially reducing the model's ability to generalize across different clinical settings.

In summary, weakly supervised learning methods offer a highly scalable and practical solution for the development of AI diagnostic models. They leverage large, minimally annotated datasets to continuously improve and generalize well across diverse clinical environments. This scalability makes them far more suitable for widespread adoption in healthcare compared to the resource-intensive and limited approach of using pathologists' eye-tracking data.

Response:

We would like to emphasize again that one objective of this study is to develop and test a new more efficient and cost-effective medical image annotation method using an eye-tracking device combined with a novel data learning and processing model (PEAN). It significantly reduces the required clinical resources or pathologists' time in annotation of pathology images for developing new DL (or AI) models, as we have explained in our previous responses. Although we compared and demonstrated the advantages of our new methods over the traditional manual annotation methods and the existing weakly supervised learning methods, discussion of whether developing future weakly supervised learning methods can achieve more scalability is beyond the scope of this study.

Comment #4.2 Cost-Effectiveness:

Weakly supervised methods significantly reduce the cost associated with data annotation. By using high-level labels (e.g., diagnostic outcomes) instead of detailed pixel-wise annotations, these methods minimize the need for extensive manual input, thereby conserving valuable medical resources and reducing operational costs.

Response:

Image patches as input of DL models typically have a fixed size (i.e., 224 × 224 pixels). Since the image patches can involve both heterogenous tumor area and normal tissue area, only providing high-level labels (e.g., diagnostic outcomes) is not sufficient to train a highly performed and robust DL model when the size and diversity of training images are limited. Thus, currently there is no scientific evidence to support that weakly supervised learning can achieve higher and more robust performance than fully supervised learning. Additionally, distinguishing tumor and normal tissue areas depicting on one image patch is important. To address and overcome this challenge, we presented a more cost-effective image annotation method based on eye-tracking in this study. We have also demonstrated that the classification model trained using new annotation achieved the highest accuracy than the models trained using weakly supervised learning methods. In addition, as mentioned in our response to comment 1.2, the collection of eye-tracking data from pathologists does not rely on traditionally manual labor and can be seamlessly integrated into existing digital pathology workflows. When compared to weakly supervised learning, its minimal additional cost and superior performance make this approach highly competitive.

Comment #4.3 Robustness and Generalization:

Models trained with weakly supervised learning are often more robust and generalizable due to their exposure to a broader range of data variations. Unlike approaches that rely on detailed manual annotations or specific data collection methods like eye-tracking, weakly supervised learning methods can utilize large and diverse datasets that represent a wide array of clinical scenarios and patient demographics. This extensive exposure enables the models to learn more generalized and representative features, which enhances their ability to perform accurately across different datasets and clinical settings.

The inherent variability in medical data, including differences in patient populations, imaging techniques, and disease presentations, is better captured by weakly supervised learning methods. By training on such diverse datasets, these models develop a more comprehensive understanding of the underlying patterns and nuances associated with various medical conditions. This comprehensive learning process improves the model's robustness, making it less sensitive to specific biases or anomalies present in any single dataset.

Furthermore, weakly supervised models can leverage high-level labels, such as diagnostic outcomes, which are easier to obtain and less prone to subjective interpretation errors compared to detailed annotations. This reliance on broader labels reduces the risk of overfitting to specific features that may not be generalizable across different clinical environments. As a result, these models can maintain high performance even when applied to new and unseen data, ensuring their utility in real-world medical practice.

In contrast, the proposed approach of using pathologists' eye-tracking data is limited by the variability and subjectivity inherent in human visual inspection. Pathologists may focus on different regions of interest based on their individual experiences and biases, leading to inconsistencies in the training data. These inconsistencies can negatively impact the model's ability to generalize, as it may learn to rely on specific patterns that do not hold true across different clinical settings.

Additionally, the limited scope of data collected from a small group of pathologists further restricts the model's exposure to diverse diagnostic scenarios. This narrow focus can result in a model that performs well in the specific context in which it was trained but fails to adapt to broader applications. The lack of generalization is a significant drawback, as it undermines the model's reliability and effectiveness when deployed in varied clinical environments.

In summary, weakly supervised learning methods offer superior robustness and generalization due to their ability to train on large, diverse datasets and leverage high-level labels. This comprehensive exposure to a wide range of data variations enables these models to perform consistently across different clinical settings, making them a more reliable and practical choice for AI-assisted diagnostics compared to the limited and subjective approach of using eye-tracking data from pathologists.

Response:

Our previous responses have largely addressed this comment. We would like to summarize the key points here again. First, to clarify any possible confusions, we would like to reiterate that there is currently no solid scientific evidence supporting the claim that weakly supervised learning can achieve more robust performance. In fact, based on our results, the diagnostic performance of weakly supervised learning was significantly lower than PEAN and other 3 fully supervised learning methods across independent testing sets. This demonstrates the robustness of our method, which can easily adapt to a relatively large training dataset and complex clinical environments (e.g., variations in slide preparation standards in different medical centers). Additionally, while this study developed a new DL model, it is equally important to note that it also developed and tested a new, more efficient, and cost-effective image annotation method to facilitate future DL model development. This makes the study not merely a challenge to weakly supervised learning but a general enhancement to various existing methods.

Second, in this study we assembled a relatively large and diverse image dataset including 5,881 WSIs collected from two independent medical institutions. To the best of our knowledge, this is a quite unique and larger dataset compared to other existing studies reported in literature. Although only 5 pathologists were included thus far, the model has already demonstrated excellent performance. We also wish to emphasize that one of the key contributions of this research is the development of a specialized data collection scheme, which opens up possibilities for greater involvement of pathologists in the future. The temporarily limited number of participating pathologists does not affect the validity of this new image data annotation method. This will also contribute to the development of DL models capable of accurately identifying subtle patterns and underlying differences—a capability that typically relies on learning from diverse data.

Comment #4.4 Flexibility:

Weakly supervised learning methods offer a high degree of flexibility, making them easily adaptable to various types of medical imaging and pathology without the need for specialized data collection procedures, such as eye-tracking. This inherent flexibility significantly enhances their versatility and applicability across a wide range of diagnostic tasks.

One of the key advantages of weakly supervised methods is their ability to utilize broadly available high-level labels, such as diagnostic outcomes, which are common across many medical disciplines. These labels do not require intricate and time-consuming manual annotations, allowing the models to be trained on existing datasets with minimal additional effort. This adaptability means that weakly supervised models can be quickly repurposed and applied to different types of medical imaging, including radiology, histopathology, and ophthalmology, among others.

The lack of dependence on specialized data collection processes like eye-tracking further enhances the practicality of weakly supervised methods. Eye-tracking requires specific equipment and setup, as well as ongoing participation from pathologists to gather data. This not only adds complexity and cost but also limits the approach's applicability to settings where such resources are available. In contrast, weakly supervised methods can be deployed using standard medical imaging data, which is readily accessible in most clinical environments.

Moreover, weakly supervised models can easily integrate new types of medical data as they become available. This capability is particularly valuable in a rapidly evolving field like medical diagnostics, where new imaging techniques and modalities are continually being developed. Weakly supervised methods can be updated with these new data types without extensive reengineering, ensuring that the models remain relevant and effective over time.

The flexibility of weakly supervised learning also extends to its potential for hybrid approaches. These methods can be combined with other learning paradigms, such as semi-supervised or self-supervised learning, to further enhance their performance and adaptability. This combinatory approach allows for the leveraging of both labeled and unlabeled data, maximizing the utility of all available information and improving the robustness of the models.

In summary, the flexibility of weakly supervised learning methods makes them exceptionally versatile and applicable to a wide array of diagnostic tasks. Their ability to utilize readily available data without the need for specialized collection procedures allows for quick adaptation to different types of medical imaging and pathology. This versatility, combined with the potential for hybrid approaches, positions weakly supervised methods as a more practical and broadly applicable solution for AI-assisted diagnostics compared to the constrained and resource-intensive approach of using eye-tracking data from pathologists.

Response:

While we appreciate the reviewer's comment, we respectfully disagree with the reviewer. We believe that weakly supervised learning does not provide more flexibility when the training dataset size of medical images is small or limited to date. It only omits the time-consuming process of pathologists' manual annotation. Due to high heterogeneity of both tumor and surrounding normal tissues, only providing final diagnostic outcome as weakly supervised learning method has severe limitation or weakness when applying to relatively small medical image datasets in current biomedical imaging field. Our new method has advantages of flexibility because the new PEAN has capability of adaptively generating classification results by effectively learning and filtering the noise or irrelevant eye-tracking patterns.

Comment #4.5 Future-Proofing:

As the availability of medical data continues to grow, weakly supervised methods are well-positioned to take advantage of this trend. They can continuously improve as more data becomes available, ensuring that their performance keeps pace with advancements in data collection and annotation technologies.

Response:

We agree with the reviewer's comment regarding future-proofing. The more advanced weakly supervised learning methods will continue being developed and emerged as more medical image data becomes available in the future. However, the new DL model, PEAN, with its excellent performance and not overly high-cost requirements, can also be a good choice for current assisted diagnosis using relatively small image datasets in current research environment. Even as reported in the response to Comment 1.1, PEAN, trained by less data, can surpass weak supervised learning trained by more data. We also believe that PEAN and the novel image annotation method reported in this manuscript will also attract more research interest to make more progress in the medical imaging AI field in the future, and shed light on a new future direction for data annotation in biomedical imaging.

==========

Responses to Reviewer #2 (R2)

Comments# In this paper the authors developed a DL system, named Pathology Expertise Acquisition Network (PEAN), to decode pathologists' expertise from eye movements recorded while they are looking at digital pathology images. This learned pathologists' expertise can then be used for precise diagnostics of whole slide image assisted diagnostics and thus reduce manual annotations effort for pathologists. The results show that their approach reduces the manual annotation time of the digital pathology images to 4%, with an average time of 36.5 seconds per annotation. Therefore, the effort of cumbersome and time-consuming manual annotation of digital pathology images can be reduced significantly.

Overall review comments:

Although the paper addresses an important topic (reduction of time-consuming manual annotation times) and the chosen approach of learning from eye movement data is a sensible and promising approach, a major revision of the paper seems sensible for the following reasons.

Response:

We thank reviewer for the thoughtful and encouraging comments. Based on the recommendation of a major revision, we have carefully considered each point raised by the reviewer and have made corresponding revisions and additions to the manuscript. Following are our point-by-point responses to reviewer's comments. The revisions made in manuscript are highlighted in Red Color.

Comment 1# - Eye Tracking (Chapter 5, 1. Data acquisition and preprocessing):

Although important parameters appear to be calculated, it is not possible to trace exactly what has been done, as:

1.) Some parameters are not explained, e.g. gaze area vs. gaze region, calculation of angular velocity)

2.) how exactly the calculated data is interpreted

3) the use of terms is sometimes confusing (e.g. eye tracking is referred to as recorded points, whereas sample data is probably meant here)

4) The setup, design and implementation of the eye tracking experiment and the eye tracker used, as well as its positioning and calibration, are also not described or shown

The same holds for the section 2 (Extraction of pathologists' expertise) and section 3(Feature distillation and classification with weakly supervised models:

Here too, the initial situation and objectives should be described before the details are discussed. All designators used should also be explained.

Response:

Following the reviewer's comments, we have revised section 4.1 (Eye Tracking) to provide more detailed information or explanation.

1) Terms of "gaze area" and "gaze region" refer to "the range of the pathological image observed by the pathologist at a given moment". To avoid confusion, all such terms have been replaced by "observation point" in the revised manuscript. The calculation process of the angular velocity of eye movement is provided by the eye tracker: "Tobii Pro Spectrum", and we called the automatically calculated result.

2) We have provided clear explanations for the involved variables, and special terms that are only applicable in this study, to avoid confusing readers.

3) We have carefully reviewed the manuscript to more clearly and consistently explain specific terms to avoid or minimize the risk of confusion. For example, we replaced "the slide-reviewing data collected by the eye tracker can be regarded as a point set $\{Points\}$" with "the slide-reviewing data collected by the eye-tracking device contains the observation points captured at a specific frequency, described as a point set $\{Points\} \in R^{points\ number \times (1,1)}$". We also replaced terms like "sample data" with more precise alternatives such as "sampled observation points" or "sampled WSIs," as "data" is too broad and can mislead readers.

4) We have revised Sections 4.1-4.3 and add more details regarding (1) eye tracking system calibration, the experimental design and setup used to collect pathologists' slide-review data in Section 4.1, (2) steps to extract pathologists' expertise from the

collected eye tracking data and data analysis using PEAN in Section 4.2, and (3) the methods and steps to conduct feature distillation and classification in Section 4.3. Hypothesis and objective of developing above study methods are added in each Section (4.1-4.3).

Comment 2# The work should also be discussed in relation to existing approaches, such as:
Deane, O., Toth, E. & Yeo, SH. Deep-SAGA: a deep-learning-based system for automatic gaze annotation from eye-tracking data. Behav Res 55, 1372–1391 (2023). https://doi.org/10.3758/s13428-022-01833-4

Response:

We thank the reviewer for recommending this recent work. We have now cited and discussed this paper in the revised manuscript. Both this study and our work explored low-cost eye tracking technology using two different DL models and for two different application purposes.

Comment 3# There is no critical discussion and Future work addressed in the paper

Response:

As recommended by the reviewer, we have added Section 3.3 in the revised manuscript's discussion to discuss the unique potential of this work in developing personalized diagnostic support models and training junior pathologists. Additionally, Section 3.4 has been included to address the limitations of this work regarding issues such as OOD (out-of-distribution) and scenarios with few pathology-reviewers.

Comment 4# p-1, l. 10+11: why AUC and seconds per annotation as time costs - why not difference to human annotators and overall annotation time? Additionally, the authors do define AUC on page 5, line 23

Response:

First, we reported average time required for pathologists to do full manual annotation and eye-tracking experiment, which are 14.2 minutes and 36.5 seconds per case (or WSI), respectively. Thus, using the annotation time required using eye-tracking method is approximately only 4% of time required to do full manual annotation by the pathologists.

Second, the area under the receiver operating characteristic curve (AUC) is a common metric for evaluating the classification performance of DL models. To maintain brevity, we intend to provide a detailed explanation of AUC in the main text rather than in the abstract. Additionally, in the abstract, we aim to highlight the difference in time demands between manual annotation and visual annotation, emphasizing that visual annotation significantly reduces labor time. If our interpretation of this comment is incorrect, we kindly ask for clarification.

In the abstract, we intended to convey that the time required for annotation using the eye-tracking method was 36.5 seconds per instance, which is only 4% of the time needed for manual pixel-by-pixel annotation (14.2 minutes per instance). The AUC = 0.984 and ACC = 0.931 refer to the validation results of our proposed DL model in the five-class classification task. Due to the word limitation, the abstract was condensed to its fullest extent to capture the essence of the work. Based on your suggestion and the latest experimental results, we have revised the abstract accordingly.

Comment 5# p 1, line 22+23: (A single WSI ... level [10]) -> it is unusual to have a whole sentence in brackets. Same on p. 2., l. 12+13.

Response:

We have checked the manuscript and removed all whole sentences in brackets to outside of brackets, as well as rewrite the corresponding sentences in the revised manuscript.

Comment 6# p 1, line 24+25: Through fine guidance ... precise diagnostics [11]. -> are there precise numbers that can be reported?

Response:

Yes, the precise diagnosis is evaluated using AUC (areas under ROC curves). To provide precise reported results, we conducted further investigation. [11] only contains a summary and cites the results reported in [12].

Comment 7# p.2, l. 16: what exactly do the authors mean by patch? - Show an example what this patch is on the WSI image

Response:

The term "patch" first appears on page 1, line 23 of the original manuscript, where referring to the process where a WSI is divided into small square tiles (224×224-pixel). In the field of pathology-assisted diagnosis, the term "patch" is commonly used to describe these smaller tiles after segmentation. To enhance the readability of the manuscript, "patch" has been replaced by "image patch" and is explained in detail upon its first appearance. Additionally, the image representing the "patch" has been reflected in Figure 4 of the original manuscript: the image input into the network is several small patches in the WSI.

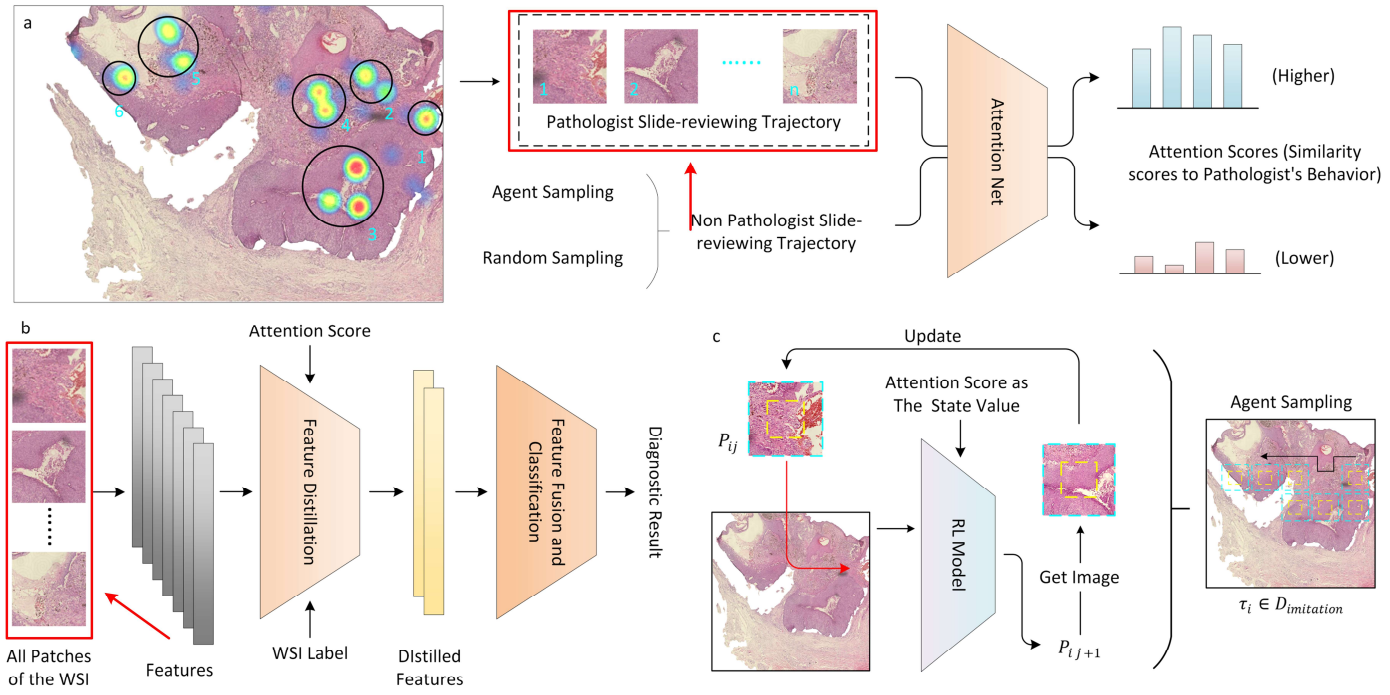


Figure 4. Model details of the three-part PEAN. a. Model for extracting pathologists' expertise; **b.** Feature distillation classification model using the extracted pathologists' expertise; and **c.** Model imitating pathologists' slide-reviewing behavior using reinforcement learning.

Comment 8# p.2, l. 19+20: pixelwise annotated map, the pathologist's visual attention map, the expertise value heatmap, and the suspicious region map selected -> the authors should provide more details how they calculated the different maps. For the heatmaps it is also important to report the parameters used to calculate the maps, also if they considered fixation durations - depending on the way and parameters chosen, the heatmaps look different and they are difficult to compare.

Response:

We have added a more detailed description of how these four types of heatmaps were generated in the revised manuscript. Specifically, 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 in the revised manuscript shows this comparison for four WSIs (corresponding to the four malignant diseases investigated in this study); their precise lesion area contours (Figure

2.a, directly drawn on the WSIs by pathologists) are shown alongside the pathologists' ROI heatmaps (Figure 2.b) and the heatmap of expertise values output by PEAN (Figure 2.c). In Figure 2.b, the observation points of pathologists, captured at 60 Hz by the eye-tracker, are mapped onto the WSIs. A circular convolution kernel is used to calculate the density of observation points within a certain range around any position in the image. For any given position on the WSI, the more observation points within a circle of 100 pixels radius centered at that, the higher the heatmap value assigned to that. Once the heatmap values for all positions on the WSI were calculated, they were normalized, resulting in the output. Figure 2.c shows the "pathology expertise value" computed by PEAN for each patch, with higher values indicating a higher potential diagnostic relevance as predicted by PEAN. The corresponding calculation of PEAN is described in detail in Section 4.2. Figure 2.d illustrates the positions of interest selected by a variant of PEAN, called PEAN-I. PEAN-I is described in detail in Section 2.6, and it autonomously selects a series of consecutive positions, resembling the movement of a pathologist's observation points, in an effort to simulate the behavior of pathologists.

Comment 9# p.2, l. 23: Internal and external test set - difference should be explained at first usage and not in line 42+43 - why is hospital G data considered as external data?

Response:

We collected digital pathology images from two independent hospitals to assemble two datasets in which we define them as the first ("internal") and the second ("external") datasets. The first dataset collected from Hospital F is defined as an "internal" dataset because this dataset is randomly divided into two subsets namely, one is a training subset, and one is a testing subset. All models are trained using data collected from Hospital F. The second dataset collected from Hospital G is considered the external dataset because WSIs from this institution are used for testing the models only. The terms "internal" and "external" are defined relative to the source of training data. In the revised manuscript's introduction, we have provided explanations for the internal and external datasets when these terms are first used. Additionally, more details are provided in Section 2.1. This is in accordance with Nature Communications' typical recommendation to include "Datasets" as the first section of the "Results" to avoid redundancy in the manuscript.

Comment 10# p.2, l. 24: Its classification performance ... -> provide more details in how far the authors' methods outperform existing ones

Response:

We have added quantitative number to describe how far our new Model outperform other existing DL models in the revised manuscript as follows: "Its classification performance and robustness significantly surpassed existing fully supervised and weakly supervised learning models. For example, PEAN-C outperformed the second-best model by 5.5% in ACC of external testing set."

Comment 11# p.2., l. 27+28: PEAN provides a human-like step-by-step search - is it because it follows the pathologists' scan path for analysis?

Response:

The "human-like step-by-step search" provided by PEAN-I (described as PEAN in the original manuscript) does not simply replicate the areas that pathologists have already focused on and then proceed with the diagnosis. Doing so would make the model's reasoning overly dependent on manual input in the real world, as the AI model would require an additional review by pathologists to make a diagnosis. Instead, PEAN-I offers an alternative approach: constructing an intelligent agent capable of autonomous exploration on the WSI. At each step, the agent captures a portion of the image centered on a specific location and determines its next move based on the image and the learned pathologist expertise. This gradually mimics the visual trajectory pathologists follow when reviewing WSIs, ultimately making a diagnosis based on the selected images. This model operation, which mimics the diagnostic process of pathologists as they browse different regions of the WSI, is referred to as "human-like pathology diagnosis". We have revised the manuscript to clarify this and prevent misunderstanding for the readers.

Comment 12# p.2., l. 29 + 30: provide more details to the statistics, also report the full results of the statistical tests, not only the p value.

Response:

In the revised manuscript, we added description of that we used a paired t test to calculate p-values reported in the manuscript, which is a commonly or widely accepted statistics method used to compare whether there is statistically significant difference or performance improvement between different DL models reported in the literature. However, to maintain the brevity of the manuscript, the detailed setup of the statistical validation is provided in Section 2.6 of the manuscript, rather than in the introduction.

Comment 13# p.3., Figure 1, 1e: the heatmaps for BCC and Melanoma look different to the others. It seems to be that there were quite a few fixations on the tumor - e.g. in BCC seems only 3 fixations.

Response:

Yes, these heatmaps look slightly different, which is expected because pathologists exhibit slightly different observation preferences for different types of lesions (due to the intra- and/or inter-reader variation). For example, Figure 1.e, shows that there are denser fixations by pathologists on the other three types of diseases. The supplementary files 1-10 include a substantial number of examples to help readers understand the preference variations from pathologists.

Comment 14# p. 2, l. 15: Heading of Chapter 2 - Results - This heading does not fit the content of section 1 because it is more on methodological descriptions. The subsection 2-6 are more on results. Better to separate this.

Response:

Thank you for the suggestion. We agree that placing the detailed description of the dataset in the first section of the "Results" chapter may create a sense of disconnection. It might be more appropriate to include the dataset information in the "Materials and Methods" chapter. However, due to the journal's editorial policy, which recommends placing "Results" as the second chapter, ahead of "Materials and Methods", readers might find it challenging to fully understand the experimental results and research significance without first knowing the dataset details. After reviewing related studies published in Nature Communications in the field of pathology-assisted diagnosis, we found that positioning the "Dataset" as the first section of the "Results" chapter appears to be a common practice. Therefore, we are inclined not to change its current placement.

Comment 15# p.2., l. 46: "EasyPathology" -> the used eye tracking system is not sufficiently described - even in Appendix 1. Which eye tracking model was used (e.g. vendor, head mounted, remote, mobile) what was the sampling rate, accuracy - how was the system prepared, set-up- calibrated? What was the distance of the participants to the screen as well as the size of the screen, how were they instructed, did they decide when a new WSI was displayed - or was there a fixed time period for each WSI. Were there re-calibrations during the experimental procedure? The reference [42] mentioned in Appendix 1 is not in the references (p. 14+15).

Response:

We have incorporated the detailed descriptions of the equipment used in the experiments (primarily including EasyPathology and the eye-tracker) and the implementation settings into Section 4.1. We then removed the redundant content in Appendix 1. In brief, the eye tracker named as "Tobii Pro Spectrum" is a screen-based eye tracker made by Tobii company. The device is calibrated before eye tracking experiment based on the device manual provided by the company. To collect eye-tracking-based WSI review data from pathologists, we also developed an "EasyPathology" slide review software, which allows pathologists to review WSIs while simultaneously using an external eye tracker to capture their eye movement signals and then are mapped onto the 2D WSI. The eye tracker records eye movements at a sampling frequency of 60Hz and uses built-in algorithms to map the pathologist's gaze position on the screen. More details can be read from the revised Section 4.1 of the revised manuscript. Additionally, we have also checked reference [49] and corrected the citation error in the revised manuscript.

Comment 16# p.4, l. 8: we determined that data could be collected continuously for 50 minutes a time (detailed in Appendix 2) - Appendix 2 does not really explain how "Initial Fixation Scale" and "Number of searches" were measured and analyzed. Also, there is a reference to Figure 1.b (Appendix 2, l. 28) which shows this data as a graphic - the figure cannot be found in the paper. Also, it is not clear how authors came to the numerous factors affecting pathologists' diagnostic accuracy, nor how they came to the correlation coefficients.

Response:

In this manuscript, we explain or analyze many factors that may enable pathologists to make accurate judgments are what we attempt to explore in the manuscript. At the beginning of attempting to incorporate these factors into the study, we did not know whether they could truly affect pathologists in making correct diagnoses. Later, we used a correlation matrix to analyze whether these factors are truly related to pathologists making correct diagnoses. The correlation matrix among these factors is shown in Appendix Figure 1.a. From the results, a considerable portion of these factors have no obvious correlation with correct diagnosis. We added explanations and measurement methods for each “first fixation scale” and “number of searches” in Appendix 1. The correlation among these factors is represented by the covariance between them.

Additionally, Appendix Figure 1.b shows a line graph that shows the changes in two key pathologist visual characteristic values over time during continuous slide review. At 50 minutes, both of these values showed significant changes.

Comment 17# p. 4, l. 8: what is meant by during the process and that the images had to be relabeled - in which situations?

Response:

In the original manuscript, we intended to convey that pathologists needed to make diagnoses during the data collection process without knowing the true labels, in order to evaluate the potential relationship between their diagnostic accuracy and visual behavior. This approach also better reflects the clinical setting where pathologists make initial diagnoses of cases. We have made the following modifications to eliminate confusion:

“During this process, the actual labels of WSIs were concealed, and the pathologists were asked to make new diagnoses. This more closely resembles the scenario in which pathologists make an initial diagnosis in a clinical setting.”

Comment 18# p.4., l. 13+14: do it mean that 5 pathologists did the manual annotations and eye tracking test? Although it may be hard to find pathologists it is a quite small number for reliable results. Also, if they do both conditions, there may be priming effects.

Response:

Yes, 5 pathologists did manual annotation and eye tracking test. We have made it clear in the revised manuscript. We agree that it is difficult to find many pathologists to participate in a research project. Despite this limitation, we believe that this is a valid study because we do not directly use raw eye tracking data, which can be quite noisy due to inter-reader variability. Instead, we developed a unique DL model (PEAN) to process the raw data by filtering noise, and finally predict the diagnosis results.

Additionally, we also conducted several tests regarding the impact of the number of pathologists participating in the study. For example, experimental results as shown in Figure 3b suggest that the more pathologists included in the study, the more accurate the diagnoses made by the DL model. Thus, our new PEAN models trained using different pathologists’ review data exhibit performance differences. But encouragingly, the PEAN model trained on five pathologists’ data has achieved highest performance in multiple tests as comparing to other existing DL models. This indicates that, despite the limited expansion of our training data sources, the high quality and expertise of the participating pathologists have endowed the model with strong diagnostic capabilities. We have added more discussions of this issue in Section 3.4 of the revised manuscript.

We have carefully considered the potential risk of bias introduced by the priming effect. A pathologist producing two types of annotations (one based on gaze-tracking and the other based on manual annotation) could introduce interference in the labeling process. It is true that the manual pixel-by-pixel annotation of WSIs was also performed by the five pathologists involved in the

study. However, the initial review of these WSIs was completed with the gaze-tracking equipment. The pathologists' visual behavior during the review process resembled their clinical workflow. Subsequently, the pathologists performed manual pixel-by-pixel annotations on these WSIs without time restrictions. Therefore, the priming effect only influenced the manual annotation process. During this process, the pathologists needed to meticulously examine the WSIs and make pixelwise annotations. Retaining some influence from the initial review is unlikely to pose a risk to the validity or authenticity of the manual annotation work. The collected manual annotations are used solely for comparison purposes and are not involved in the PEAN training process.

Comment 19# p.5, l. 19+20: Why were these fully and weakly supervised learning tools were selected as a performance comparison test? This should be motivated.

Response:

Fully and weakly supervised learning are two commonly used methods in developing DL models of medical images. Both have advantages and disadvantages. In this study, we developed and tested a new method aiming to more efficiently and cost-effectively annotate medical images and train DL model. The motivation of comparing our new PEAN model with both fully supervised learning models and weakly supervised learning models is to demonstrate whether using our new PEAN model can achieve the improved performance in skin lesion classification. This is an important comparison to support the significance of developing and applying this new image annotation method in the field of developing DL models of medical images in the future. The six models included in the original manuscript have classical architectures, providing extensive inspiration for future researchers. To further enhance our study results, we conducted new experiment by adding two novel weakly supervised learning DL models for comparison in the revised manuscript (please see Table 1).

Comment 20# p.6, Figure 3 c: The legend is not clear - what does distilled by PEAN-I and Original Image mean?

Response:

PEAN-I acts as an agent that autonomously selects a series of regions on the WSI by scanning it in a manner like the gaze patterns used by pathologists. Each step has a fixed stride and movement direction (upward, top-right, right, etc., totaling eight directions), as shown in Figure 2.d.

We further demonstrate that the regions selected by PEAN-I are diagnostically relevant and that PEAN-I can be effectively integrated with the existing weakly supervised learning models. In the first step, PEAN-I autonomously selects a subset of regions. In the second step, we trained existing weakly supervised learning models only in these selected regions, ignoring the unselected regions, which is described as "distilled by PEAN-I" in Figure 3.c. For comparison, the weakly supervised learning models were also trained on the whole region, described as "original image".

Additionally, as shown in Figure 3.c, when CLAM, ABMIL, and TransMIL are trained using the regions autonomously selected by PEAN-I, both ACC and AUC increased on two test datasets. This improvement in performance is statistically significant, with p-values of 0.0053 and 0.0161, respectively, as determined by a paired t-test. The effective enhancement of the DL model's performance demonstrates the effectiveness of mimicking pathologists' behaviors, reflecting PEAN-I's ability to "learn" from pathologists' expertise, while also providing strong evidence for the validity of this expertise.

Comment 21# p.7: The paper is missing a critical discussion of the approach and suggestions for future research

Response:

We have added a new subsection to the Discussion section, which provides a critical or forward-looking discussion on the future development of our work. In brief, our discussed work includes to (1) realize more cost-effectiveness of gaze-tracking deployment, (2) more effectively eliminate the impact of possible human error in gaze-tracking data, (3) develop a new cost-effective scheme for training junior pathologists. Please read the details presented in the revised manuscript, (4) current limitations.

Comment 22# p.8, l. 4: why do the authors present the Methods section as the last chapter of the paper. Would it not be better to explain the methods before the results are reported and discussed?

Response:

We agree that explaining the methods before reporting and discussing the results can make it easier for readers to understand. However, due to the editorial policy, we regret that we may not be able to place the “Methods” section before the "Results" section without permission. The required manuscript structure by the editorial office is as follows:

Title –Abstract – Introduction – Results – Discussion – Methods – Data and Code Availability – References.

We have adjusted the order according to the editorial office's guidelines. To mitigate the risk of readers finding it difficult to understand the experimental results without knowledge of the experimental setup, we have clarified the purpose and provided a brief description of the experimental setup before presenting each result in the “Results” section.

Comment 23# p.8, l. 15: ... as shown in Figure 1.f - there is no label f in Figure 1 on page 3.

Response:

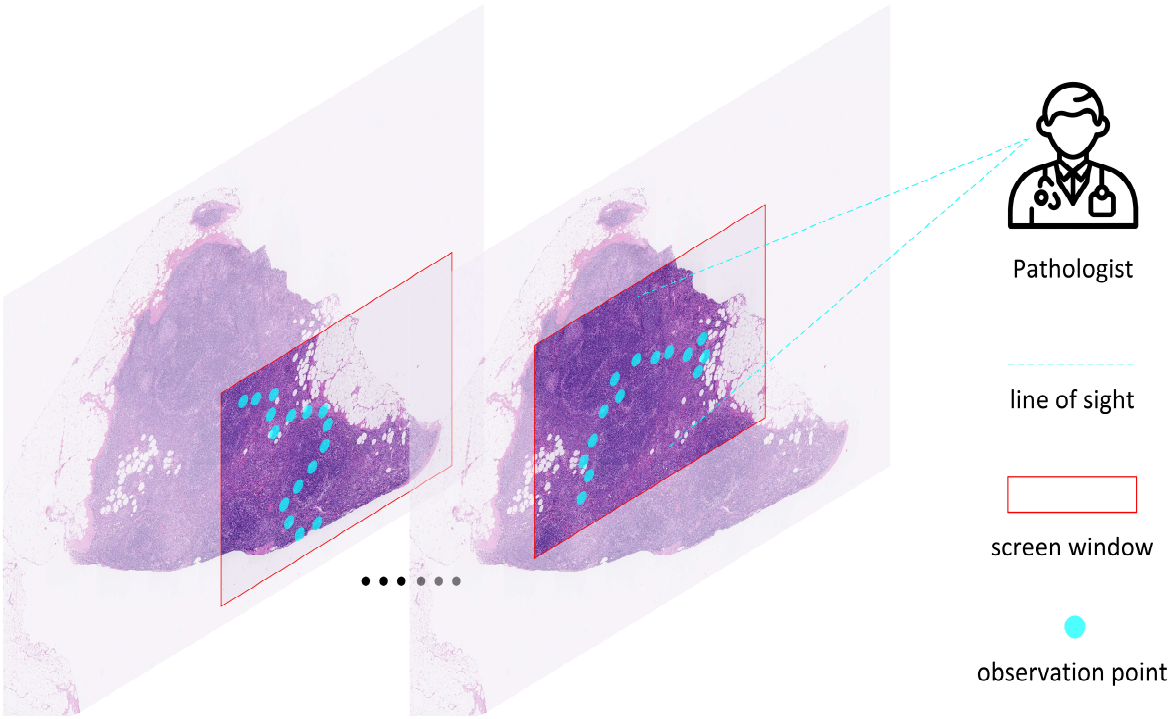
We apologize for this typo. The original Figure 1.f should be referred to as Figure 1.e. We have added a schematic diagram to describe the pathologist's review of WSIs and uploaded a video recorded during the pathologist’s review of WSIs.

Comment 24# p.8, l. 17: There should be more details on "Easy Pathology" and the used Tobii Eye Tracking System. It would be nice to add a picture showing how Easy Pathology looks like, the monitor where the software runs on and the position of the eye tracker and how the subjects operate it (e.g. what does p.8, l.22 " ... including the movement of the WSI within the software window" means? And more details on the subjects preparation and calibration/re-calibration.

Response:

Following the suggestion, we have added a new figure and a new video, with detailed explanation in figure caption in the revised manuscript. Please see Appendix Figure 3and Supplementary files 1-2. More details regarding subject preparation and device calibration have also been added in the revised manuscript.

In addition, regarding “including the movement of the WSI within the software window”, we provided a clearer explanation for readers: since the pathologist’s screen is typically not large enough to display the entire WSI, the complete review process involves a “sliding window operation” (or described as window panning over the WSI). Therefore, the eye tracker converts the signals into position coordinates on the screen, and EasyPathology further maps these coordinates onto the WSI.



Appendix Figure 3: Schematic Figure of sliding window operation and observation points

Comment 25# p.8 , l.17: there are two times the words "of the"

Response:

We have corrected typos in the revised manuscript.

Comment 26# p. 8, l. 30: The collected ... check for completeness, stability of gaze points, absence of missing data, and any occurrence of offsets"- the authors should stated in more detail what they mean by these terms in the context of their work.

Response:

Yes, we have added more details to describe the meaning and working function in the revised manuscript. For example, evaluating the integrity of observation points refers to identifying interruptions in the review process caused by autonomous behaviors of the pathologist (e.g., leaving the workstation, taking phone calls), which can result in extended “vacuum periods” where no eye-tracking data is collected. Stability refers to ensuring that the eye movement angular velocity of the pathologist does not remain at a high level for prolonged periods. Eye movement angular velocity greater than 30 degrees per second typically indicates eyelid twitch. If such anomalies occur too frequently in the review data for a particular WSI (set as more than three occurrences), it suggests that the pathologist may not be in an optimal working condition.

In addition, missing data refers to incomplete reviews, which may occur due to operational errors (e.g., repeated clicking or exiting the software prematurely) by the pathologist. Drift refers to a gradual positional shift in the eye-tracking signal as continuous viewing time increases. Although the eye-tracking device used in the study allows for a significant range of head movement and maintains long-term accuracy, there is a potential risk of drifting after extended work periods. Data cleaning for drift is performed by the pathologists themselves, who can review their recording after data collection (a feature provided by EasyPathology, which allows replay of the pathologist's observation point movements, as shown in the uploaded videos) to determine whether drift occurred. One important method to prevent drift is to set a reasonable maximum duration for each continuous working session and to recalibrate the eye tracker at the start of each session. Such details have been added to the revised manuscript.

Comment 27# p. 9, l. 3: The slide-reviewing data collected by the eye tracker can be regarded as a point set - Do you mean here sample/event data - depending on which data from the eye tracker you used for your analysis. I find the use of the term Points with

regard to eye tracking data quite confusing - because the eye tracker provides sample data every 16,6 ms (1000 Hertz/60 Hertz sampling rate of used eye tracker). You can use the eye tracker vendor software to calculate event data out of this sample data file.

Response:

We acknowledge the potential confusion in the initial draft regarding the description of the pathologist's observation points. The revisions we made are similar to those mentioned in our response to Comment 1. The “points” referred to in the comment are intended to mean “the area of the pathology image that the pathologist observes at a given moment”. These data are derived from the eye-tracker’s recordings of the pathologist’s observation positions, and in the revised manuscript, we consistently refer to these as “observation points” or “observation positions”. We replaced some inaccurately referenced terms, for example, by using “the slide-reviewing data collected by the eye-tracking device contains the observation points captured at a specific frequency, described as a point set $\{Points\} \in R^{points\ number \times (1,1)}$” instead of “the slide-reviewing data collected by the eye tracker can be regarded as a point set $\{Points\}$”. Additionally, we replaced terms like “sampled data” with more precise phrases, such as “sampled observation points” or “sampled WSIs”, to ensure clearer understanding for readers.

We understand that some of the specific terminology used frequently in this study may not be easily understood by readers. Therefore, we provided detailed explanations the first time these terms appear in the manuscript. We also conducted a thorough review to correct any inconsistencies or confusion regarding terminology throughout the manuscript.

Comment 28# p. 9, l. 3: The (1,1) in the formular means here over the whole size of W?

Response:

We apologize for this mistake. Since the screen cannot display the entire WSI at once during the review, the pathologist must use a “sliding window” to examine different regions of the WSI. In Equations 1 and 2, uppercase “W” to denote the specific WSI, while lowercase “w” referred to the portion of the image visible through the pathologist’s screen window. The equations intend to express how, during a fixed period of review with a stationary screen window, gaze time coefficients and gaze density coefficients are calculated using the coordinates and number of the pathologist’s observation points. To avoid confusion due to the visual similarity between uppercase and lowercase symbols, we replaced the lowercase symbol “w” representing the window with “win” in the revised manuscript.

Comment 29# p. 9, l. 4: the maximum tissue frame bw - is this similar to what is depicted in Figure 2a but on basis of the fixation cumulations, i.e. the tissue area including the majority of the fixation. The term "maximum tissue frame" is confusing - better to think about a more exploratory term

Response:

The blue boundary line shown in Figure 2.a was manually drawn by the pathologist to demarcate the boundary between the lesion and normal tissue. In the original manuscript, we intended to describe the boundary between the tissue in the WSI (foreground) and the white background, which we had previously defined as "the maximum tissue frame". To avoid possible confusion, we have revised this term to “the boundary of foreground”.

Comment 30# p. 9, l. 6: W is divided into windows of size [d x d] at a magnification M- I assume that this step is to find those areas in W with the highest amount of fixations - as depicted in Figure 2d. The authors should state what d and M are? Did the pathologists used different screens during the eye recording? This refers to my previous comments that more details on the "EasyPathology" should be mentioned in order to make the process (WSI checking and eye gaze recording) more clear to the reader.

Response:

We have further clarified the motivation and method for splitting the WSI into multiple $d * d$ window images. When reviewing a WSI, pathologists often need to perform multiple “sliding window” actions, meaning they can only view a portion of the WSI at any given time. To mimic this limited view during slide review and decode pathology expertise from such behavior, we sampled $d \times d$ window images from the WSI at magnification M, corresponding to the pathologist’s fixed window position during

the review, and denoted these as $\{win_i^M\}_i^K \in R^{K\times[d\times d]}$. When the pathologist performs multiple "window slides" during the review of W, all "screen images" observed by the pathologist (a total of K) are sampled and used for training PEAN. The size of d is set to 1,920 (corresponding to the screen size of 1,920×1,080) with the sampled window images extending both vertically and horizontally to satisfy the typical square image input requirement of convolutional neural networks. The magnification M corresponds to the zoom level selected by the pathologist while viewing the WSI.

Comment 31# p. 9, l. 7: in the formular i is the number of subwindows (0,...,K) W is divided into depending on chosen M and d?

Response:

As explained in our response to Comment 30, the count of "sliding windows" observed by the pathologists during their review, denoted as K, were obtained from the review data rather than sampled in a fully covered manner from the WSI.

Comment 32# p. 9, l. 8: I assume that you mean with set of points the samples in the subwindow here - see my previous comment above.

Response:

We apologize for the potentially confusing representation of these variables in the original manuscript. As explained in our response to Comment 1, this subset represents a collection of observation points recorded by the eye tracker, specifically represented as a series of coordinates.

Comment 33# p. 9, l. 9: How did you check that pathologists visual field extended beyond the window of the computer screen or the boundary of pathologists tissues in 1)

Response:

The pathologists' observation positions are recorded by the eye tracker. The screen resolution used during data collection was 1,920 × 1,080. When the recorded coordinates from the eye tracker show negative values or exceed the maximum screen resolution, it indicates that the pathologist's gaze is outside the screen, but the eye tracker is still able to calculate the coordinates of observation points that fall outside the screen. Through the segmentation method we reported in the manuscript, WSI is divided into a foreground part with only human tissue and a light-colored background part without any tissue. The boundary between the two can be computed. By comparing the pathologist's observation point coordinates with this boundary, it can be determined whether the observation point fell in the background. We have revised the related descriptions in the revised manuscript.

Comment 34# p.9, l.12: what were the values were selected for q1 and s - and why?

Response:

In this study, we set the threshold for distinguishing saccades based on the angular velocity of eye movements at $q_1°$ per second. Eye movements exceeding this threshold are considered invalid saccades. The symbol "s" means "second" and it is not a variable. We clarify this possible confusion in the revised manuscript. Specifically, when the angular velocity of the pathologist's eye movement exceeds $q_1°$ per second, the system records it as a saccade and removes the corresponding points. We have also revised this sentence and clarified the value of q_1 used in the experiment.

Comment 35# p .9. l. 15: Greater fixation - mean here higher number of fixations or also longer fixation durations?

Response:

Yes, a higher number of fixations requires a longer fixation duration. We have redefined the two types of pathologists' visual behaviors: fixation and search.

Comment 36# p. 9, l. 17: DBSCAN [43] is not in the reference list - is used here to extract fixations from the sample data - right? Or also to cluster fixations?

Response:

We apologize for the oversight that led to the missing reference list. We have now completed the reference list and provided a detailed explanation of the DBSCAN point clustering algorithm to demonstrate its suitability for classifying the pathologists' observation points into "fixation" and "search". It is an automated algorithm based on machine learning for clustering points.

Comment 37# p. 9, l. 17: Prior studies ... - state the references here - also state an explanation why the fixation zone is inadequate to localize diseased regions

Response:

Previous studies have reported that fixation zone may be inadequate to localize the diseased regions. We have cited a new reference (reference [38]) to support our viewpoint in the revised manuscript:

"Directly locating lesions by using the gaze area of pathologists is risky because pathologists cannot always focus only on those lesions. [38] shows that using the images that pathologists have focused on observing as completely equivalent to lesions for training will damage the performance of the model."

Comment 38# p. 9, l. 20: ... shifts in focus, ..., encompass underlying logical relationships ... - specify examples what this mean in the context of this work, the same with latent logical relationships p.9, line 21

Response:

We have revised description of "shifts in focus encompass underlying logical relationships" to clarify ambiguity to the readers. During the process of slide-reviewing, pathologists may not consistently focus on diagnostically relevant regions, which raises concerns about using such data to annotate WSIs. A robust DL model is needed to make reasonable inferences about the pathologist's slide-reviewing process and diagnostic logic. The contextual relationship refers to PEAN enhancing DL model training by capturing the similarities of images at different gaze locations and eliminating the interference of incorrect labels. The logic relationships mean that in the work of pathologists, WSI needs to be observed step by step, for example, starting from the epidermis and going from shallow to deep. Specifically, we refined our description of the proposed potential logical relationships and provided examples for clarity. Integrating the variable regions that pathologists have focused on and summarizing consistent and diagnostic rules from them is the key to PEAN and may attract widespread interest. Therefore, the relevant content is not only supplemented in the "Methods" but also formed into a new section and placed in the "Discussion".

Comment 39# p. 9, l. 21-23: Thus, by constructing these sequences and analyzing them as a whole, we can infer the logic and expertise of the pathologists ... - how exactly was this done and which conclusions did the authors get form the data analysis? Has this also been objectively and scientifically tested - e.g. by testing with experts at different levels of expertise?

Response:

To avoid potential confusion of the readers to the proposed method for "learning from the constructed sequence of fixation locations," we revised corresponding descriptions in the revised manuscript. In brief, we utilize a self-attention mechanism network like a Transformer to perform feature fusion on the observed images to obtain deep features containing the correlations between them. Then, the approach seeks to learn which image features which pathologists tend to focus on by comparing the differences between the features of images they focused on and those they did not, as the proposed "pathologist expertise" proposed in this

study. The purpose of PEAN is to fit out a "pathology expertise value" for each image patch in the WSI. This value is demonstrated to have a significant correlation with the ground truth (as shown in Figure 2 and discussed in Section 2.2), indicating that PEAN can reflect the images that pathologists tend to focus on. It is worth noting that a DL model is an end-to-end architecture and usually does not consider summarizing conclusions at the human language level. The trained PEAN combined with subsequent networks can obtain a WSI classifier superior to all existing methods, which can provide objective and strong support for its effectiveness.

How to compute "pathology expertise value" is the central focus of Section 4.2: decoding the expertise from the sequence of images that pathologists have focused on. It may not be appropriate to elaborate on this method in detail in the first section. Therefore, we provided a brief overview of the proposed method in the first section, followed by the introduction of necessary variables and parameters to help readers better understand the primary methods discussed in Section 4.2.

Comment 40# p.9, l. 24: How exactly is a gaze area defined and how is this different from gaze regions (p. 9, l. 26)? Sometimes Pij is stated as gaze area (e.g. p. 24) sometimes as gaze region (e.g. p. 9, l. 32 - 33). This should be consistent.

Response:

We apologize for the inconsistent use of specialized terms in the original manuscript. As explained in our response to Comment 35, we categorized the pathologists' visual behavior into "fixation" and "search" based on the density of the observed points. The term "gaze" is more appropriate for describing general visual behavior captured by the eye tracker, while "fixation" is more suitable for describing longer attention to a specific location. Therefore, regions covered by observation points classified as "fixation" are now defined as "fixation locations", replacing the terms "gaze area" and "gaze region".

Comment 41# p.9. l. 29: how are the weight coefficients set?

Response:

Two weights, β_1 and β_2, were set to 0.8 in this study. As the exponent values in formulas 1 and 2 increase, the corresponding coefficients E_{time} and $E_{density}$ decrease. We have also included these specific values in the revised manuscript for clarity.

Comment 42# p. 9, l. 39/40: the authors should explicitly state how the different magnification are incorporated into the developed model, instead of just stating that they are incorporated

Response:

The incorporation of different magnifications involves inputting two images of different sizes (corresponding to x_{ij}^{M} and x_{ij}^{2M}) from the same location into the convolutional neural network, and then concatenating the two sets of features to merge the image information at different scales. We have added the detailed description of this incorporation process in the revised manuscript.

Comment 43# p. 10 , l.2 : experience should be explained.

Response:

As described in our responses to comments 38 and 39. Pathology expertise (using "expertise" to comprehensively replace "experience") is a concept defined in this study. It is specifically manifested as the score of an image patch, usually between 0 and 1. It is used to measure the potential degree of attention of pathologists to a certain image patch. The calculation method of pathology expertise is given in Section 4.2 of the manuscript.

In this study, we have drawn the following two inferences: (1) the complete slide review process of a pathologist can be simplified as transitions between fixation locations, where search behavior helps the pathologist find the next fixation location; and (2) the pathologist derives diagnostic evidence from at least some of the images at these fixation locations.

The potential contextual relationships between fixation locations offer several advantages when analyzed as a whole rather than individually: (1) For certain diseases, pathologists need to assess multiple tissue features to make a diagnosis; (2) Multiple

fixation locations likely target images of potential lesion areas, and capturing the similarities between these images can enhance the training of DL models; (3) While pathologists may observe irrelevant areas, correct diagnoses always depend on the fixation areas, so treating these observed images as a whole helps eliminate the interference of incorrect labels.

For example, labeling an image that was heavily observed by the pathologist but unrelated to the diagnosis as a positive sample for a certain disease would impair the DL model's training. These images that the pathologists focused on are structured into a sequence containing the diagnostic evidence. By comparing this sequence to images that the pathologists never observed, we aim to learn which image features pathologists tend to focus on, representing the "expertise" learned from the pathologists.

The above content is the core idea of how this study obtains labels for training DL models from chaotic eye-tracking data and avoids the damage caused by incorrect labels. It has been fully discussed in Section 3.2 of the manuscript.

Comment 44# p.10, l. 5: The aim of the partition function Z should be explained

Response:

In Equation 3, $p(\tau)$ represents a probability distribution, and therefore the integral of this function should equal 1. The partition function Z serves this purpose. We have added further explanation of Z in the revised manuscript.

Comment 45# p.10, l. 12+13: the pretrained encoder should be explained in more detail - how was it pretrained to get which output? Also, it should be stated what the single variables of the feature vector exactly stand for

Response:

We have added a brief explanation of the pre-trained encoder to align with the setting of the pre-trained image encoder in Section 2.1 of the revised manuscript. Specifically, the convolutional neural network was pre-trained on ImageNet as the image encoder. Additionally, we included explanations of the variables representing feature vectors.

Comment 46# p.10, l. 19: Equation (5) it should be ETime

Response:

We have corrected this typo in the revised manuscript. The corrected symbol is "E_{time}".

Comment 47# Appendix Figures on page 12: why do the authors put these images on an extra page before the Appendix begins on page 13. Would it not be better that they are part of the Appendix starting on page 13?

Response:

Yes, per reviewer's suggestion and complying with the editorial policy of the journal, all Appendix Figures (including the newly added ones) have now been placed in the "Appendices" section in the revised manuscript.

Comment 48# p.10, l. 27: MIL multiple instance learning)

Response:

We are sorry that the background description of multi-instance learning appeared in the method section of original manuscript, which is not common. We have added a description of multi-instance learning in the "Introduction" section of the revised manuscript, while in Section 4.3, we focus on describing the model in this study.

Comment 49# p.10, l. 28: state references

Response:

This comment seems incomplete. In our understanding, references should be added to explain the mentioned methods in detail. However, as mentioned in the response to comment 48, we have moved the comparison of previous studies in Section 4.2 to Introduction section. There, the advantages of the method proposed in this study compared to multi-instance learning are explained in detail.

Comment 50# p.10, l. 27+ 28: state the difference between basic MIL and attention-based MIL

Response:

Comparing the advantages and disadvantages of two types of multi-instance learning basic structures seems to be beyond the focus of this study. We are sorry for such redundant descriptions in the original manuscript. Although we have moved the background information of MIL to the Introduction section, we still focus on the comparison between PEAN and MIL rather than the internal comparison of MIL.

Comment 51# p.11, l. 3: it is not clear what loss2 stands for, additionally loss3 and loss4

Response:

We have added descriptions for each loss function where they are first mentioned in the revised manuscript. In brief, (1) $loss_2$ is a cross-entropy loss function that measures the difference between predicted probabilities and true labels. Its goal is to make the model's predictions closer to the true labels. (2) $loss_3$ is a cross-entropy loss function aimed at bringing the predictions from the fully connected layer closer to the true labels. (3) $loss_4$ is a mean squared error loss function. Its purpose is to bring the expected reward as predicted by Q_{eval} closer to the actual reward, encouraging actions that maximize both present and future rewards. This allows the pathology expertise decoded from $f_{experience}$ to be transferred to Q_{eval}.

Comment 52# p.11, l. 9: Thus, which feature fusion methods the authors used, those in equation (10) / (11)?

Response:

We used a feature fusion method namely, transformer, which is currently one of the best-performing network architectures. It is based on a self-attention mechanism to merge instance-level features into WSI-level features. The information has been added to the revised manuscript.

Comment 53# p.11, l. 12: Qeval and Qtarget should be explained in more detail and also how they are updated in real time.

Response:

We have revised descriptions regarding the structure and update methods of the two networks (Q_{eval} and Q_{target}) in the revised manuscript. In brief, RL model consists of two networks, Q_{eval} and Q_{target}, both of which use fully connected layers but have different parameters. Trough an optimal training process, pathologists' expertise extracted by PEAN (represented by $f_{experience}$) can be estimated by Q_{target}. $loss_4$ is a mean squared error loss function, aimed at aligning the expected reward predicted by Q_{eval} with the actual reward, thus encouraging actions that maximize both present and future rewards. This allows the pathology expertise decoded from $f_{experience}$ to be transferred to Q_{eval}. Please read the detailed revision in the revised manuscript.

Comment 54# p.11, l. 16: fIRL should be explained

Response:

We apologize for the oversight that led to $f_{experience}$ being mistakenly written as f_{IRL}. We corrected this typo in the revised manuscript.

Comment 55# p.11, l. 41: it should be Qeval in equation (13)

Response:

We apologize for this typo again. The symbol in the revised manuscript has been corrected to "Q_{eval}".

Comment 56# Appendix Figure 1 b: this figure should be explained in more detail in the caption / text - how were the two dimensions First fixation scale (pixel squared) and searching number be calculated

Response:

Yes, we have expanded caption of Appendix Figure 1 to provide more detailed information. We have also reported calculation methods for these two variables in Appendix 1, related to Appendix Figure 1, in the revised manuscript.

==========

Responses to Reviewer #3 (R3)

Comments # This study reports on the development and performance of an AI-model (PEAN) for classifying 5 common skin lesion categories from whole slide images (WSI). Novel machine learning methods were employed to overcome the large workload required for manual pixel-wise annotation of WSI. The authors used eye-tracking software, and recorded any zooming or panning the WSI combined with final diagnosis, to label image review patterns for ML. As a result, PEAN generates 'expertise of pathologist' output in providing lesion classification, and to map out an imitation of pathologists' review trajectory of the WSI.

The quality of the manuscript is high, the study objective is novel and exciting, I believe will be of high interest to readers. Furthermore, the manuscript is well written and easy to read, with relevant research cited accordingly.

Response:

We thank the reviewer very much for taking time to review our manuscript and for providing encouraging and supportive comments. The revisions made in manuscript are highlighted in Green Color.

Comment 1# In the abstract and introduction you have referred to manual pixel-wise annotation as a 'waste' of manpower/medical resources. It would be better to describe it as having a high burden, rather than 'waste', as research to date using these methods, including your own methods for ground truth are still important to the research process, and should not be referred to as 'waste'.

Response:

We fully agree with reviewer's comment regarding the misuse or confusion of the word "waste" in the original manuscript. We have deleted the word "waste" and revised related sentences in the revised manuscript.

Comment 2# Can you provide a quantitative result of the overlap between pixel-wise annotation, eye-tracking, and PEAN outputs for the whole dataset? Currently only 4 qualitative examples (one from each lesion category) are provided. This does not allow transparency for overall overlap.

Response:

In the revised manuscript, we have added a quantitative analysis of the "pathology expertise values" fitted by the new PEAN model and the distribution of recorded pathologists' observation points. We analyzed WSIs with manual annotation from test dataset (92 WSIs) and, found a direct relationship between the manually annotated lesion areas (ground truth) and the high-attention regions identified by both the pathologists and PEAN. Referring to the data division described in Section 2.1 of the revised manuscript, these WSIs are all the available data in the testing set. Due to the high cost of manual labor, the remaining WSIs in the testing set have not been collected for manual pixelwise annotation. Calculating the overlap in the training set may not have practical benefits. The proportion of pathologists' observation points falling within the ground truth is 87.4%. Our findings revealed that the average pathology expertise value in ground truth regions was 0.822, in contrast to 0.357 in non-diagnostic regions, indicating a significant difference.

Comment 3# For the comparative models, there is no explanation on why they were selected (beyond picking 3 fully supervised, and 3 weakly supervised learning methods). Was the selection related to performance, e.g. were they the highest performing models reported? Was it related more to availability of code or other factors? Also, Table 1 should include the references with the codes listed (either in the table or in the caption).

Response:

In developing DL models in the medical image field, fully supervised and weakly supervised learning are currently two popular methods to train the models. Each has advantages and disadvantages. As reported in the previous literature, fully supervised learning models generally demonstrate higher performance, but it requires large resource (i.e., clinician's time) to manually annotate medical images. However, although using the weakly supervised learning methods can reduce labor cost, they have lower performance and

poorer robustness applying to new independent testing datasets. The goal of introducing and developing a new eye-tracking based image annotation method aims to combine advantages of both fully supervised methods (i.e., higher classification performance) and weakly supervised learning methods (i.e., reducing labor cost), while avoiding or minimizing disadvantages of these. For this purpose, we need to compare performance of our new DL model with other existing popular DL models trained using both the fully supervised and weakly supervised leaning methods reported in the literature. We have supplemented in detail the motivation for using each of the selected DL models.

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

Methods		Internal Testing Set				External Testing Set			
		ResNet50		CONCH		ResNet50		CONCH	
		ACC (%)	AUC	ACC (%)	AUC	ACC (%)	AUC	ACC (%)	AUC
Fully Supervised Learning	DLCCP [9]	89.3	0.969	91.7	0.977	74.2	0.905	83.7	0.902
	SLC [10]	90.3	0.976	92.1	0.978	75.5	0.893	87.1	0.950
	HSL [8]	91.2	0.978	93.1	0.985	76.7	0.908	87.5	0.942
Weakly Supervised Learning	CLAM [20]	86.1	0.961	88.8	0.961	70.1	0.864	77.9	0.867
	AB-MIL [16]	82.6	0.933	87.3	0.962	62.4	0.804	77.4	0.888
	TransMIL[17]	88.8	0.965	90.0	0.968	73.7	0.899	78.2	0.892
	DS-MIL [40]	88.9	0.959	90.4	0.972	73.2	0.880	80.5	0.884
	IB-MIL [41]	89.8	0.962	92.0	0.980	74.0	0.893	83.6	0.912
Ours	PEAN-C	**93.1**	**0.984**	**96.3**	**0.992**	**85.5**	**0.950**	**93.0**	**0.984**

Comment 4# Figures 3, and supplementary 1b refer to model PEAK-C, I assume this is a typo for PEAN-C?

Response:

Yes, this is a typo (sorry). “PEAK-C” should be “PEAN-C”. We have corrected this typo in the revised manuscript.

Comment 5# Did performance differ between lesion categories? There is a class distribution imbalance in the training dataset that is not discussed, and the results of accuracy per class are not reported.

Response:

Yes, classification performance differs between 5 skin lesion categories for several reasons. For example, (1) different types of lesions have different types of clinically relevant image features. (2) The number and diversity of 5 types of lesions are not uniformly distributed in our training dataset because the frequency of detecting each of these 5 types of lesion varies greatly in clinical practice or specific hospitals. All WSIs belonging to these five types of skin lesions detected from 2016 to 2022 in two hospitals were retrospectively collected, rather than being artificially selected according to the desired numbers. Therefore, the dataset distribution reported in the manuscript is close to the real world, even if it is unbalanced. We randomly divided internal dataset into training and testing subsets by maintaining the comparable ratios of these 5 types of skin lesions in both training and testing datasets to represent real patients’ distribution in two hospitals in this study. However, we believe that as increase of the size and diversity of training dataset in the future, the performance difference among 5 lesion categories will be reduced. We have added new data analysis results of performance difference of 5 lesion categories.

Supplementary Table 1. Recall of Each Disease with ResNet50

Methods	Internal Testing Set					External Testing Set				
	Nevus	BCC	Melanoma	SCC	SK	Nevus	BCC	Melanoma	SCC	SK
DLCCP	91.8%	54.1%	57.1%	64.9%	96.3%	87.5%	89.4%	58.4%	69.6%	12.9%
SLC	82.0%	78.0%	51.0%	68.9%	97.0%	93.0%	87.9%	36.4%	54.6%	19.2%
HSL	88.2%	87.4%	40.8%	72.9%	96.0%	86.8%	87.9%	54.5%	63.9%	39.9%

CLAM	77.7%	46.5%	75.5%	43.6%	97.3%	83.2%	65.9%	63.6%	39.7%	47.2%
AB-MIL	88.5%	37.7%	44.9%	15.6%	95.7%	89.4%	51.1%	46.8%	13.9%	9.1%
TransMIL	91.1%	76.7%	44.9%	51.1%	95.8%	82.0%	86.7%	66.2%	68.6%	32.5%
DS-MIL	82.6%	80.5%	67.4%	88.0%	91.6%	79.5%	94.0%	88.3%	76.8%	18.5%
IB-MIL	83.0%	79.3%	65.3%	84.9%	93.4%	44.3%	71.1%	63.3%	28.0%	86.0%
PEAN-C	86.2%	74.2%	57.1%	81.8%	97.3%	94.10%	89.7%	40.3%	45.4%	86.7%

Supplementary Table 2. Recall of Each Disease with Conch

Methods	Internal Testing Set					External Testing Set				
	Nevus	BCC	Melanoma	SCC	SK	Nevus	BCC	Melanoma	SCC	SK
DLCCP	96.1%	81.1%	69.4%	66.2%	96.0%	95.0%	93.7%	89.6%	82.5%	28.3%
SLC	91.2%	84.3%	67.4%	78.7%	95.5%	97.6%	93.7%	83.1%	59.3%	59.1%
HSL	91.2%	86.2%	65.3%	84.0%	96.1%	96.4%	93.1%	80.5%	89.2%	47.6%
CLAM	79.0%	78.6%	61.2%	85.8%	92.7%	73.5%	88.5%	87.0%	91.8%	70.3%
AB-MIL	95.4%	84.3%	26.5%	46.7%	93.2%	86.2%	94.3%	88.3%	76.8%	21.3%
TransMIL	96.1%	86.8%	28.6%	68.9%	93.7%	80.3%	84.0%	81.8%	79.4%	61.5%
DS-MIL	82.6%	80.5%	67.4%	88.0%	93.7%	84.1%	85.8%	81.8%	85.1%	57.0%
IB-MIL	94.4%	82.4%	63.3%	70.2%	96.2%	87.5%	87.9%	83.1%	84.0%	63.3%
PEAN-C	96.4%	86.2%	83.7%	88.4%	98.6%	96.8%	95.8%	88.3%	82.5%	83.9%

Comment 6# A little more information on the data acquisition will allow assessment of potential bias, for example, please clarify if the WSI selected from 5,107 represented all available WSI (with the targeted 5 classification categories)? Or were there any other relevant selection criteria? Can you also report the inter-rater agreement between the two pathologists who initially screened the diagnostic data label? i.e. what portion required arbiter review? This may raise a current 'hot-topic' regarding poor interobserver agreement between pathologists for reviewing histopathology, and should be discussed.

Response:

First, in the revised manuscript, we clarified that this study involves 5,107 patients and 5,881 WSIs because small number of patients have more than one (multiple) WSIs. We also reported that two WSIs were excluded due to low image quality. No other human control or selection criteria are added to the WSI collection process. As described in our response to Comment 5#, our image dataset distribution is kept consistent with the real-world situation to the greatest extent. Therefore, 5,881 represents all available WSIs under the target category. In addition, the possible "controversies" of WSIs by different reviewers are also further reported. This controversy includes image quality, diagnostic results, and annotation. When any pathologist judges that the image quality is unsatisfactory according to the standards reported in the manuscript, the corresponding WSI will be removed from the data set. Disputes over diagnostic results are very harmful and referees need to be introduced to discuss in detail and finally make a consistent diagnosis.

The inter-reader variability when different pathologists review WSIs and their impact have also discussed in the revised manuscript. Previous studies have shown that inter-reader variability can greatly reduce the effectiveness of manual annotation [1]. Different pathologists may have different judgments on the ground truth particularly, for diffuse tumors. This may lead to annotation confusion. For eye-tracking data, this situation may be even more severe because pathologists cannot always focus on the region most relevant to diagnosis. Thus, the collected eye tracking data or visual patterns are quite mixed. Decoding these mixed patterns into a shared embedded space to capture their consistency is difficult and is the core of this study. However, the designed structure of PEAN is quite compatible with this visual pattern to filter out the noise and extract the clinically relevant image features. We discussed in detail in Section 3.2 why the architecture of PEAN can effectively handle the confusion caused by inter-reader variability.

[1] 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

Comment 7# Section 3 of Methods discusses MIL. Please add a sentence or two at the beginning of this section to describe the relevance and use of MIL, including spelling out the acronym. Not all readers will be familiar with MIL.

Response:

There is a mistake in the description of multi-instance learning (MIL) in the original manuscript. MIL should appear in the Introduction section of the manuscript rather than the Methods section. We have added a description of multi-instance learning in the Introduction section, while in Section 4.2, we focus on describing the model.

Comment 8# Discuss limitations, especially regarding clinical transferability and what you foresee as the proceeding steps required. For example, you have not included any training methods for recognising out-of-distribution classification. This is crucial for clinical transferability – to ensure the model does not incorrectly interpret OOD classes as a false positive for other malignancy.

Response:

We agree that the current study has limitations in addressing the out-of-distribution (OOD) problem or challenge. This can inspire our further work rather than being caused by defects in the existing structure. This applies not only to PEAN but to most DL models. Novelty detection can be used as an incremental model to address the OOD problem. In addition, we have explored another effective solution: expanding the diverse training set.

Expanding the scale of training data to address the OOD problem requires fully considering the issue of data collection costs, especially since this study involves a new type of data for clinical use. In Section 3.1, we explain that collecting eye-tracking data from pathologists is time-efficient and has minimal disruption to existing clinical work. This provides good future development space for PEAN. The current shortcomings, such as OOD problem, are pointed out in Section 3.4.

==========

Responses to Reviewer #4 (R4)

Comments # I co-reviewed this manuscript with one of the reviewers who provided the listed reports.This is part of the Nature Communications initiative to facilitate training in peer review andto provide appropriate recognition for Early Career Researchers who co-reviewmanuscripts.

Response:

Sincerely thank you and another reviewer for providing valuable suggestions for this manuscript.

==========

In summary, we have carefully read and addressed each of the reviewers' comments. These thoughtful and constructive comments greatly helped us revise the manuscript to more clearly present this study and better demonstrate its significance and potential contribution to this emerging research field. We thank the reviewers again for spending time and effort to read and review our response letter and the revised manuscript. We believe that the revised version is more suitable to the broad readership of *Nature Communications*.

Sincerely,

Xiaoyu Cui

Corresponding Author

(On behalf of all other co-authors of this manuscript)

Responses to Reviewer #2 (R2)

Comments# All my comments from the first review process were answered satisfactorily and in detail and incorporated into the paper in an appropriate manner.

Response:

I sincerely express my profound gratitude for your meticulous review and esteemed acknowledgment of our work.

==========

Responses to Reviewer #3 (R3)

Comments# The authors have provided a very detailed and comprehensive response to reviewer comments, addressing all concerns with clarity and precision.
After reviewing the revised manuscript along with the detailed rebuttal letter, I have no further issues to raise. I am satisfied with the current state of the manuscript, and in my opinion, it requires no further revisions for publication.

Response:
I sincerely express my profound gratitude for your meticulous review and esteemed acknowledgment of our work.

==========

Responses to Reviewer #4 (R4)

Comments # I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Response:
I sincerely express my profound gratitude for your meticulous review and esteemed acknowledgment of our work.

==========

Responses to Reviewer #5 (R5)

Comments # This study developed a deep learning-based auxiliary diagnostic system for pathology images, utilizing eye-tracking devices to capture pathologists' image review processes, thereby providing additional supervisory signals for developing AI models. The experimental results presented in the manuscript demonstrate that this method can effectively quantify, and extract pathologists' knowledge based on gazing, thereby facilitating the training of models for tasks such as image classification and automatic annotation. With an additional set-up of a gazing tracker, collecting the eye trajectories as one new type of annotation in clinical

settings can improve performance with minimal incremental labor costs. Compared to the currently prevalent fully supervised and weakly supervised learning approaches, it has its advantages.

Four reviewers have conducted peer reviews of this manuscript:

Reviewer 1 expressed concerns about the method of relying on pathologists' eye-tracking data to train diagnostic models, raising three main issues: 1. Reliability: Pathologists' gaze patterns may not consistently target the most diagnostically relevant areas; 2. Scalability: Resource-intensive long-term collection of eye-tracking datasets; 3. Practicality: Weakly supervised methods are positioned as more practical and widely applicable AI-assisted diagnostic solutions compared to using eye-tracking data.

1. I contend that eye-tracking results need not always align with actual lesion images. The authors defined that "diagnostically relevant areas are more likely, but not absolutely, to attract the attention of pathologists." Eye-tracking data is introduced as "soft labels" to address the issue of inconsistency with the most diagnostically relevant areas, avoiding the accumulation of human annotation errors in the model. The experimental results in the manuscript also demonstrate that introducing eye-tracking data in this unique way indeed benefits AI model performance. Furthermore, the authors designed experiments showing that AI model performance improves under the guidance of multiple pathologists' visual patterns, reflecting that this method is not affected by errors in the eye-tracking data collection process.

2. While collecting eye-tracking data increases economic and time costs for clinical work, the workflow designed by the authors can only reduce cost expenditure, not eliminate it. Thus, the long-term resource consumption in promoting this research objectively exists. However, the authors also mentioned that the data collection process can be integrated into pathologists' daily work, allowing data acquisition with little extra labor time.

3. The experimental results included in the manuscript show that this method outperforms weakly supervised learning methods in performance, making it more suitable for complex clinical environments. The authors also mentioned that AI models trained under the guidance of eye-tracking data can continue training with new images and pathological diagnoses, thus the demand for eye-tracking data is relatively small. This allows annotations collected from previous pathologists to be transferred to new image datasets, reducing expansion costs.

Reviewers 2 and 3 acknowledged the promising prospects of this research and suggested revisions to multiple expressions in the manuscript, as well as supplementing relevant experimental results and discussions.

Reviewer 4 indicated that they co-authored a review with another early-career researcher and did not provide an independent review.

The authors provided detailed point-by-point responses to the comments of all four reviewers, making corresponding revisions and additions to the manuscript. Considering the manuscript and the comments from the four reviewers, I believe this study provides a relatively innovative approach. This method has its characteristics and potential application value in computational pathology, offering certain inspiration to peer researchers. After careful consideration, I recommend accepting this research for publication in Nature Communications.

Response:

I sincerely express my profound gratitude for your meticulous review and esteemed acknowledgment of our work.

==========

In conclusion, all reviewers have provided invaluable feedback and have concurred with the publication of this research. These thoughtful and constructive comments greatly helped us revise the manuscript to more clearly present this study and better demonstrate its significance and potential contribution to this emerging research field. We thank the reviewers again for spending time and effort to read and review our response letter and the revised manuscript. We believe that the revised version is more suitable to the broad readership of *Nature Communications*.

Sincerely,

Xiaoyu Cui

Corresponding Author

(On behalf of all other co-authors of this manuscript)