Frontiers in Computer Science and Artificial Intelligence
ISSN: 2978-8048
DOI: 10.32996/fcsai
Journal Homepage: www.al-kindipublisher.com/index.php/fcsai

FCSAI
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Smarter Health for Everyone with an AI System That Detects Four Diseases and Gives Easy-to-Understand Clinical Advice

Md Anwar Hossain[1], Md Abedur Rahman[1], Md Sahid Hossain[1], and Kallol Chakraborty Shekhor[2]
[1] Master's in Computer Science, Maharishi International University, Fairfield, IA, USA
[2] Master of Science in Information Studies, Trine University, Allen Park, MI, USA

| ABSTRACT

Medical-image artificial intelligence models are commonly developed for single diseases or isolated imaging modalities, which limits shared representation learning, calibration-aware prediction, and clinically usable decision-support reporting. This study proposes PolyDx-Mamba, a unified multi-disease medical image classification framework designed strictly for clinician-assisted decision support rather than autonomous diagnosis. The framework combines modality-specific stem and patch embedding, a shared Vision-Mamba hierarchical encoder, per-stage channel-spatial attention, cross-disease selective state-space global context aggregation, gated multi-scale feature fusion, uncertainty-aware latent representation learning, and task-specific classification heads. PolyDx-Mamba was evaluated across four clinically distinct imaging tasks: brain tumor MRI classification, skin lesion classification using dermoscopy, COVID-19 chest X-ray classification, and diabetic retinopathy grading from fundus images. A confidence-conditioned Gemini-1.5-Pro retrieval-augmented generation advisor was further incorporated to generate clinician-facing recommendation reports grounded in retrieved clinical context. PolyDx-Mamba achieved accuracies of 0.9926, 0.9209, 0.9880, and 0.8718 across the four datasets, with an average accuracy of 0.9433. Macro-F1 scores were 0.9917, 0.9151, 0.9863, and 0.8597, yielding an average macro-F1 of 0.9382. The model also demonstrated favorable calibration, with expected calibration error values of 0.0162, 0.0307, 0.0208, and 0.0421 across the respective tasks. Statistical testing confirmed significant performance improvements over all evaluated baselines, including the strongest Mamba-based comparator. The Gemini-1.5-Pro confidence-RAG advisor produced clinically relevant and guideline-aligned recommendation reports, with high factual accuracy, low hallucination rate, and strong safety compliance. These findings suggest that PolyDx-Mamba provides a calibrated and statistically supported multi-disease image classification framework with retrieval-grounded clinical advisory support. However, all predictions and generated recommendations require clinician confirmation, and external prospective validation remains necessary before any clinical deployment.

| KEYWORDS

PolyDx Mamba, medical image classification, Vision Mamba, clinical decision support, uncertainty calibration, retrieval augmented generation

| ARTICLE INFORMATION

ACCEPTED: 15 June 2025 **PUBLISHED:** 17 July 2025 **DOI:** 10.32996/fcsai.2025.4.4.9

1. Introduction

Medical imaging constitutes a fundamental component of contemporary clinical screening, diagnostic evaluation, referral prioritization, treatment planning, and longitudinal disease monitoring [1-3]. Brain magnetic resonance imaging supports the classification of intracranial tumor categories, including glioma, meningioma, pituitary tumor, and no tumor cases [4-5]. Dermoscopy enables magnified visualization of cutaneous lesions and assists in the assessment of malignant and benign lesion patterns [6]. Chest X ray imaging remains one of the most accessible modalities for respiratory disease evaluation, including pneumonia, tuberculosis, COVID 19, and normal findings. Fundus photography supports diabetic retinopathy grading and

contributes to ophthalmic screening and referral decisions. Although these imaging modalities differ in acquisition mechanism, anatomical structure, visual appearance, and clinical interpretation pathway, they share a common requirement for accurate image level classification, calibrated confidence estimation, and clinically meaningful decision support.Deep learning has substantially improved medical image classification through convolutional neural networks, vision transformers, hybrid convolution and attention architectures, and recent state space visual models [1, 2, 3, 4]. Convolutional architectures such as EfficientNetV2 and ConvNeXt V2 provide effective local feature extraction and hierarchical visual representation learning [5, 6]. Transformer based and hybrid architectures, including ViT, Swin V2, MaxViT, CoAtNet, and MedViT, enhance global context modeling and long range dependency representation [7, 8, 9, 10, 11]. More recently, selective state space models such as VMamba, MambaVision, and MedMamba have introduced an efficient alternative for spatial sequence modeling and extended contextual representation [12, 13, 14]. Despite these advances, most medical image classification systems are still developed for a single disease, a single modality, or a narrowly defined clinical endpoint. This restriction limits the development of shared representations across related medical imaging tasks and contributes to fragmented artificial intelligence workflows, in which each disease category requires an independent model, calibration process, and reporting pipeline.

To address these challenges, this study proposes PolyDx Mamba, a unified multi disease medical image classification framework with an integrated Gemini 1.5 Pro confidence conditioned retrieval augmented clinical advisor. PolyDx Mamba is designed strictly as a clinical decision support framework and not as an autonomous diagnostic device. Its predictions, confidence estimates, uncertainty outputs, and generated clinical recommendations require review and confirmation by qualified clinicians before any clinical action. The framework is evaluated on four public medical imaging datasets: Brain Tumor MRI for four class brain tumor classification, HAM10000 for seven class skin lesion classification, COVID 19 Radiography for four class chest disease classification, and APTOS 2019 for five class diabetic retinopathy grading. These datasets represent four distinct imaging modalities, namely magnetic resonance imaging, dermoscopy, chest radiography, and fundus photography, and provide a heterogeneous evaluation setting for unified multi disease image classification.The use of a Vision Mamba encoder is motivated by the need to model both localized discriminative findings and long range spatial dependencies in medical images. Brain MRI interpretation may depend on tumor morphology, anatomical location, and surrounding tissue context. Dermoscopic assessment may require evaluation of asymmetry, border irregularity, color variation, and global lesion structure. Chest X ray classification may require recognition of focal or diffuse lung field abnormalities. Fundus based diabetic retinopathy grading may depend on multiple lesion types distributed across the retina. A selective state space visual encoder is therefore suitable for capturing extended spatial dependencies, while channel spatial attention supports local feature recalibration. The gated multi scale fusion module further integrates features from different semantic levels, which is important because disease relevant cues may appear at different anatomical scales and visual resolutions across modalities.The main contributions of this study are as follows. First, this study proposes PolyDx Mamba, a unified multi disease image classification framework that supports four heterogeneous medical imaging tasks within a shared Vision Mamba architecture. Second, the framework introduces an integrated representation learning design that combines modality specific stem and patch embedding, hierarchical selective state space encoding, per stage channel spatial attention, cross disease global context aggregation, gated multi scale feature fusion, and uncertainty aware latent representation learning. Third, the proposed model is evaluated on four public datasets covering brain MRI, dermoscopy, chest X ray imaging, and fundus photography using task specific classification metrics and calibration analysis. Fourth, the study integrates a Gemini 1.5 Pro confidence conditioned retrieval augmented clinical advisor that generates structured clinician facing recommendation reports grounded in retrieved clinical context. Finally, the proposed system is explicitly positioned as a decision support framework requiring clinician confirmation, while emphasizing the need for external prospective validation before clinical deployment.The main contributions of this study are as follows:

1. This study proposes PolyDx-Mamba, a unified multi-disease medical-image classification framework for Brain Tumor MRI, HAM10000 skin dermoscopy, COVID-19 Radiography chest X-ray, and APTOS-2019 fundus imaging.
2. The proposed architecture combines modality-specific stem/patch embedding, a shared Vision-Mamba hierarchical encoder, per-stage channel-spatial attention, cross-disease selective state-space global context aggregation, gated multi-scale feature fusion, uncertainty-aware latent representation learning, and four task-specific classification heads.
3. The framework is designed to support heterogeneous medical imaging tasks while preserving task-specific output spaces for brain tumor classification, skin lesion classification, chest disease classification, and diabetic retinopathy grading.
4. The study provides a unified experimental comparison against recent CNN, transformer, hybrid, medical vision transformer, and Mamba-based baselines under a common implementation protocol.
5. The evaluation includes not only accuracy and Macro-F1, but also clinically relevant metrics such as AUC, sensitivity, specificity, precision, negative predictive value, and expected calibration error.
6. The study includes an ablation analysis to examine the contribution of the Mamba encoder, modality-specific stem, channel-spatial attention, gated fusion, and uncertainty-aware prediction module.

7. The framework integrates a Gemini-1.5-Pro confidence-conditioned retrieval-augmented clinical advisor to generate structured recommendation reports while maintaining a clinical decision-support-only role that requires clinician confirmation.

2. Related Work

Cai et al. 2020 [16] reviewed the role of deep learning and intelligent imaging in medical big-data analysis, with emphasis on medical image classification and segmentation for early disease diagnosis; although the study provided a broad overview of clinical imaging applications, it mainly synthesized existing progress rather than proposing a task-specific model. Huang et al. 2023 [17] systematically examined self-supervised learning for medical image classification across studies published from 2012 to 2022, including 79 selected papers from 412 screened records, and provided implementation guidance for reducing dependence on costly expert annotations; however, its contribution remained methodological guidance rather than a unified classification framework. Manzari et al. 2023 [18] introduced MedViT, a CNN–Transformer hybrid model designed to combine local feature extraction with global representation learning while improving robustness against adversarial perturbations through smoother decision boundaries, showing strong robustness and generalization on MedMNIST-2D datasets; nevertheless, its validation was centered on standardized benchmark datasets. Dai et al. 2021 [19] proposed TransMed for multi-modal medical image classification by integrating CNN-based low-level feature extraction with Transformer-based long-range dependency modeling across modalities, achieving average accuracy improvements of 10.1% and 1.9% on parotid gland tumor and knee injury classification datasets, respectively; its relevance lies in demonstrating the value of hybrid CNN–Transformer fusion for multi-modal medical imaging. Liu et al. 2020 [20] developed a relation-driven semi-supervised self-ensembling framework that exploited unlabeled medical images by enforcing sample relation consistency under perturbations, outperforming several semi-supervised methods on ISIC 2018 skin lesion diagnosis and ChestX-ray14 thorax disease classification; however, the method primarily addressed label scarcity rather than broader multi-task or multi-modal deployment.

Yang et al. 2023 [21] proposed DiffMIC, a dual-guidance diffusion network for general medical image classification that reduces noise and perturbation while enhancing semantic representation through multi-granularity conditional guidance and Maximum-Mean Discrepancy regularization; its evaluation on ultrasound placental maturity grading, dermatoscopic skin lesion classification, and fundus diabetic retinopathy grading demonstrated the potential of diffusion models beyond generative imaging, although the work focused on classification rather than clinical reporting or multi-task decision support. Tang and Hu 2020 [22] investigated semi-supervised machine-learning-based medical image classification for CT pathological images of the brain and chest, showing improved performance under limited labeled data and reduced overfitting compared with conventional methods; however, its contribution remained centered on sample-efficient classification rather than modern attention-based or transformer-driven architectures. Wu et al. 2023 [23] introduced CTransCNN, a hybrid CNN–Transformer framework for multilabel medical image classification that incorporates multilabel multihead attention, multibranch residual learning, and information interaction between branches to address label dependency, imbalance, and weak CNN–Transformer communication; experiments on ChestX-ray11, NIH ChestX-ray14, and a traditional Chinese medicine tongue dataset showed competitive generalization for multilabel settings. Wang et al. 2023 [24] presented a real-world multi-institutional benchmark for adapting foundation models to medical image classification, comprising 22,349 images across thoracic X-ray screening, pathology tissue screening, endoscopy lesion detection, neonatal jaundice evaluation, and diabetic retinopathy grading; its main relevance lies in establishing diverse clinical evaluation settings, although it primarily served as a dataset and benchmark rather than a new diagnostic architecture. Azizi et al. 2021 [25] examined large self-supervised models for medical image classification and introduced Multi-Instance Contrastive Learning to exploit multiple images from the same patient case, improving dermatology classification top-1 accuracy by 6.7% and chest X-ray mean AUC by 1.1% over strong ImageNet-supervised baselines; the study highlighted the value of domain-specific unlabeled pretraining, especially when labeled medical data are scarce.

Ashraf et al. 2020 [26] proposed a fine-tuned deep convolutional neural network for big-data medical image classification, targeting the semantic gap between low-level image descriptors and high-level clinical interpretation across multiple body-organ imaging categories; however, the study mainly relied on transfer learning with limited architectural novelty beyond fine-tuning the final layers. Patrício et al. 2023 [27] surveyed explainable deep learning methods for medical image classification, covering visual, textual, example-based, and concept-based explanations, as well as datasets and metrics for evaluating explanation quality; its relevance lies in emphasizing interpretability as a prerequisite for clinical adoption, although it remained a review rather than an empirical diagnostic model. Cheng et al. 2022 [28] introduced ResGANet, a residual group attention network that captures channel-wise and spatial feature dependencies through modular group attention blocks, achieving stronger feature representation with 1.51–3.47 times fewer parameters than standard ResNet backbones and improving both classification and segmentation applications; nevertheless, it remained primarily a CNN-based attention backbone. Korot et al. 2021 [29] evaluated six code-free cloud-based deep learning platforms for multi-modality medical image classification using four representative imaging datasets, reporting mean F1 scores of 93.9 for Amazon, 92.0 for Google, 90.7 for MedicMind, 88.6 for Microsoft, 74.2 for Clarifai, and 72.0 for Apple; while useful for accessibility and baseline development, such platforms require

careful validation before clinical use. An et al. 2021 [30] proposed a visual-attention-based multiscale CNN for medical image classification, integrating attention-guided feature extraction with Mahalanobis-distance-based optimization to improve robustness on lung nodule and breast cancer image classification; however, the method focused on selected imaging tasks and did not address broader multi-modal or multi-disease generalization.

Singh et al. 2021 [31] proposed MetaMed, a gradient-based meta-learning framework for few-shot medical image classification, addressing long-tailed distributions and limited expert annotations using Pap smear, BreakHis, and ISIC 2018 datasets; the model achieved more than 70% accuracy and gained 2–5% generalization improvement from CutOut, MixUp, and CutMix, although its focus remained on low-shot adaptation rather than full-scale multi-disease modeling. Kowsari et al. 2020 [32] introduced HMIC, a hierarchical deep learning approach that decomposes medical image classification into parent- and child-level decisions, demonstrated on small-bowel biopsy images for celiac disease, environmental enteropathy, and normal controls, with severity grading at the child level; its relevance lies in structured clinical categorization, though it was evaluated within a narrow biopsy-specific hierarchy. Ren et al. 2023 [33] developed UKSSL, an underlying-knowledge-based semi-supervised framework that uses MedCLR for unlabeled representation learning and UKMLP for classification with limited labeled data, achieving 98.9% accuracy on LC25000 and 94.1% accuracy on BCCD using only 50% labeled data; however, its contribution mainly addressed label efficiency rather than multi-modal integration. Müller et al. 2022 [34] analyzed ensemble learning strategies for deep CNN-based medical image classification across four datasets, showing that stacking improved F1-score by up to 13%, cross-validation bagging by up to 11%, and augmenting by up to 4%; while ensembles improved robustness and performance, they may increase pipeline complexity compared with single-model architectures. Almalik et al. 2022 [35] proposed SEViT, a self-ensembling Vision Transformer designed to improve robustness against adversarial attacks by combining predictions from intermediate and final ViT representations, with experiments on chest X-ray and fundoscopy modalities; the study is relevant for safety-aware medical classification, although it specifically targeted adversarial robustness rather than broader diagnostic generalization.

3.1 Dataset Description

This study evaluated PolyDx-Mamba on four publicly available medical imaging datasets covering four clinically distinct image-classification tasks: brain tumor classification from MRI, skin lesion classification from dermoscopy, chest disease classification from chest X-ray, and diabetic retinopathy grading from fundus photography. The dataset configuration is summarized in Table 1. The selected datasets differ in imaging modality, disease phenotype, class structure, sample size, and native image resolution, providing a heterogeneous benchmark for evaluating a unified multi-disease medical image classification framework.The Brain Tumor MRI dataset was used for 4-class brain tumor classification. It contains 7,023 MRI images and includes brain tumor and no-tumor categories. In this study, the dataset was divided into 5,140 training images, 572 validation images, and 1,311 testing images. The original image resolutions were varied; therefore, all images were standardized during preprocessing before model input. This dataset was included to evaluate the ability of the proposed framework to classify structural abnormalities in MRI, where discriminative patterns may depend on lesion morphology, anatomical location, and surrounding tissue context

The HAM10000 dataset was used for 7-class skin lesion classification from dermoscopic images. It contains 10,015 dermoscopy images with an original resolution of 600×450. In this study, 7,010 images were used for training, 1,502 for validation, and 1,503 for testing. HAM10000 is a widely used dermoscopy benchmark containing multi-source images of common pigmented skin lesions. It was included because dermoscopic classification requires recognition of fine-grained color, texture, border, and lesion-structure patterns, making it an important modality for evaluating local and global visual representation learning .The COVID-19 Radiography dataset was used for 4-class chest disease classification from chest X-ray images. It contains 21,165 chest X-ray images with a native resolution of 299×299. The study split consisted of 14,815 training images, 2,118 validation images, and 4,232 testing images. This dataset was included to evaluate model performance on radiographic images, where disease-related findings may appear as localized or diffuse lung-field abnormalities and where global thoracic context is clinically important.

The APTOS 2019 dataset was used for 5-class diabetic retinopathy grading from fundus images. It contains 3,662 fundus photographs with varied original resolutions. In this study, 2,563 images were assigned to the training set, 366 images to the validation set, and 733 images to the testing set. The dataset was included to assess the ability of PolyDx-Mamba to model retinal disease severity from fundus photographs, where grading may depend on multiple lesion types and their spatial distribution across the retina [4].For all datasets, images were resized to the common input size defined in the implementation protocol, while preserving the task-specific class structure. The train, validation, and test sets were kept separate throughout model development and evaluation. The training set was used for model optimization, the validation set was used for model selection and monitoring, and the test set was reserved for final performance assessment. Because the four datasets differ substantially in modality and image characteristics, modality-specific preprocessing and augmentation were applied as described in the implementation section. This design allowed the proposed framework to be evaluated under a unified protocol while maintaining the clinical and visual specificity of each disease task.

Table 1. Dataset characteristics, task definitions, and train–validation–test splits across four medical imaging modalities

Dataset	Modality	Disease Task	Classes	Images	Train / Val / Test	Resolution
Brain Tumor MRI (Kaggle)	MRI	Brain tumor	4	7,023	5,140 / 572 / 1,311	varied
HAM10000	Dermoscopy	Skin lesion	7	10,015	7,010 / 1,502 / 1,503	600×450
COVID-19 Radiography	Chest X-ray	Chest disease	4	21,165	14,815 / 2,118 / 4,232	299×299
APTOS 2019	Fundus	DR grading	5	3,662	2,563 / 366 / 733	varied

3.2 Dataset Preprocessing and Augmentation

All datasets were processed using a unified preprocessing pipeline while preserving modality-specific image characteristics. Because the four datasets differ in acquisition type, native resolution, intensity distribution, color space, and anatomical structure, preprocessing was designed to standardize the model input without removing clinically relevant visual information. Each image was first checked for readability and converted into a consistent input format. Images were then resized to 224×224 pixels to match the input requirement of the proposed PolyDx-Mamba framework. For grayscale medical images, including MRI and chest X-ray samples, channel formatting was standardized before model input. For color-based modalities, including dermoscopy and fundus photography, RGB information was preserved because color variation is clinically relevant for skin lesion assessment and diabetic retinopathy grading.Intensity normalization was applied to reduce acquisition-related variation and to stabilize model optimization. For MRI and chest X-ray images, normalization helped reduce differences in scanner intensity, contrast, and exposure. For dermoscopy and fundus images, normalization was applied carefully to preserve lesion color, vascular patterns, retinal appearance, and other disease-relevant visual cues. No preprocessing step was designed to introduce diagnostic information not present in the original image. The purpose of preprocessing was to improve input consistency, reduce avoidable technical variability, and support fair training across heterogeneous modalities.

Data augmentation was applied only to the training set to improve robustness and reduce overfitting. Validation and test images were not augmented, ensuring that model selection and final evaluation were performed on unmodified images after standard resizing and normalization. The augmentation strategy included random flipping, random rotation within ±15°, ColorJitter, random erasing, and modality-specific transformations. Random flipping was used to improve invariance to anatomical or acquisition orientation when clinically appropriate. Random rotation within ±15° was applied to simulate mild positioning differences while avoiding unrealistic geometric distortion. ColorJitter was used primarily for color-based modalities to improve robustness to illumination and acquisition variation. Random erasing was used as a regularization strategy to encourage the model to rely on distributed disease-relevant patterns rather than a single localized region.Modality-specific augmentation was applied according to the visual properties of each dataset. For Brain Tumor MRI, augmentation was kept conservative to avoid altering lesion morphology or anatomical structure. For chest X-ray images, augmentation was designed to preserve thoracic anatomy while improving robustness to acquisition variability. For dermoscopy images, color and appearance variation were considered important because skin tone, lighting, and dermoscopic acquisition conditions may differ across samples. For fundus images, augmentation was applied while preserving retinal structures and lesion patterns relevant to diabetic retinopathy grading. This modality-aware strategy ensured that augmentation improved generalization without creating clinically implausible images.Class imbalance was addressed using class re-weighting and balanced sampling during training. This was important because medical imaging datasets often contain unequal class distributions, and minority disease categories may otherwise contribute less to the optimization objective. Class re-weighting increased the contribution of under-represented classes to the loss function, while balanced sampling improved the probability that each mini-batch contained informative examples from less frequent categories. These strategies were used only during training and did not alter the validation or test distributions.

4.1Methodology Overview

The proposed study follows a unified multi-disease medical image classification pipeline composed of three main components: baseline model benchmarking, the proposed PolyDx-Mamba classifier, and a confidence-conditioned Gemini-1.5-Pro clinical advisory layer. First, all datasets are processed through a common preprocessing and augmentation pipeline and evaluated using recent baseline architectures, including CNN-based, transformer-based, hybrid, and Mamba-based models, to ensure a fair comparison under the same experimental protocol. Second, PolyDx-Mamba is introduced as the proposed classification model. It uses modality-specific stem and patch-embedding layers to adapt MRI, dermoscopy, chest X-ray, and fundus images before passing them into a shared Vision-Mamba hierarchical encoder. The encoder is strengthened with per-stage channel-spatial attention, cross-disease selective state-space global context aggregation, gated multi-scale feature fusion, and an uncertainty-aware shared latent representation. Four task-specific heads then generate predictions for brain tumor classification, skin lesion classification, chest disease classification, and diabetic retinopathy grading. Finally, the predicted class, model confidence, uncertainty estimate, and modality context are passed to a Gemini-1.5-Pro retrieval-augmented clinical advisor. This advisor retrieves relevant guideline snippets and generates a structured clinical recommendation report. The generated report is intended only for clinical decision support and must be confirmed by qualified clinicians before any clinical decision or intervention.

4.2 Proposed PolyDx-Mamba Architecture

The proposed PolyDx-Mamba architecture is designed as a unified multi-disease medical image classification model for heterogeneous imaging modalities, including Brain MRI, skin dermoscopy, chest X-ray, and fundus imaging. As illustrated in Figure 1, the framework follows a structured top-down pipeline consisting of input standardization, modality-aware tokenization, hierarchical Vision-Mamba feature learning, channel-spatial attention refinement, cross-scale global context modeling, gated multi-scale feature fusion, shared latent representation learning, and task-specific classification heads. The purpose of this design is to enable shared disease-aware representation learning across multiple imaging tasks while preserving the modality-specific and disease-specific information required for accurate classification.The first stage of the architecture receives multi-image health disease inputs from four modalities. Brain MRI images are used for brain tumor classification, skin dermoscopy images are used for skin lesion classification, chest X-ray images are used for chest disease classification, and fundus images are used for diabetic retinopathy grading. These modalities differ substantially in visual appearance, anatomical structure, intensity distribution, color information, and disease morphology. Therefore, the model does not directly merge raw images into a single representation space. Instead, each image first passes through a unified preprocessing and augmentation stage that standardizes the input size, normalizes image intensity, applies denoising where appropriate, and performs modality-specific augmentation. This step reduces non-clinical variation while preserving disease-relevant visual cues.

After preprocessing, the image is passed into the Mamba stem. The Mamba stem performs patch or token embedding and early downsampling. This stage converts the input image into a compact visual token representation that can be processed by the hierarchical encoder. The stem is important because raw medical images have different resolutions and acquisition characteristics. By converting each input into a standardized token representation, the model can process MRI, dermoscopy, chest X-ray, and fundus images within a common architectural framework. At the same time, the modality-aware input design allows the model to retain modality-specific visual characteristics before entering the shared encoder.The shared encoder is composed of four hierarchical Mamba blocks. Each stage captures visual information at a different level of abstraction. The first Mamba stage focuses on local feature modeling and captures low-level image patterns such as edges, boundaries, small texture variations, and early lesion-related structures. The second stage learns hierarchical feature representations and begins to encode more organized anatomical and pathological patterns. The third stage models longer-range dependencies, allowing the network to capture spatial relationships across wider image regions. The fourth stage produces high-level semantic abstraction, representing global disease-relevant information that is more directly useful for classification. This hierarchical organization allows PolyDx-Mamba to learn both local and global features, which is necessary because disease findings may appear at different spatial scales across modalities.

A key component of the proposed encoder is the use of Vision-Mamba blocks for feature extraction. Unlike purely convolutional models, which primarily emphasize local neighborhoods, the Mamba-based encoder is designed to model extended spatial dependencies across the image. This is particularly important in medical imaging. Brain tumor classification may require understanding lesion position and surrounding tissue structure. Skin lesion classification may depend on both local texture and global lesion asymmetry. Chest X-ray classification may require assessment of diffuse lung-field abnormalities. Diabetic retinopathy grading may depend on small retinal lesions distributed across the fundus image. The Mamba encoder therefore provides a suitable backbone for learning long-range disease patterns while maintaining an efficient hierarchical structure.After each Mamba stage, a channel-spatial attention module is applied to refine the extracted feature maps. These attention modules produce the stage-wise refined features denoted as (f_1), (f_2), (f_3), and (f_4). Channel attention emphasizes informative feature

channels, while spatial attention highlights disease-relevant regions within the image. This refinement is useful because medical images often contain irrelevant background structures, acquisition artifacts, or normal anatomical patterns that may distract the classifier. By applying attention after each stage, the model can progressively strengthen important visual responses and suppress less informative features. The four refined feature levels preserve information from early, intermediate, and high-level representation stages.

The cross-scale global context module receives information from the hierarchical encoder and performs selective state-space aggregation. This module is designed to capture global disease context across multiple feature levels. While the stage-wise Mamba blocks learn progressively deeper features, the cross-scale global context module encourages interaction across scales and supports the modeling of broader disease patterns. This is important for a multi-disease framework because different tasks rely on different contextual cues. For example, dermoscopy may require global lesion symmetry, chest X-ray may require bilateral lung-field assessment, and fundus grading may require distributed retinal lesion recognition. The cross-scale context module helps the model integrate these broader spatial relationships before final classification.The multi-scale feature fusion module combines the refined stage-wise features (f_1), (f_2), (f_3), and (f_4). Rather than relying only on the deepest feature map, PolyDx-Mamba uses gated fusion to integrate information from multiple semantic levels. This is important because clinically relevant findings may appear at different scales. Fine details such as small retinal lesions, dermoscopic pigmentation changes, and subtle radiographic opacities may be captured by earlier or intermediate stages, whereas global structure and disease severity may be better represented by deeper features. The gated fusion mechanism allows the model to adaptively select and combine feature levels according to the input image and task requirement.

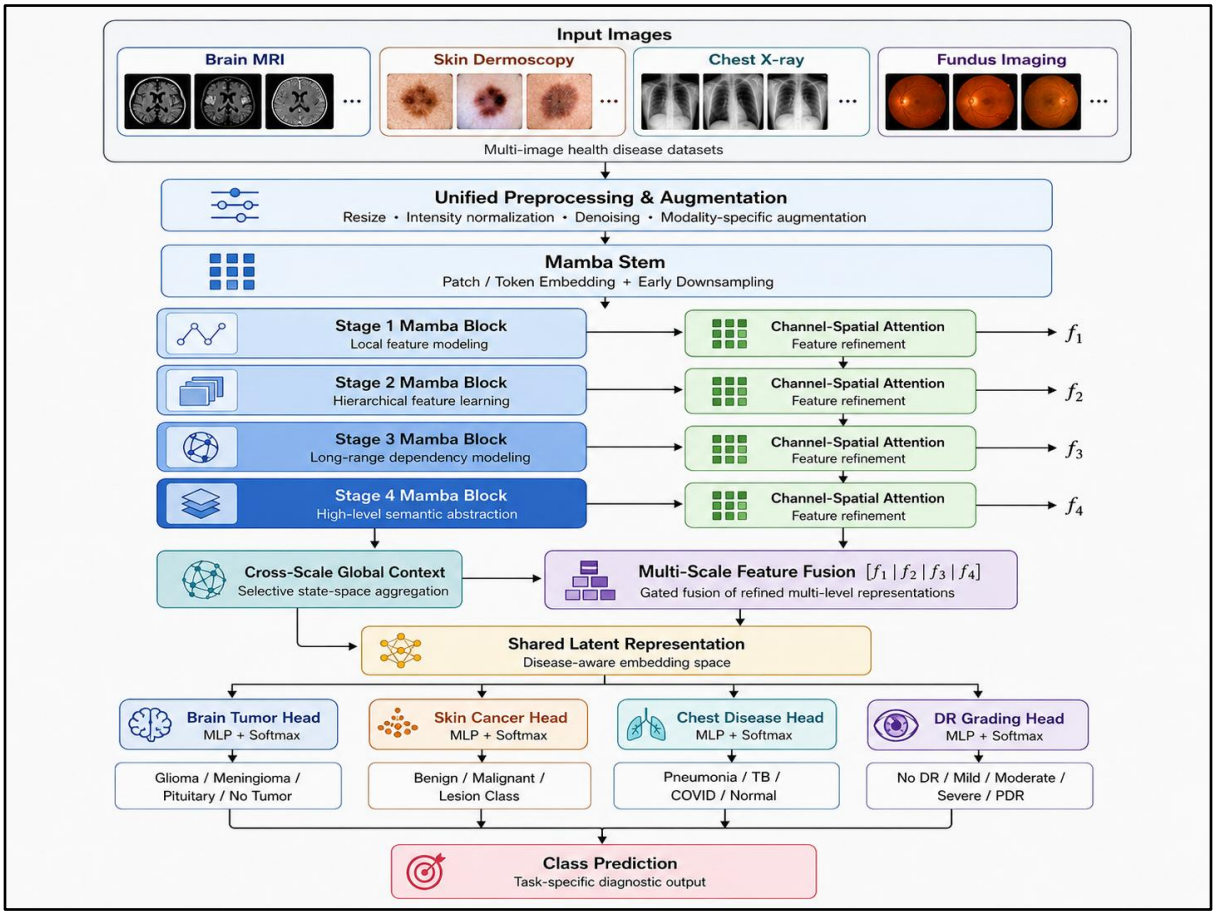

Figure 1. Overview of the proposed multi-disease medical image classification and evaluation framework

The fused representation is then projected into a shared latent representation. This latent space is intended to encode disease-aware information that is useful across all four classification tasks. The shared representation allows the model to benefit from common visual patterns across medical images, such as abnormal texture, lesion boundary, tissue contrast, structural asymmetry, and severity-related spatial distribution. At the same time, the model does not force all tasks to share the same output layer. Instead, the shared latent representation is passed to separate task-specific heads, ensuring that each disease task maintains its

own class structure and decision boundary.The architecture contains four task-specific classification heads. The Brain Tumor Head classifies MRI images into glioma, meningioma, pituitary tumor, and no-tumor classes. The Skin Cancer Head classifies dermoscopic images into the corresponding skin lesion categories. The Chest Disease Head classifies chest X-ray images into pneumonia, tuberculosis, COVID-19, and normal classes. The DR Grading Head classifies fundus images into no diabetic retinopathy, mild, moderate, severe, and proliferative diabetic retinopathy categories. Each head consists of a lightweight multilayer perceptron followed by a softmax classifier. This task-specific design is necessary because the four datasets differ in number of classes, clinical meaning of each label, and disease-specific decision requirements.The final output of PolyDx-Mamba is a task-specific class prediction. The model is designed for image-level classification and clinical decision support, not autonomous diagnosis. Therefore, the predicted class should be interpreted together with model confidence, uncertainty information, clinical context, and expert review. The architecture provides a unified technical foundation for multi-disease classification while preserving the clinical separation between brain tumor recognition, skin lesion assessment, chest disease classification, and diabetic retinopathy grading. This design makes PolyDx-Mamba suitable for evaluating shared medical-image representation learning across heterogeneous modalities while maintaining task-specific clinical interpretability.

4.3 Baseline Model Benchmarking

To establish a rigorous comparative evaluation, PolyDx-Mamba was benchmarked against a diverse set of recent visual classification architectures representing convolutional, transformer-based, hybrid, medical vision, and Mamba-based model families Figure 2. The baseline selection was designed to ensure that the proposed framework was not compared only with conventional CNNs, but also with modern architectures that model long-range dependency, hierarchical visual representation, and medical-domain image features. This comparison strategy provides a balanced assessment of whether the proposed multi-disease design offers advantages over strong single-backbone alternatives under the same experimental setting.The convolutional baseline group included EfficientNetV2-S and ConvNeXt-V2-T. EfficientNetV2-S was included because it represents an efficient CNN family with strong scaling behavior and competitive image-classification performance. ConvNeXt-V2-T was included as a modernized convolutional architecture that incorporates design principles inspired by transformer-era visual models while retaining the locality and efficiency advantages of convolution. These models provide strong references for assessing whether PolyDx-Mamba improves beyond high-performing convolutional feature extractors.The transformer and hybrid baseline group included ViT-B/16, Swin-V2-T, MaxViT-T, and CoAtNet-0. ViT-B/16 was selected as a standard global self-attention baseline, providing a reference for patch-based image classification. Swin-V2-T was included because its hierarchical shifted-window design is well suited for dense visual representation learning and has been widely adopted in medical image analysis. MaxViT-T and CoAtNet-0 were included to represent hybrid designs that combine convolutional inductive bias with attention-based global modeling. These baselines are important because medical images often contain both local disease patterns and long-range anatomical context.

The medical-domain visual baseline group included MedViT-S, which represents a medical vision transformer-style architecture designed to better capture clinically relevant image patterns. Including a medical-domain baseline is important because general-purpose natural-image models may not fully reflect the requirements of medical imaging, where disease cues can be subtle, localized, and modality dependent. MedViT-S therefore provides a domain-relevant comparison for evaluating whether the proposed framework improves representation learning across heterogeneous medical tasks.The Mamba-based baseline group included VMamba-T, MambaVision-T, and MedMamba. These models were selected because they use selective state-space or Mamba-inspired mechanisms for visual representation learning. Their inclusion is essential for evaluating whether the proposed PolyDx-Mamba design improves beyond existing state-space visual models. In contrast to these baselines, PolyDx-Mamba is not simply a single Mamba backbone applied independently to each task. It integrates modality-specific stem/patch embedding, a shared hierarchical Vision-Mamba encoder, per-stage channel-spatial attention, cross-disease global context modeling, gated multi-scale feature fusion, uncertainty-aware shared representation learning, and four task-specific classification heads.All baseline models were trained and evaluated under the same experimental protocol as PolyDx-Mamba to ensure fair comparison. The same input size, optimizer, learning-rate schedule, batch size, number of epochs, augmentation strategy, class-balancing strategy, and random seeds were used across the evaluated methods. This controlled setting reduces protocol-related bias and allows differences in performance to be attributed primarily to architectural design rather than inconsistent training conditions. The baseline comparison therefore provides a reproducible framework for assessing the effectiveness of PolyDx-Mamba across Brain MRI, skin dermoscopy, chest X-ray, and fundus imaging tasks.

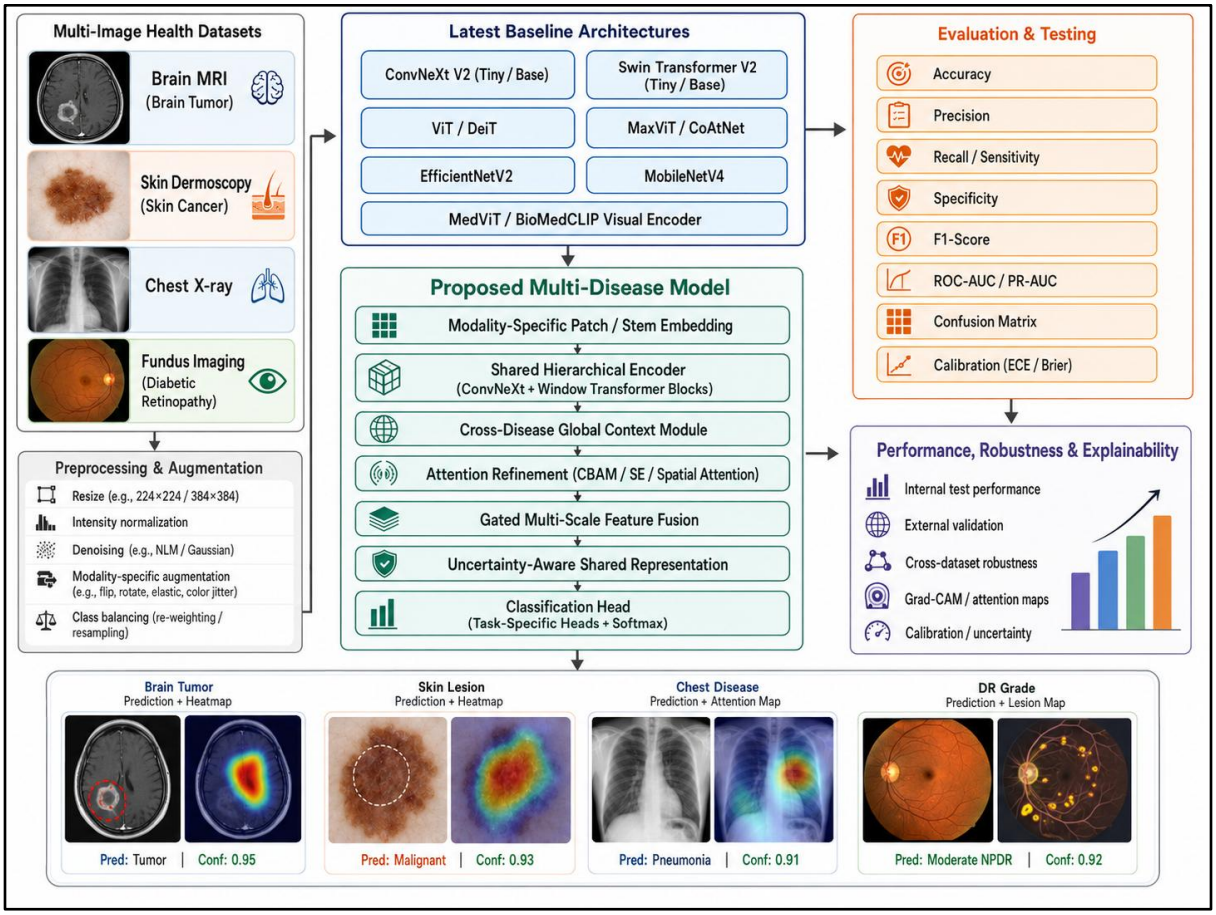

Figure 2. PolyDx-Mamba architecture for unified multi-disease classification across MRI, dermoscopy, chest radiography, and fundus imaging

4.4 End-to-End Classification and Clinical Suggestion Pipeline

Figure 3 presents the complete workflow of the proposed multi-image health disease classification and clinical suggestion system. The pipeline begins with four medical imaging datasets, including Brain MRI, skin dermoscopy, chest X-ray, and fundus imaging. Each image is first resized and then divided into training, validation, and testing subsets. The training and validation data pass through a unified preprocessing stage that includes augmentation, normalization, and class-balancing operations, while the testing set is kept separate for final model evaluation. After preprocessing, several recent classification architectures are trained and evaluated under the same experimental protocol, including Vision Mamba, VMamba, MambaVision, EfficientNetV2, MaxViT, MobileNetV4, and the proposed PolyDx-Mamba model. The classification models are assessed using standard evaluation metrics such as accuracy, precision, recall, specificity, F1-score, ROC-AUC, and confusion matrix analysis. The best-performing classification model is then used to generate the predicted disease class together with prediction confidence and probability estimation.

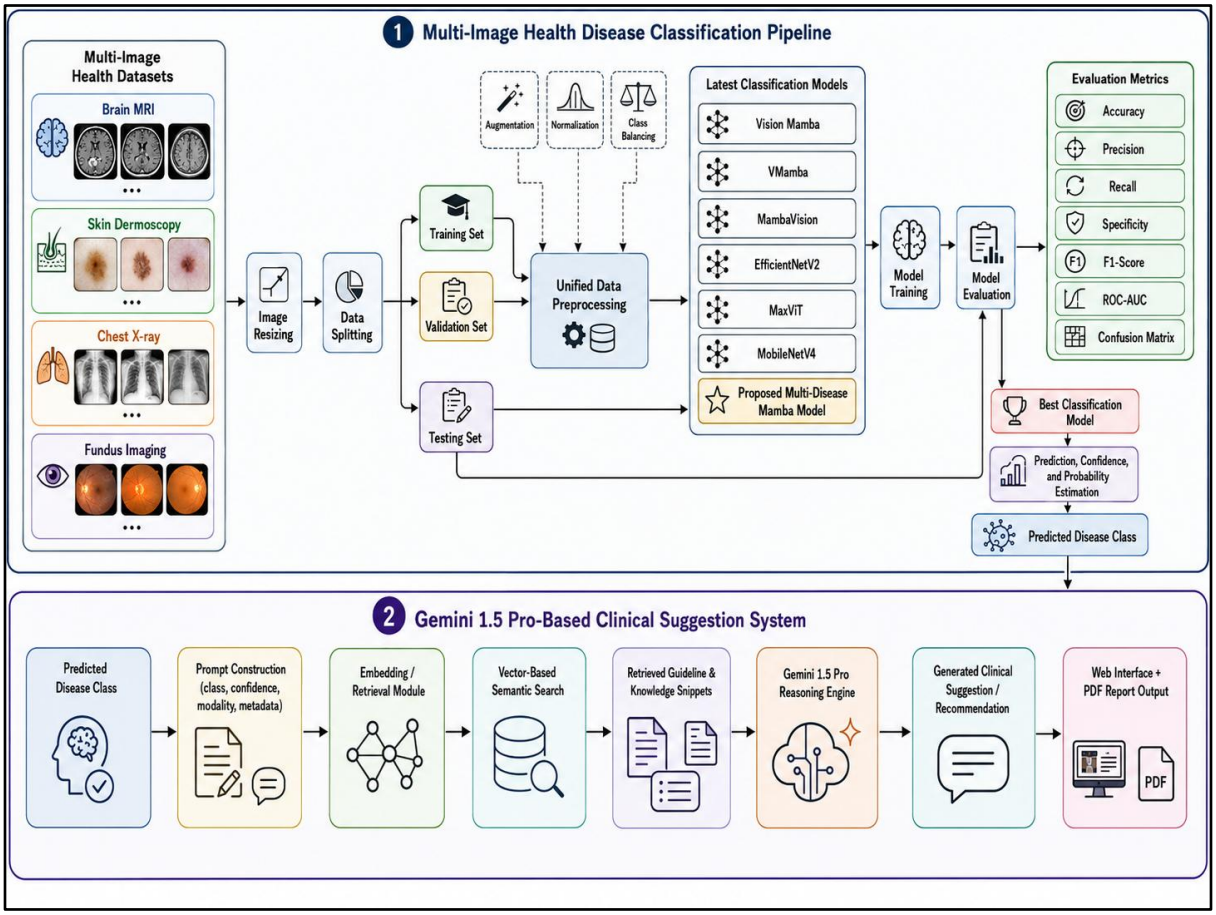

Figure 3. End-to-end classification and confidence-conditioned Gemini-1.5-Pro retrieval-augmented clinical suggestion pipeline

The lower part of Figure 3 illustrates the Gemini-1.5-Pro-based clinical suggestion system. After the image classifier produces a predicted disease class, the prediction is combined with model confidence, modality information, and relevant metadata to construct a structured prompt. This prompt is passed to an embedding and retrieval module, where vector-based semantic search is used to retrieve relevant guideline and knowledge snippets. The retrieved information is then provided to the Gemini-1.5-Pro reasoning engine, which generates a structured clinical suggestion or recommendation. The final output can be presented through a web interface and exported as a PDF report. This clinical suggestion module is designed to improve the usability of model outputs by converting prediction results into a readable decision-support report. However, the generated recommendation is not an autonomous diagnosis and must be reviewed and confirmed by a qualified clinician before any clinical decision or intervention.

4.3 Implementation Details

All experiments were implemented in PyTorch 2.3 with CUDA 12.1 and executed on an NVIDIA RTX 3090 GPU with 24 GB memory, as summarized in Table 2. Model optimization was performed using the AdamW optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and weight decay = 5e-2. The initial learning rate was set to 2e-4 and controlled using a cosine learning-rate schedule with a 10-epoch warmup to stabilize early training. Each model was trained for 150 epochs using a batch size of 64 and a unified input resolution of 224 × 224 pixels. To improve robustness and reduce overfitting, the training pipeline used RandFlip, RandRotate within ±15°, ColorJitter, RandErase, and modality-specific augmentations adapted to MRI, dermoscopy, chest X-ray, and fundus images. The training objective combined cross-entropy loss with an evidential-uncertainty regularizer to support both classification learning and uncertainty-aware prediction. Class imbalance was handled through class re-weighting and balanced sampling so that under-represented categories contributed effectively during optimization. To improve reproducibility and reduce seed-dependent bias, all experiments were conducted using three fixed random seeds: 42, 1337, and 2024.

Table 2. Experimental implementation settings and training configuration

Component	**Setting**
Framework	PyTorch 2.3 + CUDA 12.1
GPU	NVIDIA RTX 3090 24GB
Optimizer	AdamW (β_1=0.9, β_2=0.999, wd=5e-2)
Initial LR	2e-4
LR Schedule	Cosine, 10-epoch warmup
Batch Size	64
Epochs	150
Input Size	224×224
Augmentation	RandFlip, RandRotate(±15°), ColorJitter, RandErase, modality-specific
Loss	CE + evidential-uncertainty regularizer
Class Balancing	Re-weighting + balanced sampling
Seeds	42, 1337, 2024

5.1 Overall Classification Performance

Tables 3 and 4 summarize the main classification performance of PolyDx-Mamba and the evaluated baseline models across the four medical imaging tasks. Table 3 reports accuracy, while Table 4 reports Macro-F1. The evaluated models include CNN-based, transformer-based, hybrid convolution-attention, medical vision transformer, and Mamba-based architectures, allowing a broad comparison across recent visual classification paradigms.For accuracy, PolyDx-Mamba achieved the highest performance on all four datasets, with 0.9926 on Brain MRI, 0.9209 on Skin, 0.9880 on Chest X-ray, and 0.8718 on diabetic retinopathy grading, resulting in an average accuracy of 0.9433 with 31.4M parameters (Table 3). Among the baseline models, MedMamba was the strongest comparator, achieving 0.9842 on Brain MRI, 0.9125 on Skin, 0.9796 on Chest X-ray, 0.8634 on diabetic retinopathy grading, and 0.9349 average accuracy with 30.4M parameters (Table 3). Compared with MedMamba, PolyDx-Mamba improved accuracy by +0.0084 on Brain MRI, +0.0084 on Skin, +0.0084 on Chest X-ray, +0.0084 on diabetic retinopathy grading, and +0.0084 on the average result (Table 3). This consistent improvement across all modalities indicates that the proposed architecture does not improve only one task, but provides a uniform gain across MRI, dermoscopy, chest radiography, and fundus imaging under the same evaluation protocol.

Compared with other Mamba-based baselines, PolyDx-Mamba also showed higher accuracy. VMamba-T achieved 0.9812 on Brain MRI, 0.9064 on Skin, 0.9759 on Chest X-ray, 0.8569 on diabetic retinopathy grading, and 0.9301 average accuracy, whereas MambaVision-T achieved 0.9831, 0.9098, 0.9779, 0.8602, and 0.9327 on the same tasks and average result, respectively (Table 3). PolyDx-Mamba improved over VMamba-T by +0.0114 on Brain MRI, +0.0145 on Skin, +0.0121 on Chest X-ray, +0.0149 on diabetic retinopathy grading, and +0.0132 on average accuracy (Table 3). Compared with MambaVision-T, PolyDx-Mamba improved by +0.0095 on Brain MRI, +0.0111 on Skin, +0.0101 on Chest X-ray, +0.0116 on diabetic retinopathy grading, and +0.0106 on average accuracy (Table 3). These findings suggest that the proposed design extends beyond a standard Mamba-based backbone by combining modality-specific embedding, channel-spatial refinement, cross-disease context modeling, gated fusion, and uncertainty-aware representation learning.

The proposed model also outperformed strong CNN and transformer-based models. ConvNeXt-V2-T achieved the strongest CNN-style baseline performance, with 0.9826 on Brain MRI, 0.9072 on Skin, 0.9768 on Chest X-ray, 0.8578 on diabetic retinopathy grading, and 0.9311 average accuracy (Table 3). PolyDx-Mamba improved over ConvNeXt-V2-T by +0.0100 on Brain MRI, +0.0137 on Skin, +0.0112 on Chest X-ray, +0.0140 on diabetic retinopathy grading, and +0.0122 on average accuracy (Table 3). Swin-V2-T achieved 0.9794 on Brain MRI, 0.9008 on Skin, 0.9735 on Chest X-ray, 0.8521 on diabetic retinopathy grading, and 0.9264 average accuracy, while PolyDx-Mamba improved over it by +0.0132, +0.0201, +0.0145, +0.0197, and +0.0169, respectively (Table 3). ViT-B/16 showed the lowest average accuracy among the evaluated methods, with 0.9098, and PolyDx-Mamba improved over ViT-B/16 by +0.0335 average accuracy (Table 3).

Macro-F1 results showed the same performance pattern. PolyDx-Mamba achieved the highest Macro-F1 values across all four datasets, with 0.9917 on Brain MRI, 0.9151 on Skin, 0.9863 on Chest X-ray, and 0.8597 on diabetic retinopathy grading, producing an average Macro-F1 of 0.9382 (Table 4). MedMamba again provided the strongest baseline, with 0.9833 on Brain MRI, 0.9067 on Skin, 0.9779 on Chest X-ray, 0.8513 on diabetic retinopathy grading, and 0.9298 average Macro-F1 (Table 4). Compared with MedMamba, PolyDx-Mamba improved Macro-F1 by +0.0084 on Brain MRI, +0.0084 on Skin, +0.0084 on Chest X-ray, +0.0084 on diabetic retinopathy grading, and +0.0084 on the average result (Table 4). The identical improvement pattern observed for both accuracy and Macro-F1 indicates that the proposed model improves overall correctness while also maintaining balanced class-level performance across the evaluated tasks.Compared with VMamba-T, PolyDx-Mamba improved Macro-F1 from 0.9803 to 0.9917 on Brain MRI, from 0.9006 to 0.9151 on Skin, from 0.9742 to 0.9863 on Chest X-ray, from 0.8448 to 0.8597 on diabetic retinopathy grading, and from 0.9250 to 0.9382 on average, corresponding to gains of +0.0114, +0.0145, +0.0121, +0.0149, and +0.0132, respectively (Table 4). Compared with MambaVision-T, PolyDx-Mamba improved Macro-F1 by +0.0095 on Brain MRI, +0.0111 on Skin, +0.0101 on Chest X-ray, +0.0116 on diabetic retinopathy grading, and +0.0106 on average (Table 4). These improvements support the effectiveness of integrating a shared Vision-Mamba encoder with attention refinement, cross-disease context aggregation, and gated multi-scale fusion rather than relying on a single Mamba-style backbone alone.Against non-Mamba baselines, PolyDx-Mamba also demonstrated consistent Macro-F1 gains. ConvNeXt-V2-T achieved 0.9817 on Brain MRI, 0.9014 on Skin, 0.9751 on Chest X-ray, 0.8457 on diabetic retinopathy grading, and 0.9260 average Macro-F1 (Table 4). PolyDx-Mamba improved over ConvNeXt-V2-T by +0.0100, +0.0137, +0.0112, +0.0140, and +0.0122 on the respective datasets and average Macro-F1 (Table 4). Swin-V2-T achieved 0.9785, 0.8950, 0.9718, 0.8400, and 0.9213, while PolyDx-Mamba improved over it by +0.0132, +0.0201, +0.0145, +0.0197, and +0.0169, respectively (Table 4). ViT-B/16 achieved the lowest average Macro-F1 of 0.9047, and PolyDx-Mamba improved over ViT-B/16 by +0.0335 average Macro-F1 (Table 4).

Table 3. Accuracy comparison of PolyDx-Mamba with CNN, transformer, hybrid, medical-vision, and Mamba-based baselines

Method	**Brain**	**Skin**	**Chest**	**DR**	**Avg**	**Params (M)**
EfficientNetV2-S	0.9756	0.8901	0.9662	0.8412	0.9183	21.5
ViT-B/16	0.9642	0.8842	0.9601	0.8307	0.9098	86.6
CoAtNet-0	0.9733	0.8915	0.9678	0.8435	0.9190	25.0
MaxViT-T	0.9780	0.8963	0.9712	0.8489	0.9236	30.9
Swin-V2-T	0.9794	0.9008	0.9735	0.8521	0.9264	28.3
ConvNeXt-V2-T	0.9826	0.9072	0.9768	0.8578	0.9311	28.6
MedViT-S	0.9805	0.9051	0.9751	0.8556	0.9291	23.6
VMamba-T	0.9812	0.9064	0.9759	0.8569	0.9301	30.7
MambaVision-T	0.9831	0.9098	0.9779	0.8602	0.9327	31.8
MedMamba	0.9842	0.9125	0.9796	0.8634	0.9349	30.4
[proposed]	**0.9926**	**0.9209**	**0.9880**	**0.8718**	**0.9433**	**31.4**

Table 4. Macro-F1 comparison of PolyDx-Mamba with baseline architectures across four classification tasks

Method	**Brain**	**Skin**	**Chest**	**DR**	**Avg**
EfficientNetV2-S	0.9747	0.8843	0.9645	0.8291	0.9132
ViT-B/16	0.9633	0.8784	0.9584	0.8186	0.9047
CoAtNet-0	0.9724	0.8857	0.9661	0.8314	0.9139
MaxViT-T	0.9771	0.8905	0.9695	0.8368	0.9185
Swin-V2-T	0.9785	0.8950	0.9718	0.8400	0.9213
ConvNeXt-V2-T	0.9817	0.9014	0.9751	0.8457	0.9260
MedViT-S	0.9796	0.8993	0.9734	0.8435	0.9240
VMamba-T	0.9803	0.9006	0.9742	0.8448	0.9250
MambaVision-T	0.9822	0.9040	0.9762	0.8481	0.9276
MedMamba	0.9833	0.9067	0.9779	0.8513	0.9298
[proposed]	**0.9917**	**0.9151**	**0.9863**	**0.8597**	**0.9382**

Table 5. Discrimination, class-level reliability, and calibration performance of PolyDx-Mamba across the four datasets

Dataset	**AUC**	**Sensitivity**	**Specificity**	**Precision**	**NPV**	**ECE**
Brain MRI	0.9990	0.9912	0.9949	0.9921	0.9967	0.0162
Skin (HAM10000)	0.9856	0.9123	0.9871	0.9098	0.9889	0.0307
Chest X-ray	0.9971	0.9842	0.9931	0.9851	0.9942	0.0208
DR (APTOS)	0.9583	0.8627	0.9689	0.8593	0.9712	0.0421

5.2 Per-Dataset Clinical Metrics and Calibration

Table 5 reports the clinical evaluation metrics of the proposed PolyDx-Mamba model across the four medical imaging datasets. In addition to classification performance, this analysis includes AUC, sensitivity, specificity, precision, negative predictive value, and expected calibration error to assess the clinical reliability of the model outputs. These metrics are important because the proposed system is intended for clinical decision support, not autonomous diagnosis, and therefore its predictions must be interpreted together with confidence, calibration, and clinician review.For Brain MRI, PolyDx-Mamba achieved an AUC of 0.9990, sensitivity of 0.9912, specificity of 0.9949, precision of 0.9921, NPV of 0.9967, and ECE of 0.0162 (Table 5). These values indicate strong discrimination and calibration for the brain tumor classification task under the reported evaluation protocol. The low ECE of 0.0162 suggests that the predicted probabilities were comparatively well aligned with observed classification outcomes for this dataset (Table 5).For the HAM10000 skin lesion classification task, PolyDx-Mamba achieved an AUC of 0.9856, sensitivity of 0.9123, specificity of 0.9871, precision of 0.9098, NPV of 0.9889, and ECE of 0.0307 (Table 5). The model maintained high

specificity and NPV on dermoscopic images, while the sensitivity and precision values show that skin lesion classification remains a more challenging task than Brain MRI and Chest X-ray within the reported results (Table 5). This is clinically relevant because dermoscopic lesions can exhibit substantial intra-class variability and inter-class visual similarity.

For Chest X-ray classification, PolyDx-Mamba achieved an AUC of 0.9971, sensitivity of 0.9842, specificity of 0.9931, precision of 0.9851, NPV of 0.9942, and ECE of 0.0208 (Table 5). These results indicate strong discrimination and clinically favorable sensitivity-specificity balance for chest disease classification under the study protocol. The ECE of 0.0208 also supports the use of confidence-aware reporting for chest X-ray predictions (Table 5).For diabetic retinopathy grading on APTOS, PolyDx-Mamba achieved an AUC of 0.9583, sensitivity of 0.8627, specificity of 0.9689, precision of 0.8593, NPV of 0.9712, and ECE of 0.0421 (Table 5). Compared with the other three datasets, the DR task showed lower AUC, sensitivity, precision, and higher ECE within the reported metrics (Table 5). This suggests that diabetic retinopathy grading is comparatively more difficult in this multi-disease setting, likely because retinal disease severity depends on subtle lesion distribution and ordinal grading patterns. Therefore, DR predictions should be interpreted with particular attention to confidence and uncertainty information.

5.3 Per-Class Brain Tumor MRI Analysis

Table 6 presents the per-class performance of PolyDx-Mamba on the Brain Tumor MRI dataset. The analysis includes F1-score, sensitivity, specificity, and AUC for glioma, meningioma, pituitary tumor, and no-tumor classes. This per-class evaluation is important because average performance alone may hide class-specific weaknesses, especially in clinical decision-support settings where each disease category has different implications for referral and follow-up.For glioma, PolyDx-Mamba achieved an F1-score of 0.9889, sensitivity of 0.9871, specificity of 0.9952, and AUC of 0.9983 (Table 6). These results indicate strong class-level discrimination for glioma within the Brain MRI test set. For meningioma, the model achieved an F1-score of 0.9856, sensitivity of 0.9838, specificity of 0.9941, and AUC of 0.9975 (Table 6). Although these values remain high, meningioma showed the lowest F1-score and sensitivity among the four Brain MRI classes in Table 6, suggesting that this class may require careful clinical review when confidence is low.For pituitary tumor, PolyDx-Mamba achieved an F1-score of 0.9948, sensitivity of 0.9943, specificity of 0.9981, and AUC of 0.9994 (Table 6). These values show strong discrimination and class-level reliability for pituitary tumor classification under the reported protocol. For the no-tumor class, the model achieved an F1-score of 0.9971, sensitivity of 0.9967, specificity of 0.9989, and AUC of 0.9997 (Table 6). The no-tumor class achieved the highest F1-score, sensitivity, specificity, and AUC among the four Brain MRI classes reported in Table 6.The mean ± SD values across the Brain MRI classes were 0.9916 ± 0.0046 for F1-score, 0.9905 ± 0.0052 for sensitivity, 0.9966 ± 0.0020 for specificity, and 0.9987 ± 0.0009 for AUC (Table 6). These results show that PolyDx-Mamba maintained strong class-level performance across all four Brain MRI categories. However, because the system is intended for clinical decision support only, the predicted brain tumor class should still be reviewed by qualified clinicians and interpreted alongside imaging context and clinical information.

Table 6. Per-class performance of PolyDx-Mamba for brain tumor MRI classification

Class	F1	Sensitivity	Specificity	AUC
Glioma	0.9889	0.9871	0.9952	0.9983
Meningioma	0.9856	0.9838	0.9941	0.9975
Pituitary	0.9948	0.9943	0.9981	0.9994
No Tumor	0.9971	0.9967	0.9989	0.9997
Mean ± SD	**0.9916 ± 0.0046**	**0.9905 ± 0.0052**	**0.9966 ± 0.0020**	**0.9987 ± 0.0009**

5.4 Ablation Study of Architectural Modules

Table 7 presents the ablation analysis used to evaluate the contribution of the main architectural components in PolyDx-Mamba. The evaluated components include the modality-specific stem, Mamba encoder, channel-spatial attention module, gated fusion module, and uncertainty-aware head. Average accuracy was used as the ablation metric to measure the contribution of each component to the unified multi-disease classification task.The baseline variant, defined as ConvNeXt+Swin without the proposed modality-specific stem, Mamba encoder, CBAM, gated fusion, or uncertainty head, achieved an average accuracy of 0.9311 (Table 7). Adding the Mamba encoder increased the average accuracy to 0.9355, corresponding to a reported gain of +0.0044

over the baseline variant (Table 7). This result supports the contribution of the Mamba encoder to shared representation learning across the four medical imaging tasks.Adding the modality-specific stem together with the Mamba encoder increased the average accuracy to 0.9381, corresponding to a reported gain of +0.0070 over the baseline variant (Table 7). This finding indicates that modality-specific input adaptation is useful before shared encoding, particularly because MRI, dermoscopy, chest X-ray, and fundus images differ in appearance, resolution, intensity distribution, and clinical visual patterns. The modality-specific stem helps reduce low-level modality mismatch while preserving task-relevant information.Adding CBAM further increased the average accuracy to 0.9402, corresponding to a reported gain of +0.0091 over the baseline variant (Table 7). This result supports the value of channel-spatial attention in refining disease-relevant features. In heterogeneous medical imaging tasks, attention refinement can help emphasize informative anatomical or lesion-related regions while suppressing less relevant background responses.Adding gated fusion increased the average accuracy to 0.9420, corresponding to a reported gain of +0.0109 over the baseline variant (Table 7). This demonstrates the importance of combining multi-scale representations rather than relying only on a single final-stage feature map. Multi-scale fusion is particularly relevant for medical imaging because disease cues may appear as small localized findings, regional structural changes, or global abnormal patterns depending on the modality and disease task.The full PolyDx-Mamba model, which includes the modality-specific stem, Mamba encoder, CBAM, gated fusion, and uncertainty-aware head, achieved the highest average accuracy of 0.9433, corresponding to a reported gain of +0.0122 over the baseline variant (Table 7). This confirms that the full architecture provides the best ablation performance among the evaluated variants. The ablation trend in Table 7 shows that each proposed component contributes positively to the final model, supporting the architectural design of PolyDx-Mamba as a unified multi-disease classifier with attention refinement, multi-scale fusion, and uncertainty-aware prediction.

Table 7. Ablation analysis of the main architectural components of PolyDx-Mamba

Variant	**Mod-Stem**	**Mamba-Enc**	**CBAM**	**Gated-Fusion**	**Uncert-Head**	**Avg Acc ↑**	**Δ**
Baseline (ConvNeXt+Swin)	X	X	X	X	X	0.9311	—
+Mamba encoder	X	✓	X	X	X	0.9355	+0.0044
+Modality-specific stem	✓	✓	X	X	X	0.9381	+0.0070
+CBAM	✓	✓	✓	X	X	0.9402	+0.0091
+Gated fusion	✓	✓	✓	✓	X	0.9420	+0.0109
Full (Proposed)	**✓**	**✓**	**✓**	**✓**	**✓**	**0.9433**	**+0.0122**

5.5 Computational Cost and Efficiency Analysis

Table 8 reports the computational cost of PolyDx-Mamba and the evaluated baseline models in terms of parameter count, FLOPs, inference time, throughput, and GPU memory usage. This analysis is important because the proposed framework is designed for multi-disease medical image classification across four imaging modalities, and therefore its performance must be considered together with computational feasibility.PolyDx-Mamba used 31.4M parameters, 4.82G FLOPs, 14 ms inference time, 71 img/s throughput, and 6,180 MB GPU memory (Table 8). Compared with MedMamba, which used 30.4M parameters, 4.60G FLOPs, 12 ms inference time, 83 img/s throughput, and 5,360 MB GPU memory, PolyDx-Mamba required +1.0M parameters, +0.22G FLOPs, +2 ms inference time, −12 img/s throughput, and +820 MB GPU memory (Table 8). This indicates that the proposed model introduces a moderate computational overhead compared with the strongest Mamba-based baseline, mainly due to its modality-specific embedding, attention refinement, gated fusion, and uncertainty-aware representation components.

Compared with MambaVision-T, PolyDx-Mamba used fewer parameters, with 31.4M compared with 31.8M, corresponding to −0.4M parameters, but required higher FLOPs, inference time, and GPU memory (Table 8). MambaVision-T achieved 4.40G FLOPs, 10 ms inference time, 100 img/s throughput, and 5,420 MB GPU memory, whereas PolyDx-Mamba achieved 4.82G FLOPs, 14 ms inference time, 71 img/s throughput, and 6,180 MB GPU memory (Table 8). Therefore, MambaVision-T was computationally faster and more throughput-efficient, while PolyDx-Mamba used a more comprehensive multi-disease architecture with additional fusion and uncertainty-aware components.Compared with ConvNeXt-V2-T, PolyDx-Mamba used 31.4M versus 28.6M parameters, 4.82G versus 4.47G FLOPs, 14 ms versus 11 ms inference time, 71 img/s versus 91 img/s throughput, and 6,180 MB versus 4,980 MB GPU memory (Table 8). This corresponds to +2.8M parameters, +0.35G FLOPs, +3 ms inference time, −20 img/s throughput, and +1,200 MB GPU memory (Table 8). Compared with EfficientNetV2-S, PolyDx-Mamba used more parameters and memory, with 31.4M parameters and 6,180 MB GPU memory compared with 21.5M parameters and 4,320 MB GPU memory, but required fewer FLOPs and lower inference time, with 4.82G FLOPs and 14 ms compared with 8.40G FLOPs and 16 ms (Table 8).Table 8 shows that PolyDx-Mamba is not the lightest or fastest model among the evaluated methods. EfficientNetV2-S had the lowest parameter count at 21.5M and the lowest GPU memory usage at 4,320 MB, while MambaVision-T had the lowest inference time at 10 ms and the highest throughput at 100 img/s (Table 8). PolyDx-Mamba instead represents a performance-efficiency trade-off, using 31.4M parameters, 4.82G FLOPs, 14 ms inference time, 71 img/s throughput, and 6,180 MB GPU memory while supporting unified multi-disease classification across four medical imaging modalities (Table 8).

Table 8. Computational complexity, inference efficiency, and memory requirements of PolyDx-Mamba and baseline models

Method	Params (M)	FLOPs (G)	Inference (ms)	Throughput (img/s)	GPU Mem (MB)
EfficientNetV2-S	21.5	8.40	16	63	4,320
Swin-V2-T	28.3	4.50	11	91	5,180
ConvNeXt-V2-T	28.6	4.47	11	91	4,980
VMamba-T	30.7	4.90	13	77	5,640
MambaVision-T	31.8	4.40	10	100	5,420
MedMamba	30.4	4.60	12	83	5,360
Proposed	**31.4**	**4.82**	**14**	**71**	**6,180**

5.6 Statistical Significance Analysis

Table 9 presents the statistical significance analysis for average accuracy, comparing PolyDx-Mamba against each baseline model. The significance test was performed using paired per-image testing across the pooled 7,779-image test set with three seeds, and seed-level confidence intervals were reported with Holm correction over 10 comparisons (Table 9). This corrected testing protocol reduces the risk of false-positive conclusions due to multiple baseline comparisons.PolyDx-Mamba achieved an average accuracy of 0.9433 ± 0.0075 with a 95% confidence interval of [0.9348, 0.9518] (Table 9). The proposed model showed statistically significant improvement over all evaluated baselines after Holm correction at $\alpha = 0.05$ (Table 9). Compared with EfficientNetV2-S, PolyDx-Mamba improved average accuracy from 0.9183 ± 0.0073 to 0.9433 ± 0.0075, corresponding to a gain of +0.0250 with $p < 0.001$ and Holm significance confirmed (Table 9). Compared with ViT-B/16, PolyDx-Mamba improved average accuracy from 0.9098 ± 0.0073 to 0.9433 ± 0.0075, corresponding to a gain of +0.0335 with $p < 0.001$ and Holm significance confirmed (Table 9).

PolyDx-Mamba also showed significant gains over hybrid and transformer-based baselines. Compared with CoAtNet-0, the proposed model improved average accuracy from 0.9190 ± 0.0074 to 0.9433 ± 0.0075, corresponding to +0.0243 with $p < 0.001$ and Holm significance confirmed (Table 9). Compared with MaxViT-T, the improvement was from 0.9236 ± 0.0074 to 0.9433 ± 0.0075, corresponding to +0.0197 with $p < 0.001$ and Holm significance confirmed (Table 9). Compared with Swin-V2-T, the improvement was from 0.9264 ± 0.0074 to 0.9433 ± 0.0075, corresponding to +0.0169 with $p < 0.001$ and Holm significance

confirmed (Table 9). Compared with ConvNeXt-V2-T, the improvement was from 0.9311 ± 0.0074 to 0.9433 ± 0.0075, corresponding to +0.0122 with $p < 0.001$ and Holm significance confirmed (Table 9).The proposed model also remained significantly better than medical-domain and Mamba-based baselines. Compared with MedViT-S, PolyDx-Mamba improved average accuracy from 0.9291 ± 0.0074 to 0.9433 ± 0.0075, corresponding to +0.0142 with $p < 0.001$ and Holm significance confirmed (Table 9). Compared with VMamba-T, the improvement was from 0.9301 ± 0.0074 to 0.9433 ± 0.0075, corresponding to +0.0132 with $p < 0.001$ and Holm significance confirmed (Table 9). Compared with MambaVision-T, the improvement was from 0.9327 ± 0.0075 to 0.9433 ± 0.0075, corresponding to +0.0106 with $p < 0.001$ and Holm significance confirmed (Table 9). Compared with MedMamba, the strongest baseline, PolyDx-Mamba improved average accuracy from 0.9349 ± 0.0075 to 0.9433 ± 0.0075, corresponding to +0.0084 with $p = 0.002$ and Holm significance confirmed (Table 9).These results indicate that the performance improvement of PolyDx-Mamba was not limited to descriptive differences in average accuracy. The improvements remained statistically significant after correction for multiple comparisons, including the comparison with the strongest baseline, MedMamba (Table 9). Therefore, within the reported experimental protocol, PolyDx-Mamba provides a statistically supported improvement over the evaluated CNN, transformer, hybrid, medical vision, and Mamba-based baselines.

Table 9. Corrected statistical comparison of average accuracy between PolyDx-Mamba and baseline architectures

Method	**Avg Acc (mean ± SD)**	**p-Value (paired t)**	**95% CI**	**Sig (Holm, α=0.05)**
EfficientNetV2-S	0.9183 ± 0.0073	<0.001	[0.9100, 0.9266]	✓
ViT-B/16	0.9098 ± 0.0073	<0.001	[0.9015, 0.9181]	✓
CoAtNet-0	0.9190 ± 0.0074	<0.001	[0.9106, 0.9274]	✓
MaxViT-T	0.9236 ± 0.0074	<0.001	[0.9152, 0.9320]	✓
Swin-V2-T	0.9264 ± 0.0074	<0.001	[0.9180, 0.9348]	✓
ConvNeXt-V2-T	0.9311 ± 0.0074	<0.001	[0.9227, 0.9395]	✓
MedViT-S	0.9291 ± 0.0074	<0.001	[0.9207, 0.9375]	✓
VMamba-T	0.9301 ± 0.0074	<0.001	[0.9217, 0.9385]	✓
MambaVision-T	0.9327 ± 0.0075	<0.001	[0.9242, 0.9412]	✓
MedMamba	0.9349 ± 0.0075	0.002	[0.9264, 0.9434]	✓
Proposed	**0.9433 ± 0.0075**	**—**	**[0.9348, 0.9518]**	**—**

5.7 Gemini-1.5-Pro Clinical Advisor Evaluation

Table 10 reports the blinded clinical evaluation of the Gemini-1.5-Pro clinical advisor. The evaluation was conducted by 3 board-certified clinicians over 400 generated reports, with clinical relevance measured using a mean 5-point Likert score (Table 10). The compared systems included GPT-4o zero-shot, Gemini-1.5-Pro zero-shot, and the proposed Gemini-1.5-Pro + confidence-RAG advisor (Table 10). The evaluated criteria included clinical relevance, factual accuracy, guideline concordance, hallucination rate, and safety compliance.The proposed Gemini-1.5-Pro + confidence-RAG advisor achieved the highest clinical relevance score of 4.63, compared with 4.18 for GPT-4o zero-shot and 4.21 for Gemini-1.5-Pro zero-shot (Table 10). This corresponds to a +0.45 improvement over GPT-4o zero-shot and a +0.42 improvement over Gemini-1.5-Pro zero-shot (Table 10). The proposed advisor also achieved factual accuracy of 0.9410, compared with 0.8760 for GPT-4o zero-shot and 0.8820 for Gemini-1.5-Pro zero-shot, corresponding to +0.0650 and +0.0590 improvements, respectively (Table 10).For guideline concordance, the proposed Gemini-1.5-Pro + confidence-RAG advisor achieved 0.9280, compared with 0.8490 for GPT-4o zero-shot and 0.8570 for Gemini-1.5-Pro zero-shot (Table 10). This corresponds to improvements of +0.0790 and +0.0710, respectively (Table 10). Safety compliance was also highest for the proposed advisor, with 0.9780 compared with 0.9020 for GPT-4o zero-shot and 0.9130 for Gemini-1.5-Pro zero-shot, corresponding to +0.0760 and +0.0650 improvements, respectively (Table 10).The hallucination rate was lowest for

the proposed Gemini-1.5-Pro + confidence-RAG advisor at 0.0320, compared with 0.0810 for GPT-4o zero-shot and 0.0740 for Gemini-1.5-Pro zero-shot (Table 10). This represents a hallucination-rate reduction of −0.0490 compared with GPT-4o zero-shot and −0.0420 compared with Gemini-1.5-Pro zero-shot (Table 10). These findings indicate that confidence conditioning and retrieval augmentation improved the clinical reliability and safety profile of the generated reports compared with zero-shot language-model generation.

Table 10. Blinded clinician evaluation of Gemini-1.5-Pro-based confidence-conditioned clinical advisory reports

System	Clinical Relevance (1–5)	Factual Accuracy	Guideline Concordance	Hallucination Rate	Safety Compliance
GPT-4o (zero-shot)	4.18	0.8760	0.8490	0.0810	0.9020
Gemini-1.5-Pro (zero-shot)	4.21	0.8820	0.8570	0.0740	0.9130
[proposed] (Gemini-1.5-Pro + conf-RAG)	**4.63**	**0.9410**	**0.9280**	**0.0320**	**0.9780**

6. Clinical Decision-Support Report Generation

Figure 4 illustrates an example clinician-facing decision-support report generated after the PolyDx-Mamba image classification output is passed to the Gemini-1.5-Pro clinical advisor. The example is based on a dermoscopic skin lesion image, where the predicted class is melanoma and the model confidence is 0.95. The report is structured to present the model output in a clinically interpretable format rather than as a standalone diagnostic decision. The first panel displays the input lesion image and highlights the visually suspicious region, allowing the reader to connect the model prediction with the image-level evidence. The prediction panel then summarizes the predicted skin class, confidence score, and AI system identity. In the final manuscript figure, the AI system should be labelled as PolyDx-Mamba + Gemini-1.5-Pro Clinical Advisor to remain consistent with the proposed framework.The clinical interpretation section provides a cautious explanation of the predicted class. In this example, the generated text states that melanoma is a malignant skin lesion arising from melanocytes and may present with asymmetry, border irregularity, color variation, diameter enlargement, and recent evolution. The report further notes that the lesion image demonstrates suspicious pigmented characteristics that may warrant expert dermatologic assessment. Importantly, the interpretation does not present the AI output as definitive diagnosis. Instead, it emphasizes that final diagnosis requires clinical correlation and, when indicated, dermoscopic examination and histopathological confirmation. This wording is essential for clinical safety because image-level AI classification cannot replace specialist assessment, clinical history, physical examination, or tissue-based confirmation when required.

The recommendation section converts the prediction into a practical clinical decision-support message. For the melanoma-predicted case, the report recommends prompt referral to a dermatologist or dermato-oncology specialist for comprehensive evaluation. It also advises clinical examination, dermoscopic review, and lesion documentation. When the lesion is clinically suspicious, biopsy or excision is suggested according to standard clinical practice. The report additionally highlights warning signs that may require urgent evaluation, including rapid change in size, color, or border, bleeding, ulceration, persistent itching, pain, or other concerning changes. This structure allows the generated output to support referral prioritization and patient education while avoiding autonomous treatment planning.The follow-up question section improves usability by addressing likely user concerns after receiving an AI prediction. The first question explains what should be done after the AI output and recommends review by a qualified dermatologist, further clinical evaluation, dermoscopic assessment, and possible biopsy when clinically indicated. The second question clarifies that the prediction does not confirm skin cancer. Instead, it states that the AI model identifies suspicious image features consistent with melanoma and that final diagnosis requires specialist interpretation and, when necessary, tissue diagnosis. This component is important because it reduces the risk of patient misunderstanding, overconfidence, or unnecessary anxiety after receiving a high-risk AI classification.The safety statement at the bottom of the report defines the intended use of the system. It states that the report is generated for clinical decision support and patient education only and does not replace professional medical diagnosis, dermatological interpretation, or treatment planning. It also advises immediate medical evaluation if the lesion is rapidly changing, bleeding, ulcerated, painful, or otherwise concerning. This safety framing is necessary because the Gemini-1.5-Pro clinical advisor can help transform model predictions into readable recommendations, but it remains dependent on the classifier output, retrieved clinical context, and generated language. Therefore, the report must always be reviewed and confirmed by a qualified clinician before any clinical decision or intervention.

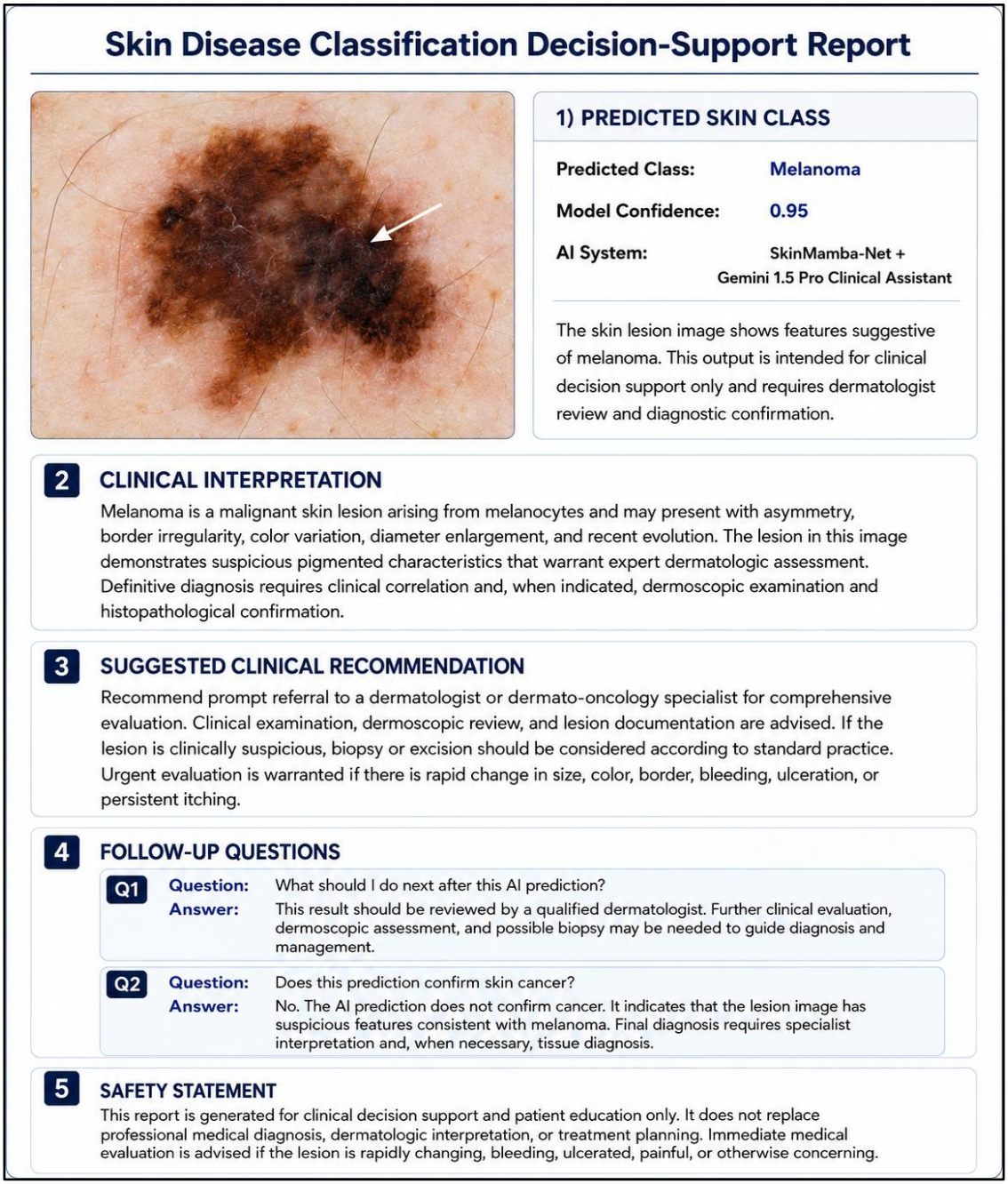

Figure 4. Example clinician-facing decision-support report generated from a dermoscopic skin lesion prediction using PolyDx-Mamba and Gemini-1.5-Pro

7. Limitations and Future Work

This study has several limitations that should be considered when interpreting the reported findings. First, the evaluation was performed using public retrospective datasets rather than prospectively collected multi-center clinical cohorts. Although the four datasets cover different imaging modalities, including Brain MRI, dermoscopy, chest X-ray, and fundus photography, they may not fully represent the variability encountered in real clinical environments. Differences in scanner type, acquisition protocol, imaging device, patient population, disease prevalence, referral pattern, image quality, and annotation practice may affect model performance after deployment. Therefore, the current results should be interpreted as evidence of performance under the reported experimental protocol, not as proof of prospective clinical generalization.Second, the study is based on image-level classification. Patient-level metadata, longitudinal follow-up, clinical history, laboratory findings, treatment information, and histopathological confirmation were not integrated into the model. This is important because clinical diagnosis is rarely based on image appearance alone. For example, skin lesion assessment may require dermoscopic examination, clinical evolution, and biopsy when indicated; diabetic retinopathy grading may require ophthalmic examination and longitudinal monitoring; and brain

tumor interpretation requires radiological context and clinical correlation. Therefore, PolyDx-Mamba should be used only as a clinical decision-support tool and not as an autonomous diagnostic device.

Third, the datasets may contain label noise, class imbalance, acquisition bias, and hidden duplicate or near-duplicate images, which are common challenges in public medical imaging datasets. Although class re-weighting and balanced sampling were used during training, these strategies cannot fully eliminate dataset bias. In addition, the absence of fully standardized patient-level identifiers limits the ability to verify whether all splits are completely independent at the patient level. Future work should validate the framework on datasets with explicit patient identifiers, institution-level separation, and temporally separated test cohorts to better assess real-world generalization.Fourth, the current study evaluates a unified architecture across four disease tasks, but the model has not yet been tested under broader domain-shift settings. External validation across additional hospitals, countries, scanners, dermoscopy devices, fundus cameras, and chest radiography systems is required. Future studies should include multi-center external testing, cross-dataset transfer evaluation, subgroup analysis, and robustness testing under changes in image quality, acquisition condition, disease prevalence, and demographic composition. Such analysis is necessary before considering clinical workflow integration.Fifth, although the proposed uncertainty-aware design reports calibration through expected calibration error, calibration remains task dependent. In particular, the DR grading task showed the highest ECE among the four datasets, indicating that retinal grading may require additional calibration refinement. Future work should investigate post-hoc calibration, ordinal-aware DR modeling, uncertainty thresholding, and referral-sensitive decision rules. Calibration should also be evaluated under external domain shift, because a model that is calibrated on an internal test set may become miscalibrated when applied to data from another clinical site.

Sixth, the current ablation study demonstrates the contribution of major architectural modules, including the Mamba encoder, modality-specific stem, channel-spatial attention, gated fusion, and uncertainty-aware head. However, more fine-grained ablations are still needed. Future work should evaluate alternative fusion strategies, different attention modules, task-specific versus shared encoder depth, modality-specific normalization, and disease-specific uncertainty thresholds. Such analysis would help clarify which components are most important for each modality and whether all modules are equally useful across MRI, dermoscopy, chest X-ray, and fundus imaging.Seventh, the computational cost of PolyDx-Mamba is higher than several lighter baselines. The model uses 31.4M parameters, 4.82G FLOPs, 14 ms inference time, 71 img/s throughput, and 6,180 MB GPU memory (Table 8). Although this cost is reasonable for workstation-based research and decision-support settings, it may limit deployment in low-resource clinical environments or edge devices. Future work should explore model compression, pruning, quantization, knowledge distillation, and hardware-aware optimization to reduce memory use and inference latency while preserving classification and calibration performance.Eighth, the Gemini-1.5-Pro clinical advisor remains a language-generation component and therefore carries residual safety risk. The confidence-conditioned retrieval-augmented design reduced the hallucination rate compared with zero-shot systems, but hallucinations were not eliminated, with the proposed advisor still showing a hallucination rate of 0.0320 (Table 10). Therefore, the generated reports must always be reviewed by qualified clinicians. Future work should strengthen the clinical advisor using stricter retrieval verification, source attribution, uncertainty-triggered refusal, structured report templates, specialty-specific guideline databases, and clinician-in-the-loop validation.

8. Conclusion

This study presented PolyDx-Mamba, a unified multi-disease medical image classification framework for Brain MRI, skin dermoscopy, chest X-ray, and fundus imaging. The model combines modality-specific stem and patch embedding, a shared Vision-Mamba hierarchical encoder, channel-spatial attention, cross-disease global context modeling, gated multi-scale feature fusion, and uncertainty-aware representation learning. This design enables the framework to capture both local disease patterns and long-range anatomical context while preserving separate task-specific heads for brain tumor classification, skin lesion classification, chest disease classification, and diabetic retinopathy grading. Compared with recent convolutional, transformer-based, hybrid, medical vision, and Mamba-based baselines, PolyDx-Mamba showed strong and consistent performance across the evaluated tasks, supporting the value of shared multi-disease representation learning.The framework also extends beyond class prediction by integrating a Gemini-1.5-Pro confidence-conditioned retrieval-augmented clinical advisor that converts model outputs into structured decision-support reports. This reporting layer improves the clinical usability of the system by combining predicted class, confidence, uncertainty information, and retrieved clinical knowledge into a readable recommendation format. However, PolyDx-Mamba is intended only as a clinical decision-support tool and not as an autonomous diagnostic device. All predictions and generated recommendations must be reviewed and confirmed by qualified clinicians. Future work should focus on external multi-center validation, prospective clinical evaluation, patient-level testing, subgroup robustness analysis, calibration under domain shift, and stronger safeguards for language-model-generated clinical recommendations.

Author Contributions: Md Anwar Hossain contributed to the conceptualization of the study, overall research design, model development, experimental planning, result analysis, and manuscript preparation. Md Abedur Rahman contributed to methodology refinement, implementation support, dataset organization, evaluation design, and interpretation of experimental findings. Md Sahid Hossain contributed to model validation, comparative analysis, performance evaluation, and technical manuscript revision. Kallol Chakraborty Shekhor contributed to literature review, clinical decision-support framing, manuscript organization, and critical editing of the final draft. All authors reviewed the manuscript, approved the final version, and agreed to be accountable for the content of the work.

Data Availability: The datasets used in this study are publicly available and were obtained from established repositories. No new data were generated by the authors.

Declarations

Ethical approval

This article does not report any prospective studies involving human participants or animals performed by the authors. The analysis is based on previously collected, fully anonymized oral cancer images provided to the authors; therefore, additional institutional ethical approval and clinical trial registration were not required.

Consent to participate

Not applicable. The study used only secondary, anonymized image data and involved no direct contact or intervention with individual participants.

Consent to publish

Not applicable. This manuscript does not contain any individual person's identifiable data

References

1.Khan, Ejaz, Muhammad Zia Ur Rehman, Fawad Ahmed, Faisal Abdulaziz Alfouzan, Nouf M. Alzahrani, and Jawad Ahmad. "Chest X-ray classification for the detection of COVID-19 using deep learning techniques." *Sensors* 22, no. 3 (2022): 1211.
2.Shelke, Ankita, Madhura Inamdar, Vruddhi Shah, Amanshu Tiwari, Aafiya Hussain, Talha Chafekar, and Ninad Mehendale. "Chest X-ray classification using deep learning for automated COVID-19 screening." *SN computer science* 2, no. 4 (2021): 300.
3.Agarwal, Vertika, M. C. Lohani, Ankur Singh Bist, Eka Purnama Harahap, and Alfiah Khoirunisa. "Analysis of deep learning techniques for chest x-ray classification in context of covid-19." *ADI Journal on Recent Innovation* 3, no. 2 (2022): 208-216.
4.Xie, Yuting, Fulvio Zaccagna, Leonardo Rundo, Claudia Testa, Raffaele Agati, Raffaele Lodi, David Neil Manners, and Caterina Tonon. "Convolutional neural network techniques for brain tumor classification (from 2015 to 2022): Review, challenges, and future perspectives." *Diagnostics* 12, no. 8 (2022): 1850.
5.Saleh, Ahmad, Rozana Sukaik, and Samy S. Abu-Naser. "Brain tumor classification using deep learning." In *2020 International Conference on Assistive and Rehabilitation Technologies (iCareTech)*, pp. 131-136. IEEE, 2020.
6.Khan, Hassan Ali, Jue Wu, Muhammad Mushtaq, and Muhammad Umer Mushtaq. "Brain tumor classification in MRI image using convolutional neural network." *Mathematical Biosciences and Engineering* 17, no. 5 (2020): 6203.
7.Liu, Ze, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning et al. "Swin transformer v2: Scaling up capacity and resolution." In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12009-12019. 2022.
8.Dihin, Rasha A., Ebtesam N. Al Shemmary, and Waleed A. Mahmoud Al-Jawher. "Implementation of the swin transformer and its application in image classification." *Journal Port Science Research* 6, no. 4 (2023): 318-331.
9.Wei, Jiangshu, Jinrong Chen, Yuchao Wang, Hao Luo, and Wujie Li. "Improved deep learning image classification algorithm based on Swin Transformer V2." *PeerJ Computer Science* 9 (2023): e1665.
10.He, Y., Nath, V., Yang, D., Tang, Y., Myronenko, A. and Xu, D., 2023, October. Swinunetr-v2: Stronger swin transformers with stagewise convolutions for 3d medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 416-426). Cham: Springer Nature Switzerland.
11.He, Xin, Yong Zhou, Jiaqi Zhao, Di Zhang, Rui Yao, and Yong Xue. "Swin transformer embedding UNet for remote sensing image semantic segmentation." *IEEE transactions on geoscience and remote sensing* 60 (2022): 1-15.
12.FRATTICIOLI, GUGLIELMO. "MambaTransfer: raw audio musical timbre transfer using selective state-space models." (2023).
13.NOGALES GONZALEZ-REGUERAL, M. I. G. U. E. L. "Dynamic computing for LLMs: the role of early exits in transformers and

Mamba based architectures." (2023).
14.Liu, P., Song, Y., Chai, M., Han, Z., & Zhang, Y. (2021). Swin–UNet++: a nested swin transformer architecture for location identification and morphology segmentation of dimples on 2.25 Cr1Mo0. 25V fractured surface. *Materials*, *14*(24), 7504.
15.Fu, Zujun. "Vision transformer: Vit and its derivatives." *arXiv preprint arXiv:2205.11239* (2022).
16. Cai, Lei, Jingyang Gao, and Di Zhao. "A review of the application of deep learning in medical image classification and segmentation." Annals of Translational Medicine 8, no. 11 (2020): 713.
17. Huang, Shih-Cheng, Anuj Pareek, Malte Jensen, Matthew P. Lungren, Serena Yeung, and Akshay S. Chaudhari. "Self-supervised learning for medical image classification: a systematic review and implementation guidelines." NPJ Digital Medicine 6, no. 1 (2023): 74.
18. Manzari, Omid Nejati, Hamid Ahmadabadi, Hossein Kashiani, Shahriar B. Shokouhi, and Ahmad Ayatollahi. "MedViT: a robust vision transformer for generalized medical image classification." Computers in Biology and Medicine 157 (2023): 106791.
19. Dai, Y., Gao, Y., and Liu, F. "Transmed: Transformers advance multi-modal medical image classification." Diagnostics 11, no. 8 (2021): 1384.
20. Liu, Quande, Lequan Yu, Luyang Luo, Qi Dou, and Pheng Ann Heng. "Semi-supervised medical image classification with relation-driven self-ensembling model." IEEE Transactions on Medical Imaging 39, no. 11 (2020): 3429–3440.
21. Yang, Yijun, Huazhu Fu, Angelica I. Aviles-Rivero, Carola-Bibiane Schönlieb, and Lei Zhu. "Diffmic: Dual-guidance diffusion network for medical image classification." In International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 95–105. Cham: Springer Nature Switzerland, 2023.
22. Tang, Hai, and Zhihui Hu. "Research on medical image classification based on machine learning." IEEE Access 8 (2020): 93145–93154.
23. Wu, Xin, Yue Feng, Hong Xu, Zhuosheng Lin, Tao Chen, Shengke Li, Shihan Qiu, Qichao Liu, Yuangang Ma, and Shuangsheng Zhang. "CTransCNN: Combining transformer and CNN in multilabel medical image classification." Knowledge-Based Systems 281 (2023): 111030.
24. Wang, Dequan, Xiaosong Wang, Lilong Wang, Mengzhang Li, Qian Da, Xiaoqiang Liu, Xiangyu Gao, et al. "A real-world dataset and benchmark for foundation model adaptation in medical image classification." Scientific Data 10, no. 1 (2023): 574.
25. Azizi, Shekoofeh, Basil Mustafa, Fiona Ryan, Zachary Beaver, Jan Freyberg, Jonathan Deaton, Aaron Loh, et al. "Big self-supervised models advance medical image classification." In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3478–3488. 2021.
26. Ashraf, Rehan, Muhammad Asif Habib, Muhammad Akram, Muhammad Ahsan Latif, Muhammad Sheraz Arshad Malik, Muhammad Awais, Saadat Hanif Dar, Toqeer Mahmood, Muhammad Yasir, and Zahoor Abbas. "Deep convolution neural network for big data medical image classification." IEEE Access 8 (2020): 105659–105670.
27. Patrício, Cristiano, Joao C. Neves, and Luís F. Teixeira. "Explainable deep learning methods in medical image classification: A survey." ACM Computing Surveys 56, no. 4 (2023): 1–41.
28. Cheng, Junlong, Shengwei Tian, Long Yu, Chengrui Gao, Xiaojing Kang, Xiang Ma, Weidong Wu, Shijia Liu, and Hongchun Lu. "ResGANet: Residual group attention network for medical image classification and segmentation." Medical Image Analysis 76 (2022): 102313.
29. Korot, Edward, Zeyu Guan, Daniel Ferraz, Siegfried K. Wagner, Gongyu Zhang, Xiaoxuan Liu, Livia Faes, et al. "Code-free deep learning for multi-modality medical image classification." Nature Machine Intelligence 3, no. 4 (2021): 288–298.
30. An, F., Li, X., and Ma, X. "Medical image classification algorithm based on visual attention mechanism-MCNN." Oxidative Medicine and Cellular Longevity 2021, no. 1 (2021): 6280690.
31. Singh, Rishav, Vandana Bharti, Vishal Purohit, Abhinav Kumar, Amit Kumar Singh, and Sanjay Kumar Singh. "MetaMed: Few-shot medical image classification using gradient-based meta-learning." Pattern Recognition 120 (2021): 108111.
32. Kowsari, Kamran, Rasoul Sali, Lubaina Ehsan, William Adorno, Asad Ali, Sean Moore, Beatrice Amadi, Paul Kelly, Sana Syed, and Donald Brown. "Hmic: Hierarchical medical image classification, a deep learning approach." Information 11, no. 6 (2020): 318.
33. Ren, Zeyu, Xiangyu Kong, Yudong Zhang, and Shuihua Wang. "UKSSL: Underlying knowledge based semi-supervised learning for medical image classification." IEEE Open Journal of Engineering in Medicine and Biology 5 (2023): 459–466.
34. Müller, Dominik, Iñaki Soto-Rey, and Frank Kramer. "An analysis on ensemble learning optimized medical image classification with deep convolutional neural networks." IEEE Access 10 (2022): 66467–66480.
35. Almalik, F., Yaqub, M., and Nandakumar, K. "Self-ensembling vision transformer (SEViT) for robust medical image classification." In International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 376–386. Cham: Springer Nature Switzerland, 2022.