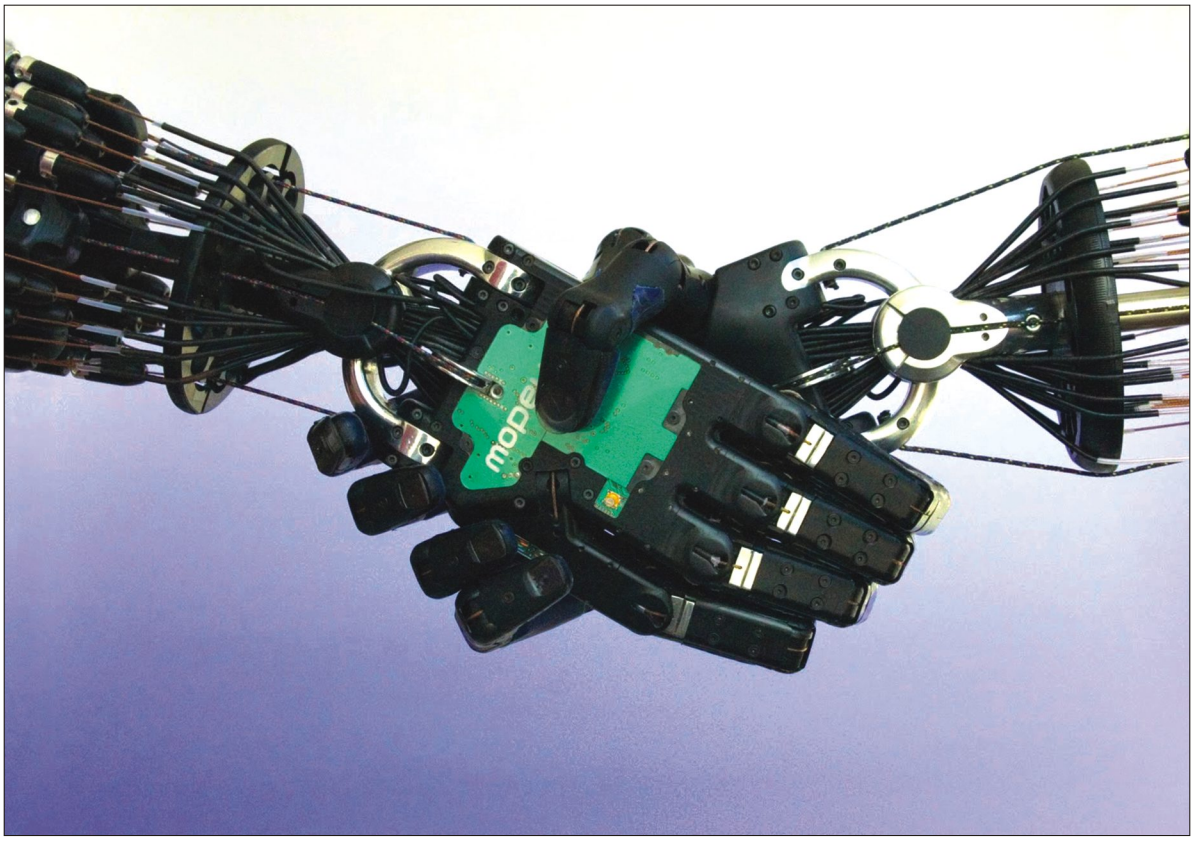

Journal of Artificial Intelligence

ISSN 1994-5450

science alert

ANSInet
an open access publisher
http://ansinet.com

OPEN ACCESS Journal of Artificial Intelligence

ISSN 1994-5450
DOI: 10.3923/jai.2026.78.93

Systematic Review
Artificial Intelligence for Thyroid Histopathology Image Analysis: A Systematic Review

[1]Gui Tao, [1]Razali Yaakob, [2]Sina Abdipoor

[1]Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, Malaysia
[2]Faculty of Engineering and Technology, Sunway University, Malaysia

Abstract

Digital pathology and whole-slide imaging have created new opportunities for computer-assisted thyroid histopathology; however, the high resolution of slides, sparsity of diagnostic regions and limited public datasets continue to make this task challenging. This systematic review highlights recent advances in artificial intelligence for thyroid histopathology within the broader context of whole-slide image analysis. The field shows a clear methodological shift from convolutional neural network-based patch approaches to weakly supervised multiple instance learning, transformer-based slide modeling, hybrid architectures and emerging paradigms such as self-supervised learning, foundation models and multimodal learning. Despite these developments, limitations in dataset availability, evaluation consistency and computational efficiency remain significant barriers. Issues of cross-institution generalization, interpretability and clinical trust are still unresolved. Overall, the evidence suggests a transition from local patch-level analysis toward more scalable, context-aware slide-level reasoning. Future progress is likely to depend not only on improved model architectures but also on richer datasets, standardized evaluation protocols and more rigorous validation frameworks.

Key words: Thyroid histopathology, digital pathology, whole-slide imaging, deep learning

Citation: Tao, G., R. Yaakob and S. Abdipoor, 2026. Artificial intelligence for thyroid histopathology image analysis: A systematic review. J. Artif. Intel., 19: 78-93.

Corresponding Author: Razali Yaakob, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, Malaysia

Competing Interest: The authors have declared that no competing interest exists.

Data Availability: All relevant data are within the paper and its supporting information files.

INTRODUCTION

Digital pathology and whole-slide imaging (WSI) have brought the field to a new level by digitizing traditional histopathology glass slides into images of very high resolution that include the complete morphological details of the whole tissue specimen. This allows for new possibilities for computer aided diagnostic systems and quantitative assessments in pathology. The analysis of WSI is not without some serious technical difficulties, since an individual slide can contain billions of pixels. In addition, regions of interest for a diagnostic are usually very rare, very variable and spread in different tissue zones rather than localized in a single spot. Such characteristics make WSI analysis different from a usual image classification task and models able to discriminate complex local morphological features while taking slide-level contextual information into account are needed[1-3].

Thyroid histopathology is a very important and challenging medical application. In clinical practice, thyroid nodules are encountered very often, but it is not always easy to distinguish between benign nodules (BN), papillary thyroid carcinoma (PTC) and follicular thyroid carcinoma (FTC). The diagnostic cues can be subtle, with a great amount of overlap and based on small-scale cytological details plus a more global tissue architecture. Also, the lack of publicly available thyroid image datasets and the fact that the interpretation can be very different for the pathologists are challenges for building robust computational models. Hence, thyroid pathology should not be seen as a simple image classification task[4-10].

The first works on deep learning applications in histopathology mostly used convolutional neural networks, in which whole slide images were cut into small patches for CNNs to learn particular local features. These studies showed that deep neural networks can find interesting histological features and help with pathological classification. Nevertheless, patch-based CNN methods have strong limitations, since predicting independently for each patch and then aggregating them with basic methods does not well capture long spatial dependencies between distant areas of the tissue. To reduce the labeling workload and to make learning at the slide level, multiple instance learning (MIL) has become a popular weakly supervised approach in computational pathology. More recently, transformer-based techniques have been adopted to more directly capture the relations among patch embeddings and thus WSI analysis is shifted from simple weighting of the instances toward slide-level reasoning that takes into account correlations among patches[4,11-15].

The use of transformer-based approaches in whole slide image analysis also brings new computational difficulties, since self-attention is costly when slides are represented by hundreds or thousands of patch tokens. This has motivated research on hybrid and computationally efficient architectures for long sequences, including CNNs for local feature extraction, weakly supervised methods for key instance selection, transformer-like structures for relation modeling, graph based frameworks for spatial reasoning and Mamba-based state-space models[16-20]. Besides architectural design, recent developments such as self-supervised learning, foundation models for pathology, graph-based reasoning and multimodal learning are also emerging. These methods can improve the transferability of learned representations, reduce dependence on annotated datasets, preserve tissue structure and mix imaging data with clinical records, text descriptions, or molecular profiles. In this sense, the present systematic review summarizes recent advances in artificial intelligence for thyroid histopathology image analysis within the wider field of WSI analysis and discusses the main methodological approaches, commonly used datasets, standard assessment methods, existing problems and future research directions in computational thyroid pathology[18,21-26].

METHODOLOGY OF THE SYSTEMATIC REVIEW

This review was conducted following the PRISMA 2020 reporting framework for systematic reviews. The PRISMA 2020 provides a 27-item checklist and revised flow diagrams intended to improve transparency in how systematic reviews are identified, screened, assessed and reported[27].

Review scope and research focus: Our aim with this review is to provide a systematic summary of the state-of-the-art in deep learning methods for WSIs, with a focus on the specific application of interest (thyroid) as well as other methods developed for thyroid WSI classification. The literature search was focused on the transition from CNN based patch modelling to MIL, transformer-based slide modelling, efficient hybrid long-sequence architecture; novel paradigms such as self-supervision, pathology foundation models.

To guide the review, the following questions were considered:

- How do the deep learning methods evolve for WSI analysis from CNN based pipeline to MIL, transformers or hybrids?
- What are the key methodological advantages and limitations for each approach in CP?
- What are the remaining challenges of thyroid WSI analysis with respect to data availability, efficiency, robustness and interpretation?

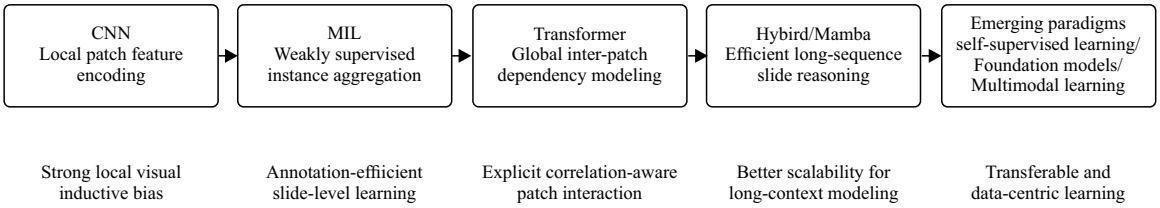

Fig. 1. Methodological evolution of whole-slide image analysis in computational pathology

To avoid repeated explanations in later sections, this review uses a unified conceptual framework for WSI analysis. First, WSIs introduce a scale problem because a single slide may contain a very large number of pixels and patch tokens. Second, CNN-based methods provide effective local morphology representation but are limited in slide-level contextual reasoning when patch predictions are treated independently. Third, MIL reduces the need for dense annotations by learning from slide-level labels and identifying informative instances. Fourth, transformer-based methods strengthen inter-patch dependency modeling but increase computational cost for long WSI token sequences. Finally, hybrid, graph-based and state-space models seek to balance local feature extraction, global contextual reasoning, computational efficiency and interpretability. The subsequent sections therefore discuss each methodological family with reference to this framework rather than repeating these foundational concepts.

Figure 1 summarizes the methodological evolution of WSI analysis from CNN-based patch modeling to MIL, transformer-based dependency modeling, hybrid/Mamba-based efficient reasoning and emerging paradigms such as self-supervised learning, foundation models and multimodal learning.

Information sources: The databases searched were Scopus (as the main source) plus Google Scholar to trace citations back to other articles that may have been missed or could be potentially relevant but are not included within the first search. Scopus was chosen since this is an abstract and citation database containing information on a wide range of disciplines.

In addition to database searching, backward and forward citation checks were performed on highly influential articles, benchmark datasets and recent review papers in computational pathology.

Search strategy: The search strategy aimed at capturing studies on thyroid histopathology, WSIs and the main methodological paradigms used for computation pathology. The search terms are built by combining disease-, domain-specific and method-specific keywords[27,28].

The main concepts included:

- domain terms: “Thyroid”, “histopathology”, “pathology”, “whole-slide image”, “WSI”, “computational pathology”
- method terms: “Deep learning”, “convolutional neural network”, “CNN”, “multiple instance learning”, “MIL”, “transformer”, “vision transformer”, “state space model”, “Mamba”, “hybrid model”, “foundation model”, “self-supervised learning”

A representative Scopus query was structured using a title-abstract-keyword search format, for example:

- TITLE-ABS-KEY ((thyroid OR histopathology OR pathology OR "whole-slide image" OR WSI OR "computational pathology") AND ("deep learning" OR CNN OR "multiple instance learning" OR MIL OR transformer OR "vision transformer" OR Mamba OR "state space model" OR "foundation model" OR "self-supervised learning"))

To improve relevance, targeted searches were also run for specific subtopics such as thyroid tumor classification, MIL-based WSI classification, transformer-based pathology methods, pathology foundation models and public WSI benchmark datasets.

Eligibility criteria: Studies were included if they met one or more of the following criteria:

- Focused on histopathology image analysis (digital pathology), or whole slide image analysis
- Investigated deep learning methods relevant to WSI classification or slide-level representation learning
- Presented one or some of the following types of methods: CNN based methods, MIL, Transformer-based methods, Hybrid architectures, self-supervised learning, foundation models, graph-based reasoning, or multimodal pathology learning
- Were published in English
- Were peer-reviewed journal articles, journal-published benchmark studies, or highly relevant review articles

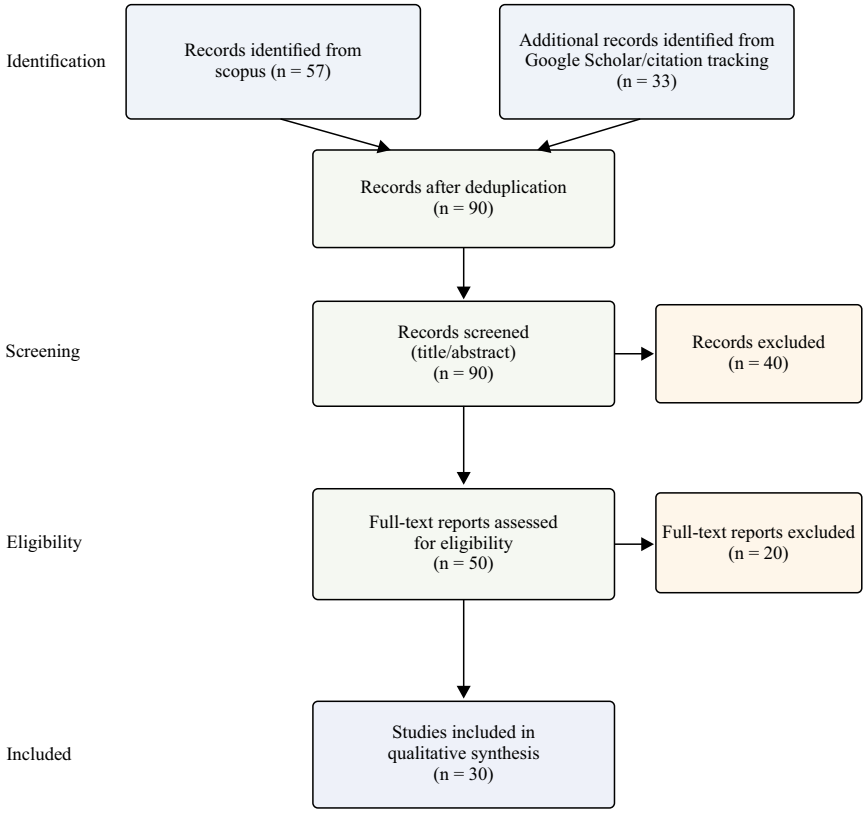

Fig. 2: PRISMA 2020 flow diagram of study selection

Studies were excluded if they:

- Focused exclusively on non-pathology medical imaging tasks without methodological relevance to WSI analysis
- Addressed only patch-level image classification without meaningful connection to slide-level pathology modeling, unless they represented an important early baseline
- Were editorials, short abstracts, slides, non-scholarly commentary, or duplicate records

Lacked sufficient methodological detail for interpretation.

Study selection: All retrieved records were first exported to a spreadsheet for screening and deduplication. Duplicate records were removed manually and, where possible, by matching titles, authors, year and DOI. The remaining records were screened in two stages.

In the first stage, titles and abstracts were reviewed to remove clearly irrelevant studies. In the second stage, full texts of potentially relevant papers were assessed according to the eligibility criteria. Additional influential studies identified through citation chaining were considered. Figure 2 presents the PRISMA 2020 flow diagram summarizing the identification, screening, eligibility assessment and final inclusion of studies in this review.

DATASETS AND BENCHMARKING SCHEMES IN WHOLE-SLICE IMAGE ANALYSIS

High-quality datasets and a robust benchmarking scheme are key elements in developing and evaluating computational pathology AI models. Whole-slice image (WSI) datasets lay the groundwork for training DL models and comparing the performances among various approaches. Given that WSI dataset is usually large-sized and can be derived from multiple sources with different institutes, dataset characteristics (e.g. annotation strategies, staining difference, scan resolution etc.) greatly affect the performance and generalization ability of ML models. Therefore, it is important to understand the common data sets and evaluation scheme that are widely utilized for studying existing advances in this domain.

Publicly available whole-slice image datasets: Several publicly available WSI datasets have been widely adopted as benchmarks for computational pathology studies. Most of these datasets include the set of digitized histopathology

slides accompanied by their diagnostic label or region-level annotation, which is essential to enable repeatable experiments and facilitate algorithmic comparison among each other[29-31].

The CAMELYON series of challenges are one of the most influential data sets on this area. The purpose of the CAMELYON16 data set is to test the automatic detection of lymph node metastases in breast cancer. It contains hundreds of whole-slice images scanned from sentinel lymph node biopsy and fine-grained labels for the presence of metastasis. CAMELYON16 is now a popular benchmark dataset for WSI analysis algorithm evaluation due to its large slice size and complicated tissue structure, especially for the detection/classification problems[29].

The second big data set widely used for computational pathology is PANDA (Prostate Cancer Grading Assessment). The data consists of thousands of prostate biopsy slices across multiple institutes, annotated by pathologists. PANDA is released with the goal of a large international challenge for improving automated prostate cancer grading. The dataset contains various types of samples and a large amount of WSI data which is suitable to train a DL model that has good generalizability[30].

Besides challenging datasets, large-scale biomedical database such as The Cancer Genome Atlas (TCGA) has also become an important source of data in WSI-based study. TCGA includes thousands of pathological slices with various cancer types and associated abundant genomic and clinical data. Because of the size and diversity, TCGA is often used to build ML model correlating histopathologic features with clinical outcome/molecular features[31].

However, many public WSI datasets remain limited by institutional bias, staining variability and the lack of dense annotations, which motivates weakly supervised approaches such as MIL[32].

Thyroid histopathology datasets: Unlike with some other cancers, e.g. breast or prostate cancer, there is not a large amount of public data for thyroid histopathology. However, there is a lack of medical image dataset that can be used in the field of thyroid tumors classification and diagnosis. Some researchers have started filling that void with building up the thyroid tumor classification data set[4-10].

Conventional thyroid histopathology studies often rely upon local datasets with digital microscopy images, or image patches sampled from whole slides, that usually contain classes of benign nodules, papillary thyroid carcinoma and follicular thyroid carcinoma. Although these data are useful as training sets to develop a classifier, they have the disadvantages that they may be too small in size or there is no standard protocol so that results from one study cannot be compared with another.

In recent years, some attempts were made in building a big benchmark data set for thyroid pathology research. For instance, the ThyHisTer dataset, which is focused on ternary classification of thyroid tumors, this dataset contains histopathological images that represent three most common diagnostic classes: Benign nodules, papillary thyroid carcinoma and follicular thyroid carcinoma. With standardized data and evaluation setting, the dataset enables to make a more systematic comparison for various DL methods[33].

Although there are some progresses in this field, WSI of thyroid still have a few issues. Firstly, the number of public sections is small when comparing with dataset from other kinds of cancer. Secondly, the morphological similarity of some thyroid tissues makes it hard to acquire thyroid tissue section information.

The subtypes of thyroid tumors may cause a large variation between different pathologists (inter-observer variation), making it difficult to generate robust annotations; this highlights the importance of having larger and more representative thyroid pathology data sets in the future.

Evaluation metrics and benchmark protocols: In addition to data availability, evaluation protocols and metrics also matter when evaluating the WSI classification models. Since histopathological datasets tend to be imbalanced in classes with complicated diagnostic classes, reliance of only one metric, e.g., accuracy, might not be sufficient to capture the model performance.

Standard metrics applied to evaluate performance in computational pathology are: Accuracy, precision, recall, F1 score and area under the receiver operating characteristic curve (ROC-AUC), where accuracy is defined as the total fraction of correct classification. Although precision and recall tell us something about how good our model was at performing for each class individually. The F1-score is a combination of the two values into one number that might come in handy if we are dealing with unbalanced data sets. The ROC-AUC is a popular measure for evaluating the discrimination capacity of classification models over varying decision thresholds[11,13,14].

In a WSI classification task, we evaluate on the slide level instead of the patch level. In most experiments, we consider each WSIs as one instance and aggregate the predictions of different patches together for the final prediction on slide level. To prevent leaking of information, it is also necessary to make sure that patches coming from a single slide do not appear together in the train and test set[11,13,14].

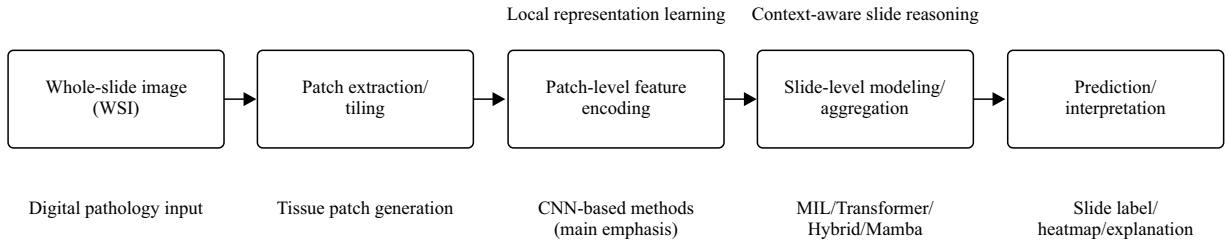

Fig. 3: Generic pipeline of whole-slide image analysis in computational pathology

The second aspect we would like to consider is the cross-institution generalization problem, i.e., models learned by using data only from one institute can perform well for an internal testing set but cannot be generalized to slides that were collected at other hospitals because of differences between their staining procedures and scanning equipment. Accordingly, recent works have paid more attention to the multi-center validation or cross dataset testing for evaluating the robustness of models[32].

In conclusion, we believe that dataset selection as well as evaluation protocol are key factors influencing the validity of experiments performed on computational pathology tasks. Public benchmarks with standardised evaluation criteria allow comparison between various algorithms and identification of future avenues to explore.

CNN-BASED METHODS FOR HISTOPATHOLOGICAL IMAGE ANALYSIS

Convolutional neural networks (CNNs) are one of the first types of deep learning that showed promising results for digital pathology. The good performances achieved by CNNs on natural images rapidly encouraged its use in histopathological image classification, where they showed success in learning hierarchical visual features from the tissue images. For WSI classification, CNN models were the first viable solution to achieve automatic tumoral area identification, tissue types and diagnostically relevant morphology. Even as newer paradigms such as MIL and transformers become more prominent, CNNs continue to be an integral part of many pathology pipelines, especially in the step where patches are converted into features[2,11].

The sustained relevance of CNNs for computational pathology is explainable by their advantages with respect to local representation learning: Histopathological diagnosis often relies upon visual information like nuclear shape, chromatin texture, glandular pattern, stromal features and local architecture of the tissue. Convolutional layers are naturally well-suited on such fine structure as convolution layers make use of local spatial regularities and have good inductive bias when working with images. This is why so many of the subsequent WSI methods, despite being built upon either MIL or a transformer aggregation scheme, continue to use CNN encoders as the first step in feature extraction[4,11-14]. As shown in Fig. 3, whole-slide image analysis typically involves patch extraction, patch-level feature encoding, slide-level aggregation and final prediction, with different methodological paradigms emphasizing different stages of the workflow.

Patch-based CNN pipelines in early WSI analysis: Since whole-slide images are much larger than the capacity of conventional CNNs, many early pathologists' pipelines followed an approach based on patches, i.e., where each WSI was segmented into many small patches with a fixed size and the CNN is run on every patch individually. Predictions will be made either on patch level or following any kind of slide-level pooling[11].

This design made WSI analysis computationally feasible and allowed existing CNN backbones such as VGG, ResNet and DenseNet to be transferred to pathology tasks.

In the case of thyroid pathology, this approach was also observed at first DNN works where CNNs architectures are applied for classifying local patches extracted from thyroid tumour slides, demonstrating that learned visual features may capture useful difference between the diagnostic categories. These studies were important since they proved the viability of AI-assisted thyroid pathology, but they also highlighted a significant shortcoming: slide-level diagnosis cannot simply be reduced to independent patch classification[4].

Slide-level aggregation in CNN-based frameworks: However, as the study of WSI advanced, researchers realized that a patch-level prediction is not enough to address most clinical tasks In pathology, a diagnostically meaningful slide typically consists of a mix of relevant tissue and background as well as normal structure, artifact and morphologically ambiguous regions. A high performing patch classifier is no guarantee for a high performing slide classifier, if and only if the patch predictions could be summarized to reflect the real diagnostic structure of the slide.

Early CNN based WSI systems often tackled this problem by rather straightforward aggregation schemes, for example via majority vote, simple averaging of the patch scores (avg), max pooling or using some heuristics for selecting regions. These methods were too simplistic and did not take into consideration that potentially only a subset of patches may have true diagnostic content in them. In addition, they generally did not model interaction between patches was thus approached more like a set of independent local predictions rather than a cohesive tissue level entity[11,13].

This limitation is particularly pertinent to thyroid pathology. The differentiation among BN, PTC and FTC may rely on an interplay of minor local findings along with wider structural context. An approach which simply takes average over patch-wise scores can potentially fail to capture the correlation between sparse but collectively informative patches. Therefore, the weakness of early CNN-based slide aggregation was more than just an inconvenient technical issue: It indicated that there is indeed some fundamental incompatibility between local feature extraction and slide-level pathological reasoning[4-10].

Transfer learning and the role of CNN encoders: In spite of these shortcomings in earlier patch-based pipelines, CNNs were still very useful because of the success of transfer learning. Pretrained CNN backbones on large scale natural image datasets can be finetuned/adapted for pathology tasks, often showing good empirical performance. Particularly crucial for this in the field of medical imaging were small, labelled data sets compared to other fields of computer vision[2,21,22].

Transfer learning in computational pathology enabled re-use of strong image encoders for the representation of patches without the need of huge amounts of data specific to each task. Thus, CNNs were no longer just end-to-end classifiers but could be used as a generic feature extractor to feed downstream slide level models. This has been an important role till date. In most MIL, transformer and hybrid WSI pipelines, this initial step remains passing image patches through a CNN to get compact embeddings prior to more sophisticated aggregation or sequence modelling being carried out[12,14].

On the other hand, we do not understand CNNs to merely represent a prior generation of WSI models which have now become obsolete. Instead, they still form an important building block of representation learning in digital pathology although the general slide-level network architecture may not be strictly based on pure convolutions anymore.

Limitations of CNN-based WSI analysis: As outlined in review scope and research focus the major limitation of CNN-based WSI analysis is not local feature extraction itself, but the difficulty of converting many independent patch-level representations into a reliable slide-level decision. CNNs are effective for learning nuclear morphology, glandular patterns and local tissue texture, but they do not explicitly model long-range relationships among distant tissue regions. They also often require patch- or region-level annotation, which is expensive and difficult to obtain in routine pathology.

Therefore, the main bottleneck of early CNN-based WSI pipelines lies in slide-level evidence integration. Simple aggregation strategies such as averaging, voting, or max pooling may ignore the sparse and heterogeneous distribution of diagnostic evidence. This limitation motivated the shift toward weakly supervised slide-level modeling methods such as MIL, which can identify informative patches and aggregate them more adaptively under slide-level supervision[11-14].

MULTIPLE INSTANCE LEARNING FOR WHOLE-SLIDE IMAGE ANALYSIS

The big size and the weak annotation characteristic of WSIs bring great challenges to traditional supervised learning method. In most histopathological data sets, only provides slide-level annotations and lacks pixel/region level label because manually annotating those is expensive, which hinders us from applying conventional deep learning model requiring complete label in their training process. To overcome this challenge, multiple instance learning (MIL) has become the most popular approach in weakly supervised WSI classification[13,14].

In the MIL paradigm, we regard every WSIs as a bag containing many image patches (instances) while assigning only one label for an entire bag, where it needs to learn the correlation between instance level feature and slide-level label. For histopathology. This formulation is natural for clinical practice as diagnosis is often performed at the slide level from only a few diagnostically relevant regions. As a result, MIL allows the model to be learned on large scale of WSI data with no need for exhausting annotation[13,14].

Attention-based multiple instance learning: In early MIL methods, the bag-level feature was usually obtained by simply aggregating the instance level features using a max pooling operation (or other similar operations like mean pooling). However, these approaches consider that each case contributes similarly or only the most active case matters, both of which are unlikely to hold true for histopathology images.

An important milestone in MIL field was an appearance of attention based pooling mechanism. Here a model is trained to assign adaptive weights to instances, enables the model to attend over the patches that are diagnostically meaningful and suppress other irrelevant ones; this helps improve both classification accuracy as well as interpretation[12].

A recent example is AttriMIL, which introduces attribute-aware instance scoring together with region-wise and slide-wise constraints to improve WSI classification and disease-positive region localization. Although attention maps can help identify influential tissue regions, they should be interpreted as model-associated evidence rather than as definitive causal explanations[12].

Representative MIL architectures for WSI analysis: Following the introduction of attention-based MIL, several representative architectures have been proposed to further improve feature aggregation and modeling capacity in WSI analysis.

One representative extension is clustering-constrained attention multiple instance learning (CLAM), which introduces instance-level clustering constraints to separate diagnostically informative instances from non-informative ones. This design improves class-specific representation learning and provides an interpretable and widely used baseline for WSI classification[13].

Another recent advancement is E^2-MIL, an explainable and evidential multiple instance learning framework for WSI classification. The method combines attention refinement, spatial-structure modeling and uncertainty-aware instance classification to improve both the interpretability and reliability of slide-level predictions[34].

More recently, transformer-based MIL models such as FR-MIL have been proposed to improve relationships among WSI instances. The FR-MIL combines transformer-based feature modeling with distribution recalibration, enabling the model to account for both inter-instance dependencies and feature-distribution variations across slides[14].

The above-mentioned examples show how the MIL models have evolved from very simplistic pooling approaches to increasingly complex methods that can learn complicated interactions between samples.

Advanced MIL extensions: Although MIL has become a practical weakly supervised framework for WSI classification, its conventional forms still have several limitations. As noted in review scope and research focus many MIL models treat patches as unordered instances and therefore underuse spatial organization, although tissue arrangement is clinically meaningful in histopathology. In addition, large WSI bags may contain hundreds or thousands of instances, which increases computational cost and makes training less efficient.

Another limitation is interpretability. Attention weights can highlight influential patches, but they do not always provide a causally reliable explanation of the prediction. Moreover, conventional attention-based MIL mainly estimates instance importance, while many pathology decisions depend on relationships among multiple regions rather than on isolated high-scoring patches. These limitations have also encouraged stronger representation learning and self-supervised transformer-based extensions for WSI analysis[21,24].

Limitations of conventional MIL methods: However, there are some limitations of existing MIL methods that need to be addressed for efficient large-scale WSI analyses.

To begin with, most of the MIL models consider instance as an unordered set without taking full advantage of spatial relationship among patches. Practically, the spatial relationship between tissue structures usually plays an important role in diagnosing and ignoring such relations would reduce the capability for capturing complicated pathology pattern.

Secondly, the number of instances in a WSI could be very large (e.g., up to thousands of patches for one slide), which poses a computational challenge when dealing with such large bags of instances leading to less efficient training/inference processes.

Thirdly while attention is more interpretable, the attention weights are not necessarily exactly pointing at those diagnostic regions; sometimes, models could attend on uninformative regions and this might lead to less trustworthy predictions for clinicians.

Finally, conventional attention-based MIL mainly estimates instance-wise importance without considering explicit modeling of inter-instance interactions which is crucial in scenarios where a diagnosis decision relies on relationship across several regions from the slide[14].

Transition toward global dependency modeling: The limitations of conventional MIL motivated methods that model explicit relationships among patch embeddings. Transformer-based approaches extend MIL-style slide modeling by allowing patch representations to interact through self-attention, thereby supporting more direct global dependency modeling. However, this benefit also introduces scalability concerns for long WSI token sequences[14,15].

TRANSFORMER-BASED METHODS FOR WHOLE-SLIDE IMAGE ANALYSIS

Transformer based architecture has been becoming a promising research topic on computational pathology and WSI analysis. Compared with traditional convolutional neural network, transformers offer a more flexible way to model the long-range dependency by using self-attention. This property is particularly appealing to histopathology, as diagnostically relevant information on a WSI are frequently dispersed over multiple faraway tissue areas instead of being localized into one patch only[14,15].

But extending the use of transformer for WSI analysis isn't simply an extension of common vision tasks. Unlike natural images, whole-slide images are very huge and they are typically stored in form of hundreds or thousands of patches. Therefore, analysis is not only an image classification problem, but also a long-sequence modeling problem. Therefore, we believe that the key merit of the transformer-based approach in such a setup is not to replace convolution by self-attention but instead modeling relations between patches[14,24].

From patch encoding to slide-level correlation modeling: In most of the WSI systems based on transformers, the raw gigapixel slides are not given to be processed. Rather, the patches are first encoded with CNNs or other feature extraction methods. Then the embeddings are given to the transformer layers. In this setting, the transformer is treated as a sequence model on the level of a slide, which operates on the representations of patches[14,24].

This approach overcomes a major restriction of the multiple instance learning approach in the framework, by being able to model the significance of the patch and also the relation between the patches. In the case of thyroid histopathology, an analysis based on such correlation aware is very useful, as the diagnosis determination is usually based on the knowing by the combined analysis of the local characteristics of the cells and the more global pattern of the tissue architecture[4,6-10,14].

Representative transformer-based approaches in computational pathology: A representative recent approach in this direction is FR-MIL, which combines multiple instance learning with transformer-based feature modeling and distribution recalibration for WSI classification. Rather than considering only instance importance, the method also accounts for feature-distribution variations and relationships among instances, thereby improving slide-level representation learning[14].

Another representative direction is self-supervised transformer pretraining for WSI analysis. MaskHIT uses masked patch reconstruction and contrastive learning to pretrain transformer representations on large WSI collections. The resulting model learns contextual information from patch positions and visual patterns and can be transferred to downstream tasks such as survival prediction, cancer subtype classification and tumor grading[24].

In addition, there are some other works that explore a transformer-based feature encoder trained by either self-supervised or contrastive learning. They do not always employ a transformer for the last slide level classifier but show that transformer-based representations are able to improve downstream WSI classification performance. In pathology, where labeled data are often scarce, such pre-trained encoders offer an essential means of improving generalization[21,24].

More recently, transformer-based WSI research has expanded toward vision-language and multi-modal settings. Although these methods are still emerging, they suggest that transformer architectures may serve as a common backbone for integrating morphology, text and molecular information in future pathology systems[23,25].

Advantages of transformer models in WSI analysis: The Transformer based approach brings some key benefits to analyze WSIs[14,15,24].

Their first is that it provides an even more explicit means of modelling longer range dependencies between patches as in the field of histopathology abnormalities can be found across multiple locations; rather, a diagnosis can rely on how different tissue regions are distributed and related with each other. Self-attention enables models to explicitly learn these types of relations in contrast to standard convolutions/pooling.

Second, transformer is intrinsically friendly to the representation in varying-length sequences. Since we convert a slide into a sequential collection of patch embeddings, the same transformer structure could also adaptively switch to other types of aggregation such as global tokens, position embedding, hierarchical clustering, or combination with MIL[15].

Third, transformers can enhance representational richness for weakly supervised tasks. Going beyond estimating importances, these models can learn more informative bag level embeddings that could potentially benefit from a fine-grained classification task in the presence of subtle context differences[14].

This ability is especially applicable to thyroid histopathology where the differentiation between benign nodules, papillary thyroid carcinoma and follicular thyroid

carcinoma often require more than just local recognition of cell or nucleus shape, but rather knowledge about the larger structure of the tissue. Transformers based architectures are attractive as they offer a way to model such distributed visual clues jointly[4-10].

Limitations of transformer-based WSI models: Transformer-based WSI models improve inter-patch correlation modeling, but several limitations remain. First, self-attention scales quadratically with the number of tokens, making direct modeling of long WSI patch sequences computationally expensive[15,24]. Second, transformers are often data-hungry, whereas pathology datasets are usually smaller and more heterogeneous than natural image datasets[21,22]. Third, pure transformer models lack the local inductive bias of CNNs, which remains useful for fine-grained histological texture and morphology[2,15].

Another WSI-specific issue is token ordering. Unlike language sequences, WSI patch sequences can be constructed by spatial coordinates, raster scanning, attention ranking, clustering, or feature selection. Different ordering strategies may affect how contextual information is modeled, especially in long-sequence or state-space architectures[19,20]. These limitations explain the growing interest in hybrid designs that combine local encoders, weakly supervised selection and efficient global reasoning.

From transformer bottlenecks to efficient long-sequence modeling: The computational bottleneck of transformer-based WSI modeling has encouraged the use of efficient long-sequence architectures. Mamba-based state-space models provide a possible alternative because they can model long contexts more efficiently than full self-attention. Their applications to medical imaging and WSI analysis are discussed in the following section[16,19,20].

HYBRID AND EFFICIENT LONG-SEQUENCE ARCHITECTURES FOR WHOLE-SLIDE IMAGE ANALYSIS

With the evolution of WSI analysis techniques, it became clear that there is an inherent problem at the core of computational pathology. On the one hand, accurate diagnosis requires not only fine-grained local morphology but also higher-level tissue-level context. On the other hand, modeling such information at slide level is computationally expensive as an WSI usually has to be represented with a very long sequence of patch embeddings. This trade-off is the reason for a growing trend in more hybrid and effective long sequence designs, as seen recently.

In previous work, it has been demonstrated that CNNs excel in the extraction of local texture/morphology information, while multiple instance learning offers a realistic weakly-supervised paradigm to aggregate the patch level information and transformer-based models further refine the slide level representations with explicit correlation modeling between different patches. But with the quadratic complexity for self-attention, it is hard to scale up pure transformer solutions if the number of instances grows too much. Thus, recent work have started to explore the new architectures, which could preserve global modelling ability and reduce computation cost at the same time. Here, hybrid models, as well as state-space based sequential models are now very important[11,14,16,17].

Why hybrid architectures are needed in computational pathology? Hybrid architectures are motivated by the complementary strengths and weaknesses summarized in review scope and research focus. CNNs provide strong local morphology representation, MIL supports weakly supervised instance selection, transformers model inter-patch relationships and state-space models offer efficient long-context processing. For computational pathology, combining these components is attractive because diagnostic decisions often depend on both fine-grained cellular details and broader tissue-level organization.

Therefore, the value of hybrid architectures lies not in architectural novelty alone, but in balancing local representation, slide-level reasoning, computational feasibility and clinical interpretability. This balance is particularly important for thyroid histopathology, where subtle local cytological changes and larger tissue patterns may both contribute to diagnosis.

CNN-transformer hybrids for WSI analysis: A common hybrid strategy in WSI analysis is to use a CNN encoder for patch-level feature extraction and a transformer or attention-based module for slide-level aggregation[14,24]. This design avoids applying transformers directly to raw gigapixel slides, preserves the local visual bias of convolutional encoders and introduces global reasoning after compact patch embeddings have been obtained.

In practice, many modern WSI pipelines are already hybrid in this sense: local features are first encoded from patches and higher-level modules then model instance importance or inter-patch dependency. However, CNN-Transformer hybrids do not fully solve the scalability problem, because the number of patch embeddings can still be very large. This motivates more efficient long-sequence alternatives such as Mamba and other state-space models.

State-space models and mamba for long-sequence WSI modeling: Recent Mamba-based state-space models have emerged as efficient alternatives to attention-based sequence modeling. They propagate information through structured state representations and can capture long-range dependencies without explicitly computing pairwise attention between all tokens. Medical-image studies have further shown that Mamba-based and hybrid CNN-Mamba architectures can support long-context representation learning with reduced computational cost[16,17].

This property is particularly relevant to WSI analysis, where a slide is represented by a long collection of patch embeddings. WSI-specific state-space approaches use selective scanning, graph construction, or local-aware processing to integrate spatial structure and long-range dependencies across instances. GraphMamba, for example, combines graph modeling with selective state-space modules to capture both local tissue connectivity and global bag-level context[19,20].

For thyroid histopathology, such models may help combine subtle local cytological features with more distributed tissue-level patterns. However, their benefit must be validated on thyroid-specific and multi-institutional datasets and efficiency gains should be reported together with diagnostic accuracy, robustness and interpretability.

Mamba-based hybrid designs: A promising direction: Mamba-based architectures should not be viewed as exclusive alternatives to CNNs, transformers, or graph models. Instead, recent studies support hybrid designs that combine local feature extraction with efficient long-range modeling. MambaMIM demonstrates the use of hybrid convolutional and Mamba representations in medical imaging, GCNet-Mamba combines convolutional modules with state-space layers and GraphMamba integrates graph-based spatial modeling with selective state-space processing[16,17,19].

Within a WSI pipeline, weakly supervised instance selection can first reduce redundant or non-informative patches, after which graph- or state-space-based modules can model spatial and long-range relationships among the retained instances. Such combinations aim to preserve local morphology and slide-level context while avoiding the full computational cost of global self-attention[19,20].

Current limitations and open questions: There are a few open questions. One of the main problems is that there is no natural ordering for the sequence of the patches taken out of a whole-slide image. All those patches can be ordered with several strategies: By raster scanning, by the spatial coordinates, by the computed attention score, using a clustering technique and so on. For the case of sequential state-space models, the order taken can affect a lot the flow and the aggregation of information in the whole sequence. Therefore, the way to construct those sequences is a very important and central problem for the building of Mamba based model in the context of whole-slide image[19,20].

Second, one cannot guarantee the diagnostic accuracy by the efficiency only. If the important local visual feature or the clinically important spatial pattern is destroyed, the model which is faster cannot be useful in the clinical practice. Third, compared to the large body of work for CNNs, MIL and transformers, the pathology specific validation for the Mamba model is more recent. So, there is a need for more specialized WSI benchmarks and external validation studies. Last, the interpretability of state-space models is still limited and their use in clinical practice will require clearer and more transparent explanatory mechanisms[35].

EMERGING PARADIGMS BEYOND CONVENTIONAL ARCHITECTURE DESIGN

Computational pathology's evolution is historically marked by advancements in its model architectures: From CNNs to MILs. Transformers or lately hybrid long-sequence models. Yet recent studies suggest that there is no need to design even more complicated backbones for additional improvements on the analysis of WSIs. In contrast, there is increasing interest in more general approaches for dealing with data sparsity, representation transfer, structured reasoning and multi-modal fusion. In that sense we are slowly leaving behind pure architectural innovation for more datacentric and pretraining centric approaches on the space of transformers[21-23].

This shift is especially critical for histopathology. Analysis of whole slide images diverges from most traditional computer vision applications in a few ways: Annotations are expensive, clinically curated data are scarce; there is significant variation between institutions; and the diagnostic information of interest can be rare and spread out over extremely large volumes. All these issues have motivated emerging paradigms like self-supervised learning, foundation models, graph based reasoning and multimodal learning[1,2].

Self-supervised learning in computational pathology: With self-supervised learning, models are able to learn meaningful histopathological features from the images themselves, without any need of manual annotations. Instead of using only

the diagnostic label, the techniques of SSL are learnt with some pretext tasks or with a contrastive learning objective. After acquiring such representations, they can be transferred for use in different downstream applications, as the classification, the segmentation, the information retrieval and the survival analysis. In the workflow of the processing of whole slide images, the SSL is often used to pretrain the patch encoders before they are used in some multiple instance learning framework, in some transformer architectures or in other hybrid models for the analysis at the slide level[21,24].

The advantages of SSL are also more apparent in cases of a small number of annotated data. Even powerful methods of aggregation of the information over a slide cannot fully compensate for the drawbacks of the poor quality of the patch level features. Thus, using more properly pretrained encoders can improve the performance of the whole WSI analysis pipeline. In the case of the thyroid pathology, SSL can have the potential to reduce dependence on the small, labeled data set. It can also make the model more robust against different variations of the tissue staining protocols or between different institutions. Nevertheless, these potential advantages need to be carefully and specifically validated in the setup of the thyroid pathology research[21,22].

Foundation models for pathology: Pathology foundation models are trained using large and diverse pathology datasets before being further fine-tuned for an application. The goal of such foundation models is to move the discipline from learning features for a single task to provide more general histopathological representations. For the case of whole-slide image (WSI) analysis, such foundations models could improve the quality of the patch level feature embeddings, enable the performance on a diagnostic task with few or rare samples and provide more robust initial parameters for the model that works at the level of the slide , e.g. multiple instance learning (MIL), transformer architectures, Mamba models, or multimodal frameworks[22,23,25,26].

This is particularly of interest for the thyroid histopathology because there are not many publicly available thyroid datasets and obtaining annotations for such specific tasks is also expensive. A good, developed foundation model can potentially transfer some of the histomorphology knowledge from the bigger more general pathology collections to the thyroid related ones. However, we should not expect the successful knowledge transfer for granted. As the foundation models are usually trained on different and heterogeneous data, we should check how well it works in the cases of detailed thyroid diagnosis, for the patients of other cohort and for the cases that are clinically tricky[22,26].

Structured and graph-based reasoning: In graph-based WSI modeling, image patches or cells are represented as nodes, while spatial adjacency or feature similarity defines the edges between them. This representation preserves tissue organization more explicitly than unordered MIL bags or simple token sequences. Therefore, it seems a more natural way of representing the architecture of the tissue than using an unordered multi-instance learning (MIL) bags or a seq of tokens. With the graph neural network, we can spread information over the connected tissue regions in a way of modeling some important structure of the tissue like local neighborhood, the morphological boundaries, the overall pattern of the tissue organization[18].

Therefore, graph-based techniques represent a different strategy to the MIL, the transformer architectures and the state-space models. They are more useful in a diagnostic case in which the spatial ordering and the relation contexts of the components of the tissue are diagnostically useful rather than on the isolated look of a given patch. One of the difficult point is in the graph construction itself, which is a big design problem. Moreover, models on graph can have more computational demands and more methodological difficulties.

Multimodal and vision-language directions: The WSI analysis is also heading towards the direction of multimodal learning, where histopathology images will be combined with other data sources like clinical metadata, genomic profiles, pathology report or text description. That reflects the real world in which decisions are taken seldom just from looking at images[23,25].

Vision-language approaches are of particular interest since it provides the bridge from an image representation to textual medical knowledge In computational pathology, such models could facilitate cross-modal retrieval, report-assisted representation learning, or even more semantically-aware slide classification; and they might aid in the development of explainability by mapping visual regions to clinically relevant description[23,25].

Multimodality might also be particularly useful for thyroid pathology in the future. Integrating slide morphology and cytology molecular test results, or formalized pathology reports could enhance the ability to make a diagnosis on challenging cases when there is subtlety and ambiguity with respect to the visual pattern[5,6,23].

CHALLENGES AND FUTURE DIRECTIONS

However, despite this rapid advancement there remain some very significant hurdles that impede the further development and clinical implementation of artificial intelligence approaches in WSIs. Some of these hurdles are not simply technicalities; they reflect more fundamental issues within the field of computational pathology related to data quality, model design, efficiency and robustness and explainability. It is important to understand such problems to determine what areas of research are lacking (especially about thyroid histopathology)[1,3].

Data and benchmark limitations in thyroid WSI analysis: The most obvious shortcoming to studies involving thyroid histopathology is that there are few large standardized and publicly available WSI data sets. In contrast to other, better explored fields like the analysis of breast or prostate pathology, the number and variety of thyroid-specific digital pathology resources is still very limited. As a result, it becomes difficult for researchers to make model comparisons and achieve reproducibility[4-10,33].

Furthermore, the task of thyroid diagnosis is also a very difficult classification issue. It is hard to distinguish between benign nodules, papillary thyroid carcinoma and follicular thyroid carcinoma is much harder compared with conducting binary tumor detection, since class appearance variations may be minor and context dependent. Hence, we argue that the design of benchmarks in thyroid pathology needs to focus on both data set size as well as clinically relevant class granularities. Diagnostic confusion and observer variation[4-10,33].

In future studies, we believe that bigger multi-center thyroid WSI data sets with clear evaluation protocol, more balanced class distribution and a greater need of outside-of-the-box verification. If we don't have this reference point then how can we be sure that the progress is due to an improved grasp of the disease process rather than simply fitting to our small data set[6-10,33].

From patch importance to structured slide reasoning: As discussed in earlier sections, WSI analysis has gradually moved from patch-level classification to instance selection, inter-patch dependency modeling and efficient long-sequence reasoning. The remaining challenge is to make slide-level reasoning more structured rather than simply identifying a few high-scoring patches[11-14,16,17,34].

In many pathology tasks, diagnosis depends on how multiple tissue regions relate to each other in structure, distribution and context. Future WSI systems should therefore incorporate explicit spatial modeling, hierarchical context, graph-based representation, or efficient sequence modeling to preserve relationships among tissue regions[14,18,24]. The key question is no longer only which patch is important, but how the selected evidence is organized across the whole slide[17,18,24-26].

Efficiency versus diagnostic fidelity: The second significant problem is a tradeoff between computational efficiency and fidelity of diagnosis. WSIs, by nature, are big and slide level models must handle long patch series in realistic resource limitations, which leads to a constant tug-of-war of more expressive model structures against scalability[14,16,17].

Transformer-based methods enhanced the global dependency modeling but brought a huge computational burden due to quadratic self-attention. Hybrid and state space-based methods were proposed to tackle this issue; however, the bigger problem persists: How do we keep WSI models computationally feasible while not losing clinically relevant detail[14,16,17,19,20].

This question is particularly relevant to deploying in real world pathologies. An approach achieving a little bit better accuracy at the expense of unreasonably large inference time/memory might not be as helpful as an alternative, faster one yielding slightly worse results. Accordingly, future works should provide besides the diagnostics scores also some computational metrics like number of parameters or FLOPs, inference latency, inference throughput and model size. In that sense, we consider efficiency as one of the most important criteria to evaluate instead of considering it as a side note issue[17,19,20].

Generalization and robustness across institutions: A long-standing challenge of computational pathology is the domain variability. Slides from different hospitals can be stained differently, scanner type, tissue preparation, section thickness and annotation style. It is common to observe that models trained on a single dataset performs very well in an internal validation setting while deteriorating significantly for external test data (from another source)[6,26,32].

In the case of thyroid histopathology, such issue might be even more critical as existing public datasets are scarce and potentially do not reflect adequate representation of all the variability present in real life. Thus, reported improvements in terms of one benchmark can be an overestimation of the actual robustness[4-10,33].

To address this problem, there is a need for more systematic evaluation on the cross-center generalization, domain adaptation and stain invariant representation learning

and good pretraining. Self-supervised learning and foundation models can help, but we must test the benefits explicitly in a multi-institution setting instead of assuming it will happen. Finally, strong clinical AI systems should be developed and validated both in terms of their performance in each setting (accuracy) and the robustness of this performance when applied to different settings (reliability)[6,21-23,26,32].

Interpretability and clinical trust: Interpretability is one of the most important barriers for clinical adoption. For pathology, automated predictions will only be trusted if they come along with highlighting of diagnostically relevant areas or an explanation in line with expert knowledge. MIL attention maps, class activation maps, or transformer attention visualization only partially address this problem[12,35].

One key issue here is that attention doesn't mean explanation, high attention weights/highlighted region might not always be the real reason for the model's prediction. For a clinical context, this difference is significant. Plausible, but non-causally trustworthy explanations can give us the illusion of confidence instead of trust[35].

There is need for future work to go beyond qualitative heatmap visualizations, as well as the development of a more rigorous protocol for evaluating interpretability, such as comparisons with pathologist-defined ROIs, repeated run stability analyses, manual grading by experts on the highlighted areas and/or clinician-driven explanation benchmarking. In case of thyroid pathology in particular, the model's interpretation needs to be evaluated with respect to fine-grained diagnostic criteria, not just coarse tumor localization[6-10,35].

CONCLUSION

This systematic review synthesizes recent advancements in artificial intelligence tailored for thyroid histopathology, contextualizing these breakthroughs within the broader paradigm of whole-slide imaging (WSI). The surveyed literature reveals a distinct methodological evolution in computational pathology: transitioning from localized, patch-based convolutional neural network (CNN) pipelines toward weakly supervised multiple instance learning (MIL) frameworks, transformer-driven global correlation modeling and highly efficient hybrid or selective state-space architectures (such as Mamba) designed for comprehensive slide-level reasoning. The primary contribution of this work is the consolidation of these rapid developments into a unified methodological framework that directly bridges thyroid-specific diagnostic clinical challenges with operational machine learning parameters. Despite substantial algorithmic progression, the translation of AI tools into routine clinical workflows remains constrained by a pronounced scarcity of public datasets, inconsistent benchmarking protocols, heavy computational burdens and explainability mechanisms that lack sufficient reliability for clinical trust. To overcome these systematic bottlenecks, future research must prioritize the curation of multi-institutional thyroid WSI datasets, the establishment of standardized evaluation protocols and rigorous external validation paradigms. Furthermore, technical focus should expand toward efficient long-sequence sequence modeling, structured slide reasoning, multimodal clinical data integration and explainable AI (xAI) methods explicitly aligned with pathologists' qualitative diagnostic criteria. Ultimately, meaningful maturation in this field will depend not merely on engineering incrementally stronger neural backbones, but on securing high-fidelity data, implementing rigorous evaluation schemes and fostering seamless integration with real-world clinical pathology workflows.

SIGNIFICANCE STATEMENT

Artificial intelligence has the potential to revolutionize digital pathology, yet its application to thyroid histopathology remains uniquely challenging due to subtle cytological variations and a historical deficit in large-scale public data. This systematic review addresses these barriers by organizing recent deep learning developments into a unified methodological framework, tracing the critical evolution from localized patch classification to advanced, context-aware slide reasoning models (such as Transformers and Mamba architectures). By bridging the gap between clinical diagnostic requirements and raw machine learning workflows, this work highlights systematic bottlenecks in data scarcity, multi-center validation and explainability. Ultimately, it provides an actionable roadmap for developing robust, data-centric and clinically trustworthy AI systems aligned directly with pathologists' workflows.

REFERENCES

1. Madabhushi, A. and G. Lee, 2016. Image analysis and machine learning in digital pathology: Challenges and opportunities. Med. Image Anal., 33: 170-175.
2. Litjens, G., T. Kooi, B.E. Bejnordi, A.A.A. Setio and F. Ciompi *et al.*, 2017. A survey on deep learning in medical image analysis. Med. Image Anal., 42: 60-88.

3. Bera, K., K.A. Schalper, D.L. Rimm, V. Velcheti and A. Madabhushi, 2019. Artificial intelligence in digital pathology-new tools for diagnosis and precision oncology. Nat. Rev. Clin. Oncol., 16: 703-715.
4. Wang, Y., Q. Guan, I. Lao, L. Wang and Y. Wu *et al*., 2019. Using deep convolutional neural networks for multi-classification of thyroid tumor by histopathology: A large-scale pilot study. Ann. Transl. Med., 7: 468-468.
5. Wong, C.M., B.E. Kezlarian and O. Lin, 2023. Current status of machine learning in thyroid cytopathology. J. Pathol. Inf., Vol. 14. 10.1016/j.jpi.2023.100309.
6. Slabaugh, G., L. Beltran, H. Rizvi, P. Deloukas and E. Marouli, 2023. Applications of machine and deep learning to thyroid cytology and histopathology: A review. Front. Oncol., Vol. 13. 10.3389/fonc.2023.958310.
7. Kussaibi, H. and N. Alsafwani, 2023. Trends in AI-powered classification of thyroid neoplasms based on histopathology images-a systematic review. Acta Inf. Med., 31: 280-286.
8. He, T., S. Shi, Y. Liu, L. Zhu and Y. Wei *et al*., 2024. Pathology diagnosis of intraoperative frozen thyroid lesions assisted by deep learning. BMC Cancer, Vol. 24. 10.1186/s12885-024-12849-8.
9. Nojima, S., T. Kadoi, A. Suzuki, C. Kato and S. Ishida *et al*., 2023. Deep learning-based differential diagnosis of follicular thyroid tumors using histopathological images. Mod. Pathol., Vol. 36. 10.1016/j.modpat.2023.100296.
10. Zhu, X., C. Chen, Q. Guo, J. Ma, F. Sun and H. Lu, 2022. Deep learning-based recognition of different thyroid cancer categories using whole frozen-slide images. Front. Bioeng. Biotechnol., Vol. 10. 10.3389/fbioe.2022.857377.
11. Campanella, G., M.G. Hanna, L. Geneslaw, A. Miraflor and V.W.K. Silva *et al*., 2019. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat. Med., 25: 1301-1309.
12. Cai, L., S. Huang, Y. Zhang, J. Lu and Y. Zhang, 2025. AttriMIL: Revisiting attention-based multiple instance learning for whole-slide pathological image classification from a perspective of instance attributes. Med. Image Anal., Vol. 103. 10.1016/j.media.2025.103631.
13. Lu, M.Y., D.F.K. Williamson, T.Y. Chen, R.J. Chen, M. Barbieri and F. Mahmood, 2021. Data-efficient and weakly supervised computational pathology on whole-slide images. Nat. Biomed. Eng., 5: 555-570.
14. Chikontwe, P., M. Kim, J. Jeong, H.J. Sung, H. Go, S.J. Nam and S.H. Park, 2025. FR-MIL: Distribution re-calibration-based multiple instance learning with transformer for whole slide image classification. IEEE Trans. Med. Imaging, 44: 409-421.
15. He, K., C. Gan, Z. Li, I. Rekik and Z. Yin *et al*., 2023. Transformers in medical image analysis. Intell. Med., 3: 59-78.
16. Tang, F., B. Nian, Y. Li, Z. Jiang, J. Yang, W. Liu and S.K. Zhou, 2025. MambaMIM: Pre-training Mamba with state space token interpolation and its application to medical image segmentation. Med. Image Anal., Vol. 103. 10.1016/j.media.2025.103606.
17. Zhou, Y., L. Sun, X. Xiong, G. Ti and S. Yang, 2026. GCNet-Mamba: Leveraging state space models and CNN for medical image classification. Expert Syst. Appl., Vol. 303. 10.1016/j.eswa.2025.130733.
18. Zheng, Y., R.H. Gindra, E.J. Green, E.J. Burks, M. Betke, J.E. Beane and V.B. Kolachalama, 2022. A graph-transformer for whole slide image classification. IEEE Trans. Med. Imaging, 41: 3003-3015.
19. Zheng, T., H. Yao, S. Zhao, K. Jiang and Y. Xiao, 2025. GraphMamba: Whole slide image classification meets graph-driven selective state space model. Pattern Recognit., Vol. 167. 10.1016/j.patcog.2025.111768.
20. Huang, Y., W. Zhao, Y. Fu, L. Zhu and L. Yu, 2025. Unleash the power of state space model for whole slide image with local aware scanning and importance resampling. IEEE Trans. Med. Imaging, 44: 1032-1042.
21. Wang, X., S. Yang, J. Zhang, M. Wang and J. Zhang *et al*., 2022. Transformer-based unsupervised contrastive learning for histopathological image classification. Med. Image Anal., Vol. 81. 10.1016/j.media.2022.102559.
22. Chen, R.J., T. Ding, M.Y. Lu, D.F.K. Williamson and G. Jaume *et al*., 2024. Towards a general-purpose foundation model for computational pathology. Nat. Med., 30: 850-862.
23. Ding, T., S.J. Wagner, A.H. Song, R.J. Chen and M.Y. Lu *et al*., 2025. A multimodal whole-slide foundation model for pathology. Nat. Med., 31: 3749-3761.
24. Jiang, S., L. Hondelink, A.A. Suriawinata and S. Hassanpour, 2024. Masked pre-training of transformers for histology image analysis. J. Pathol. Inf., Vol. 15. 10.1016/j.jpi.2024.100386.
25. Lu, M.Y., B. Chen, D.F.K. Williamson, R.J. Chen and I. Liang *et al*., 2024. A visual-language foundation model for computational pathology. Nat. Med., 30: 863-874.
26. Vorontsov, E., A. Bozkurt, A. Casson, G. Shaikovski and M. Zelechowski *et al*., 2024. A foundation model for clinical-grade computational pathology and rare cancers detection. Nat. Med., 30: 2924-2935.
27. Page, M.J., J.E. McKenzie, P.M. Bossuyt, I. Boutron and T.C. Hoffmann *et al*., 2021. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ, Vol. 372. 10.1136/bmj.n71.
28. Rethlefsen, M.L., S. Kirtley, S. Waffenschmidt, A.P. Ayala and D. Moher *et al*., 2021. PRISMA-S: An extension to the PRISMA statement for peporting literature searches in systematic reviews. Syst. Rev., Vol. 10. 10.1186/s13643-020-01542-z.

29. Bejnordi, B.E., M. Veta, P.J. van Diest, B. van Ginneken and N. Karssemeijer *et al*., 2017. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA, 318: 2199-2210.
30. Bulten, W., K. Kartasalo, P.H.C. Chen, P. Ström, H. Pinckaers *et al*., 2022. Artificial intelligence for diagnosis and Gleason grading of prostate cancer: The PANDA challenge. Nat. Med., 28: 154-163.
31. Weinstein, J.N., E.A. Collisson, G.B. Mills, K.R.M. SShaw and B.A. Ozenberger *et al*., 2013. The cancer genome atlas pan-cancer analysis project. Nat. Genet., 45: 1113-1120.
32. Tellez, D., G. Litjens, P. Bándi, W. Bulten, J.M. Bokhorst, F. Ciompi and J. van der Laak, 2019. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. Med. Image Anal., Vol. 58. 10.1016/j.media.2019.101544.
33. Xie, J., S. He, Y. Fu, X. Tao and S. Zhang *et al*., 2026. ThyHisTer: A new thyroid histopathology image dataset for ternary classification of thyroid cancer. Biomed. Signal Process. Control, Vol. 112. 10.1016/j.bspc.2025.108819.
34. Shi, J., C. Li, T. Gong and H. Fu, 2024. E^2-MIL: An explainable and evidential multiple instance learning framework for whole slide image classification. Med. Image Anal., Vol. 97. 10.1016/j.media.2024.103294.
35. Evans, T., C.O. Retzlaff, C. Geißler, M. Kargl and M. Plass *et al*., 2022. The explainability paradox: Challenges for xAI in digital pathology. Future Gener. Comput. Syst., 133: 281-296.