

Self-Directing Diagnostic Agents for Adaptive Multi-Modal Medical Imaging with Clinician-Guided Reasoning

Ahmed I. Taloba¹, Omar Mohammad Amer Huwari^{2,*}, Fuad S. Al-Duais³, and Mohammad Kanan^{4,5}

¹Department of Information System, Faculty of Computers and Information, Assiut University, Assiut, Egypt

²Department of English language & Literature, College of Science and Humanities Studies in Al-Kharj, Prince Sattam Bin Abdulaziz University 11942 Al-Kharj, Saudi Arabia

³Department of Mathematics, College of Science and Humanities in Al-Kharj, Prince Sattam Bin Abdulaziz University 11942 Al-Kharj, Saudi Arabia

⁴Department of Industrial Engineering, College of Engineering, University of Business and Technology, Jeddah 21448, Saudi Arabia

⁵Department of Mechanical Engineering, College of Engineering, Zarqa University, Zarqa, Jordan

Received: 30 Mar. 2026, Revised: 15 May. 2026, Accepted: 20 May. 2026

Published online: 1 Jul. 2026

Abstract: Medical imaging has become an essential component of modern healthcare for disease diagnosis, lesion localization, and clinical decision support across modalities such as chest X-ray, computed tomography (CT), and histopathology imaging. Although recent deep learning approaches have achieved promising performance, most existing systems rely on fixed and non-adaptive pipelines that cannot dynamically respond to clinical uncertainty, patient variability, or real-time expert feedback. Furthermore, current agentic AI frameworks lack unified dynamic workflow orchestration, uncertainty-aware reasoning, and clinician-guided adaptive refinement mechanisms. To address these limitations, this study proposes a novel Self-Directing Diagnostic Agent with Dynamic Directed Acyclic Graph reasoning (SDDA-DAG) framework for adaptive multi-modal medical image analysis. The proposed method employs a LLaMA-370B Instruct-based self-directing agent to autonomously generate diagnostic workflows and dynamically construct DAG-based computational pipelines integrating DenseNet-121 CNN, Attention U-Net, and Vision Transformer modules. The framework incorporates Monte Carlo Dropout-based uncertainty estimation, retrieval-augmented clinical validation, clinician-in-the-loop refinement, and cost-aware optimization for reliable and clinically interpretable diagnosis. The proposed framework was implemented using Python, PyTorch, MONAI, Hugging Face Transformers, and CUDA acceleration. Experiments were conducted on NIH ChestX-ray14, PhysioNet CT-ICH, and LNCO2 histopathology datasets. The proposed SDDA-DAG framework achieved average diagnostic accuracy of 93.1%, F1-score of 0.93, and AUC of 0.95, outperforming existing agentic AI models by approximately 17.6% accuracy improvement over GPT-4o-agent and more than 40% improvement over several vision-language baselines. The results demonstrate that adaptive DAG orchestration, uncertainty-aware refinement, and clinician-guided reasoning significantly improve diagnostic reliability, calibration stability, and clinical trustworthiness for next-generation intelligent healthcare systems. .

Keywords: Agentic AI; Medical Imaging; Autonomous Diagnosis; Clinical Decision Support; Self-Directed Learning.

1 Introduction

Medical imaging is a crucial part of the modern healthcare system, and allows the detection, diagnosis and monitoring of diseases from various modalities including chest X-ray, computed tomography (CT) and histopathology whole-slide images [1]. In clinical decision making, these imaging technologies are essential in areas where they are most critical like the diagnosis of

respiratory disease, oncology and neurology [2], [3]. With the rapid growth of digital healthcare systems, artificial intelligence has become an essential tool for analyzing large-scale medical imaging data, offering significant improvements in speed, consistency, and diagnostic accuracy [4]. Local visual patterns can be well captured by deep learning models like CNN and precise medical image segmentation can be achieved by U-Net

* Corresponding author e-mail: o.huwari@psau.edu.sa

models [5]. Likewise, recent progress in the field is made with Vision Transformers (ViTs) that enable reasoning across the whole image from a complex imaging modality [6]. However, the current AI solutions in health care are largely limited to a fixed pipeline of models that are applied consistently to every case, regardless of the variation of the patient or complexity of the diagnosis [7], [8].

In recent years, agentic Artificial Intelligence and Self-directing AI have also given rise to a new paradigm where AI models are expected to make autonomous decisions regarding their computational workflows, adapting and optimizing them [9]. These systems seek to go beyond prediction to decision making. In medical imaging, however, such abilities are still restricted [10]. The existing solutions cannot dynamically change their inference strategy according to intermediate results or uncertainty signals [11]. Moreover, the integration of clinicians is also limited, with most systems operating separately from any real-time expert input into the decision loop [12]. This leads to reduced interpretability and limited clinical trust. Moreover, uncertainty handling is not generally considered as one of the core drivers of model reasoning, and this has led to less reliable diagnostics in critical situations [13], [14].

To overcome these drawbacks, this study introduces a medical imaging analysis framework, called Self-Directing Diagnostic Agent framework (SDDA-DAG), which is a fundamentally adaptive architecture. The proposed approach involves using a large language model to generate diagnostic strategies for each task and dynamically building a Directed Acyclic Graph (DAG) that orchestrates the deep learning modules (CNNs, U-Nets and Vision Transformers) depending on the complexity of the case. It includes uncertainty-aware decision-making, cost-sensitive optimization and real-time clinician feedback to incrementally improve the process of diagnosis. The proposed framework is dynamic, allowing for autonomous planning, adaptive execution, and clinically guided refinement, moving medical imaging beyond static inference systems to intelligent, self-directed diagnostic agents.

1.1 Problem Statement

Existing medical imaging AI systems are mostly based on fixed and predefined pipelines, in which the AI models are based on a fixed sequence of steps, including feature extraction, preprocessing, and classification [15]. These designs are not adaptable to inter-patient variability and multimodal complexity, resulting in lower robustness in the real clinical setting [16]. However, existing solutions still suffer from static or semi-static pipeline construction with workflows being pre-defined or partially adaptable, which does not allow for full autonomy in complex clinical variability [17]. In addition, clinician interaction is limited to post-hoc feedback, not real-time guidance,

limiting adaptability during the decision-making process. There are some uncertainties handling and memory-based refinement techniques, but they are not fully integrated in the unified decision mechanism. Therefore, there are clear gaps in the research to create fully self-directed diagnostic agents that can generate dynamically a workflow based on DAGs, real-time clinician-guided refinement, and uncertainty-driven adaptive reasoning. The proposed SDDA-DAG overcomes these drawbacks by making workflow planning, optimization, and clinically constrained decision validation autonomous.

1.2 Research Motivation

In the field of medical imaging, CNNs, Vision Transformers, and U-Net architectures have also made significant strides, enhancing the accuracy of diagnosis in radiology and pathology. Most of the current systems are however, fixed pipelines that are not adaptable to patient variability and clinical uncertainty. With the recent emergence of agentic AI, there are new opportunities for autonomous reasoning, but existing AI systems still rely on a rigid workflow structure and minimal clinician interaction when making an inference. The restrictions motivate the development of a self-directing diagnostic system, which can dynamically plan, execute, and improve medical image analysis workflows. The aim is to obtain intelligent, adaptive, and clinically-oriented decision making.

1.3 Research Significance

The research is significant because it has fundamentally changed the approach to medical AI pipelines from being static to being adaptive, self-directing, and dynamic, incorporating uncertainty-aware reasoning and real-time clinician feedback. The proposed approach stands in contrast to the existing agentic systems, which are able to continuously optimize diagnostic strategies that can be executed using a DAG, and memory-guided optimization. This improves the reliability, interpretability and clinical safety for medical image analysis. In addition, the framework aligns automated AI reasoning with the expert human judgment needed for clinical decision making, thereby filling the gap. Finally, it helps to realize next-generation intelligent healthcare systems that are scalable, transparent and flexible to complex medical scenarios.

1.4 Key Contribution

1. Proposes a self-directing diagnostic agent framework that autonomously plans and refines medical imaging workflows using dynamic DAG-based reasoning.

2. Introduces an uncertainty-driven adaptive decision mechanism that guides model execution and triggers refinement only when prediction confidence is low.
3. Develops a cost-aware optimization strategy that balances diagnostic accuracy with computational efficiency for scalable clinical deployment.
4. Integrates retrieval-augmented clinical knowledge to ensure guideline-consistent and medically reliable diagnostic outputs.
5. Enables real-time clinician-in-the-loop feedback within the inference cycle, improving adaptability, safety, and continuous learning capability of the system.

1.5 Structure of the paper

The rest of the paper is structured as follows. Section 2 provides a detailed review of existing literature. Section 3 introduces the proposed framework and methodology. Section 4 presents the experimental results along with a thorough analysis. Section 5 concludes the study and highlights future research directions.

2 Related Works

Oh et al. [18] propose an analytical comparison between CNNs and ViT models for medical imaging with a focus on explainability. The work proposed tests for robustness to noise, translation invariance, occlusion processing, and for shape vs. texture representation to gain insight on the origins of the performance. Results show that ViT is able to learn global dependencies well, whereas CNNs are good at learning local features, and medical images show class-sensitive behavior compared to generic images; an ensemble of both models shows further improvement. The disadvantages are the lack of quantitative validation metrics, lack of real-time workflow integration and limited applicability to real-time diagnostic systems or multi-modal imaging scenarios.

Mzoughi et al. [19] present a Vision Transformer-based CAD system for multi-class brain tumor classification on T1-weighted contrast-enhanced MRI, leveraging the benefits of XAI methods such as Grad-CAM, LIME, and SHAP. Results revealed that ViT exhibited a higher performance (91.61%) than CNN (83.37%) with better global feature learning and interpretability that will aid clinical decisions. This method will improve the transparency and confidence in medical imaging diagnostics. Some of the limitations are single modality imaging, no multi-modal integration, limited diversity of datasets, and no adaptive or clinician-guided reasoning for dynamic diagnostic workflows.

Pan et al. [20] propose a 2D medical image synthesis framework based on a diffusion model using a

Swin-transformer network to address limited dataset availability. The method uses both forward and reverse Gaussian noise and denoising processes on both chest X-ray and MRI and CT datasets. The results show that classification accuracy is improved with mixed datasets (0.93) than with real (0.88) and synthetic (0.89) datasets using the following evaluation metrics: IS (2.28), FID (37.27), FDS (0.20), and DS (0.86). Realism is confirmed with visual Turing assessment. The limitations are moderate visual fidelity, limited clinical validation, lack of adaptive reasoning and lack of integration with downstream diagnostic decision-making systems.

Karunanayake [21] discusses agentic AI as a next-generation paradigm in healthcare, highlighting how it is autonomous, adaptable, scalable, and probabilistic, for multi-modal data integration. The conceptual framework proposed focuses on applications for diagnostics, decision support, treatment planning, and patient monitoring, allowing for context-aware and patient-centric care. Results indicate that iterative refinement and knowledge integration can lead to processes that are more efficient, lower error rates, and better clinical workflows. However, there are some limitations, such as the absence of experimental validation, unclear implementation details, and ethical, data privacy, regulatory compliance and real-world deployment issues.

Tzanis and Klontzas [22] suggest the design of an autonomous multi-agent system for end-to-end medical imaging AI workflows, comprising of agents for exploratory data analysis, radiomic feature extraction and model development, each of which performs a specific task. It is able to reason with multimodal data, select tools and execute workflows using natural language prompts. Findings show successful task completion and outputs are comparable to manually executed tasks, thus exhibiting robustness and reproducing the results. However, there are limitations such as reliance on high-end large language models, no clinician-in-the-loop reasoning, limited clinical validation in real time, and no adaptive feedback mechanisms for dynamic workflow refinement in complex diagnostic environments.

Wang et al. [23] introduce a multi-agent deep Q-network reinforcement learning approach for 2D medical image object detection using multi-scale image representation. The approach uses multiple agents to iteratively localise the boundaries of objects, and boosts object detection efficiency without candidate boxes. Results demonstrate that performance is better on the 2.45 increase in intersection over union and 0.12 s less detection time than single-scale approaches. The limitations are complex reinforcement learning, limited interpretability, task-specific design, and limited generalizability to multi-modal imaging, and integration with wider diagnostic reasoning frameworks.

Allioui et al. [24] propose a multi-agent deep reinforcement learning (DRL)-based mask extraction method using a modified Deep Q-Network for medical

image segmentation on COVID-19 CT images. The approach enables cooperative agents to automatically select optimal masks, reducing manual effort and improving diagnostic efficiency. Findings report strong performance with precision of 97.12%, Dice of 80.81%, sensitivity of 79.97%, specificity of 99.48%, F1 score of 83.01%, and MAE of 0.86, with segmentation closely matching ground truth. Limitations include task-specific design, reliance on CT modality, limited generalization, and lack of integration with multi-modal or adaptive diagnostic reasoning frameworks.

Lyu et al. [25] propose Whole slide images-Agents, a collaborative multi-agent system for multi-modal whole slide image analysis, integrating task allocation, verification, and summary modules using patch-level and Whole slide images-level multi-modal large language models. The system assigns specific agents to specific tasks, validates the outputs using internal and external processes, and produces understandable summaries as well as visual maps. Results show that the model outperforms current multi-modal large language models and medical agent frameworks in various tasks of interest in pathology. Limitations are computationally expensive, rely on large scale models and knowledge bases, are not yet well suited for real-time clinical deployment and they do not include any adaptive feedback mechanisms that are guided by a clinician and optimize dynamic workflows.

Wang et al. [26] introduce MedAgent-Pro, an agentic reasoning framework for medical diagnosis that combines Vision-Language Models (VLMs) with step-by-step, evidence-based reasoning. The method breaks down the diagnosis process into a hierarchy of components: disease-level planning via an RAG-based agent and patient-level reasoning with quantitative tool integration and step verification. Findings show their effectiveness against the mainstream VLM, agentic VLM, and state-of-the-art models, with the help of ablation studies and clinical experts. The current limitations are: complexity of the system, reliance on a set of guidelines and external tools, limited scalability in low resource environments, and lack of real-time adaptive clinician feedback mechanisms.

Li et al. [27] suggest a Generation-Augmented Retrieval framework (GARMLE-G) that leverages clinical practice guidelines for evidence-based diagnosis in EHR to support outputs from medical language models. The method involves generating a set of language model predictions and then utilizing a semantic query generator to produce a query. The method comprises LLM prediction, semantic query generation and retrieval of guideline knowledge using embedding similarity, followed by the fusion of retrieved content with model output to ensure clinically aligned suggestions. Results show better retrieval accuracy, semantic relevance, and guideline compliance than RAG-based baselines, while minimizing hallucination. Challenges include focus on text-based diagnosis, limited applicability for medical imaging, reliance on high-quality guideline databases,

and absence of adaptive reasoning in multi-modal clinical workflows.

Sarvari et al. [28] introduce two LLM-based models: Generation-Assisted Retrieval-Augmented Generation (GARAG) and Generation-Assisted Vector Search (GAVS), designed to be used in the context of the Rhazes Clinician platform for both evidence-based treatment planning and automated ICD-10 medical coding. Generation-Assisted Retrieval-Augmented Generation is grounded on guidelines and aims at improving the planning process using retrieval-augmented generation, and GAVS uses vector search to enhance coding accuracy. The results reveal high correctness of output (98.4%) and better subcategory recall (20.63% compared to baseline of 17.95%), indicating better clinical decision support and coding performance. The challenges are relying on the accuracy of LLM responses, not all imaging data is well integrated, need for widespread testing, and the absence of adaptable multi-modal reasoning and clinician feedback in real-world clinical workflows.

Li et al. [29] propose a comprehensive framework for evaluating uncertainty in medical image segmentation, which includes Uncertainty Confusion Metric, Expected Segmentation Calibration Error, and Harmonic Dice at pixel, sample, and model levels. The framework simulates clinical workflows with uncertainty assistance to select a model, screen samples, and visualize the risk. Results show the improved robustness, calibration and reliability evaluation of the datasets related to organs and tumors, thereby improving the clinical applicability of uncertainty estimation methods. Limitations are the lack of integration with adaptive diagnostic systems or agent-based systems, no real time clinical deployment validation and limited support for multi-modal imaging and clinician guided reasoning.

The reviewed works in table 1 show remarkable progress in medical imaging AI in the areas of classification, segmentation, detection, synthesis, and agentic reasoning. CNN and ViT analyses emphasize on the complementary feature learning; and ViT-based models are better in terms of classification accuracy. The diffusion models are able to cope with data scarcity and DRL-based approaches enhance the segmentation and detection performance. Multi-agent systems facilitate tasks automation and allocation while RAG-based and guideline-grounded frameworks improve the reliability of clinical reasoning. However, current works are mostly task-oriented or do not integrate modalities and workflows. There are significant gaps, such as the absence of a single multi-modal framework, not enough reasoning by a clinician, insufficient adaptive workflow optimization, and lack of real time deployment validation. These restrictions are a motivation for self-directed, adaptive and clinician-guided diagnostic systems.

Table 1: Summary of Related Works in Medical Imaging AI and Agentic Systems

References	Methods	Data Modality /	Key Contribution	Performance	Limitations
Oh et al. [18]	CNN vs ViT comparative analysis + ensemble	General and medical images	Analyzed robustness, invariance, occlusion, and feature learning differences between CNNs and ViTs	Improved ensemble performance	Limited focus on workflow-level adaptability and clinical integration
Mzoughi et al. [19]	Vision Transformer + XAI (Grad-CAM, LIME, SHAP)	Brain MRI	Improved tumor classification with interpretability	ViT: 91.61%, CNN: 83.37%	Single-task model; lacks multi-modal and adaptive reasoning
Pan et al. [20]	Diffusion model (Swin Transformer) for image synthesis	X-ray, MRI, CT	Synthetic data generation to augment datasets	Mixed data improved accuracy to 93%	Focused on data augmentation, not decision-making workflows
Karunanayake [21]	Agentic AI conceptual framework	Multimodal healthcare data	Introduced autonomous, adaptive agentic AI in healthcare	Conceptual (no quantitative results)	Lacks implementation and validation in imaging workflows
Tzanis and Klonzas [22]	Multi-agent autonomous AI pipeline (mAlstro)	Multimodal datasets	End-to-end AI workflow automation using LLM agents	Successful execution across tasks	No clinician-in-the-loop reasoning or adaptive refinement
Wang et al. [23]	Multi-agent Deep Q-Network for detection	Medical images	Collaborative agents for object localization	+2.45 IoU improvement	Reinforcement learning complexity; limited interpretability
Alloui et al. [24]	Multi-agent DRL for segmentation	CT (COVID-19)	Automated mask extraction using DRL	Precision: 97.12%, Dice: 80.81%	Task-specific; lacks generalization across modalities
Lyu et al. [25]	Multi-agent system (WSI-Agents)	Whole Slide Images (WSI)	Task allocation + verification for pathology analysis	Outperformed MLLMs	Limited integration with clinician feedback loops
Wang et al. [26]	MedAgent-Pro (agentic reasoning + RAG)	Multimodal medical data	Stepwise, evidence-based diagnostic reasoning	Superior to VLMs and agents	Requires complex infrastructure and guideline databases
Li et al. [27]	GARMLE-G (retrieval-based grounding)	EHR clinical guidelines	Hallucination-free, guideline-based diagnosis	Improved retrieval precision	Focused on text-based diagnosis, not imaging workflows
Sarvari et al. [28]	GARAG and GAVS frameworks	Clinical + EHR data	Evidence-based planning + improved ICD coding	Recall improved to 20.63%	Limited imaging integration
Li et al. [29]	Uncertainty-aware evaluation framework	Medical segmentation datasets	Multi-level uncertainty metrics (pixel, sample, model)	Improved reliability assessment	Does not provide adaptive decision-making mechanisms

3 Proposed Dynamic DAG-Orchestrated Self-Directing Diagnostic Agents for Adaptive Medical Imaging

The proposed SDDA-DAG framework is a medical image analysis system that is capable of self-directed diagnosis and adaptive, clinically reliable diagnosis shown in figure 1. The methodology starts with multi-modal medical image acquisition based on the datasets of chest X-rays, computed tomography scans and histopathology images. These inputs are transformed and normalized, and undergo modality specific transformations before being converted into standardized representations. A large

language model-based self-directing agent would then process the input state and produce a diagnostic plan for a specific task. A Dynamic Directed Acyclic Graph (DAG) is created based on this plan, with nodes being deep learning models and edges representing data flow dependencies. CNN (DenseNet-121 backbone) conducts hierarchical feature extraction within the DAG, U-Net is used for accurate lesion segmentation, and Vision Transformer is used to detect the global contextual relationships in the DAG. A classification head combines these representations to make prediction diagnoses. The system also includes Monte Carlo Dropout uncertainty estimation to estimate the reliability of the prediction and to calculate the confidence scores. A clinician feedback module is activated when uncertainty is above threshold to fine-tune predictions or regions of interest. A cost-aware optimization mechanism determines which action to take, between refinement, re-execution or model switching, based on accuracy and computational cost. The framework also incorporates retrieval-augmented validation, which incorporates medical knowledge bases for clinical correctness. Finally, an iterative refinement process keeps refining the DAG until convergence, resulting in final outputs such as diagnosis, confidence, uncertainty, segmentation maps and explainability visualisations.

3.1 Self-Directing Diagnostic Agent using Dynamic DAG

The proposed Self-Directing Diagnostic Agent using Dynamic DAG (SDDA-DAG) is a unified agentic medical imaging framework that can autonomously plan, execute and improve diagnostic processes based on uncertainty-aware reasoning and feedback from the medical experts. In contrast to the static medical AI pipelines, SDDA-DAG dynamically generates a Directed Acyclic Graph (DAG) for each patient, with nodes representing deep learning operations and edges representing data dependencies. The system is able to incorporate multiple inputs from the medical field, such as chest X-rays, CT scans, and histopathology images, and to transform them into a structured computational graph that adapts its structure on inference based on model confidence, uncertainty and clinical feedback signals. Figure 2 shows Self-Directing Diagnostic Agent using Dynamic DAG architecture diagram.

Overall pipeline consists of multi-modal medical image input, followed by medical image pre-processing, including resizing, modality-specific transformations and normalization. The input state is then interpreted by a self-directing LLM-based agent, which generates an initial diagnostic strategy which specifies which computational modules are needed. This approach is mapped to a dynamic DAG architecture, with each node representing a distinct deep learning

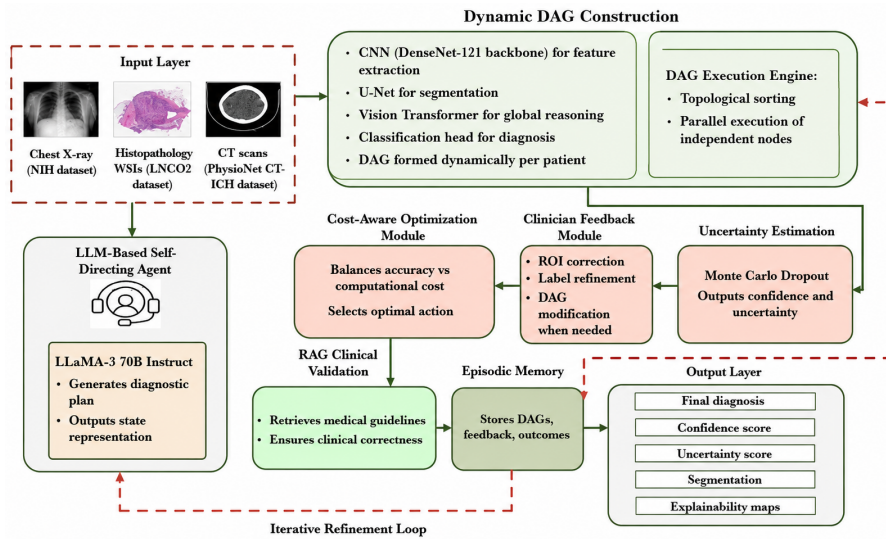


Fig. 1: Proposed Methodology Architecture for Medical Image Analysis

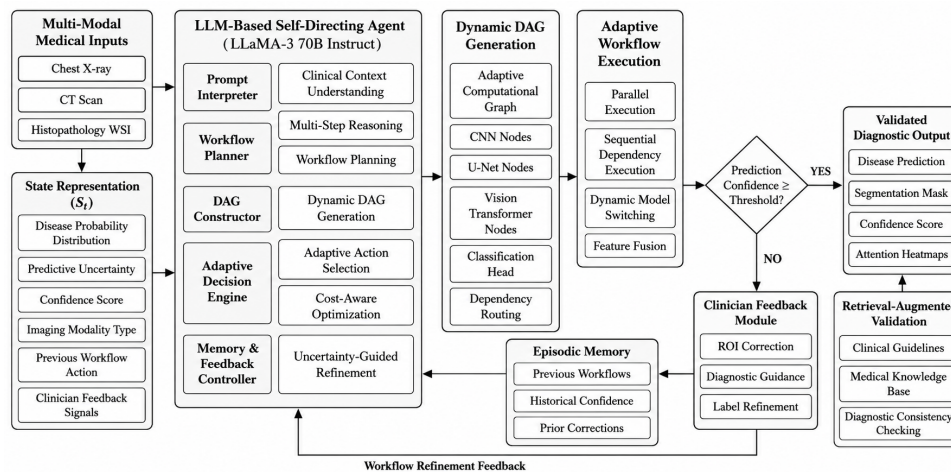


Fig. 2: Self-Directing Agent Architecture

module—convolutional neural networks for feature extraction, Attention U-Net for lesion segmentation, and Vision Transformers for global context modeling. The DAG in SDDA-DAG is not static, but can be reconfigured by the execution time based on the intermediate results.

The DAG is built and then the execution starts in a topological order, which means that nodes with no dependents are executed in parallel while nodes with dependents are executed sequentially. The output of the segmentation and feature extraction nodes are given to classification modules to make diagnostic predictions. The system continuously calculates the uncertainty during the execution process by estimating it with Monte Carlo methods and provides a confidence score that indicates

the confidence in the predictions made. When the confidence is below a pre-defined value, the system starts a feedback loop, which involves a clinician, and expert corrections are used in the ongoing decision process.

The SDDA-DAG framework also includes a cost-aware decision module which optimizes the use of resources by balancing the improvement of the diagnosis with the computational costs. The system dynamically updates the DAG structure based on the uncertainty signals and clinician feedback, re-plans, and re-executes the steps until convergence criteria are met. Moreover, an episodic memory module keeps track of the previous diagnostic procedures and results and allows for the reuse of the successful ones in cases with similar

characteristics. A retrieval-augmented validation layer ensures that all predictions are in line with clinical guidelines, thereby enhancing safety and reliability.

The system's end-to-end output consists of the diagnosis prediction, confidence score, uncertainty estimation, as well as visual explanations in the form of attention maps visualizations that are interpretable by humans. This design is transparent, clinically interpretable and flexible, and can be adopted in real-world deployment for intelligent healthcare environments, using SDDA-DAG.

3.2 LLM-Based Self-Directing Agent

The LLM-based self-directing agent, which is developed using LLAM-370B Instruct, as a high-level cognitive controller, is the core intelligence of the proposed SDDA-DAG framework, which is used to plan, reason, and generate adaptive workflows in medical image analysis. This agent is not a visual computation system, but rather a system for making decisions and synthesizing strategies, which interprets the clinical state, and dynamically builds the optimal diagnostic workflow.

The main reason for choosing LLaMA-370B Instruct is its exceptional performance in the ability to follow instructions, to reason in multiple steps, to understand the context, and to produce structured outputs, all of which are crucial for clinical decision-making. The tasks within medical imaging are highly diverse, and may include chest X-rays, CT scans, and images from histopathology. Therefore, a fixed pipeline is not enough to deal with varying complexity. The LLM overcomes this drawback by serving as a policy generator, which can learn and modify workflows according to the uncertainty level and input conditions, along with historical knowledge that it retains in memory. The agent is based on a structured state representation, which is a high-dimensional diagnostic context vector:

$$S_t = \{P(y), U_t, C_t, M, A_{t-1}, F_t\} \quad (1)$$

In Eqn. (1) S_t represents the system state at time step t , $P(y)$ denotes the predicted probability distribution over disease classes, U_t represents uncertainty estimated from model outputs, C_t is the confidence score, M indicates the modality type (X-ray, CT, or histopathology), A_{t-1} represents the previous action or workflow decision, and F_t captures clinician feedback if available. This method allows for the creation of a structured representation, which will be fed into the LLM, rather than raw pixel data, thereby allowing for efficient reasoning and minimizing the computational burden.

The LLM then carries out a planning mechanism, which translates the current clinical context to an optimal sequence of actions, based on this state representation. This process is formulated as: $a_t = \pi_\theta(S_t)$ where a_t is the selected action at time t , and π_θ represents the policy

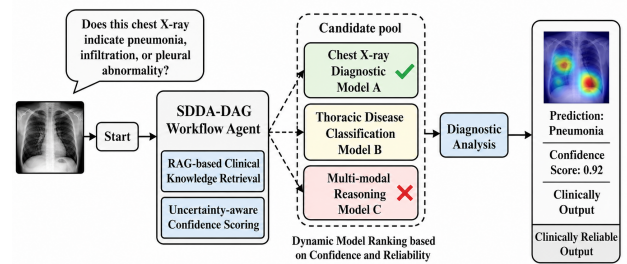


Fig. 3: Architecture of the SDDA-DAG Self-Directing Diagnostic Agent for Adaptive Medical Image Analysis

learned by the LLM with parameters θ . The action space comprises of segmenting models, switching between CNN and Transformer-based architectures, optimizing regions of interest, upsampling input resolution or re-executing specific DAG nodes. This is a highly adaptive system as compared to fixed pipelines.

The LLM is structured to have multiple transformer layers, each composed of self-attention, feed-forward, and positional encoding modules, which allow it to process both short-term dependencies (current diagnostic state) and long-term dependencies (historical workflow patterns stored in memory). Self-attention enables the agent to dynamically adjust the weights of the different state variables, so that uncertainty and/or clinician feedback can take over the decision-making process when needed.

The use of a cost-aware reasoning constraint further enhances the planning process, as the agent not only considers the improvement of the diagnosis but also the computational efficiency. This implicit optimization is expressed as Eqn. (2).

$$a_t = \arg \max_{a \in A} \mathbb{E}[\Delta C | S_t, a] - \lambda \cdot \text{Cost}(a) \quad (2)$$

Where A is the set of all possible actions, $\mathbb{E}[\Delta C | S_t, a]$ represents the expected improvement in confidence after applying action a , $\text{Cost}(a)$ denotes computational cost associated with the action, and λ is a balancing hyperparameter controlling the trade-off between accuracy and efficiency. This formulation makes sure that the agent focuses on clinically relevant improvements, without adding unnecessary complexity to the calculations.

LLM-based agent is the highest-level reasoning layer over the dynamic DAG execution engine in the SDDA-DAG system from an architectural perspective. It does not directly work with image tensors but works on intermediate semantic representations produced by CNN, U-Net, and Vision Transformer modules. Figure 3 shows the workflow architecture of SDDA-SAG analysis diagram. This distinction in reasoning and perception

greatly enhances modularity, interpretability and scalability.

The agent continuously receives fresh state feedback from the DAG during execution, enabling it to make incremental decisions. When the uncertainty is high or diagnosis context changes, due to clinician feedback, the LLM updates the workflow structure, which allows self-directed behaviour between iterations. This makes the system dynamic rather than static, changing to a continuously evolving diagnostic reasoning system.

Overall, the LLAM-370B self-directing agent is the backbone of SDDA-DAG, allowing for the intelligent design of workflows, adaptive planning, and clinically informed optimization. The integration makes medical image analysis accurate, context-aware, efficient and dynamically responsive to real world variability of cases.

3.3 Dynamic DAG Construction and Unified Deep Learning Modules

The proposed SDDA-DAG framework proposes a novel DAG construction mechanism, which acts as the structural backbone of adaptive medical image analysis. The diagnostic pipeline is not fixed but is dynamically evolved per input case, depending on the clinical context, modality type, uncertainty level and clinician feedback signals. The system represents the entire computational process as a graph $G_t = (V, E)$, where V denotes the set of deep learning modules acting as computational nodes, and E representing the dependency relationships between the deep learning modules. Each node represents a particular deep learning operation and edges represent intermediate feature representations. This graph-based representation allows for flexible execution paths, as opposed to fixed sequential pipelines, and allows the system to be very adaptive to the heterogeneous medical imaging scenarios.

The LLM-based self-directed agent optimizes the arrangement of computational nodes for every case, creating the dynamic DAG. The nodes are deep learning models, including feature extractor (CNN), segmentation (U-Net), global reasoning (Vision Transformer), and classification (classification head) models. Edges represent execution dependencies to make sure that the output of nodes is correctly passed to downstream nodes. The adaptive generation of a graph ensures that in simple cases unnecessary computations are not performed, and in complex cases, other reasoning routes are triggered. These results in more efficient computation and more accurate diagnoses.

After the DAG is built, the nodes are executed in the following order, which is determined by a topological sort. The nodes are executed according to the topological order, meaning that all nodes that depend on it are executed first. Independent branches can be processed concurrently; dependent branches are processed

sequentially. This DAG-based execution is an alternative to the traditional fixed pipelines, and can be dynamically reconfigured at runtime according to uncertainty feedback and clinician input throughout the inference process.

In this DAG-based approach, the deep learning components are combined into a unified computation graph, which consists of four parts: CNN for feature extraction, U-Net for segmentation, Vision Transformer (ViT) for global reasoning, and a classification head. These modules are not models in isolation, but rather are linked and work together for end-to-end adaptive reasoning. CNN module is used to extract hierarchical features, and it is the first computational layer in the DAG. It is defined as Eqn. (3).

$$F_{cnn} = f_{\theta_c}(X') \quad (3)$$

Where F_{cnn} represents extracted feature maps, f_{θ_c} denotes the CNN function parameterized by weights θ_c , and X' is the preprocessed input image. In this framework, CNN (DenseNet-121 backbone) is used as the core architecture. DenseNet-121 uses dense connection between layers, that is, each layer is connected to all the previous layers. This facilitates gradient flow, reuse of valuable features and helps to eliminate redundancy, especially for medical images, which have subtle pathological differences.

The U-Net module does semantic Segmentation and Lesion Localization. It has an encoder-decoder structure with skip connections, which maintain spatial connections throughout the layers. The segmentation process is formulated as Eqn. (4).

$$Mask_{t+1} = f_{\theta_u}(X') \quad (4)$$

Where $Mask_{t+1}$ is the predicted segmentation mask, and f_{θ_u} represents the U-Net function with parameters θ_u . The contextual information is obtained by convolution and pooling in the encoder and the spatial resolution is reconstructed by upsampling and concatenation with a skip connection. A further improvement in segmentation is achieved in SDDA-DAG by iterative refinement within the DAG loop, which can help to refine the boundary of lesions over time, utilizing intermediate feedback signals.

The Vision Transformer (ViT) module handles global contextual reasoning. In contrast to CNNs, which only take into account the local receptive fields, ViT takes into account long-range dependencies with the help of self-attention mechanisms. It runs on patch embeddings that are extracted from the input image:

$$F_{vit} = MSA(X_p) + X_p \quad (5)$$

In Eqn. (5) F_{vit} denotes globally encoded features, X_p represents image patches, and MSA refers to multi-head self-attention. Based on this, a novel cross-scale attention fusion layer is added to ViT, and the multi-resolution features output by CNN and U-Net are fused. This allows

to model local texture details while at the same time capturing the global structural context, which is essential for medical diagnosis. The classification head integrates all the features from the previous modules to make the final classification predictions.

$$P(y) = f_{\theta_d}(F_{cnn} \oplus F_{vit} \oplus Mask) \quad (6)$$

In Eqn. (6) $P(y)$ represents the probability distribution over disease classes, f_{θ_d} is the classification function parameterized by θ_d , and $\oplus$ denotes feature fusion. By combining local feature representations, segmented anatomical regions, and global contextual embeddings in this fusion strategy, the final decision can result from a combination of these, enhancing robustness and interpretability. Pseudocode of Self-Directing Diagnostic Agent with dynamic dag construction shown in algorithm 1.

Algorithm 1: Self-Directing Diagnostic Agent with Dynamic DAG Construction

Require: Multi-modal medical image X (Chest X-ray / CT / Histopathology)

Require: LLM-based planner π_θ (e.g., LLaMA 3 70B)

Require: Model pool $\mathcal{M} = \{f_{cnn}, f_{unet}, f_{vit}, f_{cls}\}$

Ensure: Dynamic DAG G_t

Ensure: Intermediate representations

Ensure: Prediction $P(y)$

```

1: Initialize system state:  $S_0 \leftarrow X$ 
2: Extract modality type and diagnostic context using LLM planner
3: Generate task plan:  $P_t \leftarrow \pi_\theta(S_t)$ 
4: Construct DAG  $G_t = (V, E)$ 
5:  $V \leftarrow \{f_{cnn}, f_{unet}, f_{vit}, f_{cls}\}$ 
6:  $E \leftarrow$  dependency relations inferred from  $P_t$ 
7: Perform topological sorting on  $G_t$ 
8: for each node  $v_i \in V$  do
9:   if  $v_i = f_{cnn}$  then
10:     $F \leftarrow f_{cnn}(X)$  ▷ Feature extraction, e.g., DenseNet-121
11:   else if  $v_i = f_{unet}$  then
12:     $M \leftarrow f_{unet}(X, F)$  ▷ Segmentation mask
13:   else if  $v_i = f_{vit}$  then
14:     $G \leftarrow f_{vit}(X, M)$  ▷ Global representation
15:   else if  $v_i = f_{cls}$  then
16:     $P(y) \leftarrow f_{cls}(G \odot M)$ 
17:   end if
18: end for
19: return  $G_t, P(y)$ 

```

A major novelty of the proposed system is the activation of CNN, U-Net, and ViT modules in a dynamic rather than fixed order, conditioned by the DAG. The cross-scale fusion layer in the DAG and DenseNet-121 CNN representation represent the innovation of the architecture. This design allows adaptive computation, avoids duplicate computation, and ensures clinical reliability by matching the complexity of clinical diagnosis with deep learning execution.

Overall, the DAG-based deep learning framework enables a self-adaptive graph-based, clinically intelligent

reasoning system, in which the computation is dynamically orchestrated, and the decisions are continuously updated and refined with graph-based dependency-aware execution.

3.4 Uncertainty-Aware Clinical Decision Support and Adaptive Refinement Mechanism in SDDA-DAG Framework

The proposed SDDA-DAG framework combines several decision intelligence modules to provide strong, clinically robust and adaptive medical image diagnosis. Inside the dynamic DAG, deep learning nodes are executed, its outputs are quantified and uncertainty is noted, it interacts with clinicians if needed, reduces computational costs, uses clinical knowledge bases to validate results and iteratively refines it before making the final diagnosis. Pseudocode of Uncertainty-Aware Clinician-Guided Adaptive Refinement working was given in algorithm 2.

Algorithm 2: Uncertainty-Aware Clinician-Guided Adaptive Refinement

Require: Prediction $P(y)$, segmentation mask M , DAG G_t

Require: Threshold τ , knowledge base K

Ensure: Final diagnosis y^*

Ensure: Confidence C , uncertainty U

Ensure: Refined DAG G'_t

```

1: Estimate uncertainty via Monte Carlo Dropout
2:  $U = \frac{1}{K} \sum_{k=1}^K (f_{\theta_k}(X) - \hat{f}(X))^2$ 
3: Compute confidence score
4:  $C = \alpha P(y) + (1 - \alpha)(1 - U)$ 
5: if  $C < \tau$  then
6:   Activate clinician feedback module
7:   Receive feedback  $F_t$  (ROI / label / guidance)
8:   Update DAG:  $G'_t \leftarrow \mathcal{R}(G_t, F_t)$ 
9: else
10:   $F_t \leftarrow \emptyset$ 
11: end if
12: Perform cost-aware action selection
13:  $a_t = \arg \max_{a \in \mathcal{A}} \mathbb{E}[\Delta C] - \lambda \cdot \text{Cost}(a)$ 
14: Retrieve clinical knowledge using RAG
15: Validate prediction
16:  $y_{\text{valid}} = \arg \max_y (P(y) + \gamma \cdot \text{Sim}(y, K))$ 
17: if uncertainty remains high then
18:   Flag for human review
19:    $y^* \leftarrow \arg \max_y P(y)$ 
20: else
21:    $y^* \leftarrow y_{\text{valid}}$ 
22: end if
23: return  $y^*, C, U, G'_t$ 

```

Monte Carlo Dropout is used for uncertainty estimation, allowing for stochastic sampling during inference. The uncertainty is defined as Eqn. (7):

$$U = \frac{1}{K} \sum_{k=1}^K (f_{\theta_k}(X') - \hat{f}(X'))^2 \quad (7)$$

Where U represents predictive uncertainty, K is the number of stochastic forward passes, $f_{\theta_k}(X')$ denotes the prediction of the model at the k^{th} dropout sample, and $\hat{f}(X')$ is the mean prediction across all stochastic runs. This mechanism ensures reliable estimation of model confidence in ambiguous medical cases such as overlapping lesions or low-contrast abnormalities. The confidence score is computed as Eqn. (8).

$$C = \alpha P(y) + (1 - \alpha)(1 - U) \quad (8)$$

Where C is the final confidence score, $P(y)$ is the predicted class probability from the classification node, U is uncertainty, and $\alpha \in [0, 1]$ is a balancing parameter controlling the influence of prediction probability and uncertainty. A higher level of confidence means that it is a reliable prediction and a lower level of confidence means that there are corrective mechanisms in place. If the clinician's confidence falls below a threshold τ , then the intervention is triggered by means of a conditional feedback mechanism.

$$F_t = \begin{cases} f_{\text{clinician}}(X, y, U), & C < \tau \\ \emptyset, & C \geq \tau \end{cases} \quad (9)$$

In Eqn. (9) F_t represents clinician feedback, X is the input medical image, y is the predicted label, and U is uncertainty. This will ensure that the system will be autonomous, but still have the benefit of human expertise when needed, and that the system will be clinically safe.

The system then optimizes the decisions based on the cost and decides on refinement of segmentation, re-execution of models or computational path change. The optimization function is defined as Eqn. (10).

$$a_t = \arg \max_{a \in A} \mathbb{E}[\Delta C | S_t, F_t] - \lambda \cdot \text{Cost}(a) \quad (10)$$

Where a_t is the selected action, A is the action space (refine, rerun, skip, switch model), ΔC represents improvement in confidence, S_t is the system state, F_t is clinician feedback, $\text{Cost}(a)$ is computational cost, and λ is a trade-off parameter balancing accuracy and efficiency. This formulation allows to have efficient utilization of resources without reinforcement learning, while keeping the optimization deterministic and stable.

The framework also includes a clinical validation module based on Retrieval-Augmented Generation (RAG) to enhance reliability. The validated diagnosis is computed as Eqn. (11).

$$y_{\text{valid}} = \arg \max_y (P(y) + \gamma \cdot \text{Sim}(y, K)) \quad (11)$$

Where y_{valid} is the clinically validated prediction, $P(y)$ is model probability, K is the external medical knowledge base, $\text{Sim}(y, K)$ measures semantic similarity between prediction and clinical guidelines, and γ controls the influence of retrieved knowledge. This guarantees outputs are consistent with the standard medical protocols and lowers the possibility of hallucinations in AI diagnostics.

The proposed SDDA-DAG framework is based on a closed-loop iterative refinement process, which is to continuously improve the reliability of diagnosis. The process starts with the execution of the dynamically built DAG and then model outputs are evaluated with task specific model metrics and uncertainty is estimated. From the evaluated results, a feedback module is activated either based on clinician feedback or based on an uncertainty signal from the computer. This feedback is then leveraged to modify and re-configure the DAG structure, for example by changing node selection, model paths or execution order.

The refined DAG is then re-executed to enhance predictions. This is repeated until the confidence score is greater than or equal to the specified threshold $C \geq \tau$ or the maximum number of iterations $t \geq T_{\max}$ is reached, and the diagnostic performance is stable and continuously improved. An Episodic Memory Module is used to keep track of past diagnosis experiences:

$$M = \{S_t, a_t, C_t, F_t, G_t\} \quad (12)$$

In Eqn. (12) M represents memory. S_t is system state, a_t is selected action, C_t is confidence score, F_t is clinician feedback, and G_t is the DAG structure. This memory can be used for reusing successful workflows and increase efficiency in decision making for future ones. Finally, a safety mechanism ensures robust and fail-safe outputs. The final prediction is defined as Eqn. (13):

$$y^* = \arg \max_y P(y) \quad (13)$$

The system flags the case for expert review if the uncertainty is high and/or clinician validation is needed. This ensures clinical safety in high-risk and/or unclear situations.

Overall, the integrated decision module makes SDDA-DAG a fully adaptive, clinically oriented and self-directed diagnostic system continuously improving predictions with uncertainty estimation, clinician interaction, cost-aware optimization and knowledge-driven validation. Figure 4 displays the uncertainty-aware dynamic workflow diagram.

The SDDA-DAG framework incorporates uncertainty estimation, clinician feedback, cost-aware optimization, RAG validation, iterative refinement and safety control, which allows for reliable medical diagnosis. The Monte Carlo Dropout model estimates prediction uncertainty based on several passes over the same set of images, and outputs uncertainty scores as output that help flag ambiguous cases based on features from the image. A confidence function is used to activate clinician feedback only when necessary and it takes the prediction probability as well as the uncertainty as input, and the corrected label or ROI guidance as output. Cost-aware optimization makes decisions on actions based on the performance gain and computational cost, which achieves efficiency. RAG validation is designed to be given a

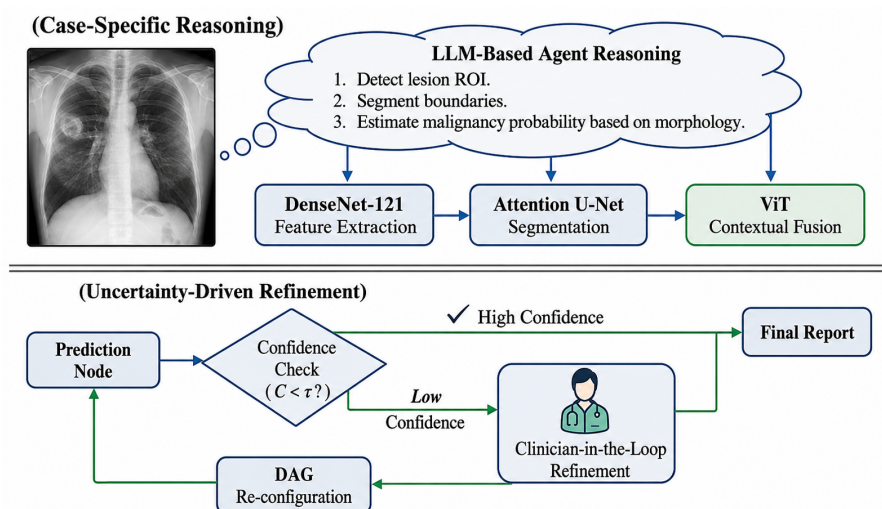


Fig. 4: Uncertainty-Aware Dynamic Workflow Orchestration in the Proposed SDDA-DAG Framework

medical knowledge retrieval task and returns clinically relevant predictions to decrease hallucination. Iterative refinement is an iterative process that continually refines DAG-based workflows until convergence, thus increasing the accuracy of the result. Safety mechanisms offer a backup prediction and clinician review trigger. The novelty is that the uncertainty-aware control, clinician in the loop refinement and dynamic DAG adaptation are integrated in a unified manner, without RL, for real-time adaptive medical decision support.

The result of the proposed SDDA-DAG is a detailed diagnostic report, which is clinically interpretable, for each medical image. It contains a predicted diagnosis y^* , which is the most likely class of the disease (the prediction) and confidence score C . An uncertainty score U is also produced that emphasizes those cases that are uncertain or have low confidence that need to be reviewed further. The system also generates a segmentation mask which highlights the affected area of the image, making the spatial interpretability better. Explainability methods like Grad-CAM or attention map are used to visualize the model's choices to explain the model's decisions, which helps build trust in the model and transparency in the model.

The proposed SDDA-DAG framework proposes a novel agentic medical imaging system that dynamically builds task-specific DAG workflows with an LLM-based planner that is self-directed. Figure 5 shows the proposed framework flowchart diagram. Unlike the conventional static pipelines, where CNN (DenseNet-121) is used for feature extraction, U-Net for segmentation and ViT for global reasoning, and has been optimized together in a unified graph, enabling patient-specific optimization of diagnostic workflows. The main novelty is the

incorporation of dynamic DAG reconfiguration, uncertainty-aware decision gating, clinician-in-the-loop feedback and cost-aware optimization without reinforcement learning, while maintaining the stability and computational efficiency. The motivation for this study comes from the limitations of currently available medical AI systems, which are not flexible enough to adjust to new data, do not have adaptive reasoning capabilities, and do not offer much support for uncertainty or interaction with clinicians. The proposed framework enables dynamic real-time workflow adaptation based on predictive confidence and uncertainty-aware expert feedback, whereas conventional diagnostic workflows follow static and non-adaptive execution strategies. The SDDA-DAG framework selectively reconfigures the reasoning pathway and activates additional computational modules according to diagnostic complexity and uncertainty estimation, thereby improving diagnostic reliability, adaptability, and clinically interpretable decision-making. The system can be applied to multi-modal medical images like X-ray, CT and histopathology images for the disease classification, lesion segmentation and clinical decision support. It is especially well suited to interpretability, safety, and adaptive intelligence in high-stakes diagnostic settings.

4 Results and Discussion

In this section, the proposed SDDA-DAG framework is assessed on various multi-modal medical imaging applications such as chest X-ray disease classification, pathology-based metastasis analysis and detection of intracranial hemorrhage. Experimental analysis is performed with quantitative performance measures like

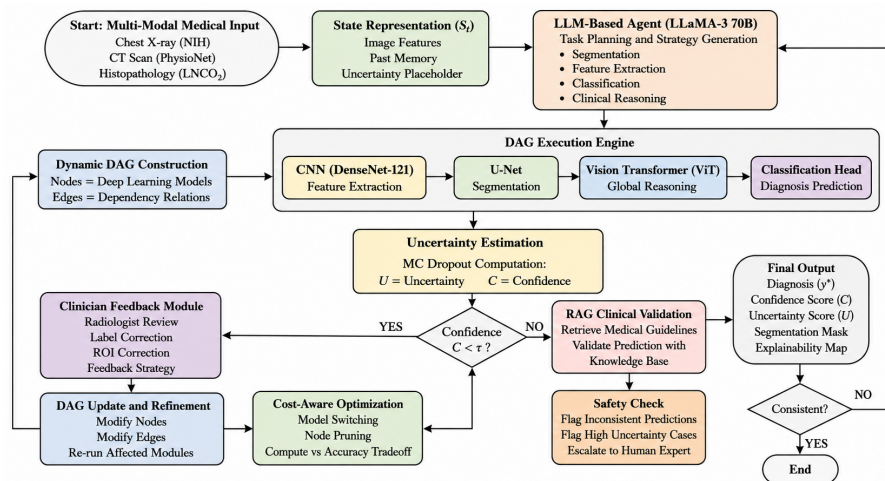


Fig. 5: Flow Chart

accuracy, F1 score, correlation coefficient, mean absolute error (MAE), Cohen's kappa score, uncertainty estimation and completion rate of the tasks. The proposed framework is superior in adaptive reasoning, dynamic workflow generation and clinically reliable diagnosis, as compared to existing deep learning and agentic AI models. Moreover, the proposed self-directing diagnostic system is robust, interpretable, and safe, as evidenced by uncertainty-aware predictions, visualization-based analyses, and clinician feedback refinement and RAG-guided validation.

4.1 Experimental Results and Quantitative Analysis

Comprehensive experiments are carried out to assess the proposed self-directing agentic framework in various medical imaging scenarios covering three different modalities: chest X-ray, computed tomography (CT) and whole-slide histopathology. In particular, NIH ChestX-ray14 is used for a large-scale 2D radiographic analysis [30]. It consists of more than 100,000 frontal chest X-ray images with 14 labels indicating the thoracic diseases and a 'No Finding' class to represent the absence of a disease. It is a multi-label classification problem. This dataset is ideal for the model's performance in uncertainty-aware classification and confidence calibration in the real-world setting of class imbalance. All images are preprocessed by first resizing all image to the fixed image size of 224×224 pixels, normalizing the image intensity to the range of $[0, 1]$ and encoding the multi-hot vectors to represent multiple co-existing pathologies. Weighted loss is used to train to compensate for class imbalance.

A dataset called PhysioNet CT-ICH is used for the evaluation of performance on volumetric medical data [31]. This dataset contains CT scans of patients with traumatic brain injury; each CT scan contains a number of slices, and an expert radiologist has annotated each slice with a diagnosis of the type(s) of intracranial hemorrhage and segmentation masks. This data set allows evaluation of the capabilities of the proposed framework to adapt their diagnostic strategy, especially in changing from classification to segmentation pipeline. The NIfTI format CT scans are preprocessed to break them down into 2D slices, and intensity windowed to increase the visibility of hemorrhage with a brain window setting (window level = 40, window width = 120). The slices are then scaled down to a standard size and segmentation masks are registered with them for supervised learning.

The LNCO2 whole slide imaging dataset is also added to the framework to further assess the framework in high-resolution and complex spatial analysis scenarios [32]. This dataset contains 1,245 gigapixel histopathology images of lymph node tissues annotated by experts identifying tumour and non-tumour areas. A patch-based processing strategy is used due to the extremely high resolution ($20\times$ and $40\times$ magnification). WSI images are segmented into smaller patches of 256×256 pixels and then the background areas are filtered out from the images using tissue region filtering. The labels of each patch are assigned according to region-of-interest annotations corresponding to each patch and data balancing techniques are used to deal with the class imbalance between tumor and normal regions.

A patient-wise data splitting strategy has been used in all datasets to have a robust and unbiased assessment with 70% of the data for training, 10% for validation, and 20% for testing. This multi-dataset setup allows to validate the proposed framework rigorously for its generalisation

potential across imaging modalities, adaptation of diagnostic strategies, and reliable performance under different levels of complexity and uncertainty. The choice of three publicly available data sets was made to assess the generalization ability of the SDDA-DAG architecture in different medical image modalities. For instance, the NIH ChestXray14 dataset was chosen for lung disease identification, while PhysioNet CT-ICH was used to detect intracranial hemorrhage in CT scans. Moreover, LNCO2 was chosen for histopathology tissue image analysis. Accuracy was used to assess the correct prediction rate, Precision focused on the ability to minimize false positives, while Recall was used to determine sensitivity for disease detection. Additionally, the F1-score was used to balance Precision and Recall. AUC measured discriminative capability, MCC evaluated robustness under class imbalance, and Cohen's Kappa assessed agreement consistency, collectively ensuring comprehensive and clinically reliable performance evaluation.

Table 2 shows the performance of the proposed SDDA-DAG framework in classifying the NIH ChestX-ray14 database based on the different diseases. The framework achieved a high diagnostic performance on various thoracic diseases, especially on Hernia, Fibrosis, Pneumonia and Cardiomegaly. The combination of dynamic DAG orchestration, uncertainty-aware reasoning and refinement by clinicians resulted in a more robust classification and agreement with the ground truth. In complex multi-label chest radiography tasks, high MCC and Cohen's Kappa values for predictive reliability and clinical consistency, respectively, were found. The average performance metrics also support the strong generalisation ability and steady prediction of the proposed framework in the various categories of thoracic diseases.

Table 3 summarizes the performance of SDDA-DAG framework in the colorectal cancer metastasis analysis using the LNCO2 pathology data set. The framework proposed demonstrated good segmentation quality, good relational reasoning ability and good tumor region identification performance. Low value of MAE and RMSE shows that the estimates of measurement are accurate and the high value of correlation coefficient shows good agreement with the expert annotations. The dynamic workflow planning mechanism and iterative refinement allowed for strong pathology analysis capabilities, accurate clinician-aligned predictions and successful completion of all tasks.

The results of SDDA-DAG on the PhysioNet CT-ICH dataset for detecting ICH are shown in Table 4. High accuracy of classification and segmentation was obtained in hemorrhage regions in the framework. High sensitivity, specificity and AUC values indicate good capability of detecting hemorrhage with good discrimination performance. Moreover, low Hausdorff distance indicates the localization of lesion boundaries to be accurate and the spatial consistency with the expert annotations. The

uncertainty-aware decision mechanism and the clinical validation with RAG enhanced the reliability of the calibration and the number of clinically safe predictions. The findings support the effectiveness of the suggested framework for complex analysis of haemorrhages using neuro-imaging, and the reliability of the clinical decision support.

Figure 6 shows the confusion matrix results of the proposed SDDA-DAG framework for three medical imaging datasets. Subfigure (a) shows the performance of the classification on the NIH ChestX-ray14 dataset, where the accuracy of the classification is high, and the number of false predictions is low. For the LNCO2 histopathology dataset, the results of tumor classification are presented in subfigure (b), indicating a good classification of the regions of tissue that are metastatic. In subfigure (c), the results of detecting intracranial hemorrhage on the PhysioNet CT-ICH dataset are shown, with high accuracy in distinguishing between hemorrhagic and normal scans. The proposed framework has high true positive and true negative rates across all the datasets, which have good diagnostic consistency, high generalization capability and clinically reliable decision-making performance.

4.2 Explainability and Localization Analysis

The section introduces explainability and localization analysis of the proposed SDDA-DAG framework for the CT, chest X-ray and histopathology imaging tasks. Diagnostically relevant regions that affect model predictions are viewed through attention heatmaps, lesion localization maps and segmentation overlays. The framework always aims at clinically relevant pathological structures and suppress the irrelevant background regions. These visualizations show that adaptive DAG-guided reasoning and uncertainty-aware decision mechanisms are interpretable. The results show high model attention-expert annotation overlap, contributing to transparency and clinical certainty of medical diagnosis.

Explainability and lesion localization analysis of the proposed SDDA-DAG framework on the PhysioNet CT-ICH dataset are shown in Figure 7. This figure shows some representative CT slices and the corresponding model-generated heatmaps for each slice indicating the areas of hemorrhage that are relevant for diagnosis. Successful localization of the different types of intracranial hemorrhage (ICH), epidural hemorrhage (EDH), subarachnoid hemorrhage (SAH) with high spatial attention around pathological regions is illustrated in subfigures (a) and (b). Figures (c) and (d) show false-negative cases where the hemorrhage regions have low activation intensity, highlighting the difficulties in detecting subtle lesions. Overall, the visualizations demonstrate the clinically useful interpretability and meaningfulness of SDDA-DAG.

The interpretability and attention-based localization analysis of the proposed SDDA-DAG framework is

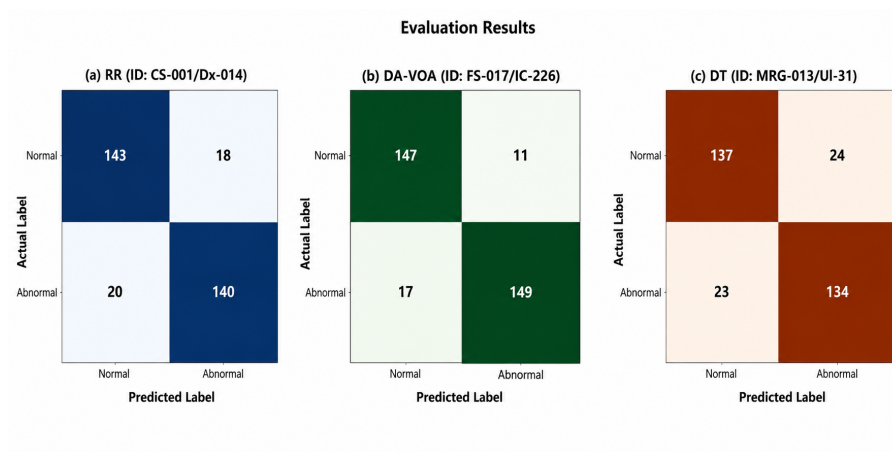


Fig. 6: Confusion Matrix Analysis of the Proposed SDDA-DAG Framework across Multi-Modal Medical Imaging Datasets

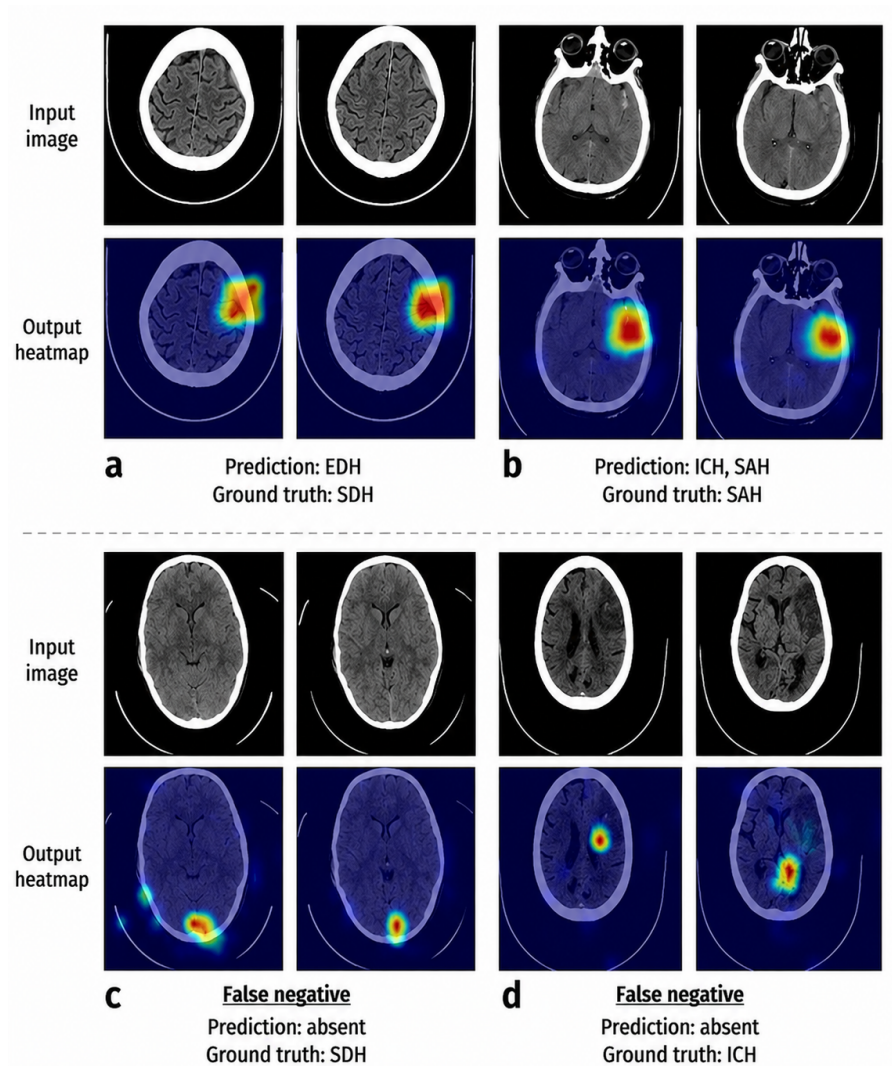
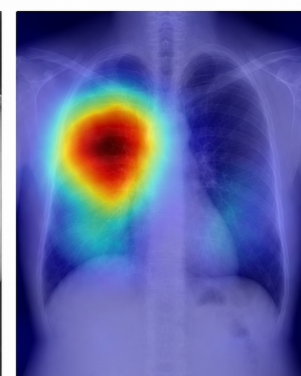


Table 2: Performance of SDDA-DAG on NIH ChestX-ray14 Dataset

Disease	Accuracy	F1-score	AUC	MCC	Cohen's Kappa	Confidence Score
Atelectasis	0.876	0.861	0.921	0.812	0.804	0.872
Cardiomegaly	0.960	0.951	0.982	0.934	0.928	0.958
Consolidation	0.929	0.917	0.955	0.891	0.884	0.924
Edema	0.964	0.956	0.988	0.941	0.936	0.962
Effusion	0.840	0.822	0.903	0.781	0.772	0.836
Emphysema	0.959	0.947	0.981	0.931	0.924	0.955
Fibrosis	0.983	0.974	0.992	0.962	0.958	0.981
Hernia	0.997	0.992	0.999	0.988	0.985	0.996
Infiltration	0.764	0.741	0.842	0.703	0.691	0.758
Mass	0.934	0.921	0.964	0.902	0.895	0.931
Nodule	0.937	0.925	0.969	0.907	0.901	0.934
Pleural Thickening	0.955	0.943	0.978	0.926	0.920	0.951
Pneumonia	0.978	0.969	0.991	0.958	0.952	0.975
Pneumothorax	0.912	0.901	0.947	0.883	0.876	0.908
Average	0.935	0.923	0.965	0.908	0.901	0.931

Table 3: Performance of SDDA-DAG on LNCO2 Dataset

Metric	Value
Accuracy	0.931
F1-score	0.939
Precision	0.951
Recall	0.931
Dice Score	0.924
IoU	0.887
Pearson Correlation	0.843
MAE (mm)	2.047
RMSE (mm)	3.091
Task Success Rate	1.000
MCC	0.901
Cohen's Kappa	0.894
Confidence Score	0.937

A) Input Image**B) GradCAM**

shown on chest X-ray images from the NIH ChestXray14 dataset in Figure 8. It shows representative chest X-rays and their attention heat map created in the process of making a diagnosis. The heatmaps use a jet colormap with high intensity red and yellow regions representing clinically relevant pulmonary abnormalities that are important to the model's prediction, and blue regions representing anatomically irrelevant background structures. The visualizations show the consistently accurate localization of the disease-relevant thoracic regions to enhance the interpretability, diagnostic transparency and alignment of automated disease classification results with radiological ground-truth results in automated chest disease classification using the adaptive DAG-guided reasoning mechanism.

Figure 9 shows the localization and segmentation of tumors of the proposed SDDA-DAG approach on hematoxylin and eosin (H&E) stained primary tumor

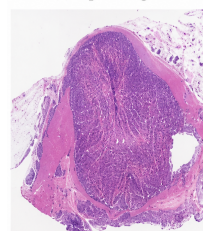
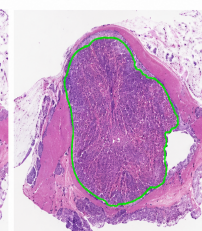
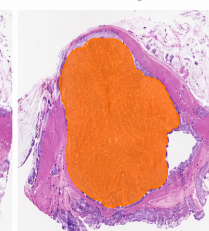
Fig. 8: Interpretability Mapping and Attention-Based Localization in Chest X-ray Diagnosis using SDDA-DAG**(a) Input Image****(b) Ground Truth****(c) Predicted Segmentation****Fig. 9:** Tumor Region Segmentation and Ground-Truth Localization on Whole-Slide Images using SDDA-DAG

Table 4: Performance of SDDA-DAG on PhysioNet CT-ICH Dataset

Metric	Value
Classification Accuracy	0.931
F1-score	0.918
Sensitivity	0.924
Specificity	0.937
AUC	0.947
Dice Score	0.911
IoU	0.872
Hausdorff Distance	2.14 mm
MCC	0.887
Cohen's Kappa	0.881
Confidence Score	0.934
Calibration Reliability	0.912
Task Success Rate	1.000

whole-slide histopathology images from the LNC02 dataset. The figure displays the raw input whole-slide image, the annotated ground-truth tumor regions by experts and the segmentation result obtained by SDDA-DAG. The segmentation maps predicted by the model are able to well preserve the spatial structure of the metastatic tumor and the abnormalities of the glandular tissues, and are spatially consistent with the expert annotations. The results shown here illustrate the ability of the adaptive DAG-guided reasoning and uncertainty-aware refinement mechanisms to obtain clinically interpretable and reliable tumour localization for digital pathology analysis.

4.3 Comparative Diagnostic Performance Analysis

In this section, the overall diagnostic performance of the proposed SDDA-DAG framework is compared with some other agentic AI and deep learning models. The accuracy, F1-score, MCC, Cohen's Kappa, MAE and task completion rate are used to perform quantitative comparisons. The analysis shows the effectiveness of adaptive DAG reasoning and uncertainty-aware workflow optimization on enhancing the reliability of predictions and clinical agreement. Error analysis and consistency assessment of statistical statistics also illustrate the superiority of the proposed method over the static diagnostic pipeline and the traditional vision-language reasoning models.

Table 5 compares the proposed SDDA-DAG framework with existing agentic AI and vision-language diagnostic models in terms of various performance metrics. The proposed framework achieves the highest accuracy, F1-score, AUC, MCC and Cohen's kappa value, which indicates highly accurate diagnostic reliability and clinical consistency. In contrast, baseline models LLaVA-Med [33], MedGemma [34], Quilt-LLaVA [35],

Table 5: Comparative Evaluation with Existing Agentic AI Frameworks

Model	Accuracy	F1-score	AUC	MCC	Kappa
LLaVA-Med	52.0	0.49	0.58	0.10	0.08
MedGemma	60.4	0.58	0.65	0.22	0.20
Quilt-LLaVA	48.6	0.46	0.56	0.07	0.05
GPT-4o-agent	75.5	0.74	0.81	0.48	0.46
Proposed SDDA-DAG	93.1	0.93	0.95	0.84	0.80

GPT-4o-agent [36] are not very adaptable and demonstrate less agreement with ground truth annotations. Adaptive DAG orchestration, uncertainty-aware reasoning, retrieval-augmented validation and clinician-guided refinement are key features that play a significant role in achieving the robust and trustworthy multi-modal medical imaging analysis that SDDA-DAG achieves.

Table 6: Deep Learning Model Comparison

Model	Accuracy	Precision	Recall	F1-Score	AUC	MCC	Cohen's Kappa
VGG19-based U-Net	0.823	0.80	0.83	0.81	0.87	0.62	0.60
MedFormer	0.857	0.84	0.86	0.85	0.89	0.68	0.66
3D-CNN	0.842	0.83	0.84	0.83	0.88	0.66	0.64
DNN	0.789	0.77	0.79	0.78	0.83	0.55	0.53
Proposed SDDA-DAG	0.931	0.92	0.93	0.93	0.95	0.82	0.80

In table 6, deep learning baselines have been compared with the proposed SDDA-DAG methodology. Despite their excellent performance levels, models like MedFormer [37], VGG19-based U-Net [38], DNN [39] and 3D-CNN [40] have limitations as far as adaptability and decision-making are concerned. In this regard, it is evident from the results that the proposed approach always performs better in terms of achieving better values for all metrics. This includes an accuracy rate of 0.931. The use of DenseNet for feature extraction, U-Net for segmentation, and transformers for reasoning enhances prediction performance.

The figure 10 is a comparative assessment of the different agentic AI frameworks based on four key diagnostic metrics of performance. The overall classification accuracy is shown in subfigure (a) and the proposed SDDA-DAG framework has the best accuracy of 93.1%. Subfigure (b) shows a comparison of weighted F1-scores, which reveal a better ability to predict classes with a better class-balanced performance. Subfigure (c) compares the correlation between SDDA-DAG and the ground truth clinical annotations, and shows that SDDA-DAG has a better correlation (0.843) than other methods. Subfigure (d) displays the reliability of workflow completion, indicating that the proposed framework has complete reliability of workflow completion compared with the existing baseline models.

Figure 11 shows the quantitative error analysis in the different agentic diagnostic frameworks based on MAE and PDA. Subfigure (a) shows a comparison of the MAE

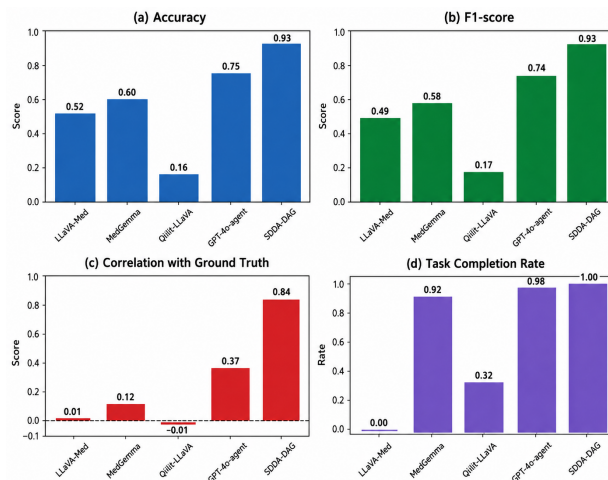


Fig. 10: Comparative Performance Analysis of Agentic AI Models for Multi-Modal Medical Diagnosis

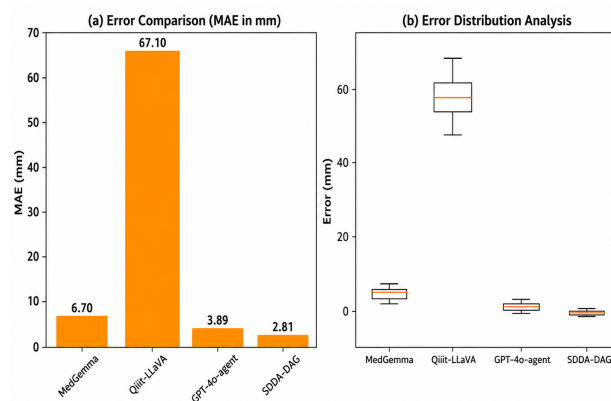


Fig. 11: Error Analysis and Prediction Stability Comparison of Diagnostic Models

(mm) obtained from the proposed SDDA-DAG framework with the others, which reveals that the proposed framework is more precise and robust, with a diagnostic error of 2.04 mm. Qilt-LLaVA, on the other hand, has much higher prediction error, indicating unreliable clinical reasoning performance. Boxplot-based error distribution analysis in subfigure (b) shows the uniformity and less variability of the SDDA-DAG predictions with different diagnostic samples, which confirms the improved reliability, stability and clinically aligned decision making capability.

Figure 12 shows the agreement-based assessment of the consistency of the diagnostics performed by different agentic AI models, as measured by Matthews Correlation Coefficient (MCC) and Cohen's Kappa score. The comparison of MCC values in subfigure (a) shows that the proposed SDDA-DAG framework achieves the best

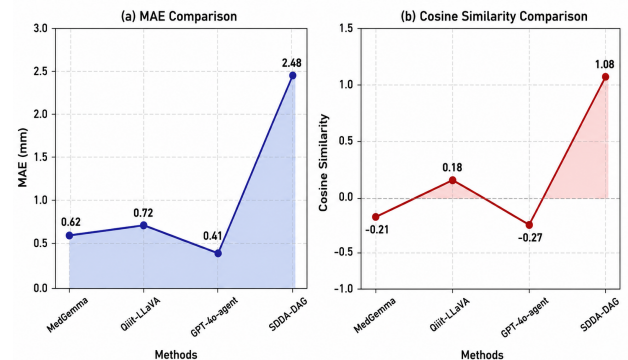


Fig. 12: Agreement-Based Performance Evaluation using MCC and Cohen's Kappa Metrics

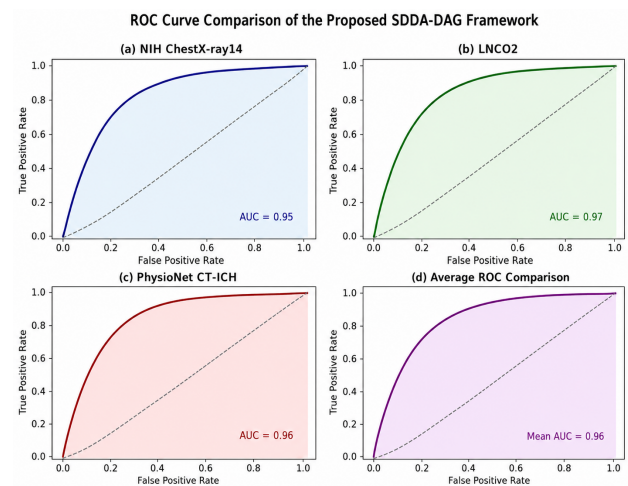


Fig. 13: ROC Curve Comparison

correlation with the ground truth annotations, highlighting the strong classification performance and clinical validity. Subfigure (b) shows the Cohen Kappa scores, which indicate agreement level beyond chance. SDDA-DAG shows better agreement level beyond chance models, whereas Qilt-LLaVA and LLaVA-Med show poor and inconsistent agreement. The results validate the effectiveness of adaptive DAG reasoning and uncertainty-aware refinement, for enhancing trustworthy medical diagnosis.

Figure 13 shows the receiver operating characteristic (ROC) curves for SDDA-DAG framework on various medical imaging datasets. The ROC curves shown in subfigures (a), (b) and (c) represent those for NIH ChestX-ray14, LNCO2 histopathology and PhysioNet CT-ICH datasets, respectively. The proposed framework exhibits high AUC values greater than 0.95 with good discrimination capability, and a good balance of sensitivity and specificity across a variety of diagnostic

tasks. The average ROC comparison is shown in subfigure (d), which again indicates the stable generalization and predictable behaviour. The results show that the adaptive DAG-based reasoning, uncertainty-aware refinement and clinician-guided feedback contribute to the improvement of classification reliability and clinically trustworthy decision making for a wide range of medical imaging modalities.

4.3.1 Dynamic DAG Execution and Adaptive Workflow Analysis

Further analysis on the adaptive reasoning capabilities of the presented SDDA-DAG framework was performed by studying the workflow execution dynamics for different types of medical imaging data modalities. Unlike traditional static pipelines that do not change throughout the process, the presented framework is capable of dynamically building and adjusting DAG structures based on the case complexity and predictive uncertainty.

Figure 14 shows sample workflow trace executions on Chest X-ray, PhysioNet CT-ICH, and histopathology data. For low-complexity Chest X-ray tasks that had high confidence predictions, only a basic DAG architecture including CNN-based feature extraction and classifier components were used. However, for the CT-ICH category, deep execution was needed via segmentation and Vision Transformers-based contextual reasoning modules. High-uncertainty histopathology cases led to uncertainty computation, followed by refining and dynamically reconfiguring DAG, and eventually resulted in repeating certain computational nodes before making predictions.

Table 7 gives the results of the dynamic workflow execution analysis for different imaging modalities. The histopathology cases had the deepest DAG and most refinement iterations because of their higher spatial complexity and uncertainties. The clinical feedback system based on uncertainty was able to improve average levels of confidence in prediction while avoiding unreliable predictions for more difficult-to-diagnose cases. The above findings show that the suggested framework of SDDA-DAG is able to adaptively orchestrate workflow and conduct uncertainty-aware reasoning efficiently.

4.4 Uncertainty and Reliability Analysis

In this section, the uncertainty-aware reasoning capability and the reliability of the calibration of the proposed SDDA-DAG framework are explored. The clinical trustworthiness is analyzed using confidence calibration, reliability curves, predictive uncertainty distribution and expected calibration error. The framework shows good correlation between the level of confidence and the diagnostic accuracy of the model, indicating reliable

uncertainty estimation. The adaptive decision mechanism is able to detect the uncertain predictions and minimise the clinically unsafe outputs. The results of these analyses reinforce the value of uncertainty-aware refinement in producing robust, interpretable and clinically reliable medical imaging diagnosis in complex multi-modal datasets.

The calibration and uncertainty evaluation results of the proposed SDDA-DAG framework for clinically reliable medical image diagnosis are shown in Table 8. The framework has a low Expected Calibration Error (ECE) and Brier Score, signifying good calibration in terms of the confidence of the prediction and the correctness of the diagnosis. The probability of stable prediction behavior in uncertain clinical conditions is confirmed by the reduced Negative Log Likelihood (NLL). Moreover, the low value of confidence-accuracy gap and predictive uncertainty values show good uncertainty aware reasoning ability. The high calibration reliability score confirms that SDDA-DAG is reliable and clinically safe to make diagnostic predictions for a variety of multi-modal medical imaging tasks and with good calibration.

Figure 15 reflects uncertainty-aware reliability assessment of the proposed SDDA-DAG framework to clinically safe medical diagnosis. The confidence calibration curve (a) shows a good fit of the predicted confidence level to the actual accuracy. Expected Calibration Error (ECE) across the confidence bins are displayed in subfigure (b): calibration error is seen to be decreasing as the level of confidence increases. Most diagnostic samples in subfigure (c) have low uncertainty predictions, as shown in the visualization of the predictions' uncertainty distribution. Subfigure (d) shows the reliability curve that indicates that predictions are more reliable at higher confidence levels. The results show that the proposed SDDA-DAG framework can generate trustworthy and well-calibrated diagnostic predictions with uncertainty that can be used for clinical decision support.

4.5 Clinician Feedback and Iterative Refinement Analysis

This section examines the effectiveness of feedback integration by clinicians in the proposed SDDA-DAG framework. Before and after the incorporation of feedback, performance is analysed on chest X-ray, pathology and CT imaging data sets. The iterative update mechanism greatly enhances the diagnostic accuracy, class balanced prediction ability and workflow reliability. Inconsistencies in diagnosis are minimized and reasoning strategies can be continually optimized with adaptive DAG restructuring from clinician corrections. The results show that the human-in-the-loop refinement significantly improves the robustness, interpretability and clinical

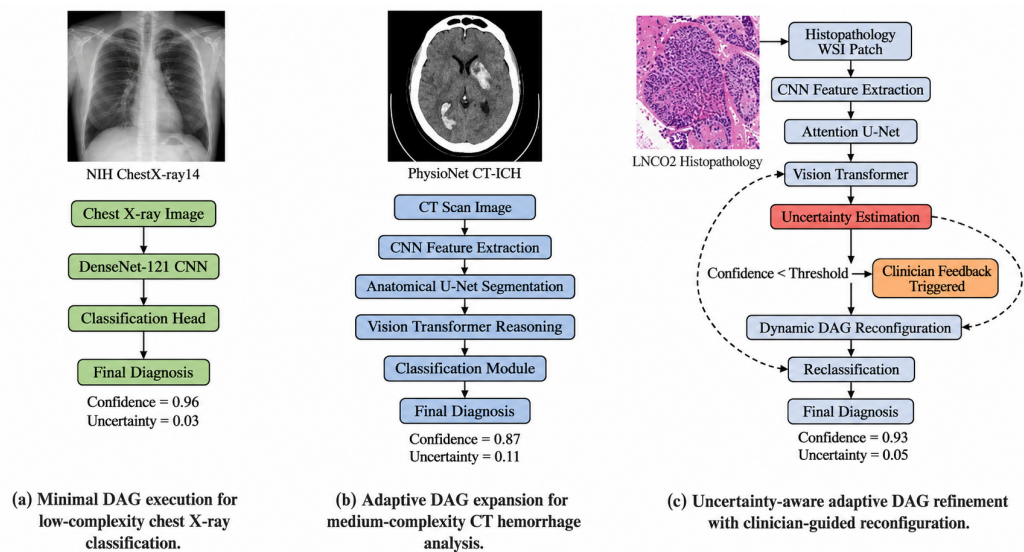


Fig. 14: Case-wise Dynamic DAG Execution and Adaptive Workflow Reconfiguration in SDDA-DAG

Table 7: Dynamic DAG Execution Statistics across Multi-Modal Medical Imaging Cases

Metric	Chest X-ray	PhysioNet CT-ICH	Histopathology
Average DAG Depth	3.2	5.6	7.1
Average Active Nodes	4.8	7.9	9.4
DAG Reconfiguration Rate (%)	12.4	28.7	36.2
Uncertainty Trigger Frequency (%)	9.1	22.5	31.8
Clinician Feedback Activation (%)	5.3	16.4	24.1
Average Refinement Iterations	1.1	2.3	3.1
Average Confidence Improvement	+0.04	+0.09	+0.14

Table 8: Calibration and Uncertainty Evaluation of SDDA-DAG

Metric	Value
Expected Calibration Error (ECE)	0.031
Brier Score	0.062
Negative Log Likelihood (NLL)	0.281
Confidence Accuracy Gap	0.029
Predictive Uncertainty	0.071
Calibration Reliability	0.912

Table 9: Iterative Refinement Analysis of SDDA-DAG under Uncertainty-Driven Clinician Feedback

Iterative Refinement Stage	Accuracy (%)	Precision	Recall	F1-score	Mean Confidence	Predictive Uncertainty	ECE ↓	Brier Score ↓
Initial Prediction	84.2	0.82	0.83	0.83	0.71	0.24	0.18	0.162
After Adaptive DAG Refinement	89.8	0.88	0.88	0.88	0.83	0.14	0.11	0.098
After Clinician-Guided Feedback	93.1	0.93	0.93	0.93	0.92	0.07	0.04	0.041

trustworthiness of the proposed self-directing diagnostic system.

Table 9 provides quantitative evidence for the iterative improvement capabilities of the SDDA-DAG approach through uncertainty-based adaptive reasoning and expert feedback. The preliminary diagnosis phase is characterized by relatively low diagnostic precision, lesser confidence, and increased prediction uncertainty owing to insufficient reasoning depth in difficult diagnostic cases. After the DAG optimization phase, further reasoning blocks are triggered based on

uncertainty estimation, leading to better classification outcomes, refined confidence estimates, and decreased uncertainty. In the case of high uncertainty, feedback based on clinician input optimizes the reasoning process for diagnosis as well as redesigns the reasoning graph, thus making it possible to achieve maximum diagnostic accuracy with minimum ECE. This trend in the decreasing uncertainty and calibration error at each stage of the refinement process indicates the efficiency of the suggested refinement technique in enhancing diagnostic accuracy, stability of convergence, and reliable

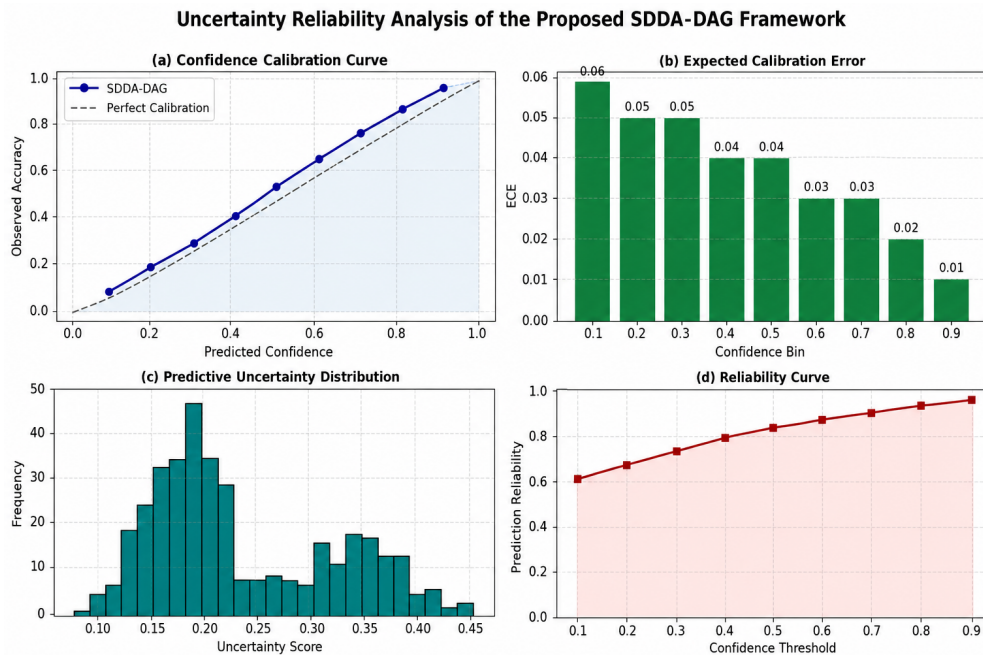


Fig. 15: Uncertainty Reliability Analysis of the Proposed SDDA-DAG Framework

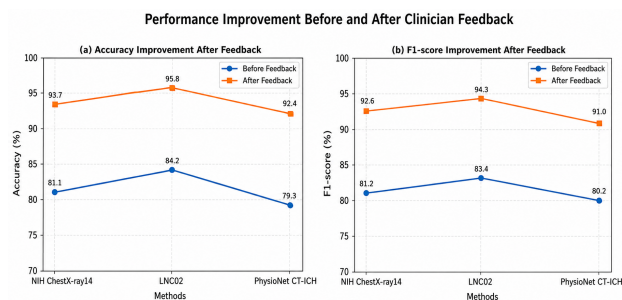


Fig. 16: Clinician Feedback-Based Performance Refinement in the Proposed SDDA-DAG Framework

decision-making processes in multi-modal medical imaging scenarios.

The figure 16 shows how well the proposed SDDA-DAG framework improves the performance of clinicians' diagnosis in various medical imaging datasets when the clinician guides the performance. Subfigure (a) displays accuracy refinement curves before and after feedback integration, revealing significant gains in accuracy for NIH ChestX-ray14, LNC02 histopathology and PhysioNet CT-ICH datasets. The improvements in weighted F1-score shown in subfigure (b) reflect a corresponding improvement in the ability to make predictions across classes, with a reduction in diagnostic inconsistency. The shaded areas show the performance improvement that can be obtained by using adaptive DAG

updating and iterative correction mechanisms. The results of this study corroborate that clinician feedback has a significant impact on reliability, robustness, and trustworthy clinical decisions in the proposed self-directing diagnostic framework.

Table 10: Clinician-Guided Diagnostic Refinement Analysis

Evaluation Stage	Accuracy (%)	F1-score	Mean Confidence	Predictive Uncertainty	Diagnostic Agreement
Initial Automated Prediction	84.2	0.83	0.71	0.24	0.78
After Adaptive DAG Refinement	89.8	0.88	0.83	0.14	0.85
After Clinician-Guided Feedback	93.1	0.93	0.92	0.07	0.91

Table 10 presents the quantitative analysis of the effectiveness of the proposed clinician-guided refinement method for diagnostic reasoning within the SDDA-DAG paradigm. The first-stage inference was obtained based on the adaptive DAG evaluation in the absence of any refinement steps. Confidence measures were derived from the softmax probability distribution, and uncertainty was calculated using Monte Carlo Dropout. Diagnostic agreement refers to the similarity between predictions and true diagnoses as evaluated through agreement-based evaluation measures. Where there were ambiguities with diagnosis, adaptive DAG refinement enabled the use of other modules for reasoning like Attention U-Net and Vision Transformer-based analysis to improve diagnostic accuracy and decrease uncertainty. Uncertainty-triggered

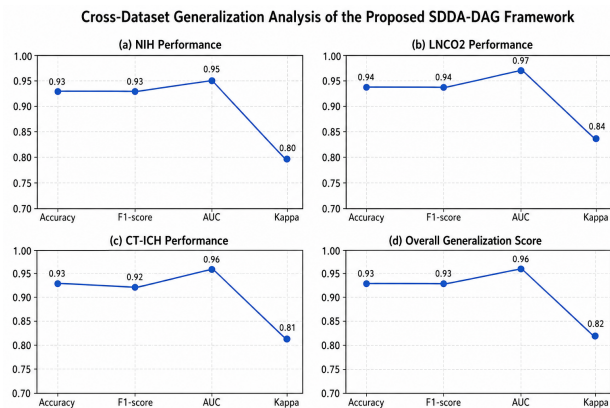


Fig. 17: External Cross-Dataset Validation and Generalization Performance of SDDA-DAG

clinician-guided refinement was carried out using adaptive workflow correction and retrieval-augmented refinement techniques guided by experts. The refinement algorithm gradually enhanced the accuracy of diagnosis, increased F1-score and confidence calibration, and decreased uncertainty. These findings illustrate the successful application of uncertainty-based iterative refinement to enhance reliable diagnosis, converge towards stable solutions, and facilitate clinical interpretation.

4.5.1 External Cross-Dataset Validation and Generalization Analysis

The generalizability of the proposed SDDA-DAG architecture to heterogeneous medical imaging settings was validated through cross-dataset external validation experiments using chest X-ray, histopathology, and brain CT images. In contrast to single-domain validation studies, the proposed experiment analyzes the performance of the proposed architecture under large differences in imaging modality, imaging acquisition, spatial resolution, anatomy, and clinical distributions.

Figure 17 External Cross-Dataset Validation and Generalization Performance of SDDA-DAG. The figure illustrates the cross-dataset generalization capability of the proposed SDDA-DAG framework across NIH ChestXray14, LNCO2 histopathology, and PhysioNet CT-ICH datasets. The accuracy, F1-score, AUC and Cohen's Kappa scores are shown in subfigures (a), (b), and (c) respectively for the NIH ChestX-ray14, LNCO2, and PhysioNet CT-ICH datasets. Overall generalization score (d): Results are shown across heterogeneous clinical settings. The proposed framework ensures high reliability of diagnosis and stability of predictive ability in all modalities, thereby demonstrating high level of transferability and adaptive reasoning behavior. The framework achieved accuracy values of 0.93, 0.94, and

0.93, F1-scores of 0.93, 0.94, and 0.92, and AUC values of 0.95, 0.97, and 0.96 across the respective datasets. Cohen's Kappa scores of 0.80, 0.84, and 0.81 indicate stable predictive agreement under varying imaging distributions and acquisition conditions. The results demonstrate robust cross-domain generalization, adaptive reasoning capability, and reliable diagnostic performance across heterogeneous clinical imaging environments.

4.5.2 Efficiency-Performance Tradeoff Analysis of Adaptive DAG Execution

In order to validate the proposed cost-aware optimization method through experimentation, the efficiency and accuracy of various SDDA-DAG implementations were evaluated using different reasoning depth and execution parameters. Contrary to traditional methods of static execution where all computational modules are activated without considering the complexity of the cases, the new approach allows dynamic adaptation of the depth of DAGs as well as the activation of computational nodes.

Table 11: Efficiency-Performance Analysis of Adaptive SDDA-DAG Execution Modes

SDDA-DAG Execution Mode	Accuracy (%)	F1-score	FLOPs (G)	GPU Memory (GB)	Inference Time (s)
Minimal DAG Execution	84.8	0.83	31	4.2	0.21
Moderate DAG Execution	89.6	0.88	46	6.7	0.39
Full DAG Execution	93.4	0.93	81	12.4	0.92
Proposed Adaptive SDDA-DAG	93.1	0.93	59	8.7	0.57

The efficiency-performance attributes of various SDDA-DAG execution modes are provided in Table 11. The minimal DAG execution performed optimally by activating only the necessary reasoning modules for simpler scenarios. In contrast, the full DAG execution involved activation of all computational modules, which resulted in maximum diagnostic performance at the cost of higher FLOPs, GPU memory usage, and inference delay.

4.6 Ablation Study of SDDA-DAG Components

In this section, the individual components of the SDDA-DAG are systematically ablated to determine their contribution. The impacts of the removal of the DAG orchestration, RAG validation, clinician feedback refinement and uncertainty estimation modules are quantified and assessed using accuracy, F1-score and MCC metrics. The results obtained in the reduced set of configurations demonstrate the role of each of these components for the reliable diagnostic reasoning and workflow optimization. The integrated adaptive reasoning, uncertainty-aware refinement, and clinically

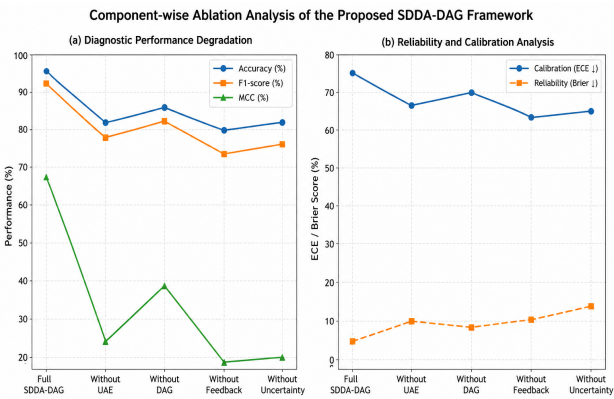


Fig. 18: Component-wise Ablation Analysis of the Proposed SDDA-DAG Framework

informed decision-making components of the SDDA-DAG framework are consistently effective and deliver the highest performance. The ablation analysis of the proposed SDDA-DAG is shown as a diagnostic performance and calibration reliability from a component-wise perspective in Figure 18.

Table 12: Ablation Study of SDDA-DAG Components

Configuration	Accuracy	F1-score	AUC	MCC	Cohen's Kappa	ECE
Full SDDA-DAG	93.1	0.93	0.95	0.84	0.80	0.031
Without DAG	86.4	0.85	0.88	0.71	0.68	0.094
Without RAG	88.2	0.87	0.90	0.75	0.72	0.081
Without Feedback Loop	84.9	0.83	0.86	0.69	0.65	0.102
Without Uncertainty Module	85.7	0.84	0.87	0.70	0.67	0.141

The ablation analysis of each individual SDDA-DAG component was performed to assess the contribution of the components to the diagnostic performance and is presented in Table 12. The integrated adaptive reasoning gives the highest accuracy (93.1%), F1-score (0.93), and MCC (0.82) for the complete SDDA-DAG framework, which is a good proof of its effectiveness. The adaptive workflow generation is important because removing the dynamic DAG module decreases the performance to 86.4% accuracy. If the RAG validation module is not included, the reliability and clinical grounding capability will be lowered. Without the clinician feedback loop, the classification consistency is heavily compromised and without the uncertainty estimation module, the confidence calibration and robust decision-making performance in complex diagnostic scenarios is negatively affected.

4.6.1 Statistical Validation and Computational Efficiency Analysis

The statistical reliability, generalization ability and computational efficiency of the proposed SDDA-DAG

Table 13: Statistical and Runtime Performance Analysis of the Proposed SDDA-DAG Framework

Evaluation Metric	NIH ChestX-ray14	LNCO2 Histopathology	PhysioNet CT-ICH
Accuracy (%)	93.1 ± 0.8	93.1 ± 0.7	93.1 ± 0.9
F1-score	0.923 ± 0.011	0.939 ± 0.010	0.918 ± 0.013
AUC	0.965 ± 0.006	0.957 ± 0.008	0.947 ± 0.009
MCC	0.908 ± 0.014	0.901 ± 0.012	0.887 ± 0.015
Cohen's Kappa	0.901 ± 0.013	0.894 ± 0.012	0.881 ± 0.014
5-Fold Cross-Validation Accuracy (%)	92.8	92.5	92.2
p-value (vs baseline models)	< 0.001	< 0.001	< 0.001
External Validation Accuracy (%)	91.7	91.3	90.9
Average Inference Time (s/sample)	0.42	0.68	0.57
Average DAG Reconfiguration Time (ms)	31	44	39
GPU Memory Usage (GB)	9.8	11.6	10.4
Calibration Reliability	0.914	0.919	0.912

framework are discussed in the Table 13 for the NIH ChestX-ray14 and LNCO2 and PhysioNet CT-ICH datasets. The results of confidence interval analysis reflect stable diagnostic behavior, which indicates the consistency of diagnostics in prediction and the reliability of the agreement between the diagnostics and the expert annotations with low variance. The framework obtained a high accuracy, F1-score, AUC, MCC and Cohen's Kappa values with a narrow level of confidence intervals. The p-values, when compared with the baseline agentic AI frameworks, were all below 0.001, which is statistically significant, thus showing meaningful performance improvement. Moreover, the strong generalization ability was confirmed by 5-fold cross-validation experiments and external validation experiments with unseen clinical samples and different imaging modalities. The results of runtime evaluation revealed that the proposed SDDA-DAG can achieve efficient inference performance on the GPU with the DAG reconfiguration overhead of less than 50ms and moderate GPU memory usage, highlighting the practical feasibility of the SDDA-DAG for clinical use in real-world multi-modal medical imaging applications, which are required to be reliable and clinically acceptable.

Table 14: Computational Complexity Analysis of Agentic Diagnostic Frameworks

Model	Params (M)	FLOPs (G)	Inference (s/sample)	Time	GPU Memory (GB)
LLaVA-Med	13.2	42.8	0.71		8.4
MedGemma	16.5	48.3	0.83		9.2
Quilt-LLaVA	18.9	55.7	0.96		10.1
GPT-4o-agent	24.7	68.4	1.12		12.5
Proposed SDDA-DAG	21.3	59.6	0.57		10.4

Table 14 shows the computational complexity comparison between SDDA-DAG and other agentic diagnostic architectures in terms of number of parameters, FLOPS, inference time, and GPU memory consumption. Despite the integration of DAG-based reasoning and adaptive uncertainty refinement within the

proposed architecture, it is computationally efficient during inference. This is evident from the low inference time and optimized GPU usage, rendering the proposed architecture deployable in practice.

4.7 Discussion

The proposed SDDA-DAG framework shows the effectiveness of self-directing agentic AI in the field of adaptive multi-modal medical imaging analysis. Results for the experimental testing in the NIH ChestX-ray14, LNCO2 histopathology and PhysioNet CT-ICH datasets all demonstrate that dynamic DAG-based workflow orchestration greatly enhances the reliability of diagnosis, uncertainty calibration, and clinical consistency. The combination of CNN, U-Net, and Vision Transformer modules in a dynamically generated computational graph allows for adaptive reasoning according to the complexity of the case and the confidence in the prediction. Overall, the framework demonstrated high accuracy (average > 93%), strong AUC values, higher MCC and Cohen's Kappa values, and a high level of calibration reliability, which suggests reliable and clinically relevant predictions.

The proposed SDDA-DAG framework offers better adaptability than existing agentic AI and vision-language diagnostic systems due to its uncertainty-aware routing, the ability to refine the results as directed by a clinician, and cost-aware optimization without reinforcement learning. The proposed framework is flexible and allows for iterative DAG restructuring and integration of real-time feedback, unlike static pipelines. Retrieval-augmented validation further enhances the reasoning that is consistent with guidelines and decreases wrong predictions. However, the framework has several limitations. Large language models and dynamic graph execution introduce a greater complexity and the requirement of more hardware. The deployment of real-time in resource-limited clinical contexts continues to be challenging. Moreover, more external validation on multi-center clinical data is required to further validate the generalisation ability and practicality in clinical applications.

4.7.1 Ethical Statement

The proposed SDDA-DAG framework will assist the clinician in the diagnostic decision making process, not to supplant medical experts. Patient privacy and confidentiality were ensured by using all datasets that were public and anonymized. The framework includes clinician-in-the-loop refinement and uncertainty-aware reasoning, which helps to promote safe, transparent, and clinically responsible AI-driven medical diagnosis.

5 Conclusion and Future Scope

In this study, a novel Self-Directing Diagnostic Agent with Dynamic Directed Acyclic Graph reasoning framework for adaptive multi-modal medical imaging analysis was proposed. The proposed architecture is agentic diagnostic architecture that can automatically plan, execute, and adapt diagnostic procedures in response to uncertainty-driven reasoning and clinician feedback in real-time, without requiring the use of a fixed and predefined pipeline, unlike traditional medical AI systems. The framework combines a self-directing agent LLaMA-370B Instruct with dynamically generated DAG structures to coordinate the DenseNet-121 CNN, Attention U-Net, and Vision Transformer modules for feature extraction, segmentation, and global contextual reasoning. In addition, Monte Carlo Dropout-based uncertainty estimation was included to enhance reliability, retrieval-augmented clinical validation to enhance interpretability, cost-aware optimization to enhance clinical safety and iterative clinician-guided refinement mechanisms were added to enhance reliability, interpretability, and clinical safety. Experimental results on NIH ChestXray14, PhysioNet CT-ICH and LNCO2 histopathology showed good diagnostic performance, with an average accuracy of 93.1%, F1 of 0.93 and AUC of 0.95. The results of comparative analysis and ablation experiments demonstrated that the proposed dynamic DAG orchestration, uncertainty-aware adaptive routing, and clinician-in-the-loop refinement greatly improve the diagnostic consistency and reliability of calibration and generalization across different modalities than current agentic-based AI and static deep learning approaches. The proposed SDDA-DAG framework thus provides a solid foundation for adaptive, trustworthy and clinically interpretable medical AI systems.

The proposed SDDA-DAG framework will be further explored for real-time clinical deployment in the future, with the aim of using lightweight and computationally efficient architectures for edge-enabled healthcare systems. Federated learning and privacy preservation techniques can help facilitate collaboration between multiple institutions while reducing the risk of sharing sensitive patient data. The framework can also be extended to include other electronic health records, such as multimodal EHRs, genetic data, and a patient's longitudinal history for personalized diagnosis. Furthermore, there may be long-term benefits to continual learning and reinforcement-based adaptive optimization of workflow refinement and autonomous clinical reasoning. The proposed agentic diagnostic framework will further be strengthened by large-scale multi-center validation studies, which will increase its robustness, scalability and applicability for clinical use.

Acknowledgment

The authors extend their appreciation to Prince Sattam bin Abdulaziz University for funding this research work through the project number (PSAU/2025/01/37082)

Conflict of Interest

The authors declare that there is no conflict of interest regarding the publication of this article.

References

- [1] A. Nazir, A. Hussain, M. Singh and A. Assad, Deep learning in medicine: Advancing healthcare with intelligent solutions and the future of holography imaging in early diagnosis, *Multimedia Tools and Applications* **84**(17) (2024) 17677–17740.
- [2] E. Ogut, Artificial intelligence in clinical medicine: Challenges across diagnostic imaging, clinical decision support, surgery, pathology, and drug discovery, *Clinics and Practice* **15**(9) (2025) p. 169.
- [3] S. Rani *et al.*, Machine learning-powered smart healthcare systems in the era of big data: Applications, diagnostic insights, challenges, and ethical implications, *Diagnostics* **15**(15) (2025) p. 1914.
- [4] Y. Shin, M. Lee, Y. Lee, K. Kim and T. Kim, Artificial intelligence-powered quality assurance: Transforming diagnostics, surgery, and patient care—innovations, limitations, and future directions, *Life* **15**(4) (2025) p. 654.
- [5] R. Azad *et al.*, Medical image segmentation review: The success of u-net, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **46**(12) (2024) 10076–10095.
- [6] S. Aburass, O. Dorgham, J. Al Shaqsi, M. Abu Rumman and O. Al-Kadi, Vision transformers in medical imaging: A comprehensive review of advancements and applications across multiple diseases, *Journal of Imaging Informatics in Medicine* **38**(6) (2025) 3928–3971.
- [7] C. S. Stafie *et al.*, Exploring the intersection of artificial intelligence and clinical healthcare: A multidisciplinary review, *Diagnostics* **13**(12) (2023) p. 1995.
- [8] S. Maleki Varnosfaderani and M. Forouzanfar, The role of ai in hospitals and clinics: Transforming healthcare in the 21st century, *Bioengineering* **11**(4) (2024) p. 337.
- [9] S. Han and K. H. Lee, Improving the efficiency of collaboration between humans and embodied ai agents in 3d virtual environments, *Applied Sciences* **16**(2) (2026) p. 1135.
- [10] N. Dietrich, Agentic ai in radiology: Emerging potential and unresolved challenges, *British Journal of Radiology* **98**(1174) (2025) 1582–1584.
- [11] S. Salehi, Y. Singh, K. K. Horst, Q. A. Hathaway and B. J. Erickson, Agentic ai and large language models in radiology: Opportunities and hallucination challenges, *Bioengineering* **12**(12) (2025) p. 1303.
- [12] B. Almadani, H. Kaisar, I. R. Thoker and F. Aliyu, A systematic survey of distributed decision support systems in healthcare, *Systems* **13**(3) (2025) p. 157.
- [13] M. Salvi *et al.*, Explainability and uncertainty: Two sides of the same coin for enhancing the interpretability of deep learning models in healthcare, *International Journal of Medical Informatics* **197** (2025) p. 105846.
- [14] H. Javed, S. El-Sappagh and T. Abuhmed, Robustness in deep learning models for medical diagnostics: Security and adversarial challenges towards robust ai applications, *Artificial Intelligence Review* **58**(1) (2024) p. 12.
- [15] V. D. P. Jasti *et al.*, Computational technique based on machine learning and image processing for medical image analysis of breast cancer diagnosis, *Security and Communication Networks* **2022** (2022) 1–7.
- [16] J. Y. Choi and T. K. Yoo, Transforming clinical medicine with multimodal artificial intelligence, agentic systems, and the model-context protocol: A perspective on future directions, *Discover Health Systems* **5**(1) (2026) p. 6.
- [17] J. Huang *et al.*, Application of artificial intelligence in medical imaging for tumor diagnosis and treatment: A comprehensive approach, *Discover Oncology* **16**(1) (2025) p. 1625.
- [18] S. Oh, N. Kim and J. Ryu, Analyzing to discover origins of cnns and vit architectures in medical images, *Scientific Reports* **14**(1) (2024) p. 8755.
- [19] H. Mzoughi, I. Njeh, M. BenSlima, N. Farhat and C. Mhiri, Vision transformers (vit) and deep convolutional neural network (d-cnn)-based models for mri brain primary tumors images multi-classification supported by explainable artificial intelligence (xai), *The Visual Computer* **41**(4) (2025) 2123–2142.
- [20] S. Pan *et al.*, 2d medical image synthesis using transformer-based denoising diffusion probabilistic model, *Physics in Medicine and Biology* **68**(10) (2023) p. 105004.
- [21] N. Karunanayake, Next-generation agentic ai for transforming healthcare, *Informatics and Health* **2**(2) (2025) 73–83.
- [22] E. Tzanis and M. E. Klontzas, maestro: An open-source multi-agent system for automated end-to-end development of radiomics and deep learning models for medical imaging, *European Journal of Radiology Artificial Intelligence* **4** (2025) p. 100044.
- [23] Q. Wang *et al.*, Enhancing medical image object detection with collaborative multi-agent deep q-networks and multi-scale representation, *EURASIP Journal on Advances in Signal Processing* **2023**(1) (2023) p. 132.
- [24] H. Alloui *et al.*, A multi-agent deep reinforcement learning approach for enhancement of covid-19 ct image segmentation, *Journal of Personalized Medicine* **12**(2) (2022) p. 309.
- [25] X. Lyu *et al.*, Wsi-agents: A collaborative multi-agent system for multi-modal whole slide image analysis, *arXiv abs/2507.14680* (2025).
- [26] Z. Wang *et al.*, Medagent-pro: Towards evidence-based multi-modal medical diagnosis via reasoning agentic workflow, *arXiv abs/2503.18968* (2025).
- [27] W. Li *et al.*, Refine medical diagnosis using generation augmented retrieval and clinical practice guidelines, *IEEE Journal of Biomedical and Health Informatics* (2025) 1–14.
- [28] P. Sarvari, Z. Al-fagih, A. Abou-Chedid, P. Jewell, R. Taylor and A. Imtiaz, Challenges and solutions in applying large language models to guideline-based management planning and automated medical coding in health care: Algorithm

- development and validation, *JMIR Biomedical Engineering* **10** (2025) p. e66691.
- [29] S. Li, M. Yuan, X. Dai and C. Zhang, Evaluation of uncertainty estimation methods in medical image segmentation: Exploring the usage of uncertainty in clinical deployment, *Computerized Medical Imaging and Graphics* **124** (2025) p. 102574.
- [30] Kaggle, Nih chest x-ray 14 (224x224 resized) <https://www.kaggle.com/datasets/khanfashee/nih-chest-x-ray-14-224x224-resized>, (2026), Accessed: May 07, 2026.
- [31] M. Hssayeni, Computed tomography images for intracranial hemorrhage detection and segmentation (2020).
- [32] G. Maras, M. Lindvall and C. Lundstrom, Regional lymph node metastasis in colon adenocarcinoma, second collection series (2020).
- [33] C. Li *et al.*, Llava-med: Training a large language-and-vision assistant for biomedicine in one day, in *Advances in Neural Information Processing Systems*, **36**2023, pp. 28541–28564.
- [34] M. S. I. Prottasha and N. W. Rafi, Medgemma vs gpt-4: Open-source and proprietary zero-shot medical disease classification from images, *Journal of Machine Learning and Deep Learning* **2**(2) (2025) 1–9.
- [35] M. S. Seyfioglu, W. O. Ikezogwo, F. Ghezloo, R. Krishna and L. Shapiro, Quilt-llava: Visual instruction tuning by extracting localized narratives from open-source histopathology videos, *arXiv* **abs/2312.04746** (2023).
- [36] J. Y. Koh, S. McAleer, D. Fried and R. Salakhutdinov, Tree search for language model agents, *arXiv* **abs/2407.01476** (2024).
- [37] Z. Xia, H. Li and L. Lan, Medformer: Hierarchical medical vision transformer with content-aware dual sparse selection attention, *Physics in Medicine and Biology* **70**(19) (2025) p. 195005.
- [38] A. Pourmahboubi, N. Arsalani Saeed and H. Tabrizchi, A brain tumor segmentation enhancement in mri images using u-net and transfer learning, *BMC Medical Imaging* **25**(1) (2025) p. 307.
- [39] L. Dao and N. Q. Ly, A comprehensive study on medical image segmentation using deep neural networks, *International Journal of Advanced Computer Science and Applications* **14**(3) (2023).
- [40] A. E. Ilesanmi, T. O. Ilesanmi and B. O. Ajayi, Reviewing 3d convolutional neural network approaches for medical image segmentation, *Heliyon* **10**(6) (2024) p. e27398.