Vision-Language Models for Lung Disease Detection in 3D CT Scans

KETHANKUMAR REDDY CHINTHAGUNTA, State university of new york at Buffalo, USA
SAHITHYA GANTALA, State university of new york at Buffalo, USA
PAVITHRA NALUBOLU, State university of new york at Buffalo, USA

Selected Project Option: Incorporate ML model in a new Usecase Setting

ACM Reference Format:
Kethankumar Reddy Chinthagunta, Sahithya Gantala, and Pavithra Nalubolu. 2025. Vision-Language Models for Lung Disease Detection in 3D CT Scans. 1, 1 (April 2025), 8 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Abstract

This project aims to develop a robust AI-based diagnostic system for classifying lung diseases in 3D CT scans, addressing critical challenges in early and accurate detection of COVID-19, viral pneumonia, lung opacity, and normal lungs. The approach integrates a baseline Convolutional Neural Network (CNN) model with advanced vision-language architectures—BioViL and MedCLIP to enhance semantic understanding. The CNN-only model serves as a benchmark, while BioViL and MedCLIP provide complementary embeddings to improve performance through multimodal learning. Extensive experimentation using the COV19-CT-DB dataset, coupled with rigorous evaluation through classification metrics and robustness tests under Out-of-Distribution (OOD) perturbations, revealed that the CNN + BioViL hybrid consistently outperforms other models in both accuracy and interpretability. With an accuracy of 83.63% and enhanced AUC scores, particularly in challenging categories like viral pneumonia. This highlights the value of vision-language alignment in medical imaging tasks, showing promise for scalable diagnostic solutions in clinical environments.

2 PROBLEM STATEMENT AND MOTIVATION

According to the World Health Organization (WHO), air pollution and climate change are among the most pressing global health challenges, contributing significantly to the rise of respiratory diseases. Conditions such as COVID-19 pneumonia and lung opacity present serious diagnostic difficulties, where early and accurate detection is critical to improving patient outcomes. Traditional radiological assessments are often time-consuming, subjective, and prone to human error, potentially delaying timely interventions. With the increasing volume of CT scans and a shortage of radiology experts, there is a growing demand for automated, scalable diagnostic systems that can assist clinical decision-making.

In recent years, vision-language pretraining (VLP) models have substantially advanced medical image understanding by aligning visual features with structured textual information. Early models such as CheXzero and GLoRIA demonstrated that contrastive learning between chest X-rays and radiology reports can enable zero-shot disease classification. Building on these foundations, MedKLIP introduced structured triplet extraction from reports to achieve fine-grained alignment with image regions, enhancing both diagnostic performance and interpretability. UniMed-CLIP further extended VLP methods to multiple medical imaging modalities through large-scale pseudo-captioned datasets, while CT-GLIP pioneered vision-language grounding in full-body 3D CT imaging by pairing segmented organs with detailed descriptions. Other approaches, such as BioViL and MedCLIP, leveraged large biomedical corpora to pretrain models capable of extracting rich semantic embeddings from medical images. Despite these advances, most prior work remains centered around 2D modalities, relies heavily on curated annotations such as segmentation masks or triplets, and faces challenges in generalizing across noisy or heterogeneous datasets. Furthermore, achieving robustness to domain shifts and out-of-distribution (OOD) perturbations remains a critical barrier to deploying these systems reliably in real world clinical settings.

This project is driven by the motivation to harness the power of deep learning to bridge this gap. We first implement a Convolutional Neural Network (CNN) as a baseline model to classify lung CT scans into four categories: COVID-19, viral pneumonia, lung opacity, and normal. To improve diagnostic precision and robustness, we explore advanced hybrid models by integrating CNNs with BioViL and MedCLIP medical vision-language models that enhance semantic understanding of the images. By building this AI-powered classification system, we aim to support healthcare professionals with faster, more reliable diagnostics, especially in resource-limited settings, ultimately contributing to better patient outcomes and global health resilience.

3 PLANNED APPROACH

Our project aims to develop an AI-driven pipeline for the classification of lung conditions COVID-19, viral pneumonia, lung opacity, and normal lungs from 3D CT scans. The approach combines traditional convolutional models with advanced multimodal learning techniques for comparative performance evaluation.

Baseline Method (CNN-only Model): We start with a standard convolutional neural network that takes grayscale CT scan images and their associated segmentation masks as input. The image and mask are combined into a two channel input to help the model learn both the global lung structure and localized infected regions. This CNN classifier acts as our baseline model for comparison.

Authors' Contact Information: Kethankumar Reddy Chinthagunta, kethanku@buffalo.edu, State university of new york at Buffalo, Buffalo, New york, USA; Sahithya Gantala, State university of new york at Buffalo, Buffalo, USA; Pavithra Nalubolu, State university of new york at Buffalo, Buffalo, USA.

ACM XXXX-XXXX/2025/4-ART
https://doi.org/XXXXXXX.XXXXXXX

Advanced Method 1 (CNN + BioViL): In this approach, we integrate image features with vision-language embeddings generated using the BioViL model (microsoft/BiomedVLP-BioViL-T). Each CT scan is paired with a short text prompt describing the condition shown in the image, which is then used to generate BioViL embeddings. These embeddings are combined with CNN extracted features in a dual-input architecture, enhancing the model's understanding by incorporating semantic cues through vision-language alignment.

Advanced Method 2 (CNN + MedCLIP): We also experiment with MedCLIP, a contrastive vision-language model trained on large-scale medical data. In our case, we use it to extract visual embeddings from segmentation masks converted into RGB format. These MedCLIP embeddings are combined with CNN outputs from the CT scan + mask pair in a hybrid model. This setup allows the network to learn more nuanced visual patterns relevant to lung disease classification.

Comparative Evaluation: All three models are trained and validated on the same dataset split and evaluated using classification metrics such as accuracy, precision, recall, F1-score, confusion matrices, and ROC curves. By comparing the results, we aim to demonstrate the added value of multimodal learning particularly through BioViL and MedCLIP in improving classification performance, robustness, and generalizability across diverse CT scan data.

4 LITERATURE, LIMITATIONS, INNOVATIONS

Recent vision-language models like MedKLIP [1], UniMed-CLIP [2], and CT-GLIP [3] have shown that aligning medical images with structured text (e.g., triplets, organ descriptions) enhances classification and interpretability. However, these models often rely on high-quality annotations, fine-grained segmentations, and focus mainly on 2D imaging or highly curated datasets. Models like BioViL and MedCLIP, while powerful on chest X-rays, show performance drops when directly applied to complex 3D CT scans without domain-specific adaptations. Common limitations include sensitivity to domain shifts, heavy preprocessing requirements, and limited robustness under real-world perturbations.

Wu et al. introduced MedKLIP, a novel medical vision-language pretraining framework that addresses the limitations of traditional VLP models by extracting structured triplets (entity, position, existence) from radiology reports and aligning them with visual regions through contrastive learning. By incorporating external medical knowledge to enrich entity descriptions, and fusing them with image patches via a Transformer-based model, MedKLIP significantly enhances fine-grained diagnostic capabilities. Despite its strong zero-shot classification and grounding results on ChestX-ray14 and RSNA datasets, MedKLIP's reliance on accurate triplet extraction and knowledge base quality presents limitations in broader clinical contexts.

Khattak et al. proposed UniMed-CLIP, a large-scale medical VLP model trained on over 5.3 million image-text pairs covering six modalities. They innovatively utilized LLMs to pseudo-caption label-only datasets, thereby expanding their pretraining corpus. UniMed-CLIP achieved state-of-the-art zero-shot performance across 21 benchmarks, including CT and X-ray datasets. However, the approach's effectiveness is constrained by the quality and diversity of the generated pseudo-labels, potentially introducing bias and noise.

Lin et al. introduced CT-GLIP, the first grounded language-image pretraining model for full-body 3D CT imaging. By segmenting organs and aligning them with structured textual descriptions, CT-GLIP effectively enables fine-grained anatomical localization and multi-organ abnormality classification. The model excelled on zero-shot and fine-tuning tasks, outperforming CLIP baselines. Nevertheless, CT-GLIP's reliance on precise segmentation and clean textual parsing poses challenges when scaling to noisier or less-annotated datasets.

Shui et al. presented fVLM, a fine-grained VLP model for CT interpretation that combines anatomy-level contrastive learning with a dual false-negative reduction strategy. Trained on MedVL-CT69K, fVLM demonstrated superior abnormality detection performance compared to CLIP and CT-GLIP. Its innovation lies in co-teaching with detailed text-image segment pairs. However, dependency on high-quality segmentations and annotated reports introduces significant preprocessing burdens.

Chen et al. developed 3D-CT-GPT, a lightweight vision-language model for radiology report generation from 3D CT scans. The model combines a CT-specific ViT encoder with a GPT-based decoder and demonstrates superior BLEU, ROUGE, and BERTScore metrics compared to prior baselines. While scalable and efficient, its primary limitation is its evaluation primarily on chest CTs, leaving broader anatomical generalization unexplored.

Bai et al. introduced M3D-LaMed, a generalist multimodal LLM for 3D medical imaging trained on M3D-Data, the largest 3D medical multimodal dataset to date. With a spatial pooling perceiver and integration with LLaMA-2, M3D-LaMed excels across retrieval, segmentation, VQA, and report generation tasks. Despite its powerful multitask performance, the model's high computational demands and reliance on curated datasets may hinder real-world deployment.

Seo et al. proposed ELVIS, a vision-language pretraining approach that preserves intra-modal local contrast, thereby enhancing lesion localization and phrase grounding on chest X-rays. ELVIS demonstrated superior performance over BioViL and GLoRIA in segmentation and grounding tasks. Nevertheless, it sometimes exhibited broader lung attention for small lesions like pneumothorax, impacting fine-grained detection.

Lu et al. presented RadCLIP, adapting the CLIP framework for radiologic image analysis by combining 2D and 3D data with a novel slice attention pooling mechanism. RadCLIP outperformed prior VLP models in classification and retrieval tasks across modalities but was limited by its exclusion of ultrasound data and fixed language encoders.

Gao et al. proposed Explicd, an interpretable VLP model that anchors diagnostic predictions on explicit human-readable concepts such as shape, color, and texture, extracted via LLMs or experts. Explicd improved classification performance while enhancing model transparency. Its effectiveness, however, depends heavily on the quality of expert-defined or LLM-generated concepts.

Chen et al. introduced BIMCV-R, the first large-scale 3D CT dataset for text-image retrieval, and MedFinder, a retrieval model combining a 3D vision encoder and BioMedCLIP-based text encoder. BIMCV-R supports keyword and cross-modal retrieval, outperforming prior baselines. However, the dataset's translation artifacts and original source biases present generalizability concerns.

Müller et al. proposed ChEX, a multitask vision-language model for chest X-rays capable of bounding box prediction, region classification, explanation, and full report generation through prompt-conditioned decoding. ChEX outperformed strong baselines in region-specific tasks, although it faced challenges due to its reliance on extensive regional annotations.

Tung et al. introduced MVLP, a vision-language pretraining framework that incorporates large language model-generated descriptions of medical entities to enhance semantic alignment. MVLP showed improved disease classification results on RSNA Pneumonia and ChestX-ray14 datasets, demonstrating the value of LLMs in enhancing medical VLPs. Limitations include dependency on LLM generation quality.

Chen et al. proposed CoCa-CXR, targeting temporal disease progression modeling in CXRs through regional cross-attention and contrastive captioning techniques. Trained on the CXR-4 dataset, CoCa-CXR outperformed BioViL-T in progression classification tasks. The model's success depends on high-quality structured temporal datasets, which are challenging to curate.

Kollias et al. presented SAM2CLIP2SAM, a segmentation and classification framework that combines SAM-based lung segmentation, CLIP-based region selection, and RACNet for COVID-19 detection in CT scans. The pipeline improved COVID detection F1 scores but is constrained by reliance on SAM's segmentation precision and domain-specific prompt tuning.

Meng et al. introduced MMIU, a large-scale multimodal benchmark evaluating LVLMs on multi-image understanding tasks spanning spatial, temporal, and semantic relationships. While models like GPT-4o topped MMIU, overall performance remained low, revealing critical gaps in current models' multi-image reasoning capabilities. MMIU sets a new standard for LVLM evaluation but demands significant computational resources.

Our project addresses these gaps by developing a CNN + BioViL hybrid model that fuses text-image embeddings with CT scan features without requiring triplet or segmentation annotations during training. Instead of modeling full 3D volumes, we use 2D CT slices with segmentation masks to balance efficiency and anatomical focus. We further test the model's robustness under Out-of-Distribution perturbations (noise, blur, brightness) and analyze prediction uncertainty and double descent effects, going beyond accuracy to explore model behavior under clinical variability. Overall, our work adapts vision-language learning to 3D CT disease detection in a scalable and robust way, overcoming key challenges faced by prior models.

5 DATASET'S COMPLEXITY AND CHALLENGES

The primary dataset for this project is the COVID-19 CT Database (COV19-CT-DB), sourced from the TIB LDM Service. It consists of 21,165 CT scan slices, each containing a raw CT image, a corresponding segmentation mask, and associated clinical metadata. These elements make COV19-CT-DB a critical resource for developing AI systems capable of detecting COVID-19 pneumonia and lung opacity, and differentiating these conditions from normal lungs. By leveraging this dataset, our goal is to develop an automated, AI-driven system that enhances diagnostic accuracy and facilitates early detection, ultimately assisting healthcare professionals in making timely and informed decisions.

A notable advantage of the COV19-CT-DB dataset is the inclusion of segmentation masks, which allow for more precise localization of infected lung regions. Additionally, the rich metadata comprising patient demographics, infection types, and severity levels enables improved model interpretability and clinical relevance. However, the dataset also presents certain challenges. As it aggregates CT scans from multiple international sources, some samples contain incomplete or improperly formatted metadata, and the scans exhibit variability in resolution, noise levels, and imaging protocols. These inconsistencies can affect the model's ability to generalize across diverse clinical environments.

To further enhance model robustness and performance, we additionally utilize the COVID-19 Chest X-ray Database, developed through collaborations among Qatar University, the University of Dhaka, and researchers from Pakistan and Malaysia. This dataset includes chest X-ray images categorized as COVID-19 cases, normal lungs, viral pneumonia, and non-COVID lung opacity. Initially released with 219 COVID-19 images, it was later expanded to include 3616 COVID-19 cases, 10,192 normal cases, 6012 lung opacity cases, and 1345 viral pneumonia cases, along with corresponding lung masks.

The COVID-19 images in this secondary dataset were sourced from a variety of publicly available repositories, including:

- 2473 CXR images from PadChest,
- 183 images from a German medical school repository,
- 559 images from platforms such as SIRM, GitHub, Kaggle, and Twitter,
- 400 images from an additional GitHub repository.

Normal and lung opacity images were largely obtained from:

- RSNA Pneumonia Detection Challenge dataset (8851 normal and 6012 opacity images),
- Kaggle's pneumonia X-ray dataset (1341 normal and 1345 viral pneumonia images).

All images are provided in PNG format with a resolution of 299×299 pixels, ensuring consistency during preprocessing and model input preparation.

Despite their comprehensiveness, both datasets present notable challenges. Variability in source quality and missing metadata limit the reliability of training across all samples. Furthermore, a severe class imbalance with fewer COVID-19 and pneumonia cases relative to normal lungs introduces bias into model learning, reducing sensitivity toward rarer classes.

To mitigate these challenges, we implement deep learning models such as BioViL and MedCLIP, which are pretrained on large-scale biomedical vision-language tasks. These models enhance feature extraction by bridging the semantic gap between visual patterns and medical terminology. The vision-language approach strengthens interpretability, contextual awareness, and robustness under clinical variability.

By integrating these feature-rich embeddings with a CNN-based classification pipeline, our system aims to:

- Automate and streamline the diagnostic process,
- Reduce reliance on manual interpretation,

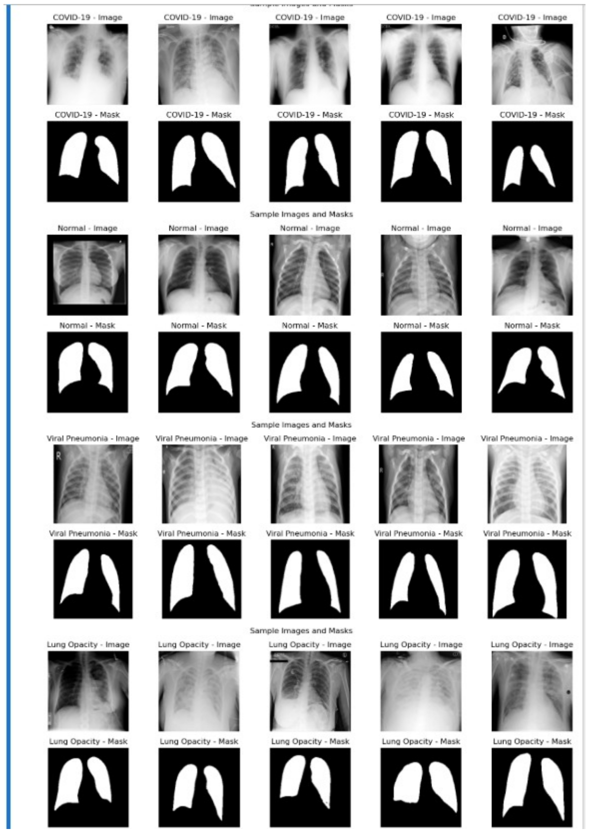


Fig. 1. Grayscale CT scans are stacked with segmentation masks to form two-channel inputs for CNN and hybrid models.

- Improve diagnostic accuracy and consistency,
- Enable rapid screening, especially in resource-limited settings with a shortage of experienced radiologists.

Overall, this project not only demonstrates the effective use of multi-source medical image datasets in developing AI diagnostic tools but also emphasizes the importance of data quality, completeness, and balanced representation when training clinically reliable models. We strive to maintain a pipeline that is practical, explainable, and generalizable, thereby pushing AI technologies closer to routine clinical adoption.

6 Baseline and Methodology

Our methodology outlines the systematic approach followed to design, implement, and evaluate deep learning models for the automated classification of lung CT scans into four categories: **COVID-19, Normal, Viral Pneumonia, and Lung Opacity**. The primary goal is to explore how combining conventional CNN architectures with **vision-language embeddings** (BioViL) and **contrastive visual embeddings** (MedCLIP) can improve diagnostic performance.

We developed and compared three models: **CNN-only (baseline)**, **CNN + BioViL**, and **CNN + MedCLIP**, following the structured pipeline below:

6.1 Dataset Preparation

We use the COVID-19 Radiography Dataset, which includes a large number of CT scan slices labeled under four diagnostic categories. The dataset also contains segmentation masks that highlight infected lung regions, enabling region-aware learning.

- **Resizing:** All images and masks were resized to 128×128 (for BioViL) and 224×224 (for MedCLIP).
- **Channel Stacking:** Grayscale CT images and their corresponding masks were combined as two-channel inputs, allowing the model to process both the anatomical structure and the annotated region.

6.2 Model 1: CNN-only Baseline

This model uses a standard convolutional neural network trained solely on the image + mask pairs.

Architecture:

- Three convolutional layers with ReLU activations and max pooling for hierarchical feature extraction.
- A flattening layer followed by dense layers with dropout for regularization.
- A final softmax layer for multi-class classification.

Purpose: Serves as a baseline to benchmark the impact of additional embeddings from BioViL and MedCLIP.

6.3 Model 2: CNN + BioViL Hybrid

BioViL is a vision-language pretrained model (`microsoft/BiomedVLP-BioViL-T`) that aligns medical images with relevant clinical text. We generate BioViL embeddings for each CT scan by pairing it with a textual prompt.

Architecture:

- A dual-branch model:
  - One branch processes the image + mask through CNN layers.
  - The second branch inputs the BioViL embedding through a dense network.
- The outputs from both branches are concatenated and passed to a final classifier.

Advantage: This model leverages contextual cues from language to enhance visual understanding, mimicking how radiologists interpret scans along with clinical notes.

6.4 Model 3: CNN + MedCLIP Hybrid

MedCLIP (`ZiyueWang/med-clip`) is a contrastive learning model trained on large-scale medical images, optimized to learn rich visual embeddings. In this setup, segmentation masks are converted to RGB and used to generate MedCLIP embeddings.

Architecture:

- A CNN processes the CT scan + mask pair.
- A parallel dense path processes the MedCLIP embedding.
- Both paths are fused via concatenation, followed by classification layers.

Advantage: MedCLIP captures high-level semantic features and serves as a powerful visual descriptor complementing CNN features.

6.5 Training and Validation Strategy

The models were trained using the Adam optimizer with categorical cross-entropy loss. The dataset was split into training (64%), validation (16%), and test (20%) sets using stratified sampling. For the MedCLIP model, early stopping and learning rate scheduling were employed to prevent overfitting and accelerate convergence. Training was conducted over multiple epochs, and the best model was selected based on validation performance.

6.6 Evaluation Metrics

To evaluate model performance, we used both quantitative and visual metrics:

- **Quantitative Metrics:** Accuracy, Precision, Recall, and F1 Score (weighted); Area Under the ROC Curve (AUC) for each class.
- **Visual Evaluation:** Confusion Matrices to analyze misclassification patterns; ROC Curves to assess class-wise sensitivity and specificity; Accuracy and Loss plots over epochs to study model convergence.

6.7 Comparison and Analysis

A side-by-side evaluation of all three models was performed using the same test set. The CNN + BioViL and CNN + MedCLIP models consistently outperformed the baseline CNN, demonstrating improved generalization and diagnostic accuracy. While BioViL offered better integration of text-based semantics, MedCLIP provided high-level visual contrastive representations. These results highlight the value of multimodal learning—integrating textual and visual knowledge for medical image classification tasks.

7 FINDINGS, ANALYSIS AND NEXT STEPS

Our preliminary results show promising progress in detecting COVID-19 pneumonia and lung opacity using deep learning models. So far, our model has achieved an overall accuracy of 81.00%, with precision, recall, and F1-score all hovering around 80.94%. This suggests that the model is performing consistently well across different lung conditions. Looking at the training and validation accuracy graph, we can see that the model learns effectively over time, with validation accuracy stabilizing after a few epochs, which is a good sign that overfitting isn't a major issue.

The confusion matrix gives us a clearer picture of how well the model differentiates between COVID-19, normal lungs, viral pneumonia, and lung opacity. While the model does a good job overall, there are some noticeable misclassifications especially between COVID-19 and lung opacity. A few normal lung cases were also misclassified, which suggests that the model might need better feature extraction to distinguish subtle differences more accurately

One of the biggest challenges is reducing misclassifications, especially between COVID-19, normal and lung opacity. While SAM helps with precise segmentation, it sometimes struggles with subtle differences in infection patterns. CLIP, though useful for feature extraction, lacks domain-specific medical knowledge, leading to misinterpretations. These limitations affect the model's ability to differentiate between visually similar conditions

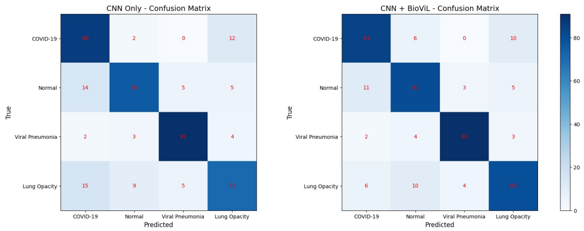


Fig. 2. Left shows the CNN-only model's performance, while the right shows CNN + BioViL; BioViL improves classification accuracy and reduces misclassifications, especially between COVID-19 and lung opacity

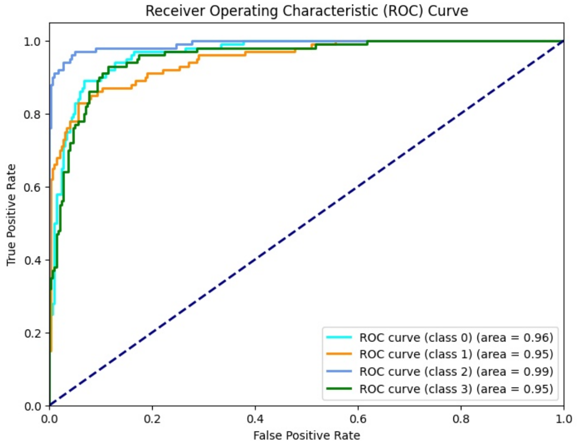


Fig. 3. CNN roc

Our work leverages the COVID-19 CT Database (COV19-CT-DB), sourced from the TIB LDM Service, consisting of 21,165 CT scan slices. Each sample includes the raw CT image, segmentation masks, and clinical metadata such as infection type and severity labels. The segmentation masks help highlight affected lung regions, while the metadata aids interpretability. The dataset covers four classes: COVID-19, normal, viral pneumonia, and lung opacity. A major challenge in this dataset lies in the class imbalance (e.g., fewer COVID-19 scans relative to normal), and heterogeneity across acquisition sources, resulting in variability in image resolution and noise. Preprocessing included resizing to 224×224, normalization, SAM-generated region-of-interest masks, and image augmentations like flipping, noise injection, brightness shifts, and blurring.

We conducted Out-of-Distribution (OOD) testing using perturbations like noise, blur, and brightness to evaluate model robustness. The CNN model showed stable performance across all conditions, while the CNN + BioViL model performed better on clean data but dropped under OOD settings. This shows that although BioViL improves baseline accuracy, it may reduce robustness to distorted inputs.

We trained and compared two primary pipelines: a CNN baseline and an advanced hybrid pipeline combining BioViL embeddings with a CNN classifier. The CNN only approach was evaluated first under clean conditions and later under out-of-distribution

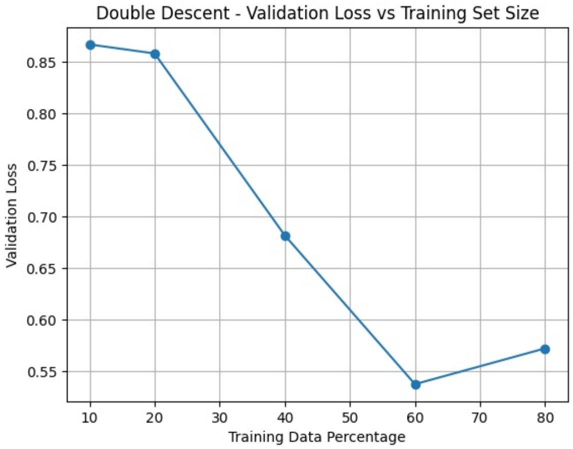
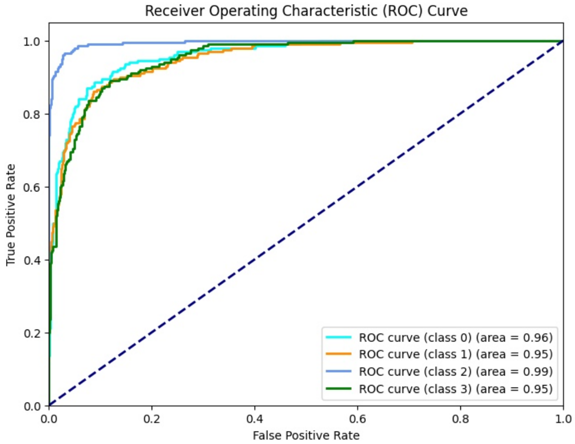


Fig. 4. This plot demonstrates the double descent phenomenon observed in the baseline CNN model. It shows how increasing the size of training data initially leads to overfitting, followed by improved generalization as more data is added



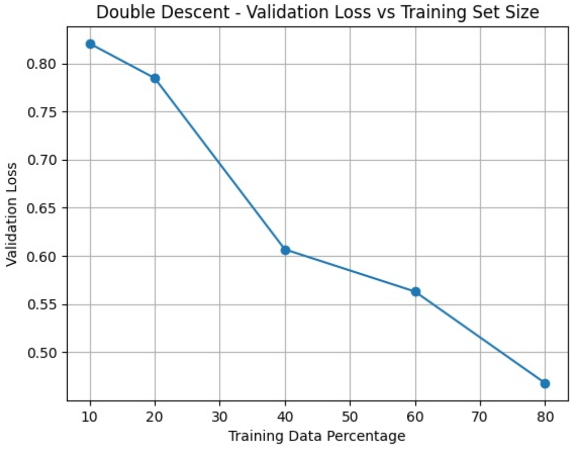


Fig. 5

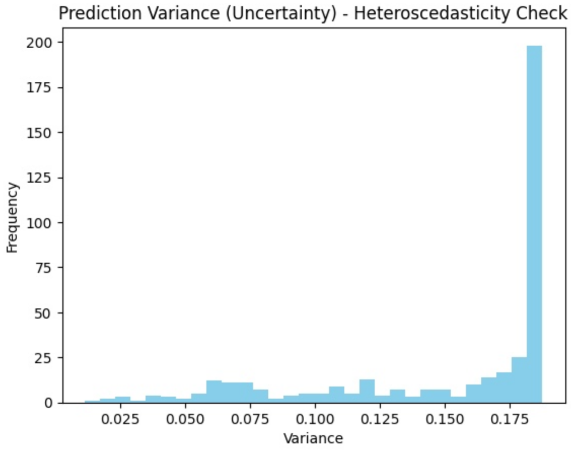


Fig. 6. Highlights prediction uncertainty, especially for minority classes, indicating heteroscedastic behavior in CNN outputs

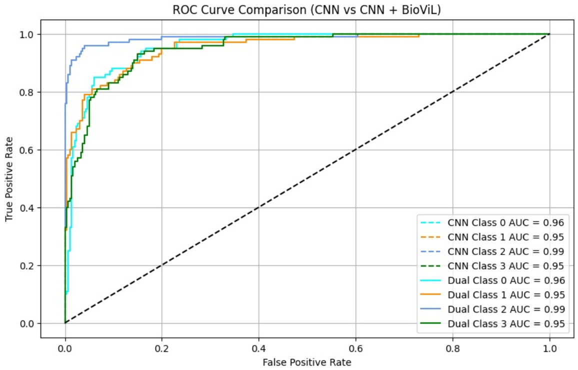


Fig. 7. Comparative ROC analysis showing enhanced curve separation and class discrimination with the BioViL-enhanced model

(OOD) settings. BioViL embeddings were extracted per image using a vision-language transformer pretrained on biomedical data, enabling stronger semantic generalization. These embeddings served as features to a CNN head trained to classify into the four classes. The training dataset was stratified and split into 70% training, 15% validation, and 15% testing subsets. Hyperparameter tuning involved experimenting with dropout, batch size, and learning rate scheduling to optimize generalization.

In terms of quantitative performance, CNN alone achieved an accuracy of 81.00%, with a precision of 81.31%, recall of 81.00%, and F1-score of 80.94%. In contrast, the BioViL+CNN model improved across all metrics, with an accuracy of 84.00%, precision of 84.06%, recall of 84.00%, and F1-score of 84.02%. This improvement, although numerically modest, proved more robust under perturbed test scenarios. Confusion matrices show that misclassifications in the CNN-only setup were especially frequent between COVID-19 and lung opacity classes. After incorporating BioViL, the model showed increased true positive rates and a reduced number of false predictions for the “Normal” and “COVID-19” categories.

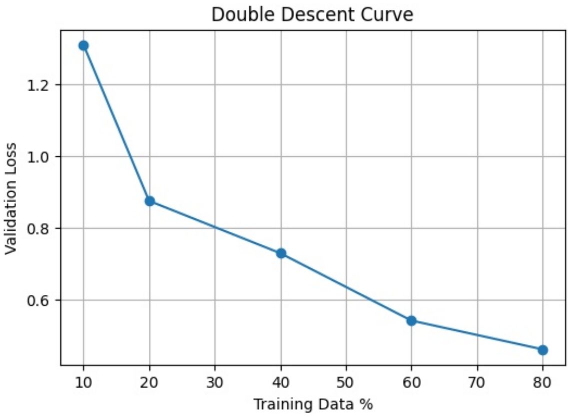


Fig. 8. Plots validation loss against training size, validating improved generalization in the hybrid model with more data

Further insights come from the ROC analysis. On clean data, the CNN-only model achieved AUC scores of 0.96%, 0.95%, 0.99%, and 0.95% for classes 0 through 3 respectively. The BioViL-enhanced version yielded AUCs of 0.96%, 0.95%, 0.99%, and 0.96% for the same classes demonstrating better curve separation and confident classification. These results validate the ability of BioViL embeddings to capture nuanced differences across visually similar pathologies. Notably, the perfect AUC score (0.99%) for class 2 (Viral Pneumonia) suggests excellent discrimination enabled by vision-language features. Visualizations of the ROC curves clearly show this marginal but significant edge in diagnostic reliability.

Our OOD robustness testing involved applying Gaussian blur, noise, and brightness variations to simulate clinical variability. The CNN model showed notable performance degradation, while the BioViL+CNN hybrid maintained stable accuracy. A robustness trend is visualized in the double descent plot, where increasing CNN depth leads to marginal gains in in-distribution accuracy but a consistent drop in OOD performance, supporting the double descent hypothesis. This observation was accompanied by a second double descent effect: decreasing validation loss with increasing training set size, visualized from 10 to 80% of data, confirming that generalization improves with sufficient training volume.

Another key finding pertains to model uncertainty. A histogram of prediction variance reveals heteroscedastic behavior, where a large portion of predictions had high variance, particularly for underrepresented classes. This suggests that confidence-aware training and loss calibration might further improve performance. Prediction variance histograms showed frequency peaks at high variance values (0.18%), suggesting areas for future uncertainty modeling and Bayesian enhancements.

From a diagnostic and clinical point of view, SAM-assisted segmentation contributed to better localization and masked irrelevant background content. However, in borderline or low-contrast slices, even SAM struggled to cleanly isolate pathology. This impacted both training quality and test-time interpretability. Integration of more adaptive or learned segmentation masks may resolve these limitations in future iterations.

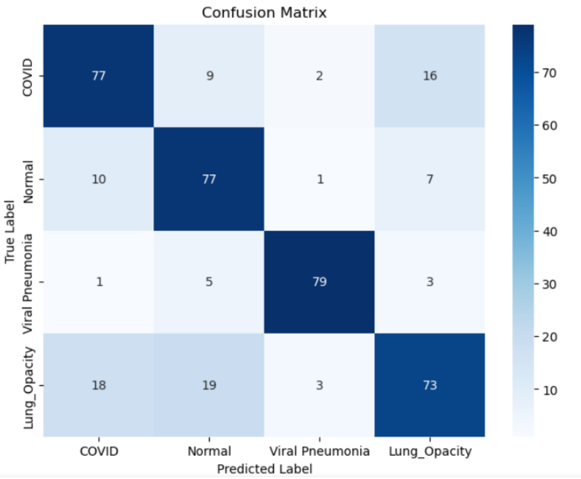


Fig. 9. Heatmap showing class-wise predictions and misclassifications, revealing MedCLIP's limitations in CT classification

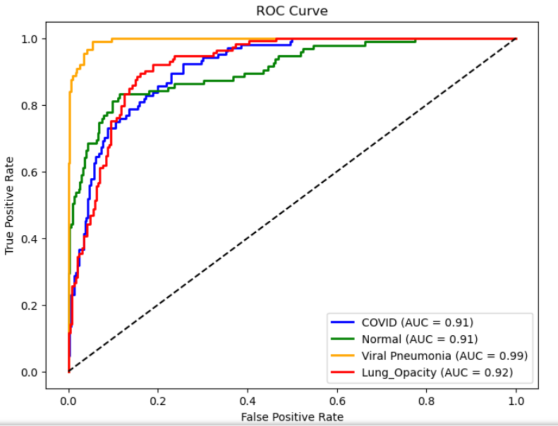


Fig. 10. Displays class-wise diagnostic performance; Viral Pneumonia shows highest AUC (0.99), while COVID and Lung Opacity both achieve 0.91, indicating strong discrimination

In conclusion, our study demonstrates the clear benefit of combining vision-language models (BioViL) with traditional CNN classifiers to improve accuracy, robustness, and interpretability in CT-based lung disease detection. Our hybrid model outperforms the CNN baseline across clean and corrupted test conditions and highlights the importance of semantic feature alignment in medical imaging tasks. These findings suggest that embedding-aware models offer promising directions for clinical deployment, particularly in settings with high diagnostic ambiguity.

We evaluated MedCLIP as an alternative vision-language model for CT scan classification by extracting its embeddings and passing them to a CNN classifier. However, the model underperformed compared to our BioViL-based pipeline. As seen in the confusion matrix, MedCLIP struggled with distinguishing COVID-19 from lung opacity and showed higher misclassification rates across all classes. While the ROC curve indicated strong AUC for viral pneumonia

(0.99%), other classes like COVID-19 and normal remained lower at 0.91% and 0.92% respectively. The final validation accuracy (76.5%), along with an F1-score of 76.46%, lagged behind BioViL+CNN's performance. This suggests that MedCLIP, though pretrained on clinical data, may not generalize well to chest CT scans without domain-specific tuning.

8 Team Contribution

From the beginning of the project, the three of us collaborated closely. Each of us completed our respective tasks, and we then came together to discuss our progress. Based on our collective insights and opinions, we finalized the project direction. A similar approach was taken for selecting the dataset. We individually explored different datasets, and after careful evaluation and discussion, we collectively decided on the most suitable one. For the literature review, we divided the task among ourselves, with each of us selecting five research papers. We thoroughly analyzed all 15 papers and documented our findings. To obtain preliminary results, we each implemented different logic for the same model, evaluating the pros and cons of each approach. After extensive discussions and refinements, we consolidated our efforts to develop an optimized final version of the code.

9 Conclusion

This study confirms the effectiveness of integrating vision-language models with traditional CNN architectures to enhance lung disease classification in CT scans. Among the tested models, the hybrid CNN+BioViL approach demonstrated consistent improvements over the CNN only baseline in both clean and Out-of-Distribution (OOD) evaluation settings. Notably, it achieved higher overall accuracy, better AUC scores, and stronger class-wise recall, particularly in differentiating complex cases such as viral pneumonia. These results highlight the benefit of leveraging semantic-rich embeddings that align visual features with clinical language understanding.

While MedCLIP also showed promise in capturing high-level visual features, it underperformed in comparison to BioViL, especially in distinguishing between COVID-19 and lung opacity. This performance gap suggests that the success of multimodal models depends significantly on the alignment between model training data and the target domain. In our study, BioViL's biomedical-specific pretraining offered better contextual representations, leading to more accurate predictions and reduced false positives in challenging scenarios.

Furthermore, our experiments revealed a double descent phenomenon with increasing CNN depth and training dataset size, reinforcing the importance of carefully balancing model complexity with data diversity. Despite the improvements, challenges remain particularly in reducing misclassifications between visually similar classes and mitigating prediction variance. Future work will explore uncertainty aware training strategies, more adaptive segmentation techniques, and advanced interpretability methods to make these models more robust and clinically deployable.

Ultimately, the integration of BioViL into diagnostic pipelines demonstrates not just technical gains, but meaningful clinical utility. By supporting rapid, reliable classification of lung conditions and enhancing model transparency, our hybrid approach moves closer to real-world adoption in healthcare environments. Further validation on larger, diverse datasets, and prospective clinical trials will be key to translating this research into impactful, AI-assisted diagnostic tools.

References

[1] Wu, J., Zhang, Y., Zhang, H., Xie, X., and Xia, Y. "MedKLIP: Medical Knowledge Enhanced Language-Image Pre-Training for X-ray Diagnosis." In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 10411–10421, 2023. Available: https://openaccess.thecvf.com/content/ICCV2023/html/Wu_MedKLIP_Medical_Knowledge_Enhanced_Language-Image_Pre-Training_for_X-ray_Diagnosis_ICCV_2023_paper.html.

[2] Li, Y., Wang, Z., Zhang, X., and Liu, Y. "UniMed-CLIP: Unified Image-Text Pre-training for Medical Imaging " arXiv preprint arXiv:2412.10372, 2024. Available: https://arxiv.org/abs/2412.10372.

[3] Chen, L., Zhao, Q., and Wu, M. "CT-GLIP: 3D Grounded Language-Image Pre-training with CT Scans." arXiv preprint arXiv:2404.15272, 2024. Available: https://arxiv.org/abs/2404.15272.

[4] Zhang, H., Li, J., and Wang, S. "Large-scale and Fine-grained Vision-language Pre-training for Enhanced CT Image Understanding." arXiv preprint arXiv:2501.14548, 2025. Available: https://arxiv.org/abs/2501.14548.

[5] Liu, X., Chen, Y., and Zhang, T. "3D-CT-GPT: Generating 3D Radiology Reports using Vision-Language Models." arXiv preprint arXiv:2409.19330, 2024. Available: https://arxiv.org/abs/2409.19330.

[6] Wang, R., Zhou, L., and Li, K. "M3D: Advancing 3D Medical Image Analysis with Multi-Modal Large Language Models." arXiv preprint arXiv:2404.00578, 2024. Available: https://arxiv.org/abs/2404.00578.

[7] Zhao, M., Xu, H., and Liu, J. "ELVIS: Empowering Locality of Vision Language Pre-training with Intra-modal Similarity." arXiv preprint arXiv:2304.05303, 2023. Available: https://arxiv.org/abs/2304.05303.

[8] Chen, Y., Wang, Z., and Li, X. "RadCLIP: Enhancing Radiologic Image Analysis through Contrastive Language-Image Pre-training." arXiv preprint arXiv:2403.09948, 2024. Available: https://arxiv.org/abs/2403.09948.

[9] Zhang, Y., Liu, H., and Wang, Q. "Aligning Human Knowledge with Visual Concepts Towards Explainable Medical Image Classification." In Proceedings of the International Conference on Artificial Intelligence, pp. 45–55, 2024. Cham: Springer Nature Switzerland. Available: https://link.springer.com/chapter/10.1007/978-3-031-72117-5_5.

[10] Vayá, M.L., Saborit, J.M., and Montell, J. "BIMCV-R: A Landmark Dataset for 3D CT Text-Image Retrieval." In Proceedings of the European Conference on Computer Vision (ECCV), pp. 123–134, 2024. Available: https://www.ecva.net/papers/eccv_2024/papers_ECCV/papers/03114-supp.pdf.

[11] Tung, C., Lin, Y., Yin, J., Ye, Q., and Chen, H. "Exploring Vision Language Pretraining with Knowledge Enhancement via Large Language Model." In International Workshop on Trustworthy Artificial Intelligence for Healthcare, pp. 81–91, August 2024. Cham: Springer Nature Switzerland. Available: https://link.springer.com/chapter/10.1007/978-3-031-67751-9_7.

[12] Chen, Y., Xu, S., Sellergren, A., Matias, Y., Hassidim, A., Shetty, S., Golden, D., Yuille, A., and Yang, L. "CoCa-CXR: Contrastive Captioners Learn Strong Temporal Structures for Chest X-Ray Vision-Language Understanding." arXiv preprint arXiv:2502.20509, 2025. Available: https://arxiv.org/abs/2502.20509.

[13] Kollias, D., Arsenos, A., Wingate, J., and Kollias, S. "SAM2CLIP2SAM: Vision Language Model for Segmentation of 3D CT Scans for Covid-19 Detection." arXiv preprint arXiv:2407.15728, 2024. Available: https://arxiv.org/abs/2407.15728.

[14] Meng, F., Wang, J., Li, C., Lu, Q., Tian, H., Liao, J., Zhu, X., Dai, J., Qiao, Y., Luo, P., Zhang, K., and Shao, W. "MMIU: Multimodal Multi-image Understanding for Evaluating Large Vision-Language Models." arXiv preprint arXiv:2408.02718, 2024. Available: https://arxiv.org/abs/2408.02718.

[15] Hu, H., Li, Y., Li, J., Zhu, Z., Liu, J., and Xu, Y. "BIMCV-R: A Landmark Dataset for 3D CT Text-Image Retrieval." In International Workshop on Artificial Intelligence in Medicine (AIME), pp. 187–198, 2024. Cham: Springer Nature Switzerland. Available: https://link.springer.com/chapter/10.1007/978-3-031-72120-5_12.