A Systematic Literature Review on Integrated Deep Learning and Multi-Agent Vision-Language Frameworks for Pathology Image Analysis and Report Generation

Usama Ali[a], Imran Shafi[b], Jamil Ahmad[b], Arlette Zarate Caceres[c,d,e], Thania Candelaria Chio Montero[c,f,g], Hafiz Muhammad Raza ur Rehman*[h], Imran Ashraf*[h]

[a]*College of Electrical and Mechanical Engineering National University of Sciences and Technology (NUST) Islamabad 44000 Pakistan; (usamaalieme40@gmail.com)*
[b]*Department of Computing Abasyn University-Islamabad Campus Islamabad Pakistan; (imran.shafi@abasyn.edu.pk; jamil@ieee.org)*
[c]*Universidad Internacional Iberoamericana Campeche 24560 Mexico; (arlette.zarate@unini.edu.mx; thania.chio@unini.edu.mx)*
[d]*Universidade Internacional do Cuanza Cuito Bie Angola.*
[e]*Universidad de La Romana La Romana Republica Dominicana.*
[f]*Universidad Europea del Atlantico. Isabel Torres 21 Santander 39011 Spain.*
[g]*Universidad Internacional Iberoamericana Arecibo Puerto Rico 00613 USA.*
[h]*Department of Information and Communication Engineering Yeungnam University Gyeongsan-si 38541 Repulic of Korea; (mrazaurrehman@ynu.ac.kr; ashrafimran@live.com)*

Abstract

This systematic literature review (SLR) investigates the integration of deep learning (DL), vision-language models (VLMs), and multi-agent systems in the analysis of pathology images and automated report generation. The rapid advancement of whole-slide imaging (WSI) technologies has posed new challenges in pathology, especially due to the scale and complexity of the data. DL techniques in general and convolutional neural networks (CNNs) and transformers in particular have significantly enhanced image analysis tasks including segmentation, classification, and detection. However, these models often lack generalizability to generate coherent, clinically relevant text, thus necessitating the integration of VLMs and large language models (LLMs). This review examines the effectiveness of VLMs and LLMs in bridging the gap between visual data and clinical text, focusing on their potential for automating the generation of pathology reports. Additionally, multi-agent systems, which leverage specialized artificial intelligence (AI) agents to collaboratively perform diagnostic tasks, are explored for their contributions to improving diagnostic accuracy and scalability. Through a synthesis of recent studies, this review highlights the successes, challenges, and future directions of these AI technologies in pathology diagnostics, offering a comprehensive foundation for the development of integrated, AI-driven diagnostic workflows.

Keywords: Deep learning, vision-language models, pathology image analysis, whole-slide imaging, multi-agent systems, automated report generation, large language models

1. Introduction

Pathology is a cornerstone of medical diagnostics, offering microscopic insight into tissue samples that reveal crucial information about disease presence, progression, and prognosis [1]. Traditionally manual inspection of glass

slides under a microscope represents an expensive, cumbersome, and unreliable diagnostic technique which is both time-consuming and subjective because of interobserver variability. These techniques represent a step change towards whole slide imaging (WSI) technology that enabled entire tissue slides to be digitized at high resolution. Digital slides allow ready sharing of data across institutions and enable telepathology and collaborative diagnostics based on greater flexibility of viewing tissue morphology [2].

The scale of pathology image data has exploded with the adoption of WSI in clinical and research environments, generating millions of gigapixel images annually. As a result of this unbridled growth, however, it has also brought about some critical challenges, namely the volume of images that require interpretation along with the complexity of visual information contained within [3]. Automated and scalable image analysis methods are in high demand to reduce pathologist workload, minimize human error, and reduce diagnostic delays critical in diseases such as cancer where early intervention is crucial [4].

1.1. Role of Deep Learning in Medical Imaging

Deep learning (DL), especially convolutional neural networks (CNNs), has revolutionized the field of medical image analysis. DL enables automatic extraction of hierarchical features directly from image data [5]. CNNs have been shown great success in cell detection, tumor classification, and tissue segmentation problems in pathology, sometimes achieving accuracy on par with or surpassing human experts. These models are very good at recognizing subtle patterns in the localized sense but have traditionally struggled with understanding the full tissue context which is vital to adopting a holistic diagnostic methodology [6].

Recent advances have introduced transformer-based architectures, which offer superior capabilities to model long-range dependencies and global context, addressing some limitations of CNNs. New hybrid models improving the performance of CNNs and transformers try to benefit from the best of both worlds, acquiring fine-scale cellular details while welding large-scale tissue structures [7]. Although these innovations address some of these issues, the whole slide image size remains enormous, staining variation is still a problem between laboratories, and robustness to diverse data sets is critical for moving to real-world clinical deployment [8].

1.2. Emergence of Vision-Language Models and Large Language Models in Medical Report Generation

While DL models have greatly enhanced the visual analysis of pathology images, translating these complex visual patterns into meaningful, clinically coherent text reports poses a distinct challenge. Pathology reports need, not only to describe findings precisely but also to put them in context, using clinical guidelines and terminologies, for clinicians to make informed decisions.

The rise of large language models (LLMs) provides a promising solution to automate this translation process. Similarly, vision-language models (VLMs) fuse image understanding and natural language processing, resulting in descriptive captions, diagnostic summaries, and complete reports consistent with clinical standards [9]. In this area,

there has been some significant progress in radiology, with models producing initial diagnostic reports based on X-rays and CT scans. However, WSIs are special in that pathology exhibits unique complexities of higher resolution and higher intricacy and that WSIs must span both fine and coarse image features [10]. These complexities are only just beginning to be taken up in current research, but are not answered by unified systems that effectively and reliably couple image analysis and text generation for pathology.

1.3. Need for Integrated Multi-Agent Systems

Conventional artificial intelligence (AI) systems are commonly applied in single tasks within medical imaging like classification or generation of reports. Nevertheless, they are unable to produce whole-potential diagnostic processes essential to clinical decision-making process because of the operations like normalization and thresholding that occur in the siloed approach [11]. An interesting solution is the multi-agent method that involves the use of several domain specific AI agents to work together to perform multi-faceted correlated tasks.

Multi-agent systems allow agents of specific expertise to be specified, one that produces detailed image features, one that produces preliminary textual descriptions and one that checks and corrects outputs according to medical ontologies and clinical guidelines. Part of this team work is iteration: the agents communicate to increase the caliber and clinical validity of the end-result pathology report. Distributed character of MAS makes MAS prone to scalability and flexibility which allows the addition of new agents to address new emerging challenges or new specific clinical requirements. This kind of integration can revolutionize the pathology operations by automating complete end-to-end workflows, removing human errors, and generating annotated data sets that can be used to accomplish research studies and development applications [12].

1.4. Objectives and Scope of the Systematic Literature Review

The purpose of this review is to critically examine the convergence between DL and multi-agent systems and large language models on pathology image analysis and report generation. The architectures of the models, datasets and training processes associated with the automated interpretation and reporting of pathology images will be reviewed and clinical uses will be discussed in the wider scope of literature. The special attention will be drawn to the combined multi-agent systems that are based on the vision language models; they are the state-of-the-art in the field.

Through a systematic evaluation of the existing literature, this review aims to determine effective methodologies, evident gaps, and crucial gaps in research. It will also look into the ways in which these technologies are being translated to clinical practice regarding scalability, interoperability as well as ethics issues. An aggregate outcome of the review will be taken as a holistic foundation to the presented hybrid multiagent design to continue developing it as a useful practical application in pathology diagnosis.

It is critical to explicitly define the scope of this review, which centers on Pathology WSI. Pathology WSI tasks involve the analysis of gigapixel, microscopic cellular images that require multi-scale zooming and navigate extreme morphological complexity. This is diffferent from radiology work (e.g. CT, MRI, or Chest X-rays), which typically

concerns macroscopic, anatomic structures at far lower resolutions. In general, though, this review contains a selective collection of more recent radiology based studies, especially those on Vision-Language Models (VLMs) and Natural Language Generation (NLG). These non-pathology structures are essential architecture blueprints since radiology AI arose earlier in the field of automated report construction and multi-agent coordination. To achieve the aims of this review, there is a need to analyze the pioneer systems of radiology because their approach to cross-modal alignment and agentic orchestration currently is being translated to address the unique, high-resolution problems of digital pathology.

1.5. Research Questions

The systematic literature review will be guided by the following key research questions:

- What are the current state-of-the-art DL approaches used in pathology image analysis, and how do they address challenges such as high-resolution data and dataset variability?
- How have vision-language models and large language models been employed to generate clinically relevant text from medical images, particularly in pathology?
- What roles do multi-agent systems play in integrating image analysis with text generation in medical diagnostics, and what are the architectures and strategies that facilitate effective inter-agent collaboration?

2. Related Work

The integration of AI in healthcare, particularly in medical imaging, has become one of the most transformative technological advances in modern medicine [13]. A potential application of AI systems, particularly DL algorithm-based, reveals an ability to greatly improve diagnostic accuracy, automate workflows, and improve clinical decision-making [14]. Their applications in medical imaging have produced some very encouraging results in improving the early detection of disease and streamlining treatment planning across a range of imaging modalities including those in radiology, pathology, and dermatology [15, 16, 17]. This integration is, however, a significant challenge, as it is inherently a multi-faceted task involving multiple types of data in the healthcare space, variability of imaging data, and population diversity [9].

Medical imaging or the interpretation of images of human body is a very significant aspect of diagnostic medicine. Images provided by assorted medical resources, including endoscopy, tissue pathology sections, and scans, including computed tomography (CT), x-rays, positron emission tomography (PET), magnetic resonance imaging (MRI), etc., can be analyzed and used to identify many diseases [18]. Last but not least, DL models, particularly CNNs and vision transformers have formed the foundation of such change that enables segmentation, classification, and detection of abnormalities on a more precise level, becoming possible. Nonetheless, even though they are now skilled pictorial

analysts, more needs to be done in linking these DL models with additional ancillary data sets accessible in the field (electronic health records (EHR)) of clinical information [19].

Regardless of the improvements in AI, the field is still characterized by a number of weaknesses, such as generalization problems, model explainability, and interpretation of findings in a clinical set-up [20]. Nonetheless, the constraints are enhanced when it comes to multi-center, multi-device use of AI models, where imaging devices, data collection mechanisms, and patient demographics may influence model performance as well all at once [21].

2.1. Deep Learning and its Applications in Medical Imaging

DL has quickly become a key tool in medical imaging, especially for tasks like diagnosing diseases, segmenting organs, and detecting abnormalities [21]. One of the most notable benefits of the DL models, including CNNs, is the automatic phrasing of hierarchical features of medical images, which is historically a highly tedious and manual procedure performed by operational professionals in the field [14]. This capability has led to dramatic advances in the accuracy of diagnostics, even surpassing the readout capability of human radiologists in some directions, sometimes, as noted, by a wide margin in some areas[19]. Nonetheless, with CNN-based models, they have been established to attain a high diagnostic accuracy of diabetic retinopathy (AUC = 0.933-1), lung cancer (AUC = 0.864-0.937) and breast cancer (AUC = 0.868-0.909) [14]. To accomplish this, they are trained on massive large quantities of data and taught to identify these very low-level patterns in medical images, which human experts could not do with repeated trials. Although the achievements here are great, the DL in medical imaging with the aim of improving the care received by the patients does have its challenges. Nevertheless, the fact that these models are affected by changes in the dataset and fluctuate when there is a change in imaging protocol, type of the scanner and patient demographics is one of the biggest limitations of these models [18]. Such inconsistency among clinical settings causes the deployment of such models to reduce model performance. An example illustrating this can include differences between the approaches that the hospitals employ in amassing data, which can restrict the use of the model to fresh settings [19]. Data augmentation, synthetic data generation, and domain adaptation have been considered to address these issues through enhancing the strength and applicability of the DL models in medical imaging practice [22].

Another important aspect of DL in medical imaging is the use of data augmentation techniques. Data augmentation is the process of making additional training data by transforming the original dataset with applications such as rotations, scaling, and flipping [22]. In medical imaging, as is often the case with annotation, we lack or have limited annotated data that is costly to obtain which makes this technique valuable [22]. Data augmentation artificially enlarges the dataset to help improve the model's performance in a scenario where the number of labels is limited. In addition, medical imaging applications have attracted interest in the subfield of machine learning known as few-shot learning, in which models are trained with very little labeled data as training data [23]. To tackle an insufficient amount of annotated medical images, we introduce a few-shot learning methods that enable models to generalize from a few examples. Studies have demonstrated that these techniques can improve the performance of DL models on different medical imaging tasks, e.g., tumor detection and organ segmentation [21]. Although they may be promising,

applying data augmentation and few-shot learning remains challenging in clinical settings. Therefore, to avoid poor accuracy caused by noisy augmented images, the quality of augmented data is also important. Second, given that few-shot learning presents a promising approach, it in general requires specialized algorithms and rigorous validation to be reliable in practical medical applications [23].

In multi-center or multi-device studies, image harmonization is critical to ensure that models trained on data from different sources can generalize effectively across varied clinical environments [23]. Inconsistencies in such model performance can arise due to variations in imaging equipment, patient demographics, and acquisition protocols [21]. To standardize medical images, often image harmonization techniques (such as grayscale normalization and image resampling) are employed to make the images more consistent and comparable within and across a variety of image centers and devices [19]. Image harmonization has been shown recently to be highly effective in increasing the diagnostic performance of DL models. In one example, in multi-center breast cancer studies of breast cancer detection, image harmonization improves the classification accuracy by up to 24.42% [21]. Furthermore, authors demonstrate that color normalization techniques can slightly improve AUC scores on external test sets and may serve as a way to improve robustness to multi-center data used in training [14].

2.2. Vision-Language Models in Medical Imaging

VLMs have emerged as a powerful tool in medical image analysis by integrating both visual and textual information. Bringing the strength of CV and the strength of natural language processing (NLP), VLMs allow AI systems to process medical images with accompanying textual information [24]. Medical report generation is one of the main usage areas of VLMs in healthcare, as the model generates textual descriptions of ratings after analyzing medical images. The value of this task is especially high when applied to radiology which suffers from an extremely large amount of imaging data that creates a barrier to manual detailed report writing for every case. VLMs have recently been shown to generate text that is coherent and relevant to the medical domain when used for automating medical report generation [25]. For example, VLM has been used to automatically generate reports from chest X-rays, and CT scans, where the reports are given to radiologists to summarize findings in aiding the diagnosis and treatment planning. They have been shown to increase efficiency and help alleviate the work of healthcare professionals. Furthermore, VLMs have been deployed as solutions to visual question answering (VQA) problems, explicitly, answering clinical questions given medical images, thereby improving clinical decision-making [9].

While VLMs excel at processing visual data, integrating medical imaging with other forms of clinical data, such as EHRs, is essential for improving diagnostic accuracy and treatment planning [19]. Patient information that is vital to the compositional context for the interpretation of medical images, including medical history, lab results, and medication details, are all contained in EHRs. By integrating this data with image analysis models, newly generated AI models can make diagnoses and treatment recommendations that equally consider imaging data as well as the patient's clinical background [26]. New advancements in DL-based data fusion techniques have also been growing to integrate medical imaging into EHR with a more accurate prediction and decision [27]. Such models, however,

encounter the problem of supporting the complexity and heterogeneity of EHR data, particularly when also the data originates in more than one place. The success of multimodal AI implementation in the healthcare environment can only be achieved when data formats are standardized and that data fusion methods are developed muscularly [21].

Cross-modal integration or the ability to query visual and text data of other sources is a severe difficulty with healthcare AI. Considering that medical imaging would obtain EHR data, advanced methods should be created to handle the various types and forms of data [28]. One significant challenge will be associated with the ability to make sure that this AI model will be able to consolidate these multifaceted sources of data into suggesting generic accessibility predictions without drowning them in noise or inaccuracies [9]. In addition, in the multimodal AI systems, patient privacy and regulatory standards require the privacy of patient data. Nevertheless, the possible benefit of cross-modal integration in healthcare is enormous. This means that with the combination of medical imaging and EHR data, more precise and tailored treatment suggestions of AI systems are enabled that eventually results in a reduction in healthcare costs and patient outcomes improvement in the end results into better patient outcomes in general [29]. Subsequent research in cross-modal AI applications in healthcare must focus on advancing the rigor and generalizability of such systems and make sure that they can be generalized to different clinical environments [30].

2.3. Multi-Agent Systems in Healthcare

Multi-agent systems are a promising approach to enhancing clinical decision-making by leveraging the expertise of multiple autonomous agents. Multi-agent systems are one of the promising solutions to the improvement of clinical decision-making when the skills of several agencies are used. These are systems wherein multiple AI models, each with a specific area of specializing in solving one of the healthcare issues, can interact to cooperate, communicate, and even interact to assist in solving the complex healthcare issues[31]. As an illustration, in a clinical decision support system, the first agent is the medical image assessor, the second agent is the patient history data assessor and the third agent produces a report upon the given analysis. Such interplay of these agents can be used to increase the accuracy of diagnosis and positively affect treatment outcomes due to leveraged optimization of decision-making. Compared to traditional clinical workflows, multi-agent systems have been found to significantly increase a workflow in a diverse, complex setting with a high degree of simultaneity in completed tasks [32]. For example, Multi-agent systems have previously been used to enhance diagnostic accuracy in pathology and radiology via the coordination of agents for image analysis, report generation, and clinical validation. In addition, these systems can optimize hospital operations by automating office tasks such as patient scheduling and resource allocation [33].

In medical imaging, segmentation is one of the most critical tasks, as it involves identifying and delineating regions of interest within an image. Automatic segmentation of anatomical structures in medical images has been automated via multi-agent systems where different agents ensemble to segment different anatomical structures [34]. For instance, one of the agents might deal with the segmentation of the brain in MRI, the other with the segmentation of tumors in CT, and so on. Through the distribution of workload among multiple agents, these systems show an ability to improve the efficiency and accuracy relative to the task which can be tedious and vulnerable to human error in some

cases. Medical image segmentation has been shown to be able to be solved better using multiagent systems instead of single-agent approaches. The diversity of expertise across agents provides these systems a robust and accurate ability to segment [33].

Recent quantitative studies indicate that collaborative multi-agent frameworks can outperform standard single-agent baselines in complex environments. For example, when using multi-agent reinforcement learning (MARL) to repeatedly optimize 3D medical image segmentation, the improvement in the Dice Similarity Coefficient (DSC), averaging a 3 percent increase in each instance (e.g., 85.56 to 88.53 on brain tumor data) was obtained over traditional single-agent convolutional networks [35]. Equally, cooperative multi-agent deep reinforcement learning models on COVID-19 CT image segmentation had high precision of 97.12 percent and high Dice scores (80.81 percent) indicating that representation of the segmentation task among specialized agents is effective in mitigating the over-segmentation and ambiguous boundaries addressed well, as compared to monolithic models [36]. Although interoperability and data consistency are still an issue in work, the heterogeneity of expertise targeted among interacting agents can offer these systems a greatly strong solution to medical image segmentation[32].

The future of multi-agent systems in healthcare is bright, and it can be applied in many spheres, such as diagnosis, treatment planning, and hospital management [37]. However, to make these systems applicable in the wide-ranging clinical practice, some obstacles should be surmounted. The capacity of multi-agent systems to operate in real time and especially speedy environments is a major challenge, particularly in clinical settings that are very time-related in nature [38]. In addition, the issues of data privacy, agent accountability, and system transparency must also be resolved in order that multi agent systems could become something safe and ethical to manage in healthcare. However, multi agent systems hold tremendous possibilities in improving the delivery of healthcare[39]. As AI technologies continue getting better, multi-agent systems are bound to play an instrumental role in automating and optimizing the clinical workflow that will lead to a higher level of efficiency and individuality of the care [33].

2.4. Limitations of Systematic Literature Reviews

The systematic literature reviews (SLRs) we have reviewed, focusing on AI and medical imaging, reveal several limitations specific to the current body of research. Those constraints are closely intertwined along with the character of the emerging AI technologies in healthcare and the existing challenges in integrating multi-modal data, model architecture, and clinical realities. Failing to ensure the standardization of the study methods is by all means one of the largest shortcomings found in the reviewed SLRs. Considering the case of DL applications in medical image analysis, there are differences in the types of neural networks taken into account, image preprocessing methods, and measures of evaluation that makes the direct comparison of the results practically impossible. The definition of success used in various studies (e.g., diagnostic accuracy, sensitivity, specificity) and the various modalities of imaging investigated (e.g., CT scan, MRI, and X-ray) pose varying obstacles that must be met in various forms, and the comparison between various studies is therefore more challenging to make [14]. Definitive conclusions could not be made due to the failure to identify an explicit outline of standardization of the methodologies of studying that have

been incorporated in the studies. As an example, during the medical image segmentation, though they execute the same procedure, one AI model is not tested according to identical protocol alongside another model, thus the results cannot be directly compared to each other directly [33].

Another limitation is the underrepresentation of real-world clinical data in the studies included in the reviews. In the majority of cases, the SLRs are focused on results of the experiments conducted in controlled conditions on datasets (e.g. TCGA or MIMIC-CXR) which can be extremely curated, though it is very uncommon that they are rich in variability encountered in the reality of a clinical setting of the experiment environment itself [20]. This constraint means that AI models may be highly performative in such data set; however, the real applicability of their words to the reality is not determinable. Currently, the research on the multi-center or multi-device modeling is limited, and thus limits the applicability of AI models to any clinical environment with different imaging devices, protocols, and patient demographics in general [40]. One of the most visible instances of such is the review of image harmonization methods where the challenge of standardization of imaging data across centers and devices highly becomes simplified, and studies reviewed do not have a commentary on how models might address real-world issues of multiple centers/clinics [21].

Finally, the deficit of transparency and ethical consideration is another reoccurring problem throughout the analyzed SLRs. Nevertheless, the information regarding what we do with the bias of the data (demographic imbalance and clinical disparity, e.g.) is lacking in many studies [18]. Also, when considering technical performance of AI models, SLRs are likely to leave ethical concerns that accompany the use of an AI model in clinical practice. The reviews casually accepted patient privacy, model accountability, and explainable AI decisions, which are pivotal to clinical adoption of AI tools (not always covered canonically in this group of papers) [33]. This lapse in the coverage of the ethical and social consequences of applying AI in medical imaging can delay the realization of the findings of research into clinical practice.

3. Methodology

3.1. Search Strategy

A comprehensive and systematic search strategy is fundamental for ensuring the inclusion of relevant and high-quality literature. To broaden and widen the scope of the research that has been reviewed around pathology image analysis, DL, multi-agent systems, and large language models, multiple reputable academic databases were chosen for this review [41]. To cover the medical imaging and artificial intelligence area, the primary databases selected for use are PubMed, IEEE Xplore, Scopus, and arXiv which include peer-reviewed journal articles, conference proceedings, and preprints.

Although the main search strategy was confined to digital pathology and WSI, additional medical imaging terms (e.g., ”medical report” and ”multimodal LLM” was important) were not strictly filtered to pathology in some query combinations. This was needed to be able to acquire very transferable underlying models and multi-agent systems

in related areas such as radiology, which offer very important methodological background to the emergent pathology processes.

The search queries were carefully constructed to balance comprehensiveness and specificity. Key search terms and phrases were combined using Boolean operators (AND, OR) to capture literature spanning the intersection of fields. Representative keywords included:

- "Pathology image analysis"
- "Whole-slide imaging"
- "Deep learning"
- "Convolutional neural networks"
- "Transformer models"
- "Vision-language models"
- "Large language models"
- "Multi-agent systems"
- "Medical report generation"
- "Image-text generation"

Searches were conducted iteratively, refining terms based on preliminary results to ensure relevant papers were not overlooked. Figure 1 shows the complete process of article selection.

3.2. Database Search Strategy and Search Strings (April–May 2025)

To ensure reproducibility, we executed database-specific search strings across four sources: PubMed, Scopus, IEEE Xplore, and arXiv. Searches were conducted between April and May 2025. The strategy targeted three intersecting concepts: (i) digital pathology / whole-slide imaging, (ii) deep learning for pathology image analysis, and (iii) vision–language / large language models and multi-agent frameworks for image-to-text generation and decision support. Where supported, we searched titles/abstracts/keywords, applied an English-language filter, and restricted publication years to 2015–2025.

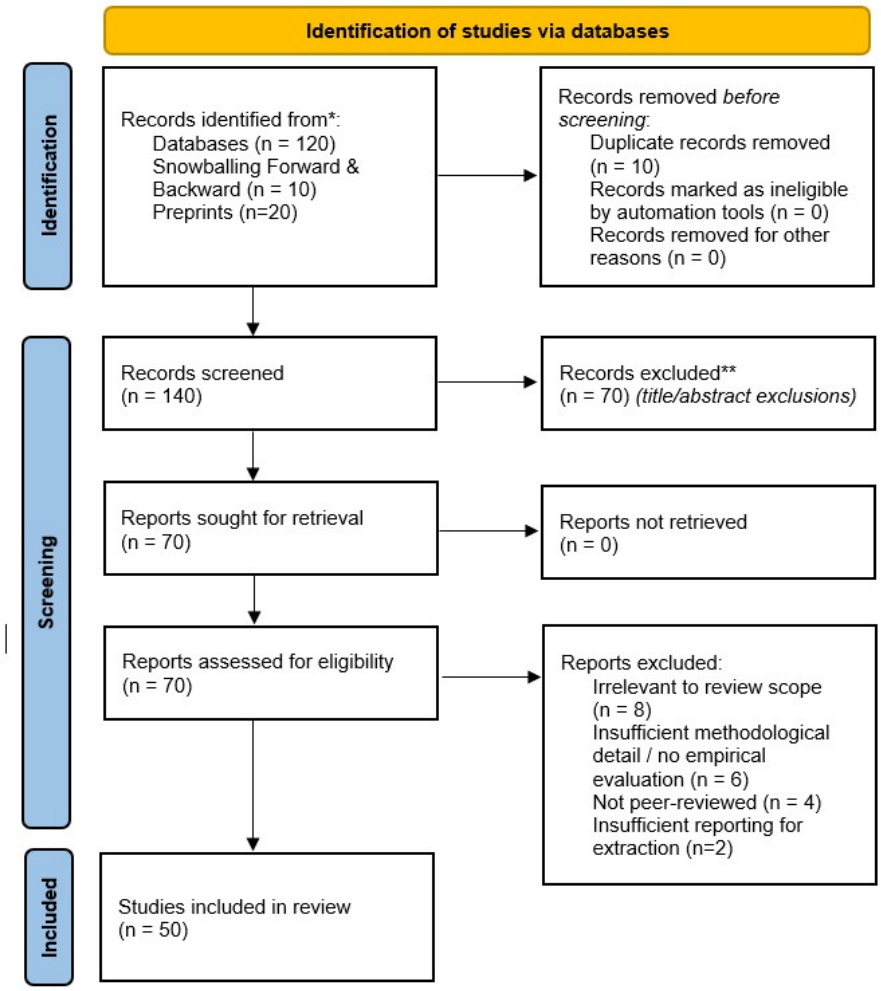

Figure 1: PRISMA study selection process flow diagram.

3.2.1. PubMed (April–May 2025)

PubMed was queried using Title/Abstract fields (and MeSH where relevant) to capture biomedical terminology. The following query was executed with filters: English; publication dates 2015/01/01–2025/12/31.

PubMed query (Title/Abstract) (n = 34):

```
(("digital pathology"[Title/Abstract] OR "whole slide imaging"[Title/Abstract] OR WSI[Title/
    Abstract]
 OR histopatholog*[Title/Abstract] OR "computational pathology"[Title/Abstract])
AND
("deep learning"[Title/Abstract] OR "convolutional neural network"[Title/Abstract] OR CNN[Title/
    Abstract]
 OR transformer[Title/Abstract] OR "vision transformer"[Title/Abstract] OR ViT[Title/Abstract]
```

```
  OR "multiple instance learning"[Title/Abstract] OR MIL[Title/Abstract] OR "foundation model"[
      Title/Abstract])
 AND
 ("vision-language"[Title/Abstract] OR "vision language"[Title/Abstract] OR multimodal[Title/
     Abstract]
  OR "image-text"[Title/Abstract] OR "image to text"[Title/Abstract] OR caption*[Title/Abstract]
  OR "report generation"[Title/Abstract] OR "large language model"[Title/Abstract] OR LLM[Title/
      Abstract]
  OR "medical report"[Title/Abstract] OR "pathology report"[Title/Abstract]
  OR "multi-agent"[Title/Abstract] OR "multi agent"[Title/Abstract] OR agentic[Title/Abstract]))
```

3.2.2. Scopus (April–May 2025)

Scopus was searched using TITLE-ABS-KEY to improve coverage of computer science venues and interdisciplinary publications. Filters were applied for: English; years 2015–2025; document types (article, conference paper, review excluded at screening stage).

Scopus query (TITLE-ABS-KEY)(n = 38):

```
TITLE-ABS-KEY(
  ("digital pathology" OR "whole slide imaging" OR WSI OR histopatholog* OR "computational
      pathology")
  AND
  ("deep learning" OR "convolutional neural network*" OR CNN OR transformer* OR "vision transformer
      " OR ViT
   OR "multiple instance learning" OR MIL OR "self-supervised" OR "foundation model*")
  AND
  ("vision-language" OR "vision language" OR multimodal OR "image-text" OR "image to text"
   OR caption* OR "report generation" OR "medical report*" OR "pathology report*"
   OR "large language model*" OR LLM OR "multimodal large language model*" OR MLLM
   OR "multi-agent" OR "multi agent" OR agentic OR "collaborative agent*")
)
AND (LIMIT-TO(LANGUAGE, "English"))
AND (PUBYEAR > 2014 AND PUBYEAR < 2026)
```

3.2.3. IEEE Xplore (April–May 2025)

IEEE Xplore was searched to capture engineering and computing studies (e.g., MIL, transformers, multimodal learning, agent-based frameworks) often not indexed uniformly in biomedical databases. Searches were applied to Metadata (Title, Abstract, Index Terms) with filters: English; years 2015–2025.

IEEE Xplore query (Metadata) (n = 26):

```
(("digital pathology" OR "whole slide imaging" OR WSI OR histopathology OR "computational pathology
    ")
 AND
 ("deep learning" OR CNN OR "convolutional neural network" OR transformer OR "vision transformer"
     OR ViT
  OR "multiple instance learning" OR MIL OR "foundation model" OR "self-supervised")
 AND
 ("vision-language" OR "image-text" OR multimodal OR "report generation" OR "medical report"
  OR "large language model" OR LLM OR "multimodal LLM" OR MLLM
  OR "multi-agent" OR agentic))
```

3.2.4. arXiv (April–May 2025)

arXiv was searched to capture rapidly emerging work in multimodal foundation models and agentic systems relevant to digital pathology. Searches were executed over title and abstract terms. We limited results to 2015–2025 and English-language manuscripts.

arXiv query (title/abstract keywords) (n = 22):

```
("digital pathology" OR "whole slide imaging" OR WSI OR histopathology OR "computational pathology
    ")
AND
("deep learning" OR CNN OR transformer OR "vision transformer" OR ViT OR "multiple instance
    learning" OR MIL
 OR "foundation model" OR "self-supervised")
AND
("vision-language" OR "image-text" OR multimodal OR "report generation" OR captioning
 OR "large language model" OR LLM OR "multimodal LLM" OR MLLM
 OR "multi-agent" OR agentic)
```

3.2.5. Search Refinement and Auditability

Searches were iteratively refined to reduce false positives while maintaining recall, primarily by (i) requiring explicit pathology/WSI terms, (ii) retaining both classical DL (CNN/MIL) and transformer/foundation-model terms, and (iii) including vision–language, LLM, and multi-agent keywords to capture integrated image-to-text and agentic frameworks. The final strings above are reported verbatim for auditability and reproducibility.

3.3. Inclusion and Exclusion Criteria

To maintain focus and relevance, specific inclusion and exclusion criteria were defined prior to the selection of studies:

3.3.1. Inclusion Criteria

- Studies published in English between 2015 and 2025 to capture recent advances.
- Peer-reviewed journal articles, conference papers, and preprints with significant technical contributions.
- Research focused on pathology or closely related medical imaging fields (e.g., radiology) involving DL or multi-agent frameworks.
- Studies addressing image analysis, report generation, or the integration of visual and textual data.
- Works discussing datasets, evaluation metrics, or clinical applications relevant to pathology image-text tasks.
- Preprints were included to capture the latest research in AI for pathology, addressing uncertainty through a quality assessment that considered peer review status and methodological rigor, prioritizing later peer-reviewed studies.
- Non-pathology studies were included only if their methodologies, such as AI models or multi-agent systems, were directly applicable to pathology image analysis, ensuring relevance to the scope of the review.

3.3.2. Exclusion Criteria

- Papers unrelated to medical imaging or pathology.
- Studies focusing solely on traditional image processing without AI or machine learning methods.
- Articles lacking sufficient methodological detail or empirical evaluation.
- Review papers, opinion pieces, or editorials (though references within these were screened for additional sources).

3.4. Study Selection Process

The selection process followed a rigorous screening protocol to ensure only the most pertinent studies were included. Initially, all search results were imported into a reference management tool, where duplicates were removed. Two rounds of screening were conducted:

The selection process followed a rigorous screening protocol to ensure only the most pertinent studies were included. The steps of the selection process were as follows:

- Title and Abstract Screening: Each paper was assessed for relevance based on its title and abstract. Studies that were clearly outside the scope or did not meet the inclusion criteria were excluded.
- Full-Text Review: The remaining papers underwent a detailed full-text review to confirm their alignment with the research objectives and methodological standards.

- Reviewer Logistics: The screening process involved two independent reviewers who assessed the studies based on predefined inclusion/exclusion criteria.

To ensure the robustness of the full-text screening phase, inter-rater reliability was quantitatively assessed. The agreement between the two independent reviewers was substantial, yielding a Cohen's kappa of 0.72 (87.1% observed agreement). The 9 remaining conflicts were resolved by a third reviewer through discussion and consensus, resulting in the final corpus of 50 studies.

3.5. Data Extraction and Synthesis

For each included study, a structured data extraction form was developed to capture essential information systematically. Key data points included:

- Publication details: authors, year, venue
- Study objectives and scope
- Technical methodologies: model architectures (CNN, Transformer, VLM, LLM), multi-agent system design, training strategies
- Datasets used: type, size, annotation details
- Evaluation metrics and results: accuracy, F1-score, BLEU score for text generation, clinical validation if available
- Challenges and limitations noted by authors
- Clinical applicability and integration aspects

The extracted data were synthesized qualitatively to identify trends, common approaches, and gaps. Quantitative meta-analysis was considered but limited due to heterogeneity in tasks, datasets, and evaluation protocols.

3.6. Quality Assessment of Included Studies

Quality assessment was conducted to evaluate the robustness and reliability of each study. Criteria included:

- Methodological rigor: clear description of algorithms, data preprocessing, and experimental setup,
- Reproducibility: availability of code, datasets, or detailed protocols,
- Evaluation thoroughness: use of appropriate metrics, baseline comparisons, and validation on multiple datasets,
- Clinical relevance: alignment with real-world pathology challenges, inclusion of domain knowledge, or clinical expert validation,

- Transparency about limitations and biases.

Studies scoring below a predetermined threshold (0-3) on these quality assessment criteria were carefully evaluated before making a final decision on their inclusion. Any study that had serious methodological limitations, including ambiguous experimental design, inadequate validation, improper description of the data, or no clinical relevance, was usually dismissed during the main synthesis to maintain a high level of rigor of the review. Nonetheless, when the research presented a new interpretation or inquiry of a new subject, even in a minor manner, it was put on the table of discussion with reservations. In these circumstances, though, the constraints and potential biases were clearly mentioned and the results were continuously interpreted in such a way that there is no overgeneralization of the findings. This moderate stance justifies the scientific soundness of the review and also formulates and brings to actual realization the thought experiment of what a comprehensive review may appear like, having a moderate perspective of potential directions toward success and simultaneous compensatory and joint failures of the extant literature.

To operationalize this quality assessment, each of the five criteria (Clarity, Reproducibility, Evaluation Rigor, Clinical Relevance, and Transparency) was evaluated using a 4-point ordinal scale: 0 (Poor), 1 (Moderate), 2 (Good), and 3 (Excellent). The overall Unweighted average of all these five dimensions resulted in the final quality score of every study. According to this rubric, the final set of 50 studies that were used had a very high overall methodological standard. The number of the studies that were categorized as "Excellent" (average score of 2.5 to 3.0, with 9 studies scoring 3.0 on the point scale) is 43. The other 7 articles were rated as Good (average score of 2.0 to 2.4). On the last point, it was important to note that none in the final corpus were classified in the moderate (1) or poor (0) category thus attesting to the fact that all the included papers comfortably met this minimum quality requirement to be included.

The dataset used in this study, including all papers reviewed, the data extraction table, and quality assessment results, is available in the repository with DOI: https://doi.org/10.5281/zenodo.18674518. The repository ensures the transparency and reproducibility of the systematic literature review process.

3.7. Deduplication Procedure (Bibcitation)

All records retrieved from the database searches and forward/backward snowballing were imported into *Bibcitation* for reference management and deduplication. Duplicates were first detected automatically using exact identifier matching (DOI, PMID/arXiv ID where available). A secondary pass was then performed using bibliographic similarity rules (exact/near-exact title matching, combined with year and first-author checks) to capture records with missing identifiers or minor formatting differences. Finally, all flagged pairs were manually verified to prevent erroneous removals (e.g., conference and journal extensions treated as distinct records when substantially different). This process removed 10 duplicate records, yielding 140 unique records for title/abstract screening.

4. Analysis

4.1. Overview of Included Studies

This section provides a summary of the key characteristics and scope of the studies included in this systematic literature review. In the selected studies, authors focus the research on applying DL and VLMs to pathology image analysis, as well as the use of multi-agent systems to generate clinical reports of the analysis. They collectively represent the latest developments, datasets, methodologies, and challenges in the domain at large.

4.1.1. Study Distribution and Research Focus

The majority of the reviewed studies fall within three main thematic areas:

- DL for Pathology Image Analysis: Several studies have developed or refined DL architectures tailored to the unique challenges of digital pathology images. The segmentation, detection, and classification of histopathological structures with CNNs, transformer-based models, or hybrid frameworks are the main focus of these works. For instance: Janowczyk and Madabhushi's extensive research on CNN applications in digital pathology, Kosaraju et al.'s HipoMap, a slide-based representation framework, and Neidlinger et al.'s EAGLE framework with a focus on efficient tile selection and feature extraction in WSIs [42, 43, 44].

- VLMs for Image-Text Pair Generation and Report Automation: Studies like those by Sun et al. [45] with PathGen-1.6M and Shi et al. [46] with ViLa-MIL highlight the integration of vision and language models to generate clinically relevant textual descriptions and diagnostic reports from pathology images. Other works include Ding et al.'s TITAN foundation model for multimodal learning including WSIs and pathology reports at the scale that achieves strong generalization across multiple tasks without fine-tuning [47].

- Multi-Agent Systems and Collaborative Frameworks: Emerging research illustrates the promise of multi-agent approaches where specialized AI agents handle distinct subtasks, such as feature extraction, text generation, and validation. Systems like LEAVS and MCCD demonstrate how collaborative agents improve annotation quality and generate synthetic training data, respectively, thereby enhancing the overall system performance and clinical utility [48, 49].

4.1.2. Data Sources and Datasets

The studies reviewed exploit diverse categories of both publicly and privately available datasets, including dedicated datasets in the form of The Cancer Genome Atlas (TCGA) and giant datasets of recently created pairs between pathology images and their text typically (PathGen-1.6M and Quilt-1M) [45, 50]. These data sets can be used to train and test models on various types of cancer, image type, and during the clinical scenario, which enables the creation of versatile and general AI tools.

4.1.3. Methodological Approaches

Approaches range in both directions between fully supervised CNN models and weakly supervised learning models (e.g. CLAM) and more sophisticated transformer and self-supervised learning models. In order to address the weakness of the quantity of labeled data, researchers use many techniques; typically data augmentation, synthetic image creation, and certain active learning methods. Moreover, domain-specific knowledge integration is provided, in order to make generated text correspond to clinical terminology and guidelines.

4.2. Evaluation Metrics and Performance

Performance evaluation consists of image-specific (e.g., F-score, accuracy, AUROC) and natural language generation (e.g., BLEU, clinical relevance measures) metrics used to evaluate performance across both modalities of image depiction and natural language generation [51]. Certain frameworks are demonstrated to be able to process at real time or near real time which is essential to clinical deployment and most studies have shown their capabilities to be better than handcrafted feature methods and older AI models.

Conventionally, the AI medical reports were evaluated against the usual measures of Natural Language Generation (NLG) such as BLEU, ROUGE and METEOR. Though these are convenient in measuring syntactic fluency and overlaps between the text and reference report against a ground-truth, the measures are basically a sham in the medical profession. A genuine report may score a high BLEU score in a situation where the resultant generated text replicates boilerplate text; thus, the generated text does not contain a very crucial diagnostic specifier that results in a calamitous clinical error[52].

4.2.1. Evaluation Metrics and Operational Definitions

In an attempt at providing formal measures of comparative heterogeneous studies in different multimodal and multi-agent constructs, the use of evaluation measures and operational terms used when conducting this review must be clearly defined. Every specific metric is preconditioned by the type of the underlying task:

- **Classification and Diagnostic Tasks:** The area under the receiver operating character parallel curve (AUC) is used to classify the performance in most cases. The main benefit of AUC is that it quantifies the power of a model to differentiate between classes (e.g., benign vs. malignant tissue) at any level of classification, thus it is much more suitable and resilient to apply to inherently imbalanced data sets such as those in pathology.

- **Segmentation and Spatial Grounding Tasks:** Tasks that must localize anatomical structures are based on the Dice Similarity Coefficient (DSC) and Intersection over Union (IoU). These measures determine the spatial overlap of the AI-predicted boundaries and the ground-truth pathologist annotation, needed to assess tumor delineation, as well as Funcitional Visual grounding of Vision-Language Models.

- **Syntactic Text Generation:** Traditional Natural Language Generation (NLG) tasks use metrics like **BLEU**, **ROUGE**, and **METEOR** to measure n-gram (word sequence) overlap between AI-generated text and reference

reports. While appropriate for assessing the syntactic fluency of a generated report, they cannot measure factual medical accuracy.

- **Semantic Report Generation:** To evaluate true medical utility, studies increasingly rely on **Clinical Correctness** and **Clinical Efficacy (CE)** metrics. These evaluate the factual accuracy of the text (e.g., successfully predicting diagnostic tags and omitting contradictory hallucinations) rather than simple word-matching.

Furthermore, several repeated qualitative claims extracted from the reviewed literature require strict operational definitions:

- **Acceptance Rate:** This review operationally defines acceptance rate as the percentage of AI-generated reports or diagnostic recommendations successfully utilized in a clinical workflow by an expert pathologist with no or only slight grammatical changes. It is based on this definition because the human-in-the-loop reader studies were removed directly out of the review process.

- **Clinically Accurate:** An AI output is considered to be clinically accurate when it can detect the main diagnostic characteristics (i.e. tumor grade, histological subtype) and provide no conflicting data or omit the most important nominal (e.g., the word no translated before a modifier). This is confirmed by grading schemes of clinicians in a structured manner and not by automated text measurements.

4.2.2. Balancing NLG with Clinical Correctness and Clinician Grading Schemes

To bridge the gap between textual similarity and medical accuracy, recent frameworks increasingly balance conventional NLG metrics with rigorous Clinical Efficacy (CE) metrics and structured clinician grading schemes. For example, the analysis of the **PathAlign** framework is performed via a special pathologist grading scheme, going beyond the word matching to narrowly evaluate whether there are any clinical significant errors or omissions present or absent predetermined issues on an individual basis[53]. Pathologists carry out a manual review of the generated text in their reader studies to verify that the generated text is able to be used to reach the diagnostic conclusion and prioritize the workflow without hallucinatory finding.

Similarly, models like the Patient-level Multi-organ Pathology Report Generation (**PMPRG**) framework emphasize the necessity of tracking clinical correctness in complex, inhomogeneous multi-organ reports. **PMPRG** uses targeted Clinical Efficacy (CE) measures to assess the ability of the formulated reports to make the right predictions and include predefined diagnostic ”tags” to supplement standard NLG measures such as METEOR [54]. Through assessing the accuracy of tag classification, the researchers can measure quantitatively whether the generated text is effective in capturing the key organ-specific diagnostic features as opposed to merely generating generic and fluent text.

4.2.3. Grounding Metrics and Spatial Interpretability

Moreover, assessing the factuality of a report would be accomplished by demonstrating that the textual claims of the AI are based on the right visual evidence, as opposed to being hallucinated by the statistical priors of the language model. This has seen the incorporation of grounding measures in the evaluation of reports. The grounded report generation process necessitates the model to match particular textual assertions with accurate X/Y coordinates or Regions of Interest (ROIs), on the entire slide image.

Measurement of this spatial correspondence means comparing model predicted attention maps or multi-scale regional features with pathologist-labeled ROI. Grounding metrics ensure an audit trail hence requiring models to explain the origin of their generated text with reference to its visible basis of evidence. This not only makes the form of the pathology report produced syntactically fluent and clinically sound, but also, correctly pegged to the underlying morphological reality of the tissue sample.

4.3. The Weak Supervision Challenge and MIL Evolution in WSI Analysis

Since WSIs are gigapixels by definition, it is virtually impossible to perform exhaustive pixel-wise annotation of such images by highly trained, expert, pathologists. This means that only one slide-level diagnostic label (i.e. malignant or benign) can be used to train models even though the morphological evidence involved in the actual diagnosis may be a minute fraction of the scanned tissue. To address this, Multiple Instance Learning (MIL) has been discovered quite useful with the field with a WSI being viewed as an instance of a bag of smaller, un-labeled image patches or instances[55].

MIL architectural developments have followed a series of different paradigms to reflect more accurately complex tissue morphology through weak supervision. In early techniques, simple pooling (max or mean pooling) was used, and this was not capable of isolating individual sections of diagnostic interest in background tissue. This shortcoming resulted in the accomplishment of Attention-based MIL whose dynamically ascribed learnable weight values to the patches allows the model to visually target regions of interest. In order to increase the unrestrictedness of discerning the extents of the spatial correlations amid the tissue structures of further areas, the field moved back to Transformer-based models with TransMIL the most eminent. More recently to further justify the large scale sequences of WSI patches with further computational power, scholars modified State Space Models that formulate infrastructures like Mamba-MIL. Meanwhile, other advanced clustering-based models have been suggested, such as MiCo, which enhances the feature representations through the contrastive and pseudo-labeling learning. This history of the simplest kind of pooling, to Mamba-MIL, to MiCo is important to comprehending why the approaches to multimodal models are effective nowadays to ground their clinical text generation in the vast scale, unsupervised visual space [56].

The evolution of attention-based and transformer-driven MIL is not merely an endpoint for slide-level classification; it serves as the critical visual engine for the advanced multimodal and multi-agent frameworks discussed in subsequent sections. For instance, modern MIL architectures produce attention weights and spatial hierarchies that are automatically used as natural heatmaps to determine the prioritization of ROI in multi-agent pipelines. PathFinder

[103] and SlideSeek [108] Systems are based on these localized signals to steer their Navigation and Explorer agents to a diagnostically rich area in the gigapixel expanse. In addition, structured patch aggregation is a fundamental requirement of Vision-Language Models (VLMs). Such multi-scale visual embeddings are used in frameworks such as PathAlign [68], which facilitates region-aware grounding, in which the generated clinical text is aligned with particular cellular morphologies. Finally, the combination of such complex representations of MIL works as one of the major protective measures against diagnostic hallucinations. These architectures achieve strict control over the text-generation process through high-attention, clinically relevant tissue patches instead of the unfiltered WSI background by imposing downstream language models with other unsupported clinical claims [94, 95].

4.4. Deep Learning Approaches for Pathology Image Analysis

DL has emerged to be a key technology in digital pathology, which has provided revolutionary advances in image processing via automated characteristic study of raw information directly at the information stage [57]. Examination of pathology images presents very special problems of very large-resolution WSIs, complex tissue structure and the extreme variability of the staining procedure and imaging equipment used [58]. The state-of-the-art of working with the DL approaches used in the analysis of pathology images is described, including the review of the model architectures, data strategies, and clinical application as outlined in the literature.

4.4.1. CNN-Based Architectures and Segmentation Frameworks

Janowczyk and Madabhushi had first performed early foundational work demonstrating the versatility of CNNs in digital pathology that was widely featured in their work [42]. They tested CNN in working with several tasks: nuclei segmentation, epithelium and tubules segmentation, lymphocytes and mitosis, and cancer classification. Their strategy involved Caffe DL framework, high F score (maximum of 0.90 in lymphocyte detection), but also aspects of the approach like choice of magnification and label quality that will be suitable in clinical translation. In this regard of a thorough assessment, CNN hegemony in pathology images was logical [42]. Huang et al. augmenting CNN capabilities combine advanced segmentation U-Net with classification EfficientNetV2 with a novel algorithm of constructing a heatmap, grounded on data enhancement, ensemble learning, and attention mechanism into a new algorithm[59]. It was revealed that the multi-component framework only reached unbelievable accuracy (precision 92.86 percent, recall 86.67 percent) on the Cancer Genome Atlas (TCGA) task of identifying cancer lesions that proves the importance of sophisticated fusion of models to enhance both the performance and scalability of analysing pathological images [59].

To overcome the limitation of data scarcity and annotation, Hossain et al. [60] added synthetic image generation using a cycle-consistent GAN to train using nuclei segmentation. The accuracy of segmentation has increased the Dice similarity coefficient (DSC) of 0.984; segmentation under models with synthetic images against models with original images also (DSC 0.805). The scarcity of the annotated pathology images is characterized by the essential role of the synthetic data creation and the advanced pipeline of preprocessing marked as the key issue to the future of

the research field [60]. In the same way, Zhang et al. [61] reported the application of CNN-based histopathological image analysis to predict the risk of further progression of oral leukoplakia, and the DL models could distinguish morphological patterns that are specific to the risk of cancer occurrence [61].

4.4.2. Slide-Level and Whole-Slide Image Representations

While patch-wise CNN analysis is common due to computational constraints, slide-level diagnosis demands aggregation across large WSIs. Kosaraju et al. [43] introduced HipoMap, a slide-based framework converting WSIs into structured representations compatible with CNNs. HipoMap is a slide-based framework introduced by Kosaraju et al. to convert WSIs into a structured representation for compatibility with CNNs. HipoMap was shown to outperform existing methods for lung cancer classification (AUC 0.96) and survival prediction (c-index 0.787) and presented a flexible, task-agnostic, slide-level pathology analysis that meets the needs of the clinical participant [43]. Slide representation learning methods are increasingly being explored in an unsupervised way. PANTHER (Song et al., [62]), is a Gaussian mixture model-based approach to derive morphological prototypes in the space of WSI patches, resulting in expressive, task-agnostic slide embeddings. On 13 datasets, their method matched or surpassed the performance of supervised MIL approaches and provided improved interpretability through prototype analysis which is a major advantage toward clinical adoption [62]. In particular, Ding et al. [47] extended multimodal integration with the development of TITAN, a large-scale foundation model pre-trained on over 335,000 WSIs that have been provided with associated pathology reports and synthetic captions. Zero-shot and few-shot generalization with this self-supervised learning and vision-language alignment to diverse clinical tasks such as retrieval of rare cancer from whole slide images, and generation of pathological reports, demonstrate a possible new frontier of DL models incorporating both image and text modalities to provide versatile clinical support [47].

4.4.3. Efficiency and Scalability in High-Resolution Image Analysis

Processing gigapixel WSIs remains a computational bottleneck. EAGLE, an efficient DL framework for emulating pathologist workflows based on two foundation models, CHIEF for tile selection and Virchow2 for feature extraction, is presented by Neidlinger et al. [44]. It is shown that EAGLE reduced the computational time by over 99% across 31 cancer-related tasks to near real time processing speed (2.27 seconds per slide) while achieving performance superior to the aforementioned leading models. This broad clinical application without requiring high-performance computing resources [44]. Meirelles et al. [63] designed an active learning (AL) framework that acquires data with diversity-aware data acquisition and network auto-reduction techniques to reduce expert annotation time without sacrificing model accuracy. Their approach is then applied to tumor-infiltrating lymphocyte classification and achieves superior predictive performance at up to 4.3× reduction in execution time for AL, while simultaneously addressing one of the key challenges in applying DL to pathology: high annotation costs [63].

4.4.4. Addressing Variability and Robustness: Data Augmentation and Domain Adaptation

The heterogeneity of pathology images, influenced by staining differences, scanning protocols, and patient populations, poses significant challenges to model generalizability. The literature uses extensive data augmentation as a strategy to overcome the issues of overfitting and robustness [42, 59] as well as synthetic data generation [60]. Similarly, Lu et al. proposed CLAM, a weakly supervised attention-based MIL model, which makes use of slide-level labels to limit the intensive annotation of the patches and guarantee model versatility in cohort and /or imaging contexts. Such an interpretable attention installation provided by CLAM focuses on diagnostically significant parts that are more reliable in the eyes of the clinician and more intelligible to the model [64]. A number of studies indicate that the incorporation of domain specific knowledge can be useful in order to leverage clinical transposability. Recently, Zhang et al. demonstrated that risk stratification models based on CNN can be improved by means of integrating the observed histological patterns with the clinical outcomes associated with them in patients with breast cancer of the second and third stages of the disease [61]. Huang et al. also enabled them to capitalize on pathological expertise and influence both attention mechanisms and ensemble strategy that contributed to interpretability and could acquire precise information on the topic [59]. Having incorporated this, the difference between the results provided by the AI and the real clinical decision-making processes is narrowed, and the results provided by the generation become valuable and practical.

4.4.5. Comparative Analysis and Limitations

Across the reviewed deep learning architectures, a clear evolutionary trade-off emerges between annotation dependency and clinical interpretability. Wholly trained CNN models are still the best at handling localized accuracy challenges, e.g. nuclei segmentation, but are fundamentally constrained by the prohibitive cost of annotating pixels pointwise [42],[60]. To avoid this bottleneck, the domain has a common architectural jump to weakly supervised MIL (e.g., CLAM) and self-supervised foundation models (e.g., TITAN, PANTHER). Nonetheless, this paradigm comes with a non-consecutive design bottleneck: as models are made less supervisory on the pixels and more move to supervisory on the slide, the problem of spatial interpretability to clinicians increases exponentially. In turn, hybrid frameworks, e.g., hybrid EAGLE [44] and active learning pipelines [63], that sensibly trade off the high-level contextual understanding of the foundation models with high-efficiency tile selection, seem to be the most viable near-term clinical solutions in that they can both be fast and interpretable.

4.5. Vision-Language Models and Large Language Models in Medical Report Generation and Image-Text Pair Generation

The integration of VLMs and LLMs has brought a paradigm shift in the medical image analysis and report generation. Integrating the capability to read and interpret intricate visual content and the capacity to write consistent and clinically useful text, these models are revolutionizing diagnostic processes in radiological, pathological, and other specialties relying heavily on imaging diagnostics, visual, and textual material [9]. This section discusses the cur-

rent state of VLMs and LLMs in the medical fields, gives a comprehensive overview of the existing methodologies, datasets, architectures and clinical applications and outlines emerging issues.

4.5.1. Large Language Models in Medical Question Answering and Clinical Dialogue

Large language models, initially designed for broad NLP tasks, have been adapted and fine-tuned to address the specialized requirements of medical question answering (MQA) and dialogue systems [5]. This field includes models such as GPT-4, ChatGPT, LLaMA (Touvron et al., [65]), PMC-LLaMA (Wu et al., [66]) and domain-tuned variants like MedPaLM (Singhal et al., [67]) and MedPaLM 2 (Singhal et al., [68]). For instance, MedPaLM 2 gains a 19% improvement in accuracy on the MedQA benchmark through fine-tuning and prompt engineering with more domain-specific knowledge [68]. Furthermore, the authors in Nanayakkara et al. [69] have shown that these LLMs can be used for ASR and transcription error correction in clinical conversations by using seq2seq approaches that use T5 and BERT architectures [69]. As such, this development caters to the practical need of accurately documenting clinical dialogue, a basis for other natural language generation tasks.

Despite impressive progress, the performance of general-purpose LLMs is often constrained by gaps in specialized medical knowledge and the inherent complexity of clinical reasoning. Duong and Solomon [70] found that the genetic question-answering performance of ChatGPT was on par with human experts, but fell short in comparison to human experts, demonstrating the weakness of a pre-trained model entirely devoid of domain-specific adaptation [70]. In response, researchers come up with systems such as ChatDoctor (Li et al., [71]) to augment LLaMA with self-contained information retrieval mechanisms, retrieving additional information from Wikipedia in real time in order to increase the accuracy and relevance of responses [71]. Clinical Camel (Toma et al., [72]) also uses dialog-based knowledge encoding with session memory and active knowledge base expansion to exceed GPT-3.5 on USMLE results and enable rich case management and generation of clinical documentation. An important trend is that hybrid systems enhance pre-trained LLMs with external knowledge bases or retrieval modules to surmount the limitations of static training corpora. As examples of culturally and linguistically sensitive medical LLMs, Chinese LLMs include DoctorGLM (Xiong et al., [73]), Zhongjing (Yang et al., [74]), BenTsao (Wang et al., [75]) and Huatuo (Li et al., [76]). Zhongjing improves complex dialogue and question handling using reinforcement learning with human feedback; and DoctorGLM provides precise symptom, diagnosis, and treatment guidance in Chinese by its prompt design and disease knowledge libraries [74, 73].

Although originally intended as a more general clinic conversation, the active knowledge retrieval tools applied in these tools can be easily transferred to pathology. They could be reconfigured to interface with external genomic databases or histopathology grading systems (e.g., Nottingham grading of breast cancer) in real-time as additional important contextual information to analyzing whole-slide images.

4.5.2. LLMs for Clinical Documentation and Medical Report Generation

Automation of clinical documentation is a prime target for LLM applications, promising to reduce clinician workload and improve report standardization. According to Cascella et al., [77], ChatGPT can generate medical notes for ICU patients with high accuracy in categorizing complex physiological parameters and LLM tends to be able to self-correct. Building on this work, Ali et al. [78] scaled this work to generate high-quality clinical letters across different scenarios with letters generated faster and with better patient satisfaction than in manual documentation. Given their larger parameter size and training data, the results reported by Waisberg et al. [79] and Abdelhady and Davis [80] demonstrate that GPT-4 can produce surgical discharge summaries, interpret medical images and handle complex clinical trial documentation. GPT-4 was also efficacious compared to ChatGPT 3.5 when tasked with EHR inbox management, with admin and chronic disease management contexts favoring GPT-4 for hospital adoption.

A number of studies have been conducted in the field of radiology to examine how GPT can be used to process unstructured free-text reports and convert them into structured formats. Mallio et al.[81] investigate applications of GPT-3.5 and GPT-4 to Italian CT reports and report up to 75 per cent word count reduction and improved clinical recall in GPT-4. Mallio et al. also examined the knowledge of GPT-4 regarding structured reporting (SR), and established that in fact GPT-4 has a highly strong knowledge of structured reporting, and can generate large and detailed structured templates. Adams et al. [82]established a 100 percent success rate of automated translation of English radiology reports to structured X-rayJSON formats and a high success rate of 100 percent on German chest X-ray datasets, and promising multilingual, multi-task promise of LLM. In distal radius fracture reporting, Bosbach et al. [83] experimented with GPT-3.5 and found high scores in grammar and style, certain problems in medical interpretation associated with realization and that the area knowledge is necessary to be sensitive enough. Wang et al. [84] argue that the output of GPT-4 structured liver ultrasound reporting was more diagnostic and efficient compared to conventional free text reporting. Then, Jiang et al. [85] compared GPT-3.5 to GPT-4 concerning thyroid ultrasound reporting, where GPT-4 was observed to be more effective over GPT-3.5 in terms of nodule hostility, as well as, in the consistency of the management recommendation.

Li et al. [86] suggested a novel pipeline, which is an interactive fusion that detects anatomical regions and prompts generated GPT-4 reports simultaneously on the first sight of a chest X-ray. According to this, their style was found to enhance anatomical and clinical detailization that is also another tendancy in terms of more interactive and understandable LLM-driven monitoring systems. Additionally, Pan et al. [87] demonstrate that GPT-4 can produce FHIR-compliant structured radiology reports for multiple modalities, with high accuracy and internal consistency, ready to be incorporated into a standardized healthcare framework for data.

The effectiveness of these models to translate free-text radiology reports to structured forms gives a first-hand guide to computation pathology. A closely related NLP pipeline can be applied to histopathology to automatically identify standardized synoptic elements of report (including tumor margins, mitotic rate, and lymphovascular invasion) in historically unstructured surgical pathology reports.

4.5.3. Vision-Language Models for Medical Report Generation and Visual Question Answering (VQA)

VLMs extend LLMs by jointly processing images and text, enabling multimodal medical report generation and visual question answering. Some of the state-of-the-art VLMs in pathology include MedViLL (Moon et al., [88]), PubMedCLIP (Eslami et al., [89]), RepsNet (Tanwani et al., [90]), BiomedCLIP (Zhang et al., [91]) and others, discussed their architectures and evaluated performances on datasets like MIMIC-CXR, Open-I, VQA-RAD and SLAKE. In particular, MedViLL stacks a ResNet-50 visual encoder with a BERT (based) textual encoder that includes positional embeddings into a unified transformer architecture. Fine-tuned on smaller datasets, pre-trained on nearly 90,000 image-report pairs, it achieves BLEU-4 scores around 0.06 and clinical label accuracy above 84%, generalizable to radiology report generation and question answering [88].

PubMedCLIP enhances contrastive learning on biomedical image-text pairs using ViT-B and Transformer text encoders, achieving superior VQA accuracy when integrated with question-conditioned reasoning frameworks [89]. By combining ResNeXt-101 image encoding, BERT text embedding, and GPT-2 decoding and jointly learning via bidirectional contrastive and cross-entropy losses, RepsNet is shown to achieve competitive BLEU scores on medical image captioning [90]. Using VQGAN tokenization and transformer generations, UniXGen (Lee et al., [92]) innovatively generates both chest X-rays and radiology reports, trained on over 200,000 studies and representing a high-performance multimodal generative method. To improve VQA on various benchmarks, RAMM (Yuan et al., 2023) introduces retrieval augmented attention mechanisms that incorporate retrieved image text pairs from large biomedical corpora [93].

X-REM by Jeong et al., [94] tackles the challenge of hallucinated information in radiology report retrieval and generation using ALBEF-based multimodal encoders and clinical label-informed ITM scoring, enhancing report relevance and semantic accuracy. To generate a radiology report that is grounded on high similarity retrieved impressions while simultaneously minimizing hallucination and improving clinical fidelity, CXR RePaiR—Gen (Ranjit et al., [95]) adopt Retrieval augmented generation (RAG) frameworks with pre-trained ALBEF models. LLaVa-Med (Li et al., [96]) apply the LLaVa multimodal foundation model to biomedicine by employing curriculum learning on PMC-15 datasets and multi-round question answering to demonstrate promising results on the VQA-RAD, SLAKE, and PathVQA datasets, showing the capacity of large pre-trained multimodal models to pathology and radiology.

Even though these methods are optimized to work with macroscopic radiological scans, retrieval-augmented generation (RAG) and hallucination-reduction methods observed in X-REM and CXR RePaiR-Gen fill a significant gap in pathology text generation. These methods may achieve substantial elimination of diagnostic hallucinations by the grounding generated text using retrieved high-similarity historical WSI patches, which are pertinent on very complex and ambiguous histopathological patterns in diagnosis.

4.5.4. Specialized Vision-Language Models in Pathology Image Analysis

Pathology presents unique challenges due to gigapixel WSIs, diverse tissue structures, and complex diagnostic criteria. In response to these issues, domain-specific VLM have thus been developed. The visual encoder uses a

pathology language image pretraining (PLIP) model trained on a large curated human pathology image text dataset, PathologyVLM, introduced by Dai et al. [97]. The method consists of two-stage training, on domain alignment that aligns various visual domains and VQA fine-tuning on the data at WSI resolution with a scale-invariant connector that keeps resolution intact. VLM significantly outperforms both zero-shot and supervised general domain and other medical VLMs specifically designed for VQA on PathVQA and PMC-VQA datasets.

Ahmed et al. [53] developed PathAlign, based on the BLIP-2 framework, trained on over 350,000 WSI-text pairs across multiple tissue types and diagnostic categories. Image-to-text retrieval, generative report, with frozen LLMs integration, and 78% accuracy on pathologist rating without significant clinical errors are supported by the model. Additionally, this large-scale dataset and model show the feasibility of language-aligned WSI embeddings for histopathological workflows. In Huang et al. [98], hospital pathology images and accompanying text posted on medical Twitter were drawn from public sharing and synthesized into OpenPath, a dataset of over 200,000 image text pairs. Authors show that their PLIP model achieves state-of-the-art performance for zero-shot classification with a variety of external pathology datasets and largely preserves the ability to search pathology images with multimodal text queries. Authors argue that these pathology-focused models require high-quality domain-specific datasets, and preserve image resolution at model levels, along with end-to-end multimodal training for clinical utility.

4.5.5. Contrasting Radiology and Pathology Reporting Paradigms

While earlier sections highlighted how radiology-based models serve as architectural blueprints for multimodal AI, it is critical to explicitly contrast the reporting paradigms of these two fields. Radiology reporting tends to be usually macroscopic and observational that usually ends up in descriptive narrative which needs to be correlated with clinical methods. Contrarily, histopathology is usually the ultimate diagnostic ground truth. As a result, a direct translation of radiology-focused Natural Language Generation (NLG) into pathology presents a high risk of clinical risks in case pathology-specific peculiarities are overlooked.

4.5.6. Pathology-Specific Nuances and Synoptic Alignment

Pathology reporting entails unique challenges that current foundational VLMs must address. First, models have to deal with high diagnostic uncertainty and within-observer error, especially in subjective tasks such as tumor grading and definition of delicate morphological thresholds. Second, in contrast to the descriptive stories, prevalent in the radiology department, contemporary oncology is based on well-organized information. The critical prognostic modifiers, including the precise tumor margins, mitotic rates, lymphovascular invasion, and the outcome of additional immunohistochemical (IHC) tests, should be reflected correctly in the AI-generated pathology report [99]. Consequently, one significant future trend of agentic AI in this area not only is moving away the origination of descriptive free text but also filling in pre-formatted, structured templates. Generative pathology models in the future should clearly be aimed at achieving the level of strict adherence to synoptic reporting, including the site entries of the College of American Pathologists (CAP) or the International Collaboration on Cancer Reporting (ICCR) [100]. The ability

of Vision-Language Models to extract these particular synoptic data points of gigapixel WSIs in an automatic, and faithful fashion is a basic requirement of their clinical implementation.

Synthesizing the advancements in VLMs and LLMs, a recurring architectural pattern is the necessary transition from generic biomedical text adaptation to pathology-native multimodal alignment. Initial generative models would simply bail out WSI patches into the regular radiological vision encoders, leading to a disastrous loss of cellular resolution. Reacting to this, higher-end systems have found the specialized architectural extensions now universally used, like scale-invariant connectors (as in PathologyVLM [97]) or multi-scale feature alignment (as in PathAlign [53]) to preserve gigapixel context. Nonetheless, one common unreliable bottleneck of most of the reviewed systems, despite the structural advantages, is that they are dependent on training on unfiltered, image-text pairings that are noisy and uncurated based on their status on public boards or unstructure historical records. Until these models are trained on structured and synoptic level data, they will have limited ability to produce consistently clinically safe and standardized pathology reports.

4.6. Multi-Agent Systems and Integrated Frameworks in Medical Imaging

Recent advances in medical imaging AI have increasingly emphasized the integration of multi-agent systems and multimodal frameworks to more closely emulate the collaborative, iterative nature of clinical decision-making [101]. While single-task DL models perform the entire complex diagnostic workflow in silico, multi-agent frameworks distribute the workflow across specialized, interacting AI agents for richer and interpretable medical image analysis [102]. This section covers some of the key developments in multi-agent and integrated artificial intelligence systems, discusses their design strategies, illustrates the benefits, and describes their clinical impact. For instance, PathFinder was designed to mimic expert pathologists' diagnostic workflow for histopathology WSIs using a multi-agent multi-modal system. To address the massive size and complexity of WSIs, PathFinder relies on four separate agents (Triage Agent, Navigation Agent, Description Agent, and Diagnosis Agent) working in sequence to triage WSIs, navigate diagnostically relevant areas, create natural language annotations from the areas and synthesize a final diagnosis [103]. This design shows the iterative nature of pathologists examining slides, where they look at salient patches of the image, write notes, and synthesize the information toward a cohesive clinical interpretation. PathFinder performed better than state-of-the-art models in melanoma classification by improving the end the average error of pathologists by 9 percent and produced inherently explainable predictions represented by the natural language descriptions of the Description Agent that are qualitatively better than GPT-4o. These findings showcase the major role MAS architectures play in ensuring that the accuracy and interpretability is enhanced, which are required to be adopted by clinicians themselves [103]. Information that a given specific model is somehow qualitatively superior to GPT-4o needs to be put into context via the evaluation rubric. In such cases the superiority was determined in terms of an LLM-as-a-judge framework that was designed with a severe 5-point Likert scale that specifically penalized models that exhibited diagnostic hallucinations and rewarded them based on clinical safety and compatibility with pathology reporting reports.

The PathGen-1.6M project also provides another example of multi-agent teamwork as it is a 1.6M pathology image-text pairs creation with the help of the ensemble of specialized AI agents that create the representative patches of WSI, generate, and refines the captions that are used in the process[45]. This large-scale multiagent pipeline, as a type of big data curation in pathology AI, aims to address one of the most notable bottlenecks: the lack of well-annotations multi-modal datasets, as well as the lack of African American specialists to annotate them which is a major limitation when training powerful VLMs. The paper demonstrated how the pairs generated may be added to existing datasets to train a pathology-specific CLIP model (PathGen CLIP) that can achieve higher accuracy on nine zero-shot image classification tasks and on a number of tasks related to whole slide image analysis. Moreover, the authors tuned PathGen-CLIP with Vicuna LLM to produce a strong multimodal model that is optimized on pathology as a platform to next-generation hepatic diagnostic systems. This method proving the strength of multi-agent models to produce and learn models with scales and data quality unaccessible to both create and previously previously unattainable describes a significant bottleneck in the development of medical AI.

Multi-agent and integrated frameworks are becoming popular in radiology for automating report generation and diagnostic assistance. RaDialog (Pellegrini et al., [104]) which includes a large vision language model that achieves clinically accurate generation of interactive dialog chest X-ray reports; the model leverages image features, structured pathology findings, and large language model capabilities. RaDialog was trained on a semi-automatically labeled instruct dataset constructed from the MIMIC-CXR database and outperformed existing models on clinical correctness while showing state-of-the-art abilities in identifying report correction opportunities and answering clinician queries. The model, significantly, also decreased a mean of 33% errors in comparison to prior reports and interactive models, for example, XRayGPT [104]. The dialog-based interactive nature of RaDialog reflects well in transferring to digital pathology workflow. A similar MAS in histopathology would enable automated output and would enable pathologists to pose interactive queries to a specialized visual agent regarding particular cells or regions of the cellular environment (e.g., “count the lymphocytes in this tumor microenvironment”) to establish a collaborative human-AI diagnostic cycle.

Complementing these, Biomed-DPT by Peng et al., [105], illustrates an innovative knowledge-enhanced dual modality prompt tuning approach that incorporates both text and visual prompts in biomedical VLMs. Domain-adapted text prompts and vision soft prompts are used in this method which uses LLM-driven prompts to focus (or ignore) the model’s attention on (or areas outside) diagnostically critical regions, improving interpretability and classification accuracy. Biomed-DPT is evaluated on 11 biomedical image datasets in three different modalities and across five organs and outperforms existing prompt optimization methods by 6-8% in classification accuracy, providing a promising example for the applicability of combined multi agent-inspired techniques in using domain knowledge to steer AI focus and improve diagnostic accuracy [105]. The dual text-and-vision prompt tuning method of Biomed-DPT is a strategic avenue that can be used to address the huge quantity of pathology data. Devoting the output of visual agents specifically to diagnostically important tissue microenvironment (like selective playing around with isolating epithelial layers and unstated benign stroma) could be promoted by domain-specific text prompts to map this

methodology to WSIs, in effect shrinking to a pivotal computational load the megapixels of image analysis.

Another LLM-based integration is presented by LEAVS, an LLM-based labeler for complex abdominal CT supervision (Lanfredi et al, [48]). LEAVS extracts structured labels of abnormalities and urgency levels for multiple organs from radiology reports using a specialized chain of thought prompting mechanism together with tree-based decision systems. LEAVS achieves an impressive average F1 score of 0.89, outperforming existing labeling tools and even human annotators on some tasks, with the capability of enriching training of vision models for the detection of abnormality across abdominal organs. When noting that systems such as LEAVS "outperformed human annotators," it is critical to specify the study design: this claim is based on a strict quantitative comparison (F1-score) against two board-certified pathologists using a predefined, structured multi-organ abnormality labeling rubric, rather than an open-ended diagnostic task. This framework demonstrates how to combine an LLM-based label extraction with downstream vision models to supervise complex, multi-organ radiological data through the strength of combining text understanding and image analysis agents all in one system [48]. Although LEAVS was designed to supervise multi-organ CT, the chain of thought prompting and decision tree logic is very relevant to the multi-stain of pathology. A text-parsing agent would be able to interact with complex biopsy texts to produce structured, weakly supervised labels that allow the training of downstream WSI vision with minimal or negligible supervision that requires pathologists to make pixel-level annotations.

Innovations in report generation via cross-modal representation learning are exemplified by the PCRL model by Zheng et al., [106], which addresses two major challenges in brain CT report generation: redundant and shifted visual representations. PCRL builds enriched cross-modal features that align better with clinical report semantics by constructing pathological clues from segmented regions as well as pathological entities. This unified framework provides for a smooth transfer, from feature representation to report generation, via fine-tuning of large language models within task-specific instruction. Experimental results show PCRL's state-of-the-art performance and prove to be a meaningful advancement to multi-agent systems that fuse radiological clinical knowledge tightly with visual data representation for radiological reporting [106]. The PCRL of building pathological clues readily induced by the segmented radiological sites could be readily transposed into frameworks of proclaiming histopathology. One way a pathology MAS would achieve this would be to run a tissue-segmentation agent (e.g. to extract glomeruli in kidney biopsies) before running the text-generation agent to assure that its text reports are strictly limited to localised cellular deviations.

Subsequently, the PathologyVLM by Dai et al., [97], leverages a two-stage training method, consisting of domain alignment from pathology image text data sets and then end-to-end fine turn-on visual question answering (VQA) tasks. PathologyVLM achieves strong cross-pathology performance on both supervised and zero-shot VQA tasks by employing a pathology-specific language image pretraining model (PLIP) as the visual encoder and a scale-invariant connector to avoid information loss due to image resizing. The core of this success is the scientific significance of domain-specific multi-agent inspired components of computational pathology; domain-specific design enables to keep more of the details of the pathological images intact in a better way [97]. ChatEXAONEPath (Kim et al.,

2024) can also be trained on expert-level multimodal large language models anywhere on top of WSIs and similar histopathology reports of the TCGA based on a retrieval-based data generation pipeline and AI-based evaluation protocols. Combining multi-agent strategies in multimodal data, such as ChatEXAONEPath to assist clinicians with complicated cancer diagnosis, has an acceptance rate of 62.9 on pan-cancer WSIs diagnosis [107].

Taken together, these multi-agent and integrated systems constitute the medical AI paradigm shift to no longer isolated task-specific frameworks, to collaboratively context-aware models that are more appropriate to clinical processes. These systems enhance accuracy, efficiency and understandability by coordinating specialized agents. They are also able to generate and annotate data in a scalable way, which is a major issue to deal with in medical AI studies. The future will be an area of further clinical integration, agent communication, and wide-range multimodal fusion-eventually aiming at artificial intelligence systems that can improve clinical decision-making via transparency and adaptability as well as reliability.

4.6.1. Detailed Multi-Agent Pipelines: PathFinder and SlideSeek

The recent developments have shifted the multi-agent systems (MAS) into a theoretical framework to practical, pathology-specific pipelines with the capability of performing autonomous diagnostic argument. The most notable instances of such development are the PathFinder and SlideSeek frameworks that use different agent designs to replicate the process of a human pathologist.

Agent Roles

PathFinder is a program based on a dedicated four-agent architecture that can process WSIs in order to extract features and processes information to calculate outputs in sequential order in real time[103]. The process starts with a Triage Agent who first undergoes a low-magnification examination to categorize the WSI as the benign or risky. In case it is considered risky, the Navigation Agent could simulates the behavior of a panning and zooming gesture of a pathologist to locate regions of interest (ROIs) and create an importance map. This guides the Description Agent that employs a vision-language model to write natural language descriptions of the localized cellular structures. Lastly, a Diagnosis Agent combines these mass descriptions to come up with a concrete diagnostic description.

In contrast, SlideSeek utilizes a hierarchical, reasoning-enabled structure built upon the PathChat+ foundation model [108]. It is based on a Supervisor Agent that serves as an overall coordinator. First hypotheses in high level WSI overviews are formulated by the supervisor who then distributes single exploration tasks to a number of Explorer Agents (or pathologist agents). These explorers will explore to assigned ROI coordinates at the same time and simultaneously extract morphological features, project them back to the supervisor. After the supervisor concludes that diagnostic evidence has been sufficiently collected, a different Reporting Agent puts the data together as interpretable and visually based diagnostic summary.

Communication Protocols The effectiveness of these systems hinges on their underlying communication protocols. PathFinder is based on a protocol that is sequential, state-passing with the output of one agent triggering and constraining the action of the subsequent agent (ex: the visual importance map of the Navigation Agent being the

direct determinant of the set of text that the Description Agent can generate) [103].

SlideSeek, on the other hand, employs a hierarchical ”hub-and-spoke” protocol. Explorer agents are not able to communicate with each other but rather run parallel and are able to communicate only with the central Supervisor Agent. The explorer agents perform the work in parallel scanning the slide at the given level of magnification and recording the morphology of the tissue in their respective reports to the supervisor with the assistance of PathChat+. The managerial aspect of this centralized state means the supervisor can make a series of adjustments of the overarching diagnostic plan by referencing incoming parallel streams of data until an evidence level is achieved [108].

Failure Modes Although medical MAS architectures are highly designed, they are all prone to special collaborative failure modes. The recent study of auditing medical multi-agent collaboration reveals the following vulnerabilities that are critical:

- **Echo-Chamber Amplification (Flawed Consensus):** In the event that an upstream agent (such as the Description Agent in PathFinder) had hallucinated a morphological feature or incorrectly interpreted visual evidence based on constraints of the base model, downstream agents can be confident and hasty about their error without checking it. This causes a bad judgment in that, the eventual diagnosis is sure to be incorrect.

- **State Synchronization Failures:** In hierarchical systems such as SlideSeek, failure of the central supervisor to correctly update state of his/her hypothesis, in response to the explorer, can cause stale state propagation or state update conflicts, which can propagate duplicate work or give conflicting diagnostic information.

- **Infinite Loops (Circular Task Delegation):** It is possible to endlessly get stuck in an unrestricted back-and-forth loop of agents without strict termination conditions. As an example, a supervisor agent could constantly demand finer grained analysis of ROI of the explorer agents but will in no manner arrive to the diagnostic confidence level necessary to cause the reporting agent to activate.

4.7. Comparative Analysis of State-of-the-Art Multimodal, Agentic, and Benchmarking Frameworks

Recent advancements in the rapid development of computational pathology have brought the new wave of multimodal-based foundation models, autonomous multi-agent pipelines, and strong clinical metrics. Table 1 is a synthesis of these state of the art systems in tabular form, to indicate architectural design of the systems and their main performance results in order to show us how fast the field is transforming.

Table 1: Comparison of the recent state-of-the-art multimodal and multi-agent pathology frameworks

Ref.	System	System Type	Dataset Scale	Supervision Paradigm	Clinical Reader Validation	Evaluation Metrics	Key Features	Key Results
[108]	PathChat+ and SlideSeek	MLLM and Multi-Agent	1.13M instruction samples, 5.49M QA turns	Instruction-tuned	Yes	VQA Accuracy, DDxBench Accuracy	Instruction-tuned pathology MLLM. Hierarchical agent framework (Supervisor/Explorer) for ROI navigation.	-Superior diagnostic reasoning on gigapixel WSIs. Highly interpretable, autonomous evidence gathering.
[47]	TITAN	Foundation Model	335,645 WSIs, 423k synthetic captions, 183k reports	Visual SSL	Yes	Zero-shot/Few-shot Classification, Cross-modal Retrieval	Pretrained on more than 335,000 WSI-report pairs. Utilizes visual self-supervised learning and vision-language alignment.	Exceptional zero-shot and few-shot classification. High accuracy in cross-modal retrieval without fine-tuning.
[109]	SlideChat	Vision-Language Assistant	4.2k WSI-caption pairs, 176k VQA pairs	Visual Instruction Learning	Yes	BLEU-1, GPT Score, VQA Accuracy	Processes entire gigapixel WSIs simultaneously. Employs a sparse-attention slide-level encoder to retain global context.	Outperforms standard patch-based models on WSI-level Visual Question Answering (VQA).

Continued on next page

Table 1 – continued from previous page

Ref.	System	System Type	Dataset Scale	Supervision Paradigm	Clinical Reader Validation	Evaluation Metrics	Key Features	Key Results
[54]	PMPRG	Report Generation	7,422 WSIs (Multi-organ dataset)	Contrastive SSL (MR-ViT) + Language Generation Loss	Yes	METEOR, BLEU, ROUGE, Clinical Efficacy (CE)	Generates structured, patient-level, multi-organ reports. Guided by predefined clinical tags and multi-scale visual features.	Significant improvements in Clinical Efficacy (CE) metrics and diagnostic tag classification accuracy.
[53]	PathAlign	Report Generation	>350,000 WSI and diagnostic text pairs	Vision-Language Alignment (BLIP-2)	Yes	Pathologist Acceptance Rate (78%), Classification Accuracy	Aligns multi-scale WSI features with clinical text. Evaluated using rigorous pathologist grading schemes.	High pathologist acceptance rates. Demonstrates a low rate of clinically significant errors or omissions.
[103]	PathFinder	Multi-Agent Pipeline	M-Path Skin Biopsy dataset (238 cases)	Binary Cross-Entropy + Text-Conditioned Visual Navigation	Yes	Diagnostic Accuracy (74%), F1-Score	4-agent sequential workflow (Triage, Navigation, Description, Diagnosis). Mimics human pathologist panning/zooming behavior.	Outperformed SOTA models in melanoma classification by 8, 9 percent. Beat human pathologist average accuracy by 9 percent.

Continued on next page

Table 1 – continued from previous page

Ref.	System	System Type	Dataset Scale	Supervision Paradigm	Clinical Reader Validation	Evaluation Metrics	Key Features	Key Results
[110]	PathMMU	Benchmark	33,428 QA pairs, 24,067 images	N/A (Evaluation Benchmark)	Yes	Zero-shot Accuracy, Human-Expert Comparison	Largest expert-validated multimodal benchmark in pathology. Contains >33,000 QA pairs testing visual grounding and reasoning.	Revealed a significant performance gap: current top LMMs still lag considerably behind expert pathologists.
[111]	PathBench	Benchmark	15,888 WSIs from 8,549 patients	N/A (Evaluation Benchmark)	None	Bootstrapped 95% CIs, Wilcoxon Signed-Rank Test, Ranking Score	Comprehensive, multi-center, multi-task evaluation framework. Implements strict data leakage prevention.	Standardized model evaluation, highlighting current generalization gaps across diverse clinical environments.

5. Discussion

5.1. Overview of Results

In this section, a synthesis of the findings of the most relevant studies reviewed in this systematic literature review is provided on the topics of DL methods, VLMs, and LLMs, as well as on integrated multi-modal and multi-agent systems being used in the analysis of pathology images and in the generation of medical reports. These studies represent various innovative approaches, including CNNs and transformer models used in the image segmentation and classification task, as well as advanced multimodal approaches, which integrate visual and textual data to provide increased support in diagnosing and to generate reports automatically.

The findings demonstrate high progress across various domains, including the increased segmentation accuracy on pathology images, strong zero-shot classification on VLMs, and improved diagnostic decision-making with multi-agent collaboration. Nevertheless, there are issues of data generalizability and incorporation of domain knowledge as well as explainability. Taking into account these summarized results allows one to reach certain conclusions about this modern situation, where an AI-based pathology analysis can live, and provide some suggestions about the direction that the future research and clinical implementation should take. Table 1, table 2 and table 3 below summarize the approach and the conclusions of notable representative studies in these three fields.

5.2. Critical Appraisal and Evidence Gaps

Although, the literature has shown high architectural developments in multimodal and agentic system of pathology, a critical criticism on the evidence underpinnings has established a high level of gaps in methodological development as well as validation. Most of the researches nowadays state excessively positive outcomes by confusing statistical significance with clinical significance, and hence more strict examination of their real deployability is required.

5.2.1. Dataset Leakage and Evaluation Protocols

A primary evidence gap lies in the structural evaluation protocols of massive VLMs and LLMs. A number of studies state extraordinary zero-shot capabilities, although such conclusions are often undermined by silent leakages of datasets [44, 47]. Since foundation models are being trained on large, usually low-quality-documented corpora on the internet, there is a possibility that public benchmark test sets (like those derived in TCGA) have been accidentally represented in the training AI pitfallsdata. In the absence of serious, verifiable isolation of multi-center, out-of-distribution, test sets, the values of AUC and accuracy values presented in most current papers are likely inaccurate reflections of the generalization properties of the models being represented.

5.2.2. Quality of Clinical Validation

Furthermore, the quality of clinical validation across the reviewed literature is highly variable. While evaluating generative outputs requires ”human-in-the-loop” reader studies, many frameworks rely on critically underpowered validation setups, often utilizing only one or two pathologists to grade the AI’s output [48, 53]. In order to develop a

true clinical trust, the next research should provide metrics of formal inter-observer agreement (e.g. Cohen and Fleiss Kappa) with a larger group of specialists certified by the board. There is further a preponderance of the literature to represent statistical significance as clinical significance and versa. The average increase in a classic NLG score such as METEOR of 2% could be statistically significant but clinically insignificant once a model continues to make sporadic hallucinations that a patient has a malignancy.

5.2.3. Clinical Maturity and Technology Readiness

Finally, it is necessary to contextualize the maturity of these systems using a Technology Readiness Level (TRL) framework. The large proportion of the multi-agent and multimodal system reviewed in this paper have TRL 3 (experimental proof-of-concept) to TRL 5 (in a simulated or relevant laboratory context) performance. They are not implementable clinical tools. The inability to expressly distinguish between a well-controlled proof-of-concept and an adequately sound regulatory-tarted system (TRL 8-9) generates an illusion of clinical preparedness, in the sphere. To cover this gap, the change in the previous testing of retrospective, well-curated datasets would have to change to prospective, real-world clinical trial [112].

5.3. Addressing the Research Questions

CNNs in addition to transformer-based architecture have shown major improvements in the analysis of pathology images, with the advancement of the DL models. These models are effective in addressing the high dimensionality and the heterogeneity of WSI which exploits the hierarchical feature extraction and attention scheme of these studies [42, 43]. The U-Net-based EfficientNetV2 hybrid models further boost the segmentation accuracy at decreased computational costs [59]. Yet variability in staining procedures, imaging protocols, and data from many clinical settings remain barriers to model generalizability and robustness [88]. To lessen annotation burden and increase model adaptability, strategies such as active learning and weakly supervised approaches are introduced [63]. Table 2 provides a summary of DL techniques used in pathology.

Table 2: Deep learning techniques used in pathology.

Ref.	Year	Methodology	Results
[42]	2016	Used CNNs (Caffe framework) for nuclei, epithelium, tubule segmentation, mitosis detection	Achieved F-scores between 0.53 and 0.90 on various segmentation/detection tasks
[43]	2023	Developed HipoMap to convert WSIs into structured maps + CNNs for lung cancer classification	AUC 0.96 for lung cancer classification; improved survival analysis with TCGA data
[59]	2024	Integrated U-Net and EfficientNetV2 with novel heat map generation algorithm	Precision 92.86%, recall 86.67%, processing time 8.26s per image; high interpretability
[61]	2023	CNN-based risk stratification model for oral leukoplakia cancer progression	High-risk group 4x more likely to develop cancer; strong prognostic value
[60]	2023	Cycle-consistent GAN for synthetic data generation; modified U-Net for nuclei segmentation	DSC improved from 0.805 to 0.984 with synthetic images; accuracy 0.97
[63]	2023	Diversity-aware data acquisition function and model simplification for tumor-infiltrating lymphocytes classification	Improved model performance and 4.3x faster active learning execution
[64]	2021	Attention-based weakly supervised learning (CLAM) on WSIs with slide-level labels	Superior data efficiency and adaptability; interpretable subregion identification
[44]	2025	EAGLE framework with tile selection (CHIEF) and feature extraction (Virchow2)	Outperformed SOTA models by up to 23%; 99% faster processing (2.27 s/slide)

5.3.1. How have vision-language models and large language models been employed to generate clinically relevant text from medical images, particularly in pathology?

An overview of VLMs and LLMs used for pathological image analysis is given in Table 3. VLMs and LLMs have shown promising capabilities in bridging the gap between image analysis and automated report generation. PathGen-1.6M and RaDialog serve as examples of exploiting visual data in conjunction with textual clinical knowledge to generate coherent and clinically relevant reports from pathology images and radiological scans [45, 104]. To fine-tune vision encoder and language models with large datasets of image caption pairs for domain-specific tasks which improves diagnostic accuracy as well as reduces the time for generation of the report [97]. These advances notwithstanding LLMs remain limited in terms of domain specificity to medical knowledge, context understanding, and potential for complex/rare cases, thus requiring future work with clinical ontologies and real-world data [92].

Table 3: Vision-language models and large language models in medical imaging.

Ref.	Domain	**Year**	**Methodology**	**Results**
[104]	Radiology and Mixed Medical Modalities	2023	Vision-language integration with LLMs, fine-tuned on chest X-ray reports (MIMIC-CXR)	33% improvement in report correction; superior clinical correctness and interactive abilities
[45]	Pathology Modalities (Whole-Slide Imaging)	2024	Multi-agent collaboration to generate pathology image-caption pairs; train pathology-specific CLIP	Significant improvement across 9 zero-shot tasks and WSI analyses
[97]	Pathology Modalities (Whole-Slide Imaging)	2024	Two-stage learning: domain alignment + VQA fine-tuning; pathology-specific visual encoder	Outperformed general domain and medical VLMs on VQA tasks; better image feature preservation
[96]	Radiology and Mixed Medical Modalities	2024	Fine-tuned LLaMA model with medical multimodal instruction-following datasets	Demonstrated strong VQA capabilities on multiple biomedical datasets
[98]	Pathology Modalities (Whole-Slide Imaging)	2023	Trained PLIP multimodal model on OpenPath dataset (208k pathology images with captions)	Achieved F1 scores 0.565–0.832 on zero-shot classification; improved knowledge retrieval
[107]	Pathology Modalities (Whole-Slide Imaging)	2024	Retrieval-based pipeline using WSIs and TCGA reports; evaluation on pan-cancer diagnosis	62.9% acceptance rate in clinical diagnosis; capable of pan-cancer WSI understanding

5.3.2. What roles do multi-agent systems play in integrating image analysis with text generation in medical diagnostics, and what are the architectures and strategies that facilitate effective inter-agent collaboration?

Multi-modal and multi-agent systems in medical imaging are described in Table 4 along with the adopted methodology and their associated results. Multi-agent systems provide a promising paradigm for collaborative AI in pathology by orchestrating specialized agents to perform distinct tasks such as image triage, feature extraction, report drafting, and validation [48]. Collaborative AI in pathology shows promise when implemented as a multi-agent system that coordinates agents specializing in different tasks (image triage, feature extraction, report drafting, and validation) as in PathFinder. For example, PathFinder provides a demonstration of how agents can iteratively traverse WSIs and produce interpretable diagnostic explanations that outperform pathologist performance in melanoma classification [103]. Unlike monolithic models, MAS architectures are more scalable, interpretable, and modular to equip more

flexible, more clinically oriented diagnostic workflows [90]. Yet, there are existing challenges in unifying agent-to-agent communication, guaranteeing agent output consistency, and combining agent outputs with current clinical decision-making processes [106].

Table 4: Multi-modal and multi-agent systems in medical imaging.

Ref.	Domain	**Year**	**Methodology**	**Results**
[103]	Pathology Modalities	2025	Multi-agent system with Triage, Navigation, Description, and Diagnosis agents	Outperformed SOTA in melanoma diagnosis by 8%; surpassed pathologists' average by 9%
[45]	Pathology Modalities	2024	Multi-agent collaboration for scalable pathology image-text pair generation	Enabled training of powerful pathology VLMs with improved zero-shot and WSI task performance
[104]	Radiology and Cross-Domain Modalities	2023	Vision-language integration + interactive dialog system with specialized instruction dataset	33% improvement in report correction; strong clinical report generation performance
[105]	Radiology and Cross-Domain Modalities	2025	Knowledge-enhanced dual prompt tuning using LLM-driven domain prompts and vision soft prompts	Classification accuracy improved by 6-8% over prior methods on 11 biomedical datasets
[48]	Radiology and Cross-Domain Modalities	2025	Chain-of-thought prompting with decision-tree system for extracting structured labels from reports	Achieved F1 score 0.89; outperformed human annotators in multi-organ abnormality labeling
[106]	Radiology and Cross-Domain Modalities	2024	Pathological clue-driven cross-modal representation learning with LLM fine-tuning	SoTA performance in brain CT report generation; improved cross-modal consistency
[107]	Pathology Modalities	2024	Multimodal LLM integrating WSIs and histopathology reports with retrieval-based data generation	Effective pan-cancer diagnostic understanding; clinical evaluation acceptance 62.9%

5.4. Clinical Applicability and Practical Considerations

While many DL and multimodal models have demonstrated promising results in controlled research settings, few have reached widespread clinical deployment. However, these fast and accurate models, including EAGLE and RaDialog, are still moving towards real time clinical application, with remaining integration barriers, for example, the use of appropriate computational infrastructures and validation in alignment with diverse patient populations [44, 104]. Furthermore, the datasets in training often lack representation of all the demographic groups which can influence clinical generalizability [45]. This remains a key requirement for AI adoption and interoperability with hospital information systems, EHRs, and the laboratory workflow is still necessary. LEAVS and PathAlign focus on structured label extraction and reporting to be aligned with clinical terminology and therefore easily connected to [48, 53]. While pathology workflows are complex and variable across institutions, therefore, it require adaptable AI frameworks that can be configured to local standards [98].

The challenges related to privacy and the compliance with the regulations are critical issues when it comes to the implementation of the AI systems that utilize patient data. Data anonymization and federated learning solutions develop as potential solutions to overcome privacy concerns without endangering model performance [47]. Secondly, the elements of ethics, such as bias alleviation, transparency, and accountability become necessary when the AI-assisted decisions may be accompanied by the possible clinical effects [53]. Multi-agent frameworks like PathFinder and LEAVS offer explainable reasoning and confidence estimate, in which clinicians can learn to trust and cooperate with the help of human agents in the loop [48, 103]. To ensure that AI results match clinic expectations and workflows, training and engagement of pathologists in the system development is required [45].

5.4.1. Defining Clinical Deployment Readiness

Although a good number of the studies reviewed show outstanding retrospective performance, the transition between a controlled laboratory environment and the clinical real-life environment needs rigorous, verifiable evidence. In support of the fusion of theory with practice, we suggest a standardized “Clinical Deployment Readiness Checklist.“This checklist is a synthesis of the required operational specifics, minimum deployment evidence, and integration specifications that the next generation of computational pathology frameworks needing to be deployment-ready must to report. Table 5 provides deployment readiness checklist for pathology AI in the context of clinical applications.

Table 5: Clinical deployment readiness checklist for pathology AI.

Readiness Criteria	**Verifiable Item / Minimum Evidence Required**	**Rationale for Pathology**
External Validation	Testing on independent, multi-center cohorts completely unseen during training.	Ensures the model generalizes across different hospital populations and laboratory protocols.
Subgroup & Stress Analysis	Performance stratified by WSI scanner type, magnification levels, and staining variations (e.g., H&E vs. IHC).	Pathology images are highly susceptible to color and hardware domain shifts; models must prove robustness to these artifacts.
Calibration Metrics	Reporting Expected Calibration Error (ECE) or Brier scores alongside accuracy.	Clinicians must know if the model's confidence scores accurately reflect the true likelihood of a diagnosis.
Compute & Latency	Explicitly stating VRAM requirements, inference time per gigapixel WSI, and hardware specifications.	Many hospital IT infrastructures cannot support massive GPU clusters; operational latency must fit standard diagnostic workflows.
Integration Touchpoints	Demonstrated compatibility with standard Laboratory Information Systems (LIS) or WSI viewing software (e.g., DICOM-WSI standards).	Standalone, fragmented AI software disrupts pathologist workflows; seamless LIS integration is mandatory for adoption.
Audit Logging & Traceability	Mechanisms to log the specific WSI patch or text prompt that triggered a specific diagnostic output.	Essential for legal compliance, regulatory audits, and retrospective failure analysis if the AI makes an error.

When applying this deployment readiness checklist to the state-of-the-art frameworks reviewed in this study, a clear gap emerges between experimental success and true clinical deployability. Although there is no single system that meets all the requirements, some of the pioneering frameworks show high compliance with certain areas. As an example, in Compute and Latency, the EAGLE framework [44] since deals with hardware constraints, processing time is minimized to about 2.27 seconds per gigapixel WSI. Regarding External Validation and cross-centre testing, traditional foundational models such as TITAN [47] or models pretrained on the OpenPath data [98] have been shown to have strong zero-shot generalization through the dissimilar and hidden cohorts. In the case of Audit Logging and Traceability, an audit trail is a natural byproduct of multi agent pipelines like Pathfinder [103] and SlideSeek [108], where the sequential spatial coordinates and texts at a local position are recorded by the navigation agents. Additionally, systems such as PathAlign [53] have taken their own initiative to inject rigorous pathologist driven reader studies

to the reasonable outcome of their contrivences. There are however critical voids that exist in the literature in general- especially in the areas of Calibration Metrics and organised Integration Touchpoints with Laboratory Information Systems (LIS). Meanwhile, benchmarking systems such as PathBench [111] will be necessary in implementing these multi center, deployment-centered appraisals.

5.5. Challenges

One of the most prominent challenges in the application of AI technologies in medical imaging is the generalizability of models across different clinical environments. CNNs and VLMs, in many cases, are trained on specific data sets that come from some set of particular institutions or imaging protocols. Therefore, these models can work well if used on their training data but have difficulty when used on other institutions' data that come from different equipment, imaging approaches or patient populations [113]. There is a lack of generalization because of variability in the type of scanners used, in the staining procedures for histopathology and in the acquisition protocols. For example, a slide scanner at one hospital might be a different model than a slide scanner at another hospital or one hospital might prepare tissues using a different method than another hospital and this can result in variations in the quality of the images received [46]. Due to a lack of consistency in data across clinical settings, the performance of the AI models degrades, and therefore the AI models become unreliable for deployment to real world settings where the data are more diverse and unpredictable. To address this challenge, we need domain adaptation strategies to train AI models to adapt to the inherent variability of clinical datasets and generalize to different clinical settings such that they are able to provide accurate results across different clinical settings [114].

The interpretability and explainability of AI models are another significant challenge that hampers their acceptance and adoption in clinical practice. CNNs and VLMs models have demonstrated impressive performance, and many clinicians regard these as black boxes. Although the initial methods of explainable AI (XAI) are based on the traditional saliency maps and generic mechanisms of attention, recent studies prove to be untrustworthy in sophisticated medical scenarios. Indicatively, research comparing saliency map techniques such as Grad-ECLIP has shown that conventional saliency maps do not deliver clinically meaningful localization, and more often tend to identify background pixels that are not spurious but actually present embedded pathologies in the image of interest instead of tracking them down and raising their salinity score [115]. The situation is even more complex with multimodal VLMs; to be able to explain anything accurately, token-to-patch grounding must be very accurate. To foster trust, XAI approaches should not be limited to high-level heatmaps whereby a particular generated clinical term (such as nuclear atypia) is anticipated to project to the right cellular structure in the gigapixel WSI in a consistent and reliable manner. Researchers are working on steps to ensure explainability by developing AI explainable (XAI) models such as attention systems and saliency maps used to identify the feature of the image the model used to make a choice [116].

Data scarcity and annotation challenges present another major obstacle in the deployment of AI in medical imaging. WSIs as high-resolution pathology images demand extensive annotations from domain experts, e.g. pathologists,

to effectively train DL models. However, the process of annotating these images is expensive and time-consuming due to the very specialized skill set required for the job and the time it takes to do it [117]. Additionally, there is a dearth of available annotated data and it is very difficult to train strong AI models with the dataset if the disease being studied is rare or the tissue type is unusual. The annotated data is also often not representative of the full range of variability seen in real-world clinical environments, even if there exists some annotated data. For example, data used for training may derive from a limited body of patients or a single institution and the data is not diverse enough. Furthermore, representation of diverse patient populations (e.g., ethnic minorities) is underrepresented further deleterious to this issue since AI models trained only on a certain demographic may appear biased towards that demographic [118]. An absence of diverse datasets can create poor-performing AI systems on populations that are underrepresented and subsequently biased diagnostic outcomes. To make progress on this challenge, larger and more diverse datasets must be created to train simpler models or learn from smaller datasets or unannotated datasets (synthetic data generation and few-shot learning) [119].

In addition to generalizability and data scarcity, the predisposition of VLMs and LLMs to hallucinate remains one of the greatest clinical safety obstacles. Whereas general medical AI hallucinations tend to have anatomical findings of fabricated claims, pathology poses a special and highly organized system of hallucination threat. These pathology-related hallucinations could be generally grouped into three different types:

- **Fabricated Grading and Subtyping:** Generative models can end up giving a false certainty assigned to a given tumor grade (such as Nottingham grading in breast cancer), or histological subtype based on spurious statistical associations instead of actual visual evidence of mitotic counts or nuclear pleomorphism [120].

- **Omission of Prognostic Modifiers:** Models can produce syntactically competent descriptions of a tumor, and be entirely silent with regard to whether or not lymphovascular invasion (LVI), perineural invasion, or whether or not the surgical margins are clear or not, which determine the kind of treatment to be given after surgery [121].

- **Hallucinated Ancillary Testing:** A hallucination of one of the most perilous forms is that a model can create its own immunohistochemical (IHC) or molecular output (e.g., pronouncing a tumor as HER2 positive) by just using standard H&E morphological priors, which are even unnecessary to actually carry out.

Because pathology reports serve as the definitive ground truth for oncology workflows, these hallucinations are not merely technical errors but severe patient safety threats. To reduce these risks, it is necessary to abandon unconstrained and free-text generation. Some effective strategies would be to base the generated text on visual evidence through high-fidelity Multiple Instance Learning (MIL) attention maps, to anchor outputs on tested historical patches with Retrieval-Augmented Generation (RAG) [94] ,[95]to force the model to produce text in standardized synoptic reporting templates (CAP or ICCR guidelines) in order to eliminate the important clinical aspects [99], [100].

5.6. Future Directions

Although the general progress of AI is vast, computational pathology directions in the future need to be prioritized explicitly, taking into account the individual bottlenecks of gigapixels WSI. The large size, context-dependence and staining variability of WSIs demand different operational habits than those of macroscopic radiology. We propose the following prioritized roadmap to translate current AI capabilities into clinical practice:

- Priority 1: High-Fidelity Multimodal Explainability: Going beyond generic, unreliable heatmaps, the future operational models will need to be implemented with VLM-specific XAI methods (like Grad-ECLIP algorithms) to ensure spatial and semantic faithfulness. The most important requirement needed to be granted regulatory approval is to ensure the generated diagnostic text is explicitly anchored on verifiable WSI tissue morphology.
- Priority 2: WSI-Optimized Federated Learning: Due to the scale of pathology WSIs- gigapixels- their standard federated learning protocols between small medical images (such as X-rays) cannot be computed in practice. The future architectures should be made absolutely optimized with the transfer of ultra-high-resolution data and complex domain adaptation of a variety of multi-center laboratory staining protocols.
- Priority 3: Standardizing Multi-Agent Operational Protocols: To translate the multi-agent systems (MAS) into clinical practice, it will be necessary to enforce standard communication and operation protocols. Subsequent studies should specify hard regulatory guardrails of autonomous multi-agent diagnostic loops, the standard format in which specialized agents (e.g., it Triage, Navigation and Description agent) transfer coordinates and hypothesis states without going into a defensive loop or hallucination echo-chambers.

Further studies will be needed to address the problem of generalizability in dissimilar datasets and clinical settings by creating stronger and more flexible AI applications. One such method is called federated learning, which allows collaborative (to be more precise, decentralized) model training on many datasets without requiring the sharing of sensitive patient data of each other, which is a promising research that is currently underway to be implemented in research practice [122]. This achieves the exposure of models to a wide variety of data in various practice settings without loss of patient privacy. Through Federated learning, models can learn on data at (different) institutions without ever having to have a patient transfer among them in the healthcare setting where data privacy is the paramount consideration. Since not every dataset involved in federated learning is generated by the same organization, federated learning may enable AI models to correct the absence of training data on the data of a single organization and achieve higher results on other data sets and platforms [123].

Moreover, data to harmony methods such as color normalization and image resampling may be utilized to reduce any variations related to use of various imaging protocols and enable models to be trained on data of different origins to be compared and implemented in a clinically significant manner across sites. These are critical guidelines on scalability and resilience of AI model in medical imaging to enable them to apply in real world clinical practice application [124].

Another key area for development is the explainability and interpretability of AI systems in medical imaging. As the adoption of AI in clinical decision-making is more completely incorporated, it is urgent that these models do not only provide the correct answer, but also clarify with clear and understandable processes how and why they get to the correct answer in the first place, it is imperative that such models ought not only to provide correct answers, but also to clarify the manner in which they do so, and the reasons behind it [116]. To ensure that the clinicians are comfortable enough to use the model to make their decisions in patient care they must feel at ease with the logic used by the model in making their predictions. This desire has led to the active advancement of XAI methods. The other method is the one that applies attention-based mechanisms, which means that in making the predictions, the AI model paid attention to the parts in the image (regions of an image) that it was interested in. That is, it makes us realize why an approach of models has categorized a region of a tissue this way and why a model is focusing on the right aspects of the image. In addition, the use of XAI with clinical ontologies (i.e. structured sets of clinical terms) will improve the interpretability of the AI systems as they will be explained by an existing body of knowledge and clinical practices that are already known about medical conditions and their treatment [20]. Moreover, such systems can also be made explainable and clinically usable through the human-in-the-loop system architectures, where clinicians can communicate with the AI models to refine the diagnoses and reports, which becomes not only accurate, but also clinically actionable as well, as well as explainable and knowledgeable by humans working with the models [125].

Finally, the issue of data scarcity and bias can be addressed through the development of large-scale, diverse, and representative datasets that include data from various demographic groups and clinical environments. The AI models that are trained today for medical imaging are frequently trained on datasets that are not sufficiently diverse, leading to biased and non-generalizable outcomes for all patient populations [125]. Crowdsourced datasets are being used to gather large amounts of annotated medical data from a wide range of collectives, including diverse patient groups and institutions to combat this. Transfer learning and unsupervised learning techniques may also make it possible to employ similar models in new tasks with less labeled training examples. AI models can be trained to perform well across different clinical environments and patient populations by using domain adaptation techniques that help improve models' generalizability [126]. Additionally, synthetic data generation methods could be leveraged to create artificial data that could be used in conjunction with empirical datasets to cater to the lack of large and underrepresented datasets. Training AI systems on diverse and high-quality datasets, including the appropriate input for the patient is important to ensure they are fair, accurate, and clinically applicable so that patient outcomes are improved [127].

To overcome the safety barriers posed by hallucinations, future research must prioritize rigorous, domain-specific evaluation frameworks and measurable mitigations. NLP measures that are generalized cannot measure the clinical severity of fabricated medical findings. Rather, the profession needs to embrace the use of specialized benchmarking instruments like Med-Hallmark that present hierarchical scoring schemes (e.g. MediHall Score) as means of establishing the precise clinical severity and underlying type of multimodal medical hallucinations. Equally, testing explicitly on multimodal reasoning in microscopy with forms of MicroVQA can give results in which researchers can

systematically isolate and quantify the errors of perception as opposed to errors of reasoning-induced hallucinations. Secondly, it will be needed to develop methods that will ensure general completeness of the facts and cross-modality of structured medical reporting, including the strategies mentioned by Moll et al. With a combination of these rigorous safety standards and grounding method into development lifecycle the future pathology MAS can make sure their outputs are closely bound to the existing visual evidence.

5.6.1. Standardizing LLM-as-a-Judge Rubrics for Pathology

As the evaluation of generative medical AI increasingly relies on "LLM-as-a-judge" frameworks, the lack of standardized rubrics makes cross-study comparisons highly subjective. Future studies must adopt and report a transparent, standardized scoring rubric to ensure reproducibility. We propose that future evaluations report against the following core dimensions:

1. **Factual Accuracy (0-5):** Does the generated text correctly identify the primary diagnosis, histological subtype, and grade present in the ground-truth data?
2. **Clinical Completeness (0-5):** Does the output include all necessary secondary prognostic markers (e.g., margins, lymphovascular invasion, mitotic rate)?
3. **Safety and Hallucination Penalty (-5 to 0):** Are there any fabricated morphological findings or contradictory statements that could lead to patient harm? Severe hallucinations should result in an automatic failure for the prompt.
4. **Reasoning Alignment (0-5):** Is the diagnostic conclusion logically supported by the described visual evidence, mimicking a pathologist's step-by-step deductive process?

6. Conclusion

This systematic literature review has explored the evolving landscape of deep learning, vision-language models (VLMs), large language models (LLMs), and multi-agent systems as applied to pathology image analysis and automated report generation. To address these challenges in high-resolution whole slide images, deep learning approaches have shown remarkable success in performing accurate segmentation, classification, and feature extraction. At the same time, VLMs and LLMs have been integrated to interpolate between visual data and clinical text to generate clinically meaningful diagnostic reports. This integration is also further enhanced by multi-agent systems which organize specialized AI agents to complete complex tasks together, improving diagnostic accuracy, scalability, and interpretability. Although major progress has been made, challenges remain including model generalization, interpretability, data standardization, and ethical issues, that continue to preclude widespread clinical adoption.

Looking ahead, the future of pathology image analysis lies in developing robust, scalable, and clinically aligned hybrid frameworks that leverage the strengths of deep learning, VLMs, LLMs, and multi-agent collaboration. To build trust and move towards integration in real-world clinical workflows, addressing research gaps that these techniques

introduce through the use of larger, more diverse datasets, improved model explainability, and privacy-preserving training methods will be critical. The convergence of these leading-edge AI technologies promises to revolutionize pathology diagnostics through the automation of labor-intensive tasks, reducing human error and hopefully ultimately leading to improved patient outcomes. This review provides a foundational roadmap to guide ongoing research and development toward the realization of fully integrated, AI-driven pathology systems.

Conflicts of Interests

"The authors declare that there is no conflict of interests."

Funding

This study is funded by the European University of Atlantic.

Ethics statement

Not applicable.

Data Availability

Not applicable.

References

[1] S. Ahuja, S. Zaheer, Advancements in pathology: Digital transformation, precision medicine, and beyond, Journal of Pathology Informatics (2024) 100408.

[2] N. Kumar, R. Gupta, S. Gupta, Whole slide imaging (wsi) in pathology: current perspectives and future directions, Journal of digital imaging 33 (4) (2020) 1034–1040.

[3] S. Zia, I. Z. Yildiz-Aktas, F. Zia, A. V. Parwani, An update on applications of digital pathology: primary diagnosis; telepathology, education and research, Diagnostic Pathology 20 (2025) 17.

[4] B. Ilhan, P. Guneri, P. Wilder-Smith, The contribution of artificial intelligence to reducing the diagnostic delay in oral cancer, Oral oncology 116 (2021) 105254.

[5] I. D. Mienye, T. G. Swart, G. Obaido, M. Jordan, P. Ilono, Deep convolutional neural networks in medical image analysis: A review, Information 16 (3) (2025) 195.

[6] X. Jiang, Z. Hu, S. Wang, Y. Zhang, Deep learning for medical image-based cancer diagnosis, Cancers 15 (14) (2023) 3608.

[7] S. Das, S. Gupta, K. Kumar, R. Sharma, Good representation, better explanation: Role of convolutional neural networks in transformer-based remote sensing image captioning, arXiv preprint arXiv:2502.16095 (2025).

[8] C. Dang, Z. Qi, T. Xu, M. Gu, J. Chen, J. Wu, Y. Lin, X. Qi, Deep learning-powered whole slide image analysis in cancer pathology, Laboratory Investigation (2025) 104186.

[9] I. Hartsock, G. Rasool, Vision-language models for medical report generation and visual question answering: A review, Frontiers in Artificial Intelligence 7 (2024) 1430984.

[10] C. Liu, Z. Wan, Y. Wang, H. Shen, H. Wang, K. Zheng, M. Zhang, R. Arcucci, Benchmarking and boosting radiology report generation for 3d high-resolution medical images, arXiv preprint arXiv:2406.07146 (2024).

[11] Y. Bian, J. Li, C. Ye, X. Jia, Q. Yang, Artificial intelligence in medical imaging: From task-specific models to large-scale foundation models, Chinese Medical Journal 138 (06) (2025) 651–663.

[12] E. Shakshuki, M. Reid, Multi-agent system applications in healthcare: current technology and future roadmap, Procedia Computer Science 52 (2015) 252–261.

[13] P. S. Aravazhi, P. Gunasekaran, N. Z. Y. Benjamin, A. Thai, K. K. Chandrasekar, N. D. Kolanu, P. Prajjwal, Y. Tekuru, L. V. Brito, P. Inban, The integration of artificial intelligence into clinical medicine: trends, challenges, and future directions, Disease-a-Month (2025) 101882.

[14] R. Aggarwal, V. Sounderajah, G. Martin, D. S. Ting, A. Karthikesalingam, D. King, H. Ashrafian, A. Darzi, Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis, NPJ digital medicine 4 (1) (2021) 65.

[15] T. Sadad, M. Safran, I. Khan, S. Alfarhood, R. Khan, I. Ashraf, Efficient classification of ecg images using a lightweight cnn with attention module and iot, Sensors 23 (18) (2023) 7697.

[16] A. Alam, A. S. Al-Shamayleh, N. Thalji, A. Raza, E. A. Morales Barajas, E. B. Thompson, I. de la Torre Diez, I. Ashraf, Novel transfer learning based bone fracture detection using radiographic images, BMC Medical Imaging 25 (1) (2025) 5.

[17] M. M. Islam, K. R. Alam, J. Uddin, I. Ashraf, M. A. Samad, Benign and malignant oral lesion image classification using fine-tuned transfer learning techniques, Diagnostics 13 (21) (2023) 3360.

[18] Y. Sun, X. Wen, Y. Zhang, L. Jin, C. Yang, Q. Zhang, M. Jiang, Z. Xu, W. Guo, J. Su, et al., Visual-language foundation models in medical imaging: A systematic review and meta-analysis of diagnostic and analytical applications, Computer Methods and Programs in Biomedicine (2025) 108870.

[19] S. K. Zhou, H. Greenspan, C. Davatzikos, J. S. Duncan, B. Van Ginneken, A. Madabhushi, J. L. Prince, D. Rueckert, R. M. Summers, A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises, Proceedings of the IEEE 109 (5) (2021) 820–838.

[20] S. S. Band, A. Yarahmadi, C.-C. Hsu, M. Biyari, M. Sookhak, R. Ameri, I. Dehzangi, A. T. Chronopoulos, H.-W. Liang, Application of explainable artificial intelligence in medical health: A systematic review of interpretability methods, Informatics in Medicine Unlocked 40 (2023) 101286.

[21] S. Seoni, A. Shahini, K. M. Meiburger, F. Marzola, G. Rotunno, U. R. Acharya, F. Molinari, M. Salvi, All you need is data preparation: A systematic review of image harmonization techniques in multi-center/device studies for medical support systems, Computer Methods and Programs in Biomedicine (2024) 108200.

[22] T. Islam, M. S. Hafiz, J. R. Jim, M. M. Kabir, M. Mridha, A systematic review of deep learning data augmentation in medical imaging: Recent advances and future research directions, Healthcare Analytics (2024) 100340.

[23] F. Hu, A. A. Chen, H. Horng, V. Bashyam, C. Davatzikos, A. Alexander-Bloch, M. Li, H. Shou, T. D. Satterthwaite, M. Yu, et al., Image harmonization: A review of statistical and deep learning methods for removing batch effects and evaluation metrics for effective harmonization, NeuroImage 274 (2023) 120125.

[24] Q. Chen, R. Zhao, S. Wang, V. M. H. Phan, A. v. d. Hengel, J. Verjans, Z. Liao, M.-S. To, Y. Xia, J. Chen, et al., A survey of medical vision-and-language applications and their techniques, arXiv preprint arXiv:2411.12195 (2024).

[25] X. Li, L. Li, Y. Jiang, H. Wang, X. Qiao, T. Feng, H. Luo, Y. Zhao, Vision-language models in medical image analysis: From simple fusion to general large models, Information Fusion (2025) 102995.

[26] Y. Si, J. Du, Z. Li, X. Jiang, T. Miller, F. Wang, W. J. Zheng, K. Roberts, Deep representation learning of patient data from electronic health records (ehr): A systematic review, Journal of biomedical informatics 115 (2021) 103671.

[27] J. Duan, J. Xiong, Y. Li, W. Ding, Deep learning based multimodal biomedical data fusion: An overview and comparative review, Information Fusion (2024) 102536.

[28] F. Krones, U. Marikkar, G. Parsons, A. Szmul, A. Mahdi, Review of multimodal machine learning approaches in healthcare, Information Fusion 114 (2025) 102690.

[29] N. Yadav, S. Pandey, A. Gupta, P. Dudani, S. Gupta, K. Rangarajan, Data privacy in healthcare: In the era of artificial intelligence, Indian Dermatology Online Journal 14 (6) (2023) 788–792.

[30] E. Shakshuki, M. Reid, Multi-agent system applications in healthcare: current technology and future roadmap, Procedia Computer Science 52 (2015) 252–261.

[31] M. G. Hanna, L. Pantanowitz, R. Dash, J. H. Harrison, M. Deebajah, J. Pantanowitz, H. H. Rashidi, Future of artificial intelligence (ai)-machine learning (ml) trends in pathology and medicine, Modern Pathology (2025) 100705.

[32] I. P. Nweke, C. O. Ogadah, K. Koshechkin, P. M. Oluwasegun, Multi-agent ai systems in healthcare: A systematic review enhancing clinical decision-making, Asian Journal of Medical Principles and Clinical Practice 8 (1) (2025) 273–285.

[33] M. T. Bennai, Z. Guessoum, S. Mazouzi, S. Cormier, M. Mezghiche, Multi-agent medical image segmentation: A survey, Computer Methods and Programs in Biomedicine 232 (2023) 107444.

[34] J. Ma, Y. He, F. Li, L. Han, C. You, B. Wang, Segment anything in medical images, Nature Communications 15 (1) (2024) 654.

[35] Q. X. X. W. B. J. X. Z. Y. Z. Y. W. Xuan Liao, Wenhao Li, Iteratively-refined interactive 3d medical image segmentation with multi-agent reinforcement learning, arXiv preprint arXiv:1911.10334 (2019).

[36] N. B. B. A.-K. K. H. A. B. G.-Z. R. D. R. M. Hanane Allioui, Mazin Abed Mohammed, A multi-agent deep reinforcement learning approach for enhancement of covid-19 ct image segmentation, Journal of Personalized Medicine (2022).

[37] S. Maleki Varnosfaderani, M. Forouzanfar, The role of ai in hospitals and clinics: transforming healthcare in the 21st century, Bioengineering 11 (4) (2024) 337.

[38] K. Gal, B. J. Grosz, Multi-agent systems: Technical & ethical challenges of functioning in a mixed group, Daedalus 151 (2) (2022) 114–126.

[39] N. Khalid, A. Qayyum, M. Bilal, A. Al-Fuqaha, J. Qadir, Privacy-preserving artificial intelligence in healthcare: Techniques and applications, Computers in Biology and Medicine 158 (2023) 106848.

[40] X. Tu, Z. He, Y. Huang, Z.-H. Zhang, M. Yang, J. Zhao, An overview of large ai models and their applications, Visual Intelligence 2 (1) (2024) 1–22.

[41] W. M. Bramer, G. B. De Jonge, M. L. Rethlefsen, F. Mast, J. Kleijnen, A systematic approach to searching: an efficient and complete method to develop literature searches, Journal of the Medical Library Association: JMLA 106 (4) (2018) 531.

[42] A. Janowczyk, A. Madabhushi, Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases, Journal of pathology informatics 7 (1) (2016) 29.

[43] S. Kosaraju, J. Park, H. Lee, J. W. Yang, M. Kang, Deep learning-based framework for slide-based histopathological image analysis, Scientific Reports 12 (1) (2022) 19075.

[44] P. Neidlinger, T. Lenz, S. Foersch, C. M. Loeffler, J. Clusmann, M. Gustav, L. A. Shaktah, R. Langer, B. Dislich, L. A. Boardman, et al., A deep learning framework for efficient pathology image analysis, arXiv preprint arXiv:2502.13027 (2025).

[45] Y. Sun, Y. Zhang, Y. Si, C. Zhu, Z. Shui, K. Zhang, J. Li, X. Lyu, T. Lin, L. Yang, Pathgen-1.6 m: 1.6 million pathology image-text pairs generation through multi-agent collaboration, arXiv preprint arXiv:2407.00203 (2024).

[46] J. Shi, C. Li, T. Gong, Y. Zheng, H. Fu, Vila-mil: Dual-scale vision-language multiple instance learning for whole slide image classification, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 11248–11258.

[47] T. Ding, S. J. Wagner, A. H. Song, R. J. Chen, M. Y. Lu, A. Zhang, A. J. Vaidya, G. Jaume, M. Shaban, A. Kim, et al., Multimodal whole slide foundation model for pathology, arXiv preprint arXiv:2411.19666 (2024).

[48] R. B. Lanfredi, Y. Zhuang, M. Finkelstein, P. T. S. Balamuralikrishna, L. Krembs, B. Khoury, A. Reddy, P. Mukherjee, N. M. Rofsky, R. M. Summers, Leavs: An llm-based labeler for abdominal ct supervision, arXiv preprint arXiv:2503.13330 (2025).

[49] M. Li, X. Hou, Z. Liu, D. Yang, Z. Qian, J. Chen, J. Wei, Y. Jiang, Q. Xu, L. Zhang, Mccd: Multi-agent collaboration-based compositional diffusion for complex text-to-image generation, in: Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 13263–13272.

[50] W. Ikezogwo, S. Seyfioglu, F. Ghezloo, D. Geva, F. Sheikh Mohammed, P. K. Anand, R. Krishna, L. Shapiro, Quilt-1m: One million image-text pairs for histopathology, Advances in neural information processing systems 36 (2023) 37995–38017.

[51] A. B. Sai, A. K. Mohankumar, M. M. Khapra, A survey of evaluation metrics used for nlg systems, ACM Computing Surveys (CSUR) 55 (2) (2022) 1–39.

[52] A. K. M. Ananya B. Sai, M. M. Khapra, A survey of evaluation metrics used for nlg systems, ACM Computing Surveys (2022).

[53] F. Ahmed, A. Sellergren, L. Yang, S. Xu, B. Babenko, A. Ward, N. Olson, A. Mohtashamian, Y. Matias, G. S. Corrado, et al., Pathalign: A vision-language model for whole slide images in histopathology, arXiv preprint arXiv:2406.19578 (2024).

[54] E. K. S. H. L. S. H. L. W.-K. J. Jing Wei Tan, SeungKyu Kim, Clinical-grade multi-organ pathology report generation for multi-scale whole slide images via a semantically guided medical text foundation model, arXiv preprint arXiv:2409.15574 (2024).

[55] L. G. A. M. V. W. K. S. K. J. B. E. B. V. E. R. D. S. K. T. J. F. Gabriele Campanella, Matthew G Hanna, Clinical-grade computational pathology using weakly supervised deep learning on whole slide images, Nat Med (2019).

[56] S. I. Daisuke Komura, Mieko Ochi, Machine learning methods for histopathological image analysis: Updates in 2024, Computational and Structural Biotechnology Journal (2024).

[57] S. Mehrvar, L. E. Himmel, P. Babburi, A. L. Goldberg, M. Guffroy, K. Janardhan, A. L. Krempley, B. Bawa, Deep learning approaches and applications in toxicologic histopathology: current status and future perspectives, Journal of Pathology Informatics 12 (1) (2021) 42.

[58] N. Dimitriou, O. Arandjelović, P. D. Caie, Deep learning for whole slide image analysis: an overview, Frontiers in medicine 6 (2019) 264.

[59] T. Huang, X. Huang, H. Yin, Deep learning methods for improving the accuracy and efficiency of pathological image analysis, Science Progress 108 (1) (2025) 00368504241306830.

[60] M. S. Hossain, L. J. Armstrong, D. M. Cook, P. Zaenker, Application of histopathology image analysis using deep learning networks, Human-Centric Intelligent Systems 4 (3) (2024) 417–436.

[61] X. Zhang, F. O. Gleber-Netto, S. Wang, R. R. Martins-Chaves, R. S. Gomez, N. Vigneswaran, A. Sarkar, W. N. William Jr, V. Papadimitrakopoulou, M. Williams, et al., Deep learning-based pathology image analysis predicts cancer progression risk in patients with oral leukoplakia, Cancer medicine 12 (6) (2023) 7508–7518.

[62] A. H. Song, R. J. Chen, T. Ding, D. F. Williamson, G. Jaume, F. Mahmood, Morphological prototyping for unsupervised slide representation learning in computational pathology, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 11566–11578.

[63] A. L. Meirelles, T. Kurc, J. Kong, R. Ferreira, J. Saltz, G. Teodoro, Effective and efficient active learning for deep learning-based tissue image analysis, Bioinformatics 39 (4) (2023) btad138.

[64] M. Y. Lu, D. F. Williamson, T. Y. Chen, R. J. Chen, M. Barbieri, F. Mahmood, Data-efficient and weakly supervised computational pathology on whole-slide images, Nature biomedical engineering 5 (6) (2021) 555–570.

[65] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al., Llama: Open and efficient foundation language models, arXiv preprint arXiv:2302.13971 (2023).

[66] C. Wu, W. Lin, X. Zhang, Y. Zhang, W. Xie, Y. Wang, Pmc-llama: toward building open-source language models for medicine, Journal of the American Medical Informatics Association 31 (9) (2024) 1833–1843.

[67] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, et al., Large language models encode clinical knowledge, Nature 620 (7972) (2023) 172–180.

[68] K. Singhal, T. Tu, J. Gottweis, R. Sayres, E. Wulczyn, M. Amin, L. Hou, K. Clark, S. R. Pfohl, H. Cole-Lewis, et al., Toward expert-level medical question answering with large language models, Nature Medicine (2025) 1–8.

[69] G. Nanayakkara, N. Wiratunga, D. Corsar, K. Martin, A. Wijekoon, Clinical dialogue transcription error correction using seq2seq models, in: Multimodal AI in healthcare: A paradigm shift in health intelligence, Springer, 2022, pp. 41–57.

[70] D. Duong, B. D. Solomon, Analysis of large-language model versus human performance for genetics questions, European Journal of Human Genetics 32 (4) (2024) 466–468.

[71] Y. Li, Z. Li, K. Zhang, R. Dan, S. Jiang, Y. Zhang, Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge, Cureus 15 (6) (2023).

[72] A. Toma, P. R. Lawler, J. Ba, R. G. Krishnan, B. B. Rubin, B. Wang, Clinical camel: An open-source expert-level medical language model

with dialogue-based knowledge encoding, CoRR (2023).

[73] H. Xiong, S. Wang, Y. Zhu, Z. Zhao, Y. Liu, L. Huang, Q. Wang, D. Shen, Doctorglm: Fine-tuning your chinese doctor is not a herculean task, arXiv preprint arXiv:2304.01097 (2023).

[74] S. Yang, H. Zhao, S. Zhu, G. Zhou, H. Xu, Y. Jia, H. Zan, Zhongjing: Enhancing the chinese medical capabilities of large language model through expert feedback and real-world multi-turn dialogue, in: Proceedings of the AAAI conference on artificial intelligence, Vol. 38, 2024, pp. 19368–19376.

[75] Q. Yang, R. Wang, J. Chen, R. Su, T. Tan, Fine-tuning medical language models for enhanced long-contextual understanding and domain expertise, arXiv preprint arXiv:2407.11536 (2024).

[76] H. Wang, C. Liu, N. Xi, Z. Qiang, S. Zhao, B. Qin, T. Liu, Huatuo: Tuning llama model with chinese medical knowledge, arXiv preprint arXiv:2304.06975 (2023).

[77] M. Cascella, J. Montomoli, V. Bellini, E. Bignami, Evaluating the feasibility of chatgpt in healthcare: an analysis of multiple clinical and research scenarios, Journal of medical systems 47 (1) (2023) 33.

[78] S. R. Ali, T. D. Dobbs, H. A. Hutchings, I. S. Whitaker, Using chatgpt to write patient clinic letters, The Lancet Digital Health 5 (4) (2023) e179–e181.

[79] E. Waisberg, J. Ong, M. Masalkhi, S. A. Kamran, N. Zaman, P. Sarker, A. G. Lee, A. Tavakkoli, Gpt-4: a new era of artificial intelligence in medicine, Irish Journal of Medical Science (1971-) 192 (6) (2023) 3197–3200.

[80] A. M. Abdelhady, C. R. Davis, Plastic surgery and artificial intelligence: how chatgpt improved operation note accuracy, time, and education, Mayo Clinic Proceedings: Digital Health 1 (3) (2023) 299–308.

[81] C. A. Mallio, C. Bernetti, A. C. Sertorio, B. B. Zobel, Chatgpt in radiology structured reporting: analysis of chatgpt-3.5 turbo and gpt-4 in reducing word count and recalling findings, Quantitative Imaging in Medicine and Surgery 14 (2) (2024) 2096.

[82] L. C. Adams, D. Truhn, F. Busch, A. Kader, S. M. Niehues, M. R. Makowski, K. K. Bressem, Leveraging gpt-4 for post hoc transformation of free-text radiology reports into structured reporting: a multilingual feasibility study, Radiology 307 (4) (2023) e230725.

[83] W. A. Bosbach, J. F. Senge, B. Nemeth, S. H. Omar, M. Mitrakovic, C. Beisbart, A. Horváth, J. Heverhagen, K. Daneshvar, Ability of chatgpt to generate competent radiology reports for distal radius fracture by use of rsna template items and integrated ao classifier, Current problems in diagnostic radiology 53 (1) (2024) 102–110.

[84] Z. Wang, R. Guo, P. Sun, L. Qian, X. Hu, Enhancing diagnostic accuracy and efficiency with gpt-4-generated structured reports: a comprehensive study, Journal of Medical and Biological Engineering 44 (1) (2024) 144–153.

[85] H. Jiang, S. Xia, Y. Yang, J. Xu, Q. Hua, Z. Mei, Y. Hou, M. Wei, L. Lai, N. Li, et al., Transforming free-text radiology reports into structured reports using chatgpt: A study on thyroid ultrasonography, European Journal of Radiology 175 (2024) 111458.

[86] H. Li, H. Wang, X. Sun, H. He, J. Feng, Prompt-guided generation of structured chest x-ray report using a pre-trained llm, in: 2024 IEEE International Conference on Multimedia and Expo (ICME), IEEE, 2024, pp. 1–6.

[87] Y. Pan, J. Fang, C. Zhu, M. Li, H. Wu, towards an automatic transformer to fhir structured radiology report via gpt-4, Available at SSRN 4717860 (2024).

[88] J. H. Moon, H. Lee, W. Shin, Y.-H. Kim, E. Choi, Multi-modal understanding and generation for medical images and text via vision-language pre-training, IEEE Journal of Biomedical and Health Informatics 26 (12) (2022) 6070–6080.

[89] S. Eslami, C. Meinel, G. De Melo, Pubmedclip: How much does clip benefit visual question answering in the medical domain?, in: Findings of the Association for Computational Linguistics: EACL 2023, 2023, pp. 1181–1193.

[90] A. K. Tanwani, J. Barral, D. Freedman, Repsnet: Combining vision with language for automated medical reports, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2022, pp. 714–724.

[91] S. Zhang, Y. Xu, N. Usuyama, H. Xu, J. Bagga, R. Tinn, S. Preston, R. Rao, M. Wei, N. Valluri, et al., Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs, arXiv preprint arXiv:2303.00915 (2023).

[92] H. Lee, D. Y. Lee, W. Kim, J.-H. Kim, T. Kim, J. Kim, L. Sunwoo, E. Choi, Vision-language generative model for view-specific chest x-ray generation, arXiv preprint arXiv:2302.12172 (2023).

[93] Z. Yuan, Q. Jin, C. Tan, Z. Zhao, H. Yuan, F. Huang, S. Huang, Ramm: Retrieval-augmented biomedical visual question answering with multi-modal pre-training, in: Proceedings of the 31st ACM International Conference on Multimedia, 2023, pp. 547–556.

[94] J. Jeong, K. Tian, A. Li, S. Hartung, S. Adithan, F. Behzadi, J. Calle, D. Osayande, M. Pohlen, P. Rajpurkar, Multimodal image-text matching improves retrieval-based chest x-ray report generation, in: Medical Imaging with Deep Learning, PMLR, 2024, pp. 978–990.

[95] M. Ranjit, G. Ganapathy, R. Manuel, T. Ganu, Retrieval augmented chest x-ray report generation using openai gpt models, in: Machine Learning for Healthcare Conference, PMLR, 2023, pp. 650–666.

[96] C. Li, C. Wong, S. Zhang, N. Usuyama, H. Liu, J. Yang, T. Naumann, H. Poon, J. Gao, Llava-med: Training a large language-and-vision assistant for biomedicine in one day, Advances in Neural Information Processing Systems 36 (2023) 28541–28564.

[97] D. Dai, Y. Zhang, Q. Yang, L. Xu, X. Shen, S. Xia, G. Wang, Pathologyvlm: a large vision-language model for pathology image understanding, Artificial Intelligence Review 58 (6) (2025) 1–19.

[98] Z. Huang, F. Bianchi, M. Yuksekgonul, T. J. Montine, J. Zou, A visual–language foundation model for pathology image analysis using medical twitter, Nature medicine 29 (9) (2023) 2307–2316.

[99] G. W. Matthew Yap, Ioana-Maria Mihai, Machine learning in biomarker-driven precision oncology: Automated immunohistochemistry scoring and emerging directions in genitourinary cancers, Current Oncology (2026).

[100] B. A. R. G. B. P. A. S. S. D. T. R. W. S. C. Jieun Hwang, Alexander K Goel, Building a standardized cancer synoptic report with semantic and syntactic interoperability: Development study using snomed ct and fast healthcare interoperability resources (fhir), JMIR Medical Informatics (2025).

[101] L. Pinto-Coelho, How artificial intelligence is shaping medical imaging technology: a survey of innovations and applications, Bioengineering 10 (12) (2023) 1435.

[102] N. Karunanayake, Next-generation agentic ai for transforming healthcare, Informatics and Health 2 (2) (2025) 73–83.

[103] F. Ghezloo, M. S. Seyfioglu, R. Soraki, W. O. Ikezogwo, B. Li, T. Vivekanandan, J. G. Elmore, R. Krishna, L. Shapiro, Pathfinder: A multi-modal multi-agent system for medical diagnostic decision-making applied to histopathology, arXiv preprint arXiv:2502.08916 (2025).

[104] C. Pellegrini, E. Özsoy, B. Busam, N. Navab, M. Keicher, Radialog: A large vision-language model for radiology report generation and conversational assistance, arXiv preprint arXiv:2311.18681 (2023).

[105] W. Peng, K. Liu, J. Hu, M. Zhang, Biomed-dpt: Dual modality prompt tuning for biomedical vision-language models, arXiv preprint arXiv:2505.05189 (2025).

[106] C. Zheng, J. Ji, Y. Shi, X. Zhang, L. Qu, See detail say clear: Towards brain ct report generation via pathological clue-driven representation learning, arXiv preprint arXiv:2409.19676 (2024).

[107] S. Kim, S. Lee, J. Jang, Chatexaonepath: An expert-level multimodal large language model for histopathology using whole slide images, arXiv preprint arXiv:2504.13023 (2025).

[108] D. F. K. W. R. J. C. T. D. B. C. A. V. L. P. L. Chengkuan Chen, Luca L. Weishaupt, A clinical-grade agentic and generative ai-driven copilot for human pathology, arXiv preprint (2025).

[109] Y. J. Y. L. J. Y. T. L. M. H. R. Y. Y. Q. J. H. Ying Chen, Guoan Wang, Slidechat: A large vision-language assistant for whole-slide pathology image understanding, arXiv preprint arXiv:2410.11761 (2024).

[110] C. Z. S. Z. Q. C. K. Z. Y. Z. D. W. X. L. M. Z. J. L. X. L. T. L. L. Y. Yuxuan Sun, Hao Wu, Pathmmu: A massive multimodal expert-level benchmark for understanding and reasoning in pathology, arXiv preprint arXiv:2401.16355 (2024).

[111] F. Z. Y. W. C. J. Z. G. J. W. O. K. T. H. Z. X. W. L. L. Z. Z. D. C. Z. G. W. W. Y. L. J. H. J. C. Z. L. J. Z. F. G. X. Z. L. L. R. C. K. C. Z. W. H. C. Jiabo Ma, Yingxue Xu, Pathbench: A comprehensive comparison benchmark for pathology foundation models towards precision oncology, arXiv preprint arXiv:2505.20202 (2024).

[112] C. A. G. Steven N Hart, Patrick L Day, Streamlining medical software development with care lifecycle and care agent: an ai-driven technology readiness level assessment tool, BMC Med Informatics Decis Making (2025).

[113] A. S. Panayides, A. Amini, N. D. Filipovic, A. Sharma, S. A. Tsaftaris, A. Young, D. Foran, N. Do, S. Golemati, T. Kurc, et al., Ai in medical imaging informatics: current challenges and future directions, IEEE journal of biomedical and health informatics 24 (7) (2020)

1837–1857.

[114] C. Mennella, U. Maniscalco, G. De Pietro, M. Esposito, Ethical and regulatory challenges of ai technologies in healthcare: A narrative review, Heliyon 10 (4) (2024).

[115] J. H. H. A. B. C. Chenyang Zhao, Kun Wang, Grad-eclip: Gradient-based visual and textual explanations for clip, arXiv preprint arXiv:2502.18816 (2025).

[116] Z. Sadeghi, R. Alizadehsani, M. A. Cifci, S. Kausar, R. Rehman, P. Mahanta, P. K. Bora, A. Almasri, R. S. Alkhawaldeh, S. Hussain, et al., A review of explainable artificial intelligence in healthcare, Computers and Electrical Engineering 118 (2024) 109370.

[117] A. Obeid, S. Boumaraf, A. Sohail, T. Hassan, S. Javed, J. Dias, M. Bennamoun, N. Werghi, Advancing histopathology with deep learning under data scarcity: A decade in review, arXiv preprint arXiv:2410.19820 (2024).

[118] A. Waqas, M. M. Bui, E. F. Glassy, I. El Naqa, P. Borkowski, A. A. Borkowski, G. Rasool, Revolutionizing digital pathology with the power of generative artificial intelligence and foundation models, Laboratory investigation 103 (11) (2023) 100255.

[119] J. L. Cross, M. A. Choma, J. A. Onofrey, Bias in medical ai: Implications for clinical decision-making, PLOS Digital Health 3 (11) (2024) e0000651.

[120] Y. Y. D. A. S. J. K. M.-G. N. S. L. Jun Seo Kim, Jeong Hoon Lee, Predicting nottingham grade in breast cancer digital pathology using a foundation model, Breast Cancer Research (2025).

[121] T. Y. K. M. Y. O. Y. O. S. W.-N. H. T. T. N. N. . M. M. Fumihiko Kinoshita, Tomoyoshi Takenaka, Development of artificial intelligence prognostic model for surgically resected non-small cell lung cancer, Scientific Reports (2023).

[122] M. Li, P. Xu, J. Hu, Z. Tang, G. Yang, From challenges and pitfalls to recommendations and opportunities: Implementing federated learning in healthcare, Medical Image Analysis (2025) 103497.

[123] M. E. van Genderen, M. Cecconi, C. Jung, Federated data access and federated learning: improved data sharing, ai model development, and learning in intensive care, Intensive Care Medicine 50 (6) (2024) 974–977.

[124] Y. Nan, J. Del Ser, S. Walsh, C. Schönlieb, M. Roberts, I. Selby, K. Howard, J. Owen, J. Neville, J. Guiot, et al., Data harmonisation for information fusion in digital healthcare: A state-of-the-art systematic review, meta-analysis and future research directions, Information Fusion 82 (2022) 99–122.

[125] G. Abgrall, A. L. Holder, Z. Chelly Dagdia, K. Zeitouni, X. Monnet, Should ai models be explainable to clinicians?, Critical Care 28 (1) (2024) 301.

[126] J. R. Wilson, L. M. Prevedello, C. D. Witiw, A. E. Flanders, E. Colak, Data liberation and crowdsourcing in medical research: The intersection of collective and artificial intelligence, Radiology: Artificial Intelligence 6 (1) (2023) e230006.

[127] V. C. Pezoulas, D. I. Zaridis, E. Mylona, C. Androutsos, K. Apostolidis, N. S. Tachos, D. I. Fotiadis, Synthetic data generation methods in healthcare: A review on open-source tools and methods, Computational and structural biotechnology journal (2024).

Computational and Structural Biotechnology Journal

A Systematic Literature Review on Integrated Deep Learning and Multi-Agent Vision-Language Frameworks for Pathology Image Analysis and Report Generation

Usama Ali, Imran Shafi, Jamil Ahmad, Arlette Zarate Caceres, Thania Candelaria Chio Montero, Hafiz Muhammad Raza ur Rehman, and Imran Ashraf

Citation: Ali U, Shafi I, Ahmad J, Caceres A, Chio Montero T, Raza ur Rehman H, Ashraf I. A Systematic Literature Review on Integrated Deep Learning and Multi-Agent Vision-Language Frameworks for Pathology Image Analysis and Report Generation. *Comput Struct Biotechnol J*.**Just Accepted Manuscript** DOI: 10.34133/csbj.0023

Computational and Structural Biotechnology Journal (ISSN) is published by the American Association for the Advancement of Science. 1200 New York Avenue NW, Washington, DC 20005.