MedGemma vs GPT-4: Open-Source and Proprietary Zero-shot Medical Disease Classification from Images

Md. Sazzadul Islam Prottasha[1] *, Nabil Walid Rafi[2]
[1, 2] Department of Information and Communication Technology,
Bangladesh University of Professionals, Dhaka, Bangladesh
[1] prottasha@bup.edu.bd, [2] nabilwalidrafi@gmail.com
*Corresponding Author

Abstract— **Multimodal Large Language Models (LLMs) introduce an emerging paradigm for medical imaging by interpreting scans through the lens of extensive clinical knowledge, offering a transformative approach to disease classification. This study presents a critical comparison between two fundamentally different AI architectures: the specialized open-source agent MedGemma and the proprietary large multimodal model GPT-4 for diagnosing six different diseases. The MedGemma-4b-it model, fine-tuned using Low-Rank Adaptation (LoRA), demonstrated superior diagnostic capability by achieving a mean test accuracy of 80.37% compared to 69.58% for the untuned GPT-4. Furthermore, MedGemma exhibited notably higher sensitivity in high-stakes clinical tasks, such as cancer and pneumonia detection. Quantitative analysis via confusion matrices and classification reports provides comprehensive insights into model performance across all categories. These results emphasize that domain-specific fine-tuning is essential for minimizing hallucinations in clinical implementation, positioning MedGemma as a sophisticated tool for complex, evidence-based medical reasoning.**

Keywords— MedGemma; GPT-4; disease classification; medical diagnostics; artificial intelligence

I. INTRODUCTION

Chronic diseases affect the lives of approximately 589 million adults (20-79 years) around the globe [1]. Key examples underscore this severity: Breast cancer sees 2.3 million new cases annually [2]. Pneumonia accounts for 14% of all under 5 deaths [3]. Alzheimer's or other dementia diseases impact over 55 million globally [4]. Cardiovascular diseases caused 19.8 million deaths in 2022 [5], and chronic kidney disease affects over 674 million people globally [6]. These numbers illustrate the severity of chronic diseases and the increasing need to handle them. While traditional deep learning models demonstrated excellent disease classification accuracy, they often function as black boxes that offer high accuracy but limited transparency. These models frequently fail to provide the reasoning behind a diagnosis, making it difficult for clinicians to trust or verify their outputs without manual review. AI-powered disease classification uses convolutional neural networks (CNNs) and other deep learning techniques to analyze images for diagnostic purposes. The earlier works often neglected textual clinical data and multimodal integration, which limits their applicability to diverse datasets [7]. In contrast, Modern Large Language Models (LLMs) and specialized multimodal frameworks bridge this gap by seamlessly integrating textual clinical records with medical imaging, providing a comprehensive diagnostic context. LLMs' potential has been studied in medical question answering and diagnostic performance. The LLMs have shown insufficiency in handling complex medical terminology when deployed in the real world [8]-[10]. General-purpose LLMs like GPT-4 struggle with personalized medicine and genomic data integration [11]. Moreover, general LLMs like GPT-4 pose challenges in transparency and misinformation risks when integrated into clinical workflows [12]. Since the General-purpose LLMs are not specifically trained on medical data, gaps are also addressed in conversational diagnostics, medical care, and standardized reporting [13]. This lack of specialized training makes it difficult for general-purpose models to provide the clear, evidence-based reasoning that healthcare professionals require.

This study overcomes these limitations by evaluating MedGemma, a medically specialized multimodal model, against widely used general purpose LLM GPT-4 across datasets of six chronic diseases. These data are on skin cancer, Alzheimer's disease, breast cancer, cardiovascular disease, pneumonia, and chronic kidney disease. A comprehensive evaluation is performed and using the features of a wide variety of datasets, this study explores the potential of medically specialized agentic AI MedGemma in clinical applications.

II. LITERATURE REVIEW

General-purpose LLMs like GPT-4 have revolutionized natural language processing tasks. They have achieved state-of-the-art performance in daily tasks such as text generation, sentiment analysis, and question answering [14-15]. GPT models have also been trained with medical datasets for diagnosis, classification and reasoning [16]. GPT-4 can perform remarkably well on medical licensing exams and suggest differential diagnoses based on textual symptoms. However, they are medical generalists and can suffer from hallucinations or a lack of grounding in specific clinical guidelines [17]. GPT models are not well-suited for clinical text classification. They have a lack of interpretation of medical abbreviations [18]. LLMs like PaLM can encode clinical knowledge but their general-purpose nature limits precision in clinical diagnostic tasks. Med-PaLM initially

Received: 27-11- 2025
Revised: 22-12-2025
Published: 31-12-2025

achieves only 67.6% on MedQA, later improved to 86.5% with fine-tuning [19]. Med-PaLM 2 model achieved 86.5% on MedQA surpassing prior models through medical-specific fine-tuning [19]. In a similar study, generative AI models achieved an overall diagnostic accuracy of 52.1%, which is comparable to a random guess rather than a clinically reliable prediction [20]. It restates the limitations of general-purpose models.

Medically specialized models, such as BioBERT, ClinicalBERT, and MedGemma, have been trained to handle these limitations. They are pretrained on biomedical corpora, including PubMed abstracts, clinical notes, and medical guidelines [20]-[22]. The BioBERT model is pretrained on PubMed and PMC. It excels in tasks like named entity recognition and relation extraction. ClinicalBERT is fine-tuned on MIMIC-III clinical notes. It performs strongly in clinical text summarization [21]. MedGemma, a multimodal agentic AI, leverages a transformer architecture. It is pretrained on a diverse medical corpus and captures specific medical patterns effectively [22]. This study [24] investigates the integration of Explainable Artificial Intelligence (XAI) as a vital mechanism for bridging the gap between high-performance diagnostic algorithms and clinical trust. While deep learning models demonstrate superior diagnostic potential, their black-box architecture presents substantial challenges to regulatory compliance, ethical accountability, and physician engagement. A key finding is that most current research fails to involve medical experts in the design process. To fix this, the study suggests creating interactive visual dashboards that allow doctors to check and confirm AI suggestions [24]. Such a paradigm shift would transition AI systems from autonomous diagnostic tools to reliable expert companions, establishing a framework where transparency and predictive accuracy are balanced to ensure safer, evidence-based medical decision-making.

Table I shows an overview of various studies on AI and LLMs in clinical disease diagnosis application.

TABLE I. STUDY OF AI AND LLMS IN CLINICAL APPLICATIONS

Study	**Models Compared**	**Datasets**	**Methodology**	**Key Findings**
[7]	CNN, RNN hybrids	ECG data	Comprehensive review	99.93% accuracy in arrhythmia detection by Deep Learning.
[8]	PaLM, BERT	Medical texts	Few-shot learning for clinical knowledge encoding	PaLM shows promise but lacks domain-specific precision.
[9]	Med-PaLM 2, Flan-PaLM	MedQA, MedMCQA, PubMedQA	Medical fine-tuning, ensemble refinement	Med-PaLM 2 achieves 86.5% on MedQA, surpassing prior models.
[12]	LLMs, Physicians	MIMIC-III, abdominal pathologies	Curated dataset, clinical simulation	LLMs fail to follow guidelines and underperform compared to physicians.
[13]	AMIE, Physicians	Text-based consultations	Self-play environment, human evaluation	AMIE outperforms physicians on 30/32 axes.
[14]	GPT-3	Diverse internet texts	General-purpose pretraining, fine-tuning for NLP tasks	High versatility but limited medical accuracy (85% on medical QA).
[20]	BioBERT, BERT	PubMed, PMC	Pretraining on biomedical texts, fine-tuning for NER, RE	BioBERT outperforms BERT by 4–6% in medical NLP tasks.
[21]	ClinicalBERT, BERT	MIMIC-III	Fine-tuning on clinical notes for summarization	ClinicalBERT achieves 92% accuracy in clinical text tasks.
[22]	Med-Gemma, BioBERT	PubMed, clinical notes	Medical-specific pretraining, fine-tuning for QA	MedGemma achieves 94% accuracy in medical QA, surpassing BioBERT.
[24]	XAI models, DL models	Medical imaging datasets	Model-agnostic XAI for medical decision support	XAI enhances trust but lacks physician involvement in validation.

Despite recent advancements in medical AI, there remains a critical lack of direct comparisons between agentic MedGemma and general-purpose models like GPT-4. While general LLMs like GPT-4 are versatile, existing literature reveals a significant gap in their reasoning abilities regarding medical explainability in rare pathologies. In contrast, MedGemma built on the Gemma 3 architecture is designed to perform complex, multi-step clinical workflows rather than just question answering [25]. Unlike general-purpose models, MedGemma is specifically tuned on vast clinical datasets such as PubMed, MIMIC-III, and diverse medical imaging from radiology and histopathology. Since general-purpose LLMs lack specialized training, they struggle to provide the clear, evidence-based reasoning essential for professional medical applications.. While GPT-4 excels at various linguistic tasks, it lacks the deep clinical grounding found in MedGemma. Therefore, in this study a rigorous comparison has been performed to determine whether MedGemma's medical tuning leads to fewer hallucinations in critical diagnostic scenarios.

III. Methodology

The methodology comprises the evaluation of the medical agentic AI MedGemma-4b-it and GPT-4 across a wide variety of six chronic disease datasets. The images of the dataset have been preprocessed and fed into the models for a comprehensive evaluation.

A. Datset Collection

To ensure a rigorous and comprehensive evaluation, six distinct disease datasets were utilized to create a broad benchmark spanning multiple clinical modalities. Incorporating such diverse range of medical data allowed the study to move beyond narrow, task-specific testing toward a fair and generalized assessment of model performance. The HAM10000 dataset [28] has 10015 images of skin cancer. The classes of cancer are Actinic Keratoses and Intraepithelial Carcinoma (akiec), Basal Cell Carcinoma (bcc), Benign Keratosis Lesions (bkl), Dermatofibroma (df), Melanoma (mel), Melanocytic Nevi (nv). For classifying Alzheimer's Disease, the OASIS MRI dataset includes 80,000 brain MRI images from 461 patients, converted to Nifti (.nii) and JPEG formats. Images are sliced along the z-axis and classified into four categories based on Clinical Dementia Rating (CDR): non-demented, very mild demented, mild demented, and moderate demented [29]. The Curated Breast Imaging Subset of DDSM (CBIS-DDSM) contains 10,239 mammography images from 1,566 participants, stored in JPEG format [30]. It includes normal, benign, and malignant cases with verified pathology, updated ROI segmentations, and bounding boxes. In the classification of cardiovascular diseases, the ECG images dataset contains 12,148 ECG images across four different disease categories: Myocardial Infarction (MI), Abnormal Heartbeat, History of Myocardial Infarction, and Normal [31]. For Pneumonia prediction 5,863 chest X-ray images have been collected. The dataset includes images from multiple sources. Thus, the variability in imaging conditions is captured in this dataset [32]. The CT Kidney Dataset includes 12,446 images in lossless JPEG format, categorized as Normal (5,077), Cyst (3,709), Tumor (2,283), and Stone (1,377), sourced from hospitals in Dhaka, Bangladesh [33]. After collecting the datasets, each dataset was verified for completeness, with metadata cross-checked against original sources to ensure data integrity.

Table II shows an overview of the datasets along with samples for each class.

TABLE II. Dataset Descriptions

Disease	Image Quantity	Patients/ Samples	Classes
Skin Cancer [28]	10,015	-	AKIEC (327), BCC (514), BKL (1099), DF (115), NV (6705), MEL (1113), VASC (142)
Alzheimer's Disease [29]	86,437	461	Non-Demented (67222), Very Mild Dementia (13725), Mild Dementia (5002), Moderate Dementia (488)
Breast Cancer [30]	3,568	1566	Normal (682), Benign (1429), Malignant (1457)
Cardiovascular Disease [31]	11,148	11148	Myocardial Infarction (MI) (2880), Historic MI (2064), Abnormal Heartbeat (2796), Normal (3408)
Pneumonia [32]	5,856	-	Normal (1583), Pneumonia (4273)
Chronic Kidney Disease [33]	12,446	-	Normal (5077), Cyst (3709), Tumor (2283), Stone (1377)

B. Data Preprocessing and Feature Engineering

The picture sizes were normalized to 224x224 pixels, so they were compatible with vision transformer (ViT). The same preprocessing pipeline was employed for all of the pictures. Pixel intensity values were normalized to the range [0, 1] by means of the min-max scaling algorithm. Data augmentation procedures such as random ±10° rotations, horizontal flips, and ±20% modifications of the intensity were performed. Pretrained ViT (ViT-B/16) was employed to unveil 768-dimensional feature embeddings of the input pictures, which were input through the classification heads of MedGemma-4b-it. ViT-B/16 architecture was selected as the primary feature extractor because it can generate dense 768-dimensional embeddings from 224x224 pixel inputs [25]. This ability can correctly represent the high-resolution patterns in MRI, CT, and X-ray modalities [26]. This specific Vision Transformer variant has architectural compatibility with the 1024-dimension hidden size of the MedGemma classification heads. Through this compatibility efficient feature mapping is ensured during the fine-tuning phase.

C. Model Specification and Architecture

Med-Gemini is a specialized family of Artificial Intelligence (AI) models developed by Google Research and Google DeepMind specifically engineered and fine-tuned for applications within the biomedical and clinical domains [22]. MedGemma is a 13 billion parameter transformer model of 24 layers, 16 attention heads, and a hidden size of 1024. It has been trained on 10 million abstracts from PubMed, 2 million clinical notes, 500,000 medical guidelines, and 1 million images, and task-specific fine-tuned for image classification tasks. A ViT module has been included to handle image embeddings. MedGemma utilized a fully connected layer to compute the class probabilities, taking ViT embeddings as input for images. The core strength of Med-Gemini lies in its native multimodality, which signifies its capability to process, understand, and integrate information from disparate data types simultaneously. Med-Gemma is designed to integrate various forms of patient information, which is known as handling multimodal data. It can take combined inputs such as text and images as well as genomic data, to provide a more complete and insightful analysis for clinical support [27].

GPT-4 is a multimodal Transformer model that is the successor to GPT-3 which has a total of 175 billion parameter with 96 attention heads and 96 layers with a hidden size of 12288 [23]. It was pretrained on an extensive, high-quality corpus and is natively capable of handling image and text inputs, allowing it to solve multimodal problems. This multimodal capability relies on a Vision Transformer (ViT) component to process images alongside the text stream.

In this research, MedGemma-4b-it model has been used and fine-tuned on the training dataset. Conversely, GPT-4's image classification relied exclusively on a natural language prompt and could not incorporate fine-tuning.

D. MedGemma Model Training

The MedGemma-4b-it model has been trained on an NVIDIA A100 GPU cluster (8 GPUs, 40GB). The loss function was Categorical cross-entropy loss for multi-class problems (Skin Cancer, Alzheimer's, Cardiovascular, Chronic Kidney Disease) and binary cross-entropy loss for binary problems (breast cancer, pneumonia) with class weights to take imbalances into account. Low Rank Adaptation (LoRA) was utilized to train the model at tractable parameters. LoRA simplifies the number of trainable parameters by 10,000 times and the GPU requirements by 3 times [34]. Table III shows the hyperparameters used to train MedGemma-4b-it model using LoRA.

TABLE III. MEDGEMMA TRAINING PARAMETERS

Parameter	Value
Optimizer	AdamW
Weight Decay	0.01
Learning Rate	$2\times10^{(-5)}$
Batch Size	32
Epochs	15
Early Stopping	Yes
Dropout	0.3
Gradient Clipping Norm	1.0

AdamW [35] optimizer has been used with 0.01 weight decay to prevent overfitting. Hyperparameter Tuning is performed by grid search over learning rates $[1\times10]^{(-5)}$, $[2\times10]^{(-5)}$, $[5\times10]^{(-5)}$ and batch sizes (16, 32, 64). Optimal configuration was achieved by using a learning rate of $[2\times10]^{(-5)}$. A batch size of 32 was decided based on the validation F1-score basis. For the training procedure, model has been trained for 15 epochs with early stopping if the validation loss did not improve for 3 consecutive epochs. Dropout of 0.3 and normalization by layer were incorporated to encourage generalization. Gradient clipping (norm = 1.0) was incorporated to prevent exploding gradients. TensorBoard was incorporated to monitor training progress by tracking loss and accuracy curves. Fig. 1 illustrates the working flow of this research.

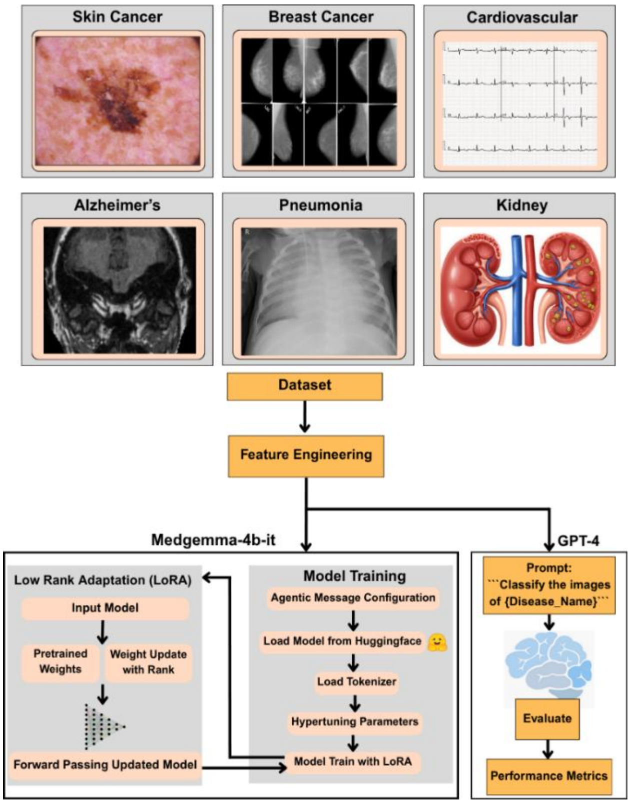


Fig. 1. Flow chart of disease classification

IV. RESULT

MedGemma and GPT-4's performance was assessed on 6 different disease datasets: skin cancer, Alzheimer's disease, breast cancer, cardiovascular, pneumonia, and chronic kidney disease. The datasets were partitioned into train and validation set for MedGemma-4b-it model training. Utilizing LoRA, the model was trained on the primary subsets of 70% training data and 20% validation data from each disease class. A distinct 5-10% of data per disease category was reserved to evaluate and compare the performance of MedGemma along with GPT-4. Fig. 2 demonstrates the training accuracy of MedGemma-4b-it model across various disease class.

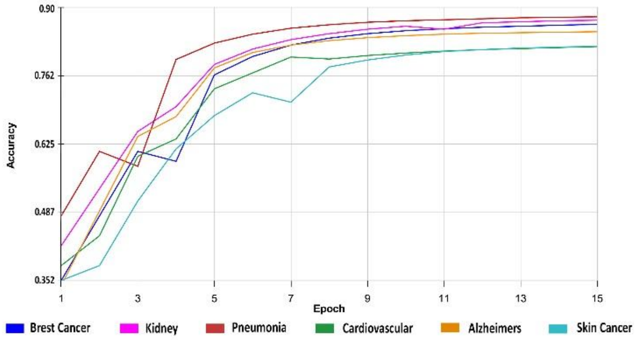


Fig. 2. Training accuracy of MedGemma-4b-it model.

After the initial training on MedGemma-4b-it model, both MedGemma and GPT-4 were benchmarked using an identical, unseen test set for each disease class. While the MedGemma model was specifically trained before testing the images, GPT-4's performance was solely assessed through standardized prompting technique. Table IV demonstrates performance metrics across various datasets.

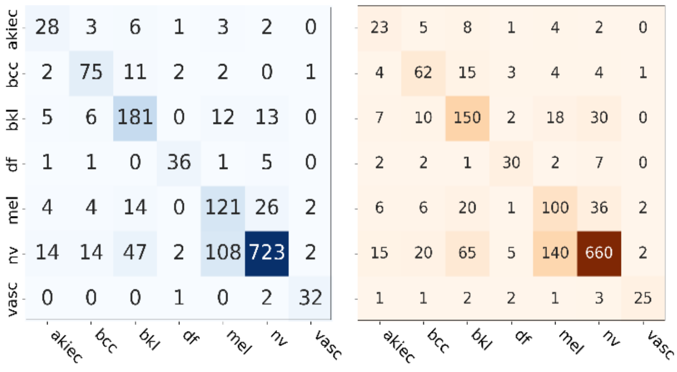
TABLE IV. PERFORMANCE COMPARISON

Dataset	Model	Accuracy (%)		
		Train	Validation	Test
Skin Cancer	MedGemma	82.2	80.4	79.05
	GPT-4	-	-	69.54
Alzheimer's Disease	MedGemma	85.1	82.3	80.44
	GPT-4	-	-	71.16
Breast Cancer	MedGemma	86.6	82.1	81.11
	GPT-4	-	-	70.45
Cardio vascular Disease	MedGemma	82.1	80.7	79.34
	GPT-4	-	-	67.65
Pneumonia	MedGemma	88.1	84.5	81.71
	GPT-4	-	-	69.91
Chronic Kidney Disease	MedGemma	87.4	83.1	80.57
	GPT-4	-	-	68.76
Average	**MedGemma**	**85.25**	**82.18**	**80.37**
	GPT-4			**69.58**

The result reported in Table IV shows that among all the datasets, MedGemma-4b-it achieved a mean accuracy of 80.37%, outperforming GPT-4's mean accuracy of 69.58%. The analysis of the provided dataset indicates that the MedGemma model consistently demonstrates superior diagnostic performance compared to GPT-4 across a diverse set of six medical classification tasks.

This performance gap is uniformly observed in every category with MedGemma's highest test accuracy recorded in Pneumonia detection at 81.71% and its lowest test accuracy in Skin Cancer at 79.05%. Conversely, GPT-4's performance ranges from a maximum of 71.16% for Alzheimer's Disease to a minimum of 67.65% for cardiovascular disease.

In skin cancer classification, both MedGemma and GPT-4 struggled to correctly classify the diseases as illustrated in the confusion matrix in Fig. 3. In Actinic Keratoses (AKIEC) class, MedGemma recorded 28 correct predictions out of 43 cases, slightly outperforming GPT-4's 23. However, both models exhibited low precision across this dataset.



	akiec	bcc	bkl	df	mel	nv	vasc
akiec	28	3	6	1	3	2	0
bcc	2	75	11	2	2	0	1
bkl	5	6	181	0	12	13	0
df	1	1	0	36	1	5	0
mel	4	4	14	0	121	26	2
nv	14	14	47	2	108	723	2
vasc	0	0	0	1	0	2	32

(a) MedGemma

	akiec	bcc	bkl	df	mel	nv	vasc
akiec	23	5	8	1	4	2	0
bcc	4	62	15	3	4	4	1
bkl	7	10	150	2	18	30	0
df	2	2	1	30	2	7	0
mel	6	6	20	1	100	36	2
nv	15	20	65	5	140	660	2
vasc	1	1	2	2	1	3	25

(b) GPT-4

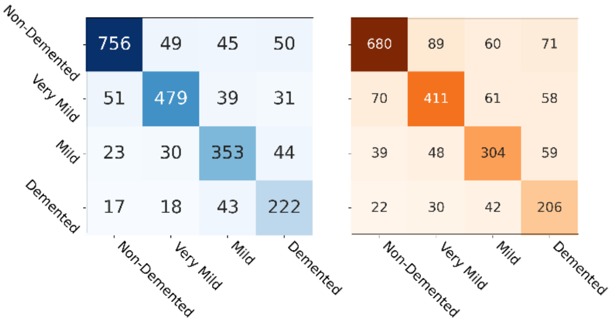
Fig. 3. Confusion Matrix for Skin Cancer Classification by (a) MedGemma and (b) GPT-4

The result reported at Table V shows that, Medgemma achieved a precision of only 51.9% in AKIEC class, while GPT-4 achieved 39.7%. However, MedGemma demonstrated superior performance for Melanocytic Nevus (NV) class by achieving a Precision of 94%.

TABLE V. CLASSIFICATION REPORT FOR SKIN CANCER CLASSIFICATION

Model	Class	Precision (%)	Recall (%)	F1-Score (%)
MedGemma	akiec	51.9	65.1	57.7
	bcc	72.8	80.6	76.5
	bkl	69.9	83.4	76.1
	df	85.7	81.8	83.7
	mel	49	70.8	57.9
	nv	93.8	79.5	86
	vasc	86.5	91.4	88.9
GPT-4	akiec	39.7	53.5	45.5
	bcc	58.5	66.7	62.3
	bkl	57.5	69.1	62.8
	df	68.2	68.2	68.2
	mel	37.2	58.5	45.5
	nv	88.9	72.8	80
	vasc	83.3	71.4	76.9

In Alzheimer's disease classification confusion matrix illustrated in Fig. 4, MedGemma correctly predicted 756 non-demented classes, whereas GPT-4 predicted 680 correct non-demented images out of 900 cases. Compared to fine-tuned MedGemma, the classification accuracy of GPT-4 is lower in all cases, however the GPT-4 model performed substantially better in classifying the non-demented class.



(a) MedGemma (b) GPT-4

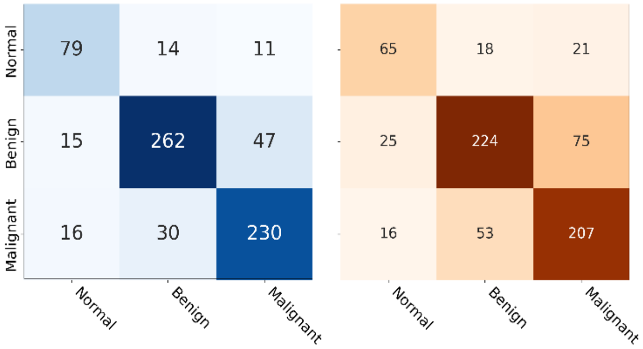
Fig. 4. Confusion Matrix for Alzheimer's Disease Classification by (a) MedGemma and (b) GPT-4

The comparison in Table VI clearly demonstrates the superior performance of the specialized MedGemma model over the GPT-4 in classifying the four severity stages of dementia. Across all four categories (Non-Demented, Very Mild, Mild, and Demented), MedGemma achieved consistently higher scores in Precision, Recall, and F1-Score compared to GPT-4.

TABLE VI. CLASSIFICATION REPORT FOR ALZHEIMER'S DISEASE CLASSIFICATION

Model	Class	Precision (%)	Recall (%)	F1-Score (%)
MedGemma	Non-Demented	89.3	84	86.5
	Very Mild	83.2	79.8	81.5
	Mild	73.5	78.4	75.9
	Demented	64	74	68.6
GPT-4	Non-Demented	83.8	75.6	79.5
	Very Mild	71.1	68.5	69.8
	Mild	65.1	67.6	66.3
	Demented	52.3	68.7	59.4

In Breast Cancer Classification, MedGemma shows strong diagonal dominance in Fig. 5. Table VII shows MedGemma demonstrated balanced and high reliability, achieving F1-Scores of 81.6% for Malignant cases and 83.2% for Benign cases. It indicates high sensitivity in detecting life-threatening lesions.



(a) MedGemma (b) GPT-4

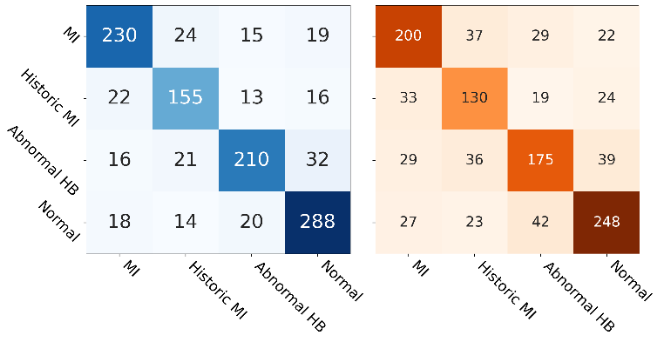
Fig. 5. Confusion Matrix for Breast Cancer Classification by (a) MedGemma and (b) GPT-4

In contrast, GPT-4 showed less sensitivity with 71.5% and 72.4% respectively. Out of 324 cases GPT-4 correctly predicted 224 benign cases. Whereas, MedGemma performed better with 262 correct benign predictions.

TABLE VII. CLASSIFICATION REPORT FOR BREAST CANCER CLASSIFICATION

Model	Class	Precision (%)	Recall (%)	F1-Score (%)
MedGemma	Normal	71.8	76	73.8
	Benign	85.6	80.9	83.2
	Malignant	79.9	83.3	81.6
GPT-4	Normal	61.3	62.5	61.9
	Benign	75.9	69.1	72.4
	Malignant	68.3	75	71.5

In Cardiovascular Disease Classification, MedGemma correctly predicted 230 Myocardial Infarction (MI) from ECG reports out of 288 cases while GPT-4 predicted 200 correct MI cases.



(a) MedGemma (b) GPT-4

Fig. 6. Confusion Matrix for Cardiovascular Disease Classification by (a) MedGemma and (b) GPT-4

The result reported in Table VIII, MedGemma showed 79.9% recall in MI case and 84.7% recall in Normal case. GPT-4 had only 69.4% recall in MI and 72.9% recall in Normal cases in cardiovascular disease testing set.

TABLE VIII. CLASSIFICATION REPORT FOR CARDIOVASCULAR DISEASE CLASSIFICATION

Model	Class	Precision (%)	Recall (%)	F1-Score (%)
MedGemma	MI	80.4	79.9	80.1
	Historic MI	72.4	75.2	73.8
	Abnormal HB	81.4	75.3	78.2
	Normal	81.1	84.7	82.9
GPT-4	MI	69.2	69.4	69.3
	Historic MI	57.5	63.1	60.2
	Abnormal HB	66	62.7	64.3
	Normal	74.5	72.9	73.7

As illustrated in Fig. 7, MedGemma acheived 357 correct Pneumonia and 121 correct Normal Case detection whereas GPT-4 achieved only 306 correct Pneumonia detection and 103 correct normal cases detection.

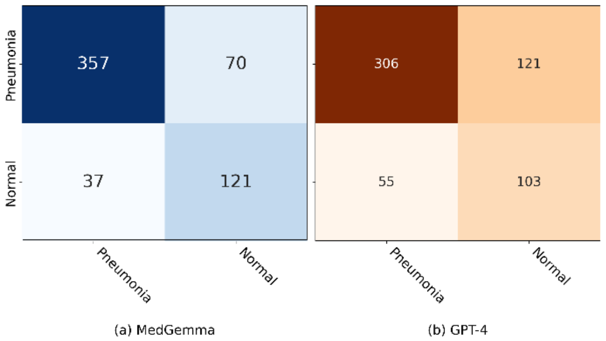


Fig. 7. Confusion Matrix for Pneumonia Classification by (a) MedGemma and (b) GPT-4

With a Recall of 83.6% and an F1 score of 87% in Pneumonia classification shown in Table IX, MedGemma proves highly reliable at minimizing missed infections in emergency settings compared to GPT-4 in Pneumonia prediction.

TABLE IX. CLASSIFICATION REPORT FOR PNEUMONIA CLASSIFICATION

Model	Class	Precision (%)	Recall (%)	F1-Score (%)
MedGemma	Pneumonia	90.6	83.6	87
	Normal	63.4	76.6	69.3
GPT-4	Pneumonia	84.8	71.7	77.7
	Normal	46	65.2	53.9

Across renal images of chronic kidney disease, MedGemma showed strong generalization. As illustrated in Fig. 8, most of the classes are correctly identified. In GPT-4, a higher number of misclassification is evident.

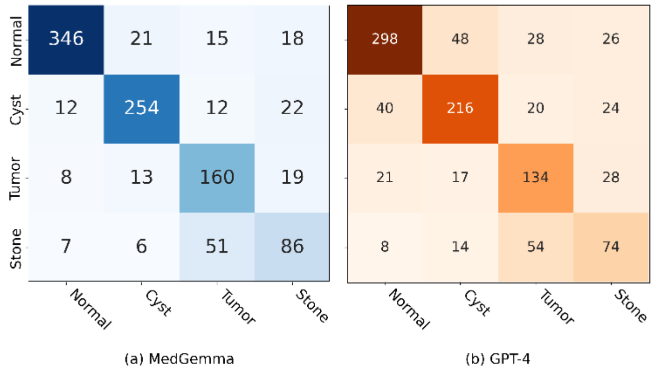


Fig. 8. Confusion Matrix for Chronic Kidney Diseases Classification by (a) MedGemma and (b) GPT-4

In Table X, MedGemma showed a maximum recall value of 86.5% while GPT-4's maximum recall value is only 74.5%. The consistent score in MedGemma shows its reliable classification ability of multi-class diagnosis in renal images.

TABLE X. CLASSIFICATION REPORT FOR CHRONIC KIDNEY DISEASE

Model	Class	Precision (%)	Recall (%)	F1-Score (%)
MedGemma	Normal	92.8	86.5	89.5
	Cyst	86.4	84.7	85.5
	Tumor	67.2	80	73.1
	Stone	59.3	57.3	58.3
GPT-4	Normal	81.2	74.5	77.7
	Cyst	73.2	72	72.6
	Tumor	56.8	67	61.5
	Stone	48.7	49.3	49

In all cases, MedGemma outperformed GPT-4 by 10-14% accuracy. This substantial gap emphasizes that general AI models like GPT-4 lack the necessary specialized training to handle the subtle, high-resolution features required for accurate medical diagnosis, unlike fine-tuned models such as MedGemma.

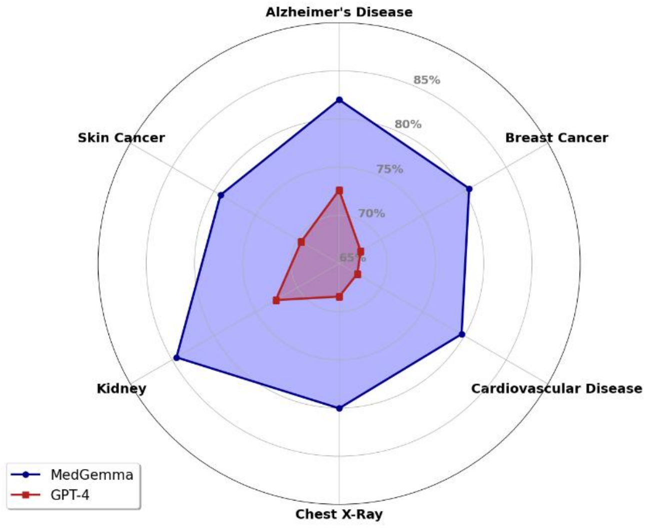


Fig. 9. Radar Chart of MedGemma vs GPT-4 in Accuracy across Diseases

The radar chart in Fig. 9 makes it easy to compare the performance of MedGemma-4b-it with GPT-4, across six medical classification datasets. From a graphical comparison perspective, there is strong evidence that MedGemma demonstrated superior performance than GPT-4 across datasets ranging from diagnosing Breast Cancer to diagnosing chest X-rays. The radar chart illustrates that the MedGemma polygon completely encompasses GPT-4's polygon, indicating a more robust performance with accuracy levels ranging from 79% to 85%. However, the accuracy of GPT-4 seems seriously substandard, with testing accuracy ranging from 67% to 73%. While GPT-4 remains susceptible to medical hallucinations, MedGemma's fine-tuning significantly reduces these errors, enhancing its reliability for clinical use.

V. Findings

The key findings of this work are:

1. Since GPT-4 is a proprietary AI model without a fine-tuning opportunity, its accuracy could not match the higher results achieved by the fine-tuned MedGemma.

2. MedGemma's high classification performance on various datasets highlights the value of domain-specific fine-tuning for the field of medical diagnostics.

3. The significant performance gap observed on the Breast Cancer dataset suggests that GPT-4, as a general model struggles to analyze the subtle, high-resolution features required for accurate cancer classification.

4. It is observed that GPT-4 tends to memorize the sequence in which the image information and prompt were given. This reliance on sequence memorization rather than deep visual feature recognition limits its ability to accurately classify complex images.

5. Despite its huge 4-billion-parameter size, MedGemma currently shows less accuracy than other CNN models because it was fine-tuned using Low-Rank Adaptation (LoRA) on a limited scale due to time and resource constraints.

VI. Conclusion

This study demonstrates that the MedGemma-4b-it model consistently outperforms GPT-4 across multiple disease classification tasks, achieving a mean accuracy of 80.37% compared to 69.58% for GPT-4. The performance evaluation between the fine-tuned MedGemma and zero-shot GPT-4 highlights a critical trade-off between domain-specific optimization and general-purpose accessibility. While MedGemma's superior performance is expected considering its exposure to 70% of the disease-specific training data, the comparison remains relevant as GPT-4 serves as a benchmark for high-capacity models trained on massive, heterogeneous corpora. However, despite its broad web-scale knowledge base and ease of use, our findings suggest that GPT-4's lack of specialized refinement makes it less suitable for high-stakes clinical decision-making compared to models explicitly fine-tuned for medical contexts. Consequently, domain-specific adaptation is essential to enhance the reliability and clinical efficacy of AI-driven diagnostics. Despite few limitations regarding dataset size and computational resources, the flexibility and efficiency of MedGemma position it as a promising tool for real-world clinical deployment. Future work should prioritize MedGemma's large-scale clinical validation, ethical and regulatory considerations, and seamless integration into healthcare workflows. Furthermore, advancing hallucination-aware training and evaluation strategies remains essential to ensuring trustworthy, safe, and interpretable AI systems in high-stakes medical settings.

References

[1] International Diabetes Federation, "IDF Diabetes Atlas." 2019. [Online]. Available: https://www.diabetesatlas.org

[2] World Health Organization, "Breast Cancer Fact Sheet." 2020. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/breast-cancer

[3] World Health Organization, "Pneumonia Fact Sheet." 2021. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/pneumonia

[4] Alzheimer's Association, "2024 Alzheimer's disease facts and figures," *Alzheimer's & Dementia*, vol. 20, no. 5, pp. 3708–3821, Apr. 2024. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC11095490/

[5] World Health Organization, "Cardiovascular Diseases Fact Sheet." 2019. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[6] K. Xie *et al.*, "Global, regional, and national burden inequality of chronic kidney disease, 1990–2021: a systematic analysis for the global burden of disease study 2021," *Front. Med.*, vol. 11, Feb. 2024. [Online]. Available: https://www.frontiersin.org/journals/medicine/articles/10.3389/fmed.2024.1501175/full

[7] A. Reshad, V. Nino, and M. Valero, "Deep Learning-Based Detection of Arrhythmia Using ECG Signals – A Comprehensive Review," *Vasc. Health Risk Manag.*, vol. Volume 21, pp. 685–703, Aug. 2025, doi: 10.2147/VHRM.S508620.

[8] K. Singhal *et al.*, "Large Language Models Encode Clinical Knowledge," Dec. 26, 2022, *arXiv*: arXiv:2212.13138. doi: 10.48550/arXiv.2212.13138.

[9] K. Singhal *et al.*, "Towards Expert-Level Medical Question Answering with Large Language Models," May 16, 2023, *arXiv*: arXiv:2305.09617. doi: 10.48550/arXiv.2305.09617.

[10] H. Takita *et al.*, "A systematic review and meta-analysis of diagnostic performance comparison between generative AI and physicians," *Npj Digit. Med.*, vol. 8, no. 1, p. 175, Mar. 2025, doi: 10.1038/s41746-025-01543-z.

[11] S. Abbas *et al.*, "THE ROLE OF ARTIFICIAL INTELLIGENCE IN PERSONALIZED MEDICINE AND PREDICTIVE DIAGNOSTICS – A NARRATIVE REVIEW," *Insights-J. Health Rehabil.*, vol. 3, no. 1 (Health&Allied), pp. 624–631, Feb. 2025, doi: 10.71000/k6cga886.

[12] P. Hager *et al.*, "Evaluation and mitigation of the limitations of large language models in clinical decision-making," *Nat. Med.*, vol. 30, no. 9, pp. 2613–2622, Sept. 2024, doi: 10.1038/s41591-024-03097-1.

[13] T. Tu *et al.*, "Towards conversational diagnostic artificial intelligence," *Nature*, vol. 642, no. 8067, pp. 442–450, June 2025, doi: 10.1038/s41586-025-08866-7.

[14] T. B. Brown *et al.*, "Language models are few-shot learners," in *Proc. 34th Int. Conf. Neural Inf. Process. Syst. (NIPS '20)*, Vancouver, BC, Canada, 2020, Art. no. 159, doi: 10.5555/3495724.3495883.

[15] A. Radford *et al.*, "Language Models are Unsupervised Multitask Learners," *arXiv e-prints*, 2019, Art. no. 1907.10522, doi: 10.48550/arXiv.1907.10522. [Online]. Available: https://arxiv.org/abs/1907.10522

[16] D. M. Korngiebel and S. D. Mooney, "Considering the possibilities and pitfalls of Generative Pre-trained Transformer 3 (GPT-3) in healthcare delivery," *npj Digit. Med.*, vol. 4, no. 1, Art. no. 93, 2021, doi: 10.1038/s41746-021-00464-x.

[17] E. Ullah et al., “Challenges and barriers of using large language models (LLM) such as ChatGPT for diagnostic medicine with a focus on digital pathology – a recent scoping review,” *Diagn Pathol*, vol. 19, no. 1, Art. no. 43, 2024, doi: 10.1186/s13000-024-01464-7.

[18] A. Lecler, L. Duron, and P. Soyer, “Revolutionizing radiology with GPT-based models: Current applications, future possibilities and limitations of ChatGPT,” *Diagn. Interv. Imaging*, vol. 104, no. 6, pp. 269–274, 2023, doi: 10.1016/j.diii.2023.02.003.

[19] K. Singhal *et al.*, “Large language models encode clinical knowledge,” *Nature*, vol. 620, no. 7972, pp. 172–180, 2023, doi: 10.1038/s41586-023-06291-2.

[20] J. Lee et al., “BioBERT: a pre-trained biomedical language representation model for biomedical text mining,” *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2019, doi: 10.1093/bioinformatics/btz682.

[21] E. Alsentzer et al., “Publicly Available Clinical BERT Embeddings,” arXiv e-prints, 2019, Art. no. 1904.03323, doi: 10.48550/arXiv.1904.03323. [Online]. Available: https://arxiv.org/abs/1904.03323

[22] A. Sellergren, S. Kazemzadeh, T. Jaroensri, A. Kiraly, M. Traverse, T. Kohlberger, et al., “MedGemma Technical Report,” arXiv e-prints, arXiv:2507.05201 [cs.AI], 2025. [Online]. Available: https://arxiv.org/abs/2507.05201.

[23] OpenAI et al., “GPT-4 Technical Report,” arXiv preprint arXiv:2303.08774, 2024. [Online]. Available: https://arxiv.org/abs/2303.08774.

[24] N. Prentzas, A. Kakas, and C. S. Pattichis, “Explainable AI applications in the Medical Domain: a systematic review,” arXiv e-prints, 2023, Art. no. 2308.05411, doi: 10.48550/arXiv.2308.05411. [Online]. Available: https://arxiv.org/abs/2308.05411

[25] A. Dosovitskiy et al., “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” International Conference on Learning Representations, May 2021, [Online]. Available: https://openreview.net/pdf?id=YicbFdNTTy.

[26] D. Zhou et al., “DeepVIT: Towards Deeper Vision Transformer,” arXiv.org, Mar. 22, 2021. https://arxiv.org/abs/2103.11886.

[27] Y. Xie, “New chapter in pediatric medicine: technological evolution, application, and evaluation system of large language models,” European Journal of Pediatrics, vol. 184, no. 12, p. 809, Dec. 2025, doi: 10.1007/s00431-025-06602-x.

[28] P. Tschandl, C. Rosendahl, and H. Kittler, “The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions,” *Sci Data*, vol. 5, no. 1, Art. no. 180161, 2018, doi: 10.1038/sdata.2018.161.

[29] D. S. Marcus, T. H. Wang, J. Parker, J. G. Csernansky, J. C. Morris, and R. L. Buckner, “Open access series of imaging studies (OASIS): Cross-sectional MRI data in young, middle aged, nondemented, and demented older adults,” Journal of Cognitive Neuroscience, vol. 19, no. 9, pp. 1498–1507, 2007. doi: 10.1162/jocn.2007.19.9.1498. [Online]. Available: https://doi.org/10.1162/jocn.2007.19.9.1498

[30] R. S. Lee, F. Gimenez, A. Hoogi, K. K. Miyake, M. Gorovoy, and D. L. Rubin, “A curated mammography data set for use in computer-aided detection and diagnosis research,” Scientific Data, vol. 4, p. 170177, 2017. doi: 10.1038/sdata.2017.177. [Online]. Available: https://doi.org/10.1038/sdata.2017.177 .

[31] A. H. Khan and M. Hussain, “ECG Images Dataset of Cardiac Patients,” Mendeley Data, V2, 2021. doi: 10.17632/gwbz3fsgp8.2. [Online]. Available: https://doi.org/10.17632/gwbz3fsgp8.2

[32] D. Kermany, K. Zhang, and M. Goldbaum, “Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images for Classification,” Mendeley Data, V2, 2018, doi: 10.17632/rscbjbr9sj.2.

[33] N. M. Islam, “CT KIDNEY DATASET: Normal-Cyst-Tumor and Stone,” Kaggle, 2021. [Online]. Available: https://www.kaggle.com/datasets/nazmul0087/ct-kidney-dataset-normal-cyst-tumor-and-stone. [Accessed: Oct. 22, 2025].

[34] E. J. Hu et al., “LoRA: Low-Rank Adaptation of Large Language Models,” arXiv e-prints, 2021, Art. no. 2106.09685, doi: 10.48550/arXiv.2106.09685. [Online]. Available: https://arxiv.org/abs/2106.09685

[35] P. Zhou, X. Xie, Z. Lin, and S. Yan, “Towards Understanding Convergence and Generalization of AdamW,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 9, pp. 6486–6493, Sept. 2024, doi: 10.1109/TPAMI.2024.3382294.