CliniQure: AI-Driven Diagnostic and Treatment Recommendations for Radiology

Moksh Shukla
mshukla4@jhu.edu

Avantika Singh
asing153@jhu.edu

Yash Singla
ysingla1@jhu.edu

Abstract

The integration of advanced machine learning techniques into medical diagnostics holds transformative potential, especially in resource-limited settings where access to expert interpretation of imaging is often constrained. This study introduces **CliniQure**, a cutting-edge, multimodal diagnostic system designed to automate disease detection, enhance interpretability, and streamline clinical workflows. Leveraging a well-annotated dataset, CliniQure adopts a multi-task learning paradigm that integrates object detection and disease classification. Object detection is performed using the Faster R-CNN architecture with a ResNet-50 backbone pre-trained on ImageNet, enabling precise localization of abnormalities. Simultaneously, disease classification employs the SE-ResNet50 model for robust multi-label classification of medical images, ensuring accurate and granular predictions.

In addition to visual analysis, CliniQure generates initial diagnostic reports using the pre-trained R2GenCMN model, which employs a cross-modal memory network (CMN) architecture and is trained on the MIMIC-CXR dataset. These crude diagnostic outputs, combined with the classification and detection results, are processed by a fine-tuned large language model (LLM) to produce final, detailed diagnostic reports. The LLM is fine-tuned with Low-Rank Adaptation (LoRA) on domain-specific clinical text data, improving coherence, relevance, and contextual understanding. This seamless integration of visual and textual data enables CliniQure to deliver comprehensive and clinically meaningful diagnostic insights.

1 Introduction

The accurate and timely interpretation of medical imaging is a cornerstone of modern healthcare, enabling the early detection and effective management of a wide spectrum of diseases. Among various imaging modalities, chest X-rays (CXRs) are particularly significant due to their widespread use as a primary diagnostic tool for conditions such as pneumonia, tuberculosis, cardiomegaly, and lung cancer. Despite their critical importance, the reliance on expert radiologists for accurate interpretation poses significant challenges in resource-constrained settings, where a shortage of trained professionals often leads to diagnostic delays, inconsistencies, and adverse patient outcomes. These challenges underscore the need for automated systems that can augment clinical workflows and improve access to reliable diagnostic services.

To address this need, this work introduces **CliniQure**, a cutting-edge system that integrates state-of-the-art machine learning techniques with radiological data to enhance diagnostic accuracy, interpretability, and workflow efficiency. At the core of CliniQure lies a multi-task learning framework designed to perform disease classification and abnormality detection simultaneously. Leveraging a robust chest X-ray dataset annotated with abnormalities, the system employs architectures such as Faster R-CNN with a ResNet-50 backbone for abnormality localization and SE-ResNet50 for multi-label disease classification. By integrating these tasks into a unified framework, CliniQure achieves both precise abnormality localization and robust predictions for classified diseases, ensuring its outputs are clinically relevant and interpretable.

Beyond imaging analysis, CliniQure incorporates advanced natural language processing (NLP) capabilities to provide a holistic diagnostic solution. Using a large language model (LLM) fine-tuned with Low-Rank Adaptation (LoRA), the system generates detailed and contextually accurate diagnostic reports based on classified diseases, detected abnormalities, and crude diagnostic reports. These crude reports, generated by a pre-trained R2GenCMN model, provide a preliminary synthesis of visual and textual data. The LLM is trained on clinical text data encompassing a wide range of

diseases, symptoms, and treatments, enabling it to refine these inputs into coherent, comprehensive, and actionable diagnostic narratives. A key innovation in CliniQure is the incorporation of structured prompting techniques to guide the LLM. The prompts are designed to integrate classified diseases, detected abnormalities, and crude reports as input, instructing the model to expand on these findings with detailed descriptions of underlying causes, associated symptoms, diagnostic methods, and treatment options. This ensures that the final outputs are not only accurate but also clinically meaningful.

This multimodal approach bridges the gap between visual and textual data interpretation, creating a comprehensive diagnostic pipeline that is both scalable and adaptable. By combining deep learning for image analysis with the contextual understanding of LLMs, and leveraging crude reports and structured prompting to maximize the relevance of the outputs, CliniQure enhances diagnostic workflows and addresses critical gaps in healthcare delivery, particularly in under-resourced regions. The integration of advanced AI technologies with clinical workflows establishes a new standard for medical imaging diagnostics, making them more accessible, accurate, and efficient.

2 Literature Review

Addressing the shortage of trained radiologists in resource-limited settings has been a major driver of automated diagnostic systems. Studies such as (Irvin et al., 2019) and (Rajpurkar et al., 2018) highlight the role of AI in augmenting diagnostic workflows to alleviate the burden on healthcare professionals. CliniQure builds on these insights by providing a scalable, efficient, and clinically relevant solution capable of delivering high-quality diagnostics in regions with limited healthcare resources.

Multi-label disease classification has been a cornerstone of medical imaging research, enabling models to identify co-occurring conditions in diagnostic images.(Wang et al., 2017) introduced the CheXNet model, a convolutional neural network trained on chest X-rays, which demonstrated state-of-the-art performance in pneumonia detection. Further, (Guan et al., 2020) utilized SE-ResNet50, an enhanced version of ResNet with Squeeze-and-Excitation blocks, to dynamically recalibrate channel-wise feature responses. These studies underscore the effectiveness of feature extraction and adaptive learning in multi-label classification tasks, forming the foundation for CliniQure's disease classification module.

Object detection models have revolutionized abnormality localization in medical imaging by providing precise bounding boxes for detected features. (Ren et al., 2015) introduced Faster R-CNN, a two-stage detection framework that integrates region proposal networks (RPNs) with convolutional feature extraction. This architecture has been extensively applied to medical datasets, as demonstrated by (Rajpurkar et al., 2018) who used it for identifying abnormalities in chest X-rays. The inclusion of a ResNet-50 backbone enhances detection accuracy by leveraging deeper feature representations. CliniQure incorporates these advancements to localize abnormalities effectively, forming a critical component of its diagnostic pipeline.

Generating preliminary diagnostic reports from medical images involves combining visual features with natural language representations.(Chen et al., 2020) introduced the R2GenCMN model, which uses a Cross-Modal Memory Network (CMN) to integrate image features with textual data, achieving superior report generation results on the MIMIC-CXR dataset. This approach enables the creation of crude diagnostic reports that provide initial interpretations of detected abnormalities and classified diseases. CliniQure leverages R2GenCMN to produce preliminary textual outputs, which serve as inputs for further refinement by the large language model.

Large language models (LLMs) such as GPT-based architectures have significantly advanced natural language understanding and generation tasks. (Hu et al., 2021) introduced Low-Rank Adaptation (LoRA) as an efficient fine-tuning method that modifies low-rank matrices within attention layers, reducing computational overhead without sacrificing performance. Fine-tuning LLMs on domain-specific data, as demonstrated by (Radford et al., 2021), enables these models to generate highly contextual and relevant outputs. In CliniQure, the LLM refines crude reports by incorporating structured prompts that include classified diseases, detected abnormalities, and preliminary reports, creating comprehensive diagnostic summaries tailored to specific clinical scenarios.

Integrating multimodal data—visual and textual—has been a key focus in recent AI research.

Works by (Zhang et al., 2022b) demonstrated the potential of combining image analysis with textual understanding to create systems that bridge the gap between visual features and natural language. These studies emphasize the importance of harmonizing multimodal inputs for generating meaningful outputs, a principle that underpins CliniQure's design. By integrating outputs from detection, classification, and report generation models, CliniQure exemplifies the advantages of multimodal systems in medical diagnostics.

Conclusively, (Wang et al., 2023) proposed an integration of large language models (LLMs) with computer-aided diagnosis (CAD) systems, exemplifying the promise of multimodal systems that synthesize visual and textual information. Their system, ChatCAD, transforms outputs from image-based networks into text inputs for LLMs, enabling richer, more intuitive diagnostics.

The success of such systems hinges on the consistency and precision of annotations across modalities. Disparities between visual annotations and textual descriptions can propagate errors, emphasizing the need for harmonized dataset preparation to support these sophisticated interactions. The study underscores that while AI systems increasingly aim to minimize annotation dependency, the alignment and completeness of available annotations remain vital for meaningful multimodal integration.

3 Dataset

The dataset identified for this project is the (VinBigData, 2020) VinDr-CXR dataset, developed as part of the VinBigData initiative and hosted on PhysioNet. This dataset is used for segmentation and classification tasks in this project, leveraging its rich annotations for abnormality localization and classification. Subsequent sections provide a detailed discussion of this dataset.

3.1 Image and Annotations

The **VinDr-CXR dataset** is a large-scale, high-quality resource designed to advance research in computer-aided diagnosis using chest radiographs. It contains 18,000 chest X-ray images corresponding to approximately 15,000 unique patients, all provided in the standard DICOM (Digital Imaging and Communications in Medicine) format. Each image is meticulously annotated by a panel of radiologists, ensuring a high degree of accuracy. The dataset includes annotations for 22 distinct abnormalities such as atelectasis, cardiomegaly, consolidation, effusion, mass, nodule, pneumonia, pneumothorax, and tuberculosis.

These annotations are stored in a **CSV file**, which provides a comprehensive and structured representation of the data. The CSV file contains following columns of interest:

- **Image ID**: A unique identifier for each image.
- **Patient ID**: A unique identifier for the patient associated with the X-ray.
- **Bounding Box Coordinates**: Includes `x_min`, `y_min`, `x_max`, and `y_max` to define the precise location of abnormalities.
- **Finding Labels**: The name of the abnormality (e.g., "Mass," "Nodule").

Figure 1 shows a chest X-ray image with bounding box annotations, demonstrating how abnormalities are localized in the scan. The corresponding details for these annotations, including bounding box coordinates, labels, and confidence scores, are illustrated in the snapshot of the CSV file shown in Table 2.

These detailed annotations make the dataset particularly suitable for various tasks, including abnormality detection, lesion localization, and region-of-interest segmentation. The structured CSV format ensures that the dataset is easy to integrate into machine learning workflows, facilitating training and evaluation.

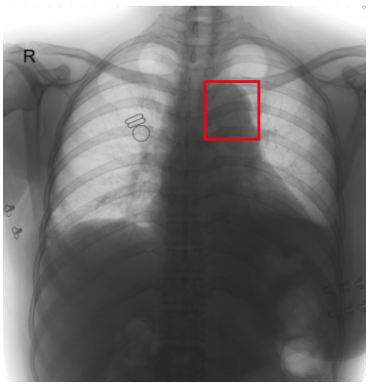

Figure 1: Chest X-ray with aortic enlargement

3.2 Clinical Metadata

Additionally, the dataset includes essential patient information, such as age, gender, and clinical indications. This metadata provides important context

Image ID	Class Name	Class ID	Rad ID	X_min	Y_min	X_max	Y_max
50a418190bc3fb1ef1633bf96789	No finding	14	R11	NA	NA	NA	NA
21a10246a5ec7af151081d0cd6d6	No finding	14	R7	NA	NA	NA	NA
9a5094b2563a1ef3ff50dc5c7ff7	Cardiomegaly	3	R10	691	1375	1653	1831
051132a778e61a86eb147c7c6f56	Aortic enlargement	0	R10	1264	743	1611	1019
063319de25ce7edb9b1c6b888129	No finding	14	R10	NA	NA	NA	NA
1c32170b4af4ce1a3030eb816775	Pleural thickening	11	R9	627	357	947	433

Table 1: VinDr-CXR Annotation Metadata

for understanding the chest X-rays, including details on the reasons why the imaging study was conducted (e.g., for suspected pneumonia, heart failure, or other clinical concerns). Since the scans are provided in DICOM format, they also include additional metadata embedded in the files, such as imaging acquisition parameters (e.g., modality, resolution, and exposure settings), timestamps, and hospital information. This embedded metadata adds another layer of contextual understanding, which can be leveraged for advanced research and model development.

All patient data is de-identified, ensuring compliance with ethical standards and data protection regulations. This guarantees that the dataset adheres to privacy guidelines while still offering a comprehensive set of information for diagnostic and machine learning tasks.

4 Methodology

This section details the comprehensive workflow for data preparation, model training, and diagnostic report generation. The primary objective of these steps is to produce **well-structured input data** tailored for a fine-tuned large language model, which is **prompted effectively** for generating detailed and accurate diagnostic reports.

The process begins with *data preprocessing*, where annotations are meticulously refined by merging overlapping bounding boxes and eliminating redundancies to ensure data consistency and quality. Subsequently, *disease detection* is carried out using Faster R-CNN, which efficiently localizes abnormalities within the input images. For *disease classification*, SE-ResNet50 is employed, leveraging its multi-label classification capabilities to accurately identify multiple disease labels associated with the detected abnormalities.

The final stage, *report generation*, employs R2GenCMN to synthesize coherent and clinically relevant diagnostic reports. This stage seamlessly integrates visual features from the image analysis with textual data, ensuring the generated reports are both contextually rich and aligned with clinical standards.

Figure 2 illustrates a pipeline that integrates chest X-rays, clinical metadata, and large language models (LLMs) for automated disease detection, classification, and report generation.

4.1 Data Preprocessing

The data preprocessing pipeline standardizes annotations from medical images, addressing inconsistencies and redundancy to create a structured, machine learning-compatible dataset. Bounding box annotations, provided by multiple radiologists, are processed to merge overlapping entries, map disease labels to numeric identifiers, and remove redundant data.

Bounding Box Merging

Overlapping bounding boxes in the same image with Intersection over Union (IoU) values above a defined threshold and similar aspect ratios are merged into a single box. The merged bounding box is computed as:

$$\text{Merged Box} = [\min(x_{\text{min}}), \min(y_{\text{min}}), \max(x_{\text{max}}), \max(y_{\text{max}})]$$

Class labels from merged bounding boxes are concatenated into a unified representation, allowing for accurate multi-disease annotations.

Redundancy Filtering

Duplicate annotations with identical `image_id` and `class_names`, or entries labeled as `"No finding"`, are removed to ensure data quality.

Table 2 shows the dataset snippet after preprocessing.

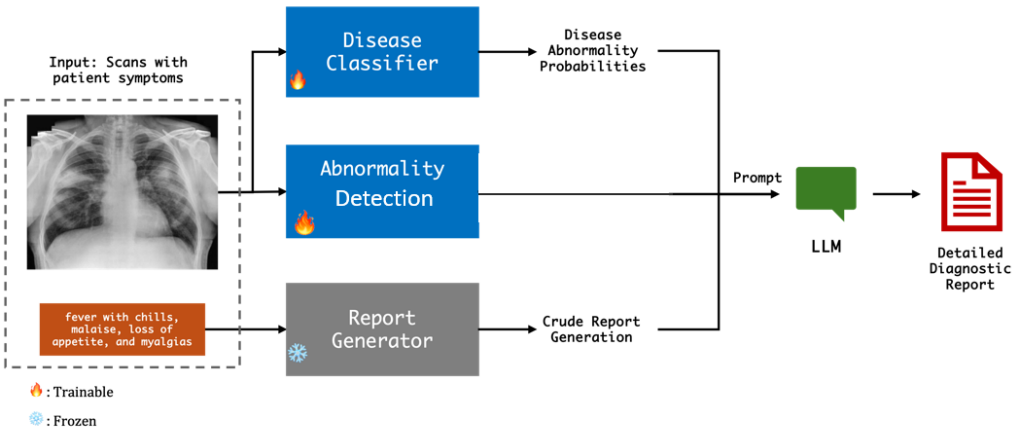

Figure 2: Proposed Training Framework

Image ID	x_min	y_min	x_max	y_max	Class Names	Class Labels
000434271f63a053c4128a0ba6352c	0.0	0.0	2336.0	2836.0	No finding	[14]
00053190460d56c53cc3e573213874	0.0	0.0	1994.0	2430.0	No finding	[14]
0005e8e3701dfb1dd93d53e2ff537b	900.0	583.0	1205.0	890.0	Infiltration, Lung Opacity	[6, 7]
0005e8e3701dfb1dd93d53e2ff537b	932.0	567.0	1197.0	896.0	Consolidation, Nodule/Mass	[4, 8]
0006e0a85696f6bb578e84fafa9a56	0.0	0.0	3000.0	3000.0	No finding	[14]

Table 2: VinDr-CXR Merged Dataset

4.2 Model Training

The training phase optimizes models for disease detection and classification in medical images. Specialized pipelines leverage pre-trained backbones, data augmentation, and robust optimization strategies to handle multi-scale detection and multi-label classification. The following sections detail the detection and classification workflows. Table 3 shows the common training params used for detection and classification.

Table 3: Transformations and Training Parameters

Parameter	Value
Image Transformations	
Resize Dimensions	[224, 224]
Scale Intensity	True
Random Rotate Probability	0.5
Random Flip Probability	0.5
Training Parameters	
Optimizer	Adam
Learning Rate	1e-4
Weight Decay	1e-5
Number of Epochs	100
Learning Rate Scheduler	Cosine

4.2.1 Disease Detection

The object detection pipeline employs the **Faster R-CNN architecture with a ResNet-50 backbone** to detect and classify objects in medical images. The dataset consists of DICOM-format images paired with bounding box annotations, split into 80% training and 20% validation data. Images are resized to 224×224, and augmentation techniques, including random rotations and random flips (each with a 50% probability), are applied to improve model generalization.

The ResNet-50 backbone, pre-trained on ImageNet, extracts feature maps with 2048 channels. The Faster R-CNN architecture integrates custom configurations, including:

- **Anchor Generator**: Supports multi-scale object detection with scales 32, 64, 128, 256, 512 and aspect ratios 0.5, 1.0, 2.0.
- **ROI Pooler**: Aligns and extracts fixed-size feature maps (7×7) from proposed regions.
- **Detection Head**: Refines bounding box coordinates and assigns class labels across 15 categories (14 object classes + background).

The training process optimizes the model using the Adam optimizer with a learning rate of 1e-4, weight decay of 1e-5, and a cosine annealing scheduler over 100 epochs.

The workflow consists of:

- **Feature Extraction**: Extracts feature maps from input images.
- **Region Proposal**: The Region Proposal Network (RPN) generates candidate bounding boxes.
- **ROI Pooling**: Extracts fixed-size feature maps from proposed regions.
- **Detection Head**: Refines bounding boxes and assigns class labels.

The Faster R-CNN model demonstrates strong performance on the medical dataset, with preprocessing and augmentation enhancing its ability to detect objects of varying sizes and shapes. This approach offers a scalable and efficient solution for object detection tasks in medical imaging. Table 4 contains the detailed detection configuration details.

Table 4: Detection Configuration Details

Parameter	Value
Model Architecture	
Backbone	ResNet-50
Feature Map Output	2048 channels, 7×7 spatial dimensions
Anchor Scales	{32, 64, 128, 256, 512}
Anchor Aspect Ratios	{0.5, 1.0, 2.0}
ROI Pooler	ROI Align, fixed output size 7×7

4.2.2 Disease Classification

This section outlines the classification pipeline developed using the **SE-ResNet50** architecture for multi-label classification, designed to handle datasets where each image can belong to multiple classes simultaneously.

The SE-ResNet50 architecture is used for multi-label classification, handling datasets where each image can belong to multiple classes. SE-ResNet enhances the traditional ResNet architecture by incorporating Squeeze-and-Excitation (SE) blocks, which apply channel-wise attention to prioritize informative feature maps. SE blocks operate through two key steps: squeeze, which aggregates global spatial information using average pooling, and excitation, which learns channel weights via fully connected layers.

Its fully connected layer is replaced to match the number of classes, with a sigmoid activation function converting logits into class probabilities. `Binary Cross-Entropy` with Logits serves as the loss function, independently applied to each class.

Training minimizes binary cross-entropy loss, with gradients optimized using Adam and a dynamic learning rate scheduler. Validation evaluates binary cross-entropy loss and Hamming loss, the latter measuring prediction accuracy across classes. We used a train/val split of 80:20, with a batch size of 4 on 15 classes.

4.3 Crude Report Generation

The next step in the input generation pipeline is **crude report generation**. The DICOM images are converted into image tensors. These image tensors are passed through the **R2GenCMN**. Inference from the chest X-ray images is extracted using the **pre-trained weights** of this model which employs a Cross-modal Memory Network (CMN) architecture to generate diagnostic reports of chest X-rays. It integrates two key components: a co-attention mechanism and cross-modal memory modules. .

The features extracted at the time of inference are stored in a **visual memory module**, which retains spatial and semantic information about different regions of the image. Simultaneously, the **textual memory module** maintains contextual information about previously generated words and sentences, ensuring linguistic fluency and coherence. The co-attention mechanism acts as a bridge, allowing the model to dynamically attend to specific regions in the visual memory while considering the context stored in the textual memory.

The output generation process is iterative, where each word or phrase in the report is produced based on the current context and the relevant visual features. For example, when describing an abnormality, the model dynamically attends to the region containing the abnormality, aligns it with corresponding clinical terminology, and generates text that accurately reflects the finding. It incorporates domain-specific language patterns learned during training on the **MIMIC-CXR dataset**.

5 LoRA Fine-Tuning for LLM

Low-Rank Adaptation (LoRA) fine-tuning was employed to adapt the `facebook/opt-1.3b` model for final report generation. The fine-tuned model is prompted with the inputs taken from the classification, detection and pre-trained model to generate detailed and contextually relevant diagnostic reports, aligning clinical descriptions with accurate disease predictions.

5.1 Finetuning Dataset

To fine-tune the large language model (LLM) using Low-Rank Adaptation (LoRA),the publicly available (FreedomIntelligence, 2024) `FreedomIntelligence/Disease_Database`, hosted on Hugging Face under the repository. was utilized. The dataset contains a total of **9,233 entries**, each providing a mapping between a disease and its associated symptoms,

The following columns of the dataset were used for the purpose of finetuning the model:

Disease	Common Symptom
Influenza	Fever, cough, sore throat
Diabetes Mellitus	Increased thirst, frequent urination
Hypertension	Headaches, dizziness

Table 5: Example entries from the Disease Database dataset.

5.2 OPT 1.3B Large Language Model

The *OPT 1.3B model* (Open Pretrained Transformer), (Zhang et al., 2022a) developed by Meta AI, is a transformer-based language model with 1.3 billion parameters, designed as an open-source alternative to GPT. It employs a standard transformer architecture with multi-head self-attention and feed-forward layers, pretrained on diverse datasets to understand and generate human-like text. This fine-tuning process for report generation enhanced the model's ability to integrate multimodal inputs and produce detailed, clinically accurate, and contextually relevant reports.

5.3 Fine-Tuning Process

For the purpose of fine-tuning, the **common symptom** column was used as the input data, while the **disease** column served as the target labels. This setup framed a supervised learning task where the model learns to generate disease predictions or respond contextually to clinical symptom descriptions. The mapping between symptoms and diseases provides a realistic training objective, enabling the model to perform tasks such as symptom-based diagnosis and clinical query handling.

The dataset's structured format, along with its wide coverage of diseases and their respective symptoms, ensured a rich semantic context for training. By fine-tuning on this dataset, the LLM was empowered to generate clinically coherent responses, support medical decision-making, and enhance healthcare-related applications.

The dataset included symptoms and corresponding diseases, which were preprocessed by tokenizing the symptom text and encoding the disease labels numerically. The tokenizer was extended with a padding token to handle variable input lengths, and the model's token embeddings were resized to match the updated vocabulary. LoRA (Low-Rank Adaptation) was utilized for efficient fine-tuning by modifying only low-rank matrices within the attention layers. The parameters for LoRA, including rank, alpha, and dropout, are detailed in Table 6. This approach significantly reduced the computational cost of training while maintaining the model's performance, making it well-suited for adapting the `OPT-1.3b` model to domain-specific tasks.

During training, the model was optimized using a learning rate of 2×10^{-5}, a batch size of 8 per device, and mixed precision training to enhance GPU efficiency (refer to Table 6 for details). The Hugging Face `Trainer` API facilitated streamlined training, evaluation, and saving of the fine-tuned model. The effectiveness of the fine-tuning process, as illustrated in Figure 3, highlights how LoRA reduces computational costs by modifying only low-rank matrices within the attention layers. The trained model and tokenizer, along with the encoded label mappings, were saved for future inference tasks that need to be done from the LLM.

Table 6: Parameters Used in Model Fine-Tuning

Parameter	Value
LoRA Rank	8
LoRA Alpha (α)	32
LoRA Dropout	10%
Learning Rate	2×10^{-5}
Batch Size (Per Device)	8
Mixed Precision Training	Enabled

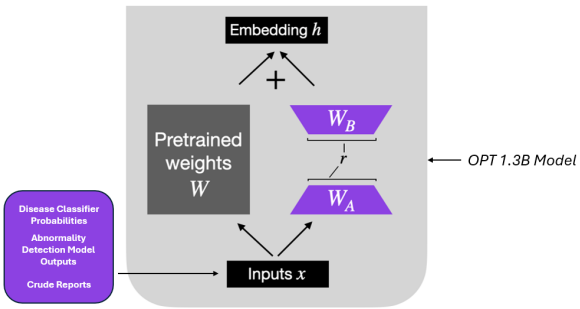

Figure 3: Illustration of LoRA Fine-Tuning: This process involves modifying only low-rank matrices within the attention layers, using the inputs obtained, while maintaining model performance.

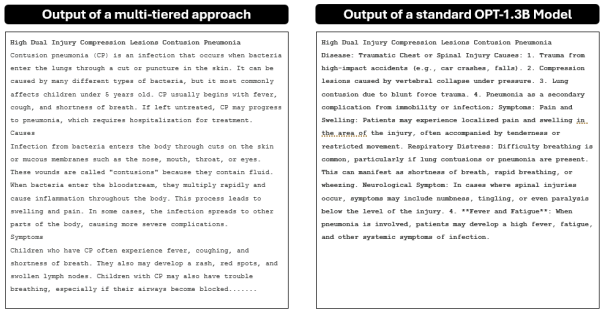

Figure 4: Comparison of Textual Outputs: The fine-tuned model produces detailed and contextually relevant medical summaries, while the standard OPT-1.3B model generates more generic and less precise outputs.

6 Results and Evaluation

6.1 Prompting for Output Generation

Textual output is obtained by designing a **structured prompt** to feed into the fine-tuned language model. This prompt is specifically designed to generate detailed medical summaries. The prompt is derived from prior classification and detection model predictions and crude input, with the explicit instruction to expand the input into a comprehensive summary.This structured prompting strategy enhances the model's ability to automate the creation of clinically relevant reports, demonstrating the utility of fine-tuned models in specialized tasks.

The textual outputs generated by the fine-tuned model are evaluated by comparing them against those produced by a standard OPT-1.3B model, which has not undergone fine-tuning but utilizes the same prompt. This comparison highlights the enhancements introduced through fine-tuning, particularly in terms of contextual relevance, specificity, and detail in the generated outputs. As illustrated in Figure 4, the fine-tuned model demonstrates a marked improvement in generating comprehensive and clinically accurate summaries, whereas the standard OPT-3B model often produces more generic and less precise results.

6.2 Evaluation Metrics

This analysis yields evaluation metrics for the classification and detection model's performance, as well as semantic similarity scores using BERT. Table 7 for the generated outputs.

The classification metrics, including Precision, Recall, and F1-Score, indicate a well-balanced model performance with an F1-Score of 0.81. For detection performance, Intersection over Union (IoU) and Dice Score are reported, with values of 0.75 and 0.76, respectively, showcasing the model's ability to localize and segment relevant features effectively, with substantial overlap.

Additionally, the high BERT scores for semantic similarity (Precision: 0.9739, Recall: 0.9940, and F1-Score: 0.9838) highlight the model's ability to produce outputs that align semantically with the ground truth labels. These ground truth labels have been obtained from GPT-4.0 which functioned efficiently as a baseline Vision Language Model (VLM), capable of examining chest X-rays and generating diagnostic reports. These metrics demonstrate the model's robust semantic understanding and its effectiveness in generating meaningful and accurate outputs.

Table 7: Evaluation Metrics for Model Performance and Generated Outputs

Metric	**Value**
Classification Performance	
Precision	0.83
Recall	0.81
F1-Score	0.81
Detection Performance	
Intersection over Union (IoU)	0.75
Dice Score	0.76
BERT Score for Semantic Similarity	
Precision	0.9739
Recall	0.9940
F1-Score	0.9838

7 Discussion

The proposed approach demonstrates a robust and systematic methodology for leveraging large lan-

guage models (LLMs) to address complex, domain-specific tasks such as medical diagnosis and report generation. By fine-tuning the OPT-1.3B model using structured, annotated datasets, the system effectively integrates multimodal inputs, such as text and visual features, to produce highly relevant and context-aware outputs. Each stage of the pipeline, from data pre-processing to fine-tuning and evaluation, contributes to the overall performance and practical applicability of the approach.

The pre-processing step ensured the dataset was well-structured, with redundant data eliminated and token embeddings aligned with the updated vocabulary. This foundation enhanced the model's ability to understand and process the domain-specific input. The integration of LoRA (Low-Rank Adaptation) during fine-tuning significantly reduced computational costs by limiting the training to low-rank matrices within the attention layers while maintaining high performance. This innovative adaptation highlights the scalability and efficiency of the approach, making it feasible for use in resource-constrained settings.

While the evaluation metrics presented underscore the effectiveness of the proposed pipeline, with strong classification and detection metrics demonstrating the model's capability to handle complex medical tasks and semantic similarity scores indicating a high degree of alignment between generated outputs and ground truth data, there remains significant scope to enhance the evaluation process. The current evaluation primarily highlights the strengths of the pipeline; however, a more comprehensive benchmarking approach could further validate its performance.

While the presented evaluation metrics validate the effectiveness of the proposed pipeline, expanding the evaluation to include benchmarking against other state-of-the-art methods and employing a more rigorous, multidimensional evaluation framework would provide a deeper and more holistic assessment of its capabilities and limitations. This approach would also pave the way for broader adoption of multi-modal data integration methodologies in complex, real-world applications.

Moreover, the reliance on high-quality datasets such as VinDr-CXR, while yielding strong results, highlights the dependence on extensive labeled data, which can be challenging to obtain in other domains. Furthermore, while LoRA reduces training overhead, performance in highly nuanced or ambiguous cases may still require additional model adaptations or ensemble approaches. Fine-tuning on additional datasets and employing advanced techniques, such as prompt engineering or domain-specific pretraining, could further enhance the system's robustness.

Overall, the proposed approach showcases significant potential for automating domain-specific tasks, particularly in medical diagnostics. Future research can focus on addressing current limitations by incorporating broader datasets, exploring alternative fine-tuning strategies, and conducting rigorous evaluations on diverse and real-world scenarios to generalize the applicability of the model further.

References

Tong Chen, Yikang Zhang, Zhangyang Wu, et al. 2020. Generating radiology reports with cross-modal memory networks. *IEEE Transactions on Medical Imaging*, 39(3):881–892.

FreedomIntelligence. 2024. Disease database dataset. https://huggingface.co/datasets/FreedomIntelligence/Disease_Database. Accessed: December 2024.

Qi Guan, Yu Huang, and Zhen Zhong. 2020. Multi-label chest x-ray image classification via squeeze-and-excitation attention mechanisms and class-specific learning. *IEEE Access*, 8:89865–89877.

Edward J Hu, Yelong Shen, Zeyuan Allen Wallis, et al. 2021. Lora: Low-rank adaptation of large language models. *Advances in Neural Information Processing Systems*, 34.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, et al. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:590–597.

Alec Radford, Jong Wook Kim, Chris Hallacy, et al. 2021. Learning transferable visual models from natural language supervision. *International Conference on Machine Learning*, pages 8748–8763.

Pranav Rajpurkar, Jeremy Irvin, Robyn L Ball, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Abhishek Bagul, Curtis P Langlotz, et al. 2018. Deep learning for chest radiograph diagnosis: A retrospective comparison of the chexnet algorithm to practicing radiologists. *PLoS medicine*, 15(11):e1002686.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.

VinBigData. 2020. Vindr-cxr: A large-scale dataset for chest x-ray abnormalities detection. Available online at https://vindr.ai/datasets/cxr. Accessed: November 15, 2024.

Sheng Wang, Zihao Zhao, Xi Ouyang, Qian Wang, and Dinggang Shen. 2023. Chatcad: Interactive computer-aided diagnosis on medical images using large language models. *arXiv preprint*.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mojtaba Bagheri, and Ronald M Summers. 2017. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106.

Susan Zhang, Stephen Roller, Naman Goyal, et al. 2022a. Opt: Open pretrained transformer language models. https://github.com/facebookresearch/metaseq. Accessed: December 2024.

Y Zhang, J Li, et al. 2022b. Multimodal ai in medical imaging: enhancing diagnostics and patient care. *Journal of Medical Imaging*, 9:031202.

A Appendix

A.1 Contributions of Team Members

Moksh worked on developing and training the disease classification and detection models. Avantika worked on getting the crude outputs from the pretrained model and the LLM finetuning. Yash worked on the finetuning script and getting the final diagnostic outputs. Equal contributions of the team members in developing the literature review, methodology and evaluation metrics.