Multimodal Chest X-ray Diagnosis using Vision Transformers and BERT

Abdullah Hani Abdellatif Al-Shobaki

AIN3002 – Deep Learning, Bahçeşehir University

Gradio Demo

Abstract

I propose a deep learning pipeline that fuses chest X-ray images and patient symptom data to diagnose five respiratory diseases. Utilizing Vision Transformers (ViT) for image analysis and BERT for textual data, the model achieves a 96.05% F1-score. The system includes a real-time Gradio interface and provides visual explanations through attention heatmaps.

Introduction

Respiratory diseases such as COVID-19, Lung Cancer, Viral Pneumonia, and Tuberculosis are leading causes of mortality worldwide. Early and accurate diagnosis is crucial. Combining imaging data with patient-reported symptoms can enhance diagnostic accuracy, but integrating these modalities poses challenges.

Motivation & Challenges

- **Limitations of X-rays**: Sole reliance on imaging may miss contextual clinical information.
- **Integration Difficulty**: Combining heterogeneous data types (images and text) is non-trivial.
- **Need for Explainability**: Clinicians require transparent AI decisions for trust and adoption.

Key Contributions

- Developed a ViT + BERT fusion model for multimodal diagnosis.
- Synthesized a multimodal dataset with realistic symptom metadata.
- Achieved a 96.05% F1-score across five classes.
- Created a Gradio-based real-time user interface.
- Implemented attention heatmaps for model explainability.

Methodology

Model Architecture:

- **Image Encoder**: ViT (Base 16x224)
- **Text Encoder**: BERT (Base Uncased)
- **Fusion**: Concatenation followed by a 2-layer MLP with LayerNorm, ReLU, and Dropout

Data Pipeline:

- **Sources**: COVID-19 Radiography Database, NIH Chest X-rays, Shenzhen TB Dataset
- **Preprocessing**: Resized images to 224x224, normalized using ImageNet statistics
- **Metadata Synthesis**: Generated age, gender, and symptom data using `generate_metadata.py`

Training Setup:

- **Loss Function**: CrossEntropyLoss
- **Optimizer**: Adam with a learning rate of 1e-4
- **Batch Size**: 16
- **Early Stopping**: Implemented to prevent overfitting

Model Comparison

Model	Accuracy	F1-Score
ViT-only (image only)	83.37%	84.93%
BERT-only (text only)	91.31%	88.81%
Fusion (ViT + BERT)	**96.0%**	**96.0%**

Fusion significantly outperforms single-modality models, validating the power of multimodal learning.

Experimental Results

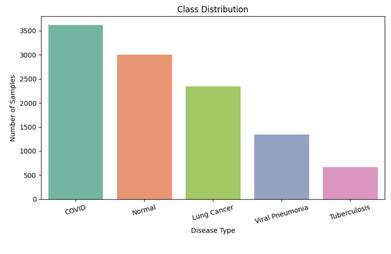

Figure 1. Class Distribution

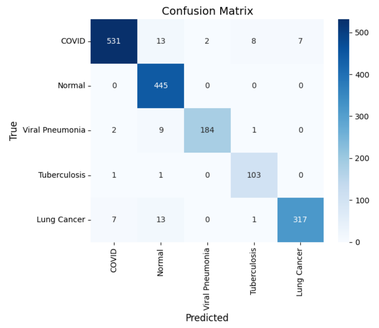

Figure 2. Confusion Matrix

Visual Explainability

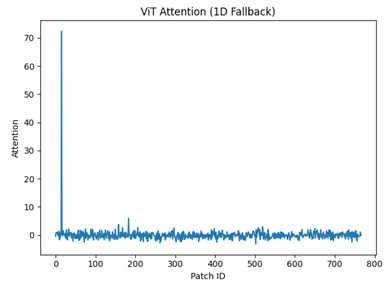

Figure 3. ViT Attention Heatmaps

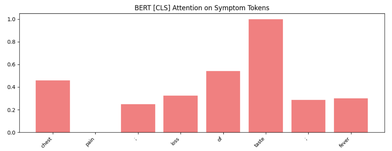

Figure 4. BERT Token Attention

Live Demo Deployment

Gradio UI Features:

- Upload chest X-ray images and input patient symptoms
- Receive top-3 predictions with confidence scores
- Visualize attention heatmaps and download detailed diagnosis reports

Training Diagnostics

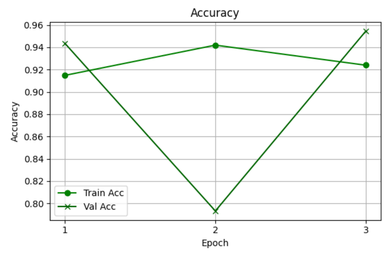

Figure 5. Training Metrics Over Epochs

```
# STEP 4: Train all models and generate results
!python main.py

Epoch [1/5]: 100% 480/480 [07:17<00:00,  1.10it/s, loss=0.183]

Epoch 1 Summary:
Train — Loss: 0.2741 | Acc: 0.9148 | F1: 0.9043
Val   — Acc: 0.9434 | F1: 0.9507
Best model updated — saved to best_model_fusion.pt
Epoch [2/5]: 100% 480/480 [07:16<00:00,  1.10it/s, loss=0.409]

Epoch 2 Summary:
Train — Loss: 0.2110 | Acc: 0.9419 | F1: 0.9383
Val   — Acc: 0.7932 | F1: 0.8326
No improvement. Patience: 1/2
Epoch [3/5]: 100% 480/480 [07:09<00:00,  1.12it/s, loss=0.025]

Epoch 3 Summary:
Train — Loss: 0.2524 | Acc: 0.9239 | F1: 0.9213
Val   — Acc: 0.9544 | F1: 0.9491
No improvement. Patience: 2/2
Early stopping triggered.
```

Figure 6. Early Stopping Visualization

Reproducibility Notes:

- Training conducted on Google Colab Pro with Tesla T4 GPU
- Random seeds set for reproducibility

Conclusion

- Successfully integrated multimodal data for respiratory disease diagnosis
- Achieved high accuracy and F1-score across all classes
- Provided model interpretability through attention mechanisms
- Developed a user-friendly interface for real-time diagnosis

Future Work

- Incorporate CT images and electronic health records (EHRs)
- Validate the model in clinical settings
- Explore integration with hospital information systems

Why Multi-Modal AI Matters

- X-rays alone are limited by ambiguity in visual patterns.
- Symptoms add valuable context — especially in early-stage or ambiguous cases.
- ViT+BERT fusion bridges visual and semantic understanding, improving diagnostic decisions.

References

1. Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers," NAACL, 2019.
2. Dosovitskiy et al., "An Image is Worth 16x16 Words," ICLR, 2021.
3. Wang et al., "ChestX-ray8: Hospital-scale Dataset," CVPR, 2017.
4. Rahman et al., "COVID-19 Radiography Database," IEEE Access, 2020.
5. Huang et al., "Fusion of Imaging and Clinical Data with Transformer," IEEE JBHI, 2021.
6. Raj et al., "TransFusion: Transfer Learning for Medical Imaging," NeurIPS, 2021.