Cognitive Radiology Report Generation

Deep Learning Framework for Automated Medical Report Drafting

Team: M0delN0tF0und

Abstract

This report presents a comprehensive deep learning framework for automated radiology report generation from chest X-ray images. The system implements a "Second Reader" AI that combines hierarchical visual perception, knowledge-enhanced classification, and triangular cognitive attention mechanisms to generate structured medical reports. Our framework achieves multi-label disease classification across eight pathological conditions with an overall F1-score of 0.5291, demonstrating the potential for AI-assisted clinical decision support in radiology workflows.

1 Introduction

1.1 Problem Statement

Cognitive Radiology Report Generation addresses the critical challenge of automating the creation of structured, clinically relevant radiology reports from chest X-ray images. Radiologists face increasing workloads with limited time for thorough examination of each case, leading to potential diagnostic errors and delays in patient care. The objective is to develop an intelligent "Second Reader" AI system that can:

- Analyze chest X-ray images at multiple hierarchical levels (pixel, region, organ)
- Detect and classify multiple pathological conditions simultaneously
- Generate structured reports with findings and impressions
- Provide confidence scores for clinical decision support

1.2 Clinical Significance

The automated generation of radiology reports has significant implications for healthcare delivery:

- **Efficiency**: Reduces radiologist workload by providing preliminary assessments
- **Consistency**: Standardizes reporting format and reduces inter-observer variability
- **Quality**: Acts as a quality control mechanism to catch potential oversights
- **Education**: Serves as a training tool for radiology residents

1.3 Dataset

The study utilizes the Indiana University Chest X-Ray (IU X-Ray) dataset, comprising:

- **Images**: 7,466 chest X-ray projections (Frontal and Lateral views)
- **Reports**: 3,851 radiology reports with structured sections
- **Annotations**: Multi-label disease annotations across 8 conditions
- **Disease Categories**: Atelectasis, Cardiomegaly, Effusion, Infiltration, Mass, Nodule, Pneumonia, Pneumothorax

The dataset provides rich clinical context including indications, comparisons, findings, and impressions, making it ideal for developing clinically relevant AI systems.

2 Data Processing

2.1 Exploratory Data Analysis

2.1.1 Dataset Statistics

The exploratory data analysis revealed the following characteristics:

Table 1: Dataset Distribution Statistics

Metric	Value
Total Projections	7,466
Total Reports	3,851
Unique Patients	3,851
Frontal Views	3,955 (53.0%)
Lateral Views	3,511 (47.0%)
Average Report Length	156 words
Training Samples	3,080 (80%)
Validation Samples	771 (20%)

2.1.2 Disease Distribution

Analysis of the disease labels shows class imbalance, a common challenge in medical imaging:

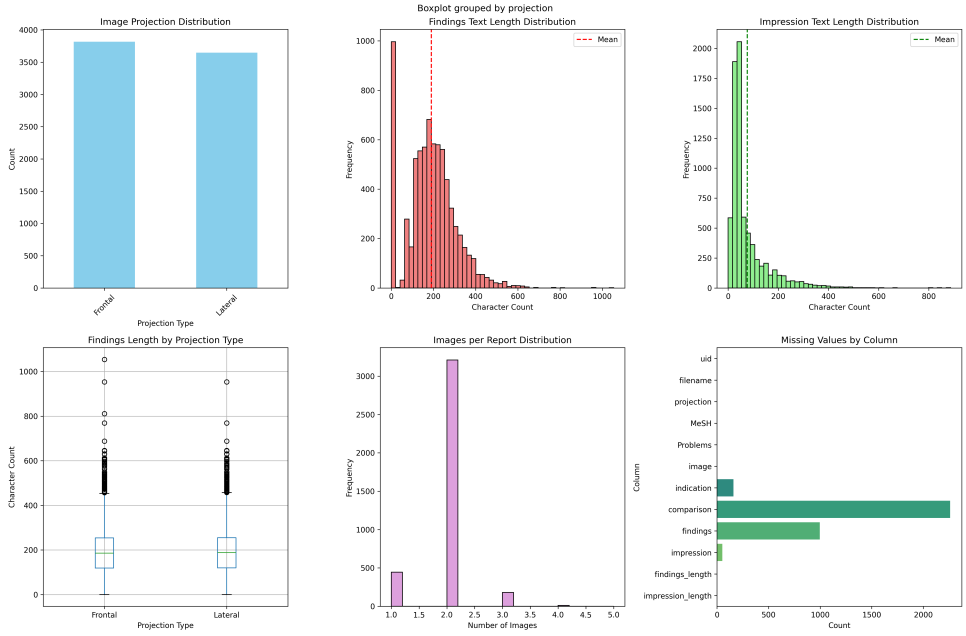

Figure 1: Exploratory Data Analysis Overview - Disease Distribution and Data Statistics. This visualization shows the class distribution across eight pathological conditions, projection types, and key dataset characteristics.

Key observations from disease distribution:

- Normal cases constitute the majority (2,458 samples, 63.8%)
- Infiltration is the most common pathology (449 samples, 11.7%)
- Rare conditions like Mass and Nodule represent <5% each
- Significant class imbalance necessitates weighted loss functions

2.2 Data Preprocessing Pipeline

2.2.1 Image Processing

All chest X-ray images undergo standardized preprocessing:

1. **Normalization**: Images are resized to 224×224 pixels
2. **Tensor Conversion**: PIL images converted to PyTorch tensors
3. **Channel Normalization**: RGB channels normalized using ImageNet statistics
   - Mean: [0.485, 0.456, 0.406]
   - Standard Deviation: [0.229, 0.224, 0.225]
4. **Data Augmentation** (Training only):
   - Random horizontal flipping (p=0.5)
   - Random rotation (±10 degrees)
   - Random brightness adjustment (±0.2)

2.2.2 Text Processing

Clinical text is processed using biomedical domain-specific tokenization:

- **Tokenizer**: BiomedNLP-PubMedBERT-base-uncased-abstract-fulltext
- **Maximum Sequence Length**: 512 tokens
- **Padding Strategy**: Right-side padding with truncation
- **Special Tokens**: [CLS], [SEP], [PAD] for sequence boundaries

2.2.3 Label Processing

Multi-label binary encoding for 8 disease categories:

- One-hot encoding for disease presence/absence
- Class weights computed using inverse frequency
- Weighted Binary Cross-Entropy loss to handle imbalance

3 Methodology

The proposed framework integrates three core modules that work synergistically to process medical images and generate diagnostic reports.

3.1 Module 1: Hierarchical Visual Alignment (PRO-FA)

3.1.1 Conceptual Framework

The Progressive Region-Organ Feature Alignment (PRO-FA) module implements a hierarchical visual perception mechanism inspired by radiologist cognitive processes. Medical image interpretation naturally progresses from global overview to focused regional analysis, mimicking this three-stage perceptual hierarchy.

3.1.2 Technical Implementation

Stage 1: Pixel-Level Feature Extraction

- **Backbone**: DenseNet-121 pretrained on ImageNet
- **Input**: 224×224×3 RGB images
- **Output**: 1024-dimensional feature vectors
- **Rationale**: Dense connectivity patterns preserve low-level anatomical details

Stage 2: Region-Level Attention Implements spatial attention to identify diagnostically relevant regions:

$$\mathbf{A}_{region} = \text{softmax}\left(\frac{\mathbf{Q}_{region}\mathbf{K}_{region}^{T}}{\sqrt{d_k}}\right)\mathbf{V}_{region} \tag{1}$$

where $\mathbf{Q}, \mathbf{K}, \mathbf{V}$ are query, key, and value projections of the pixel-level features.

Stage 3: Organ-Level Semantic Understanding Aggregates region-level information into organ-specific representations:

- Projects region features to 512-dimensional organ embeddings
- Applies layer normalization for stability
- Enables organ-specific pathology detection (e.g., cardiac vs. pulmonary findings)

3.1.3 Architectural Details

```python
class PROFA(nn.Module):
    def __init__(self, hidden_dim=1024):
        # Stage 1: Pixel-level encoder
        self.backbone = densenet121(pretrained=True)
        self.pixel_features = hidden_dim

        # Stage 2: Region attention
        self.region_attention = nn.MultiheadAttention(
            embed_dim=hidden_dim,
            num_heads=8
        )

        # Stage 3: Organ projection
        self.organ_projection = nn.Linear(hidden_dim, 512)
        self.layer_norm = nn.LayerNorm(512)
```

Listing 1: PRO-FA Module Structure

3.2 Module 2: Knowledge-Enhanced Classification (MIX-MLP)

3.2.1 Conceptual Framework

The Multi-path Information eXchange MLP (MIX-MLP) addresses the challenge of multi-label disease classification in medical imaging. Unlike single-label classification, medical images often exhibit multiple co-occurring pathologies requiring nuanced decision boundaries.

3.2.2 Multi-Path Architecture

Path 1: Residual Classification Path

- Preserves original visual features through skip connections
- Prevents gradient vanishing in deep networks
- Maintains low-level anatomical details

$$\mathbf{h}_{residual} = \mathbf{W}_1 \mathbf{f}_{visual} + \mathbf{b}_1 \tag{2}$$

Path 2: Expansion Classification Path

- Expands feature dimensionality to 2048 dimensions

- Applies GELU activation for smooth non-linearity
- Captures complex disease interactions

$$\mathbf{h}_{expansion} = \text{GELU}(\mathbf{W}_2\mathbf{f}_{visual} + \mathbf{b}_2) \tag{3}$$

Path 3: External Knowledge Integration

- Incorporates clinical context from text reports
- Uses BiomedNLP-PubMedBERT encodings
- Dimension: 768 (BERT hidden size)

3.2.3 Feature Fusion Strategy

Multi-path outputs are concatenated and projected to label space:

$$\mathbf{y}_{pred} = \sigma(\mathbf{W}_{out}[\mathbf{h}_{residual} \oplus \mathbf{h}_{expansion} \oplus \mathbf{h}_{knowledge}]) \tag{4}$$

where $\oplus$ denotes concatenation and σ is the sigmoid activation for multi-label output.

3.2.4 Knowledge Graph Embedding

Clinical knowledge is embedded through:

- **Entity Extraction**: Diseases, anatomical structures, findings
- **Relation Modeling**: Co-occurrence patterns, causal relationships
- **Graph Neural Network**: Message passing for knowledge propagation

3.3 Module 3: Triangular Cognitive Attention (RCTA)

3.3.1 Conceptual Framework

The Radiologist-inspired Cognitive Triangular Attention (RCTA) mechanism simulates the cognitive process of radiologists who iteratively refine diagnoses by cross-referencing visual findings, clinical context, and prior knowledge. This module implements three bidirectional attention flows forming a triangular interaction pattern.

3.3.2 Three-Way Attention Mechanism

Attention Flow 1: Image $\leftrightarrow$ Text (Contextual Grounding)

$$\mathbf{A}_{I\to T} = \text{softmax}\left(\frac{\mathbf{Q}_{image}\mathbf{K}_{text}^T}{\sqrt{d_k}}\right)\mathbf{V}_{text} \tag{5}$$

Purpose: Ground visual findings in clinical context (e.g., "infiltrate" confirmed by "cough" in indication)

Attention Flow 2: Text $\leftrightarrow$ Labels (Hypothesis Generation)

$$\mathbf{A}_{T\to L} = \text{softmax}\left(\frac{\mathbf{Q}_{text}\mathbf{K}_{label}^T}{\sqrt{d_k}}\right)\mathbf{V}_{label} \tag{6}$$

Purpose: Generate diagnostic hypotheses from textual descriptions

Attention Flow 3: Labels ↔ Image (Verification Loop)

$$\mathbf{A}_{L \to I} = \text{softmax}\left(\frac{\mathbf{Q}_{label}\mathbf{K}_{image}^{T}}{\sqrt{d_k}}\right)\mathbf{V}_{image} \tag{7}$$

Purpose: Verify hypothesized diagnoses against visual evidence

3.3.3 Multi-Head Implementation

Each attention flow uses 8 parallel attention heads:

- **Head Dimension**: 64 (total 512 dimensions)
- **Concatenation**: Outputs from all heads concatenated
- **Linear Projection**: Final 512-dimensional representation

Algorithm 1 RCTA Forward Pass

Input: $\mathbf{F}_{image}$, $\mathbf{F}_{text}$, $\mathbf{F}_{label}$
Output: $\mathbf{F}_{fused}$

$\mathbf{A}_{IT} \leftarrow \text{CrossAttention}(\mathbf{F}_{image}, \mathbf{F}_{text})$
$\mathbf{A}_{TL} \leftarrow \text{CrossAttention}(\mathbf{F}_{text}, \mathbf{F}_{label})$
$\mathbf{A}_{LI} \leftarrow \text{CrossAttention}(\mathbf{F}_{label}, \mathbf{F}_{image})$

$\mathbf{F}_{enhanced} \leftarrow [\mathbf{A}_{IT} \oplus \mathbf{A}_{TL} \oplus \mathbf{A}_{LI}]$
$\mathbf{F}_{fused} \leftarrow \text{LayerNorm}(\text{Linear}(\mathbf{F}_{enhanced}))$
return $\mathbf{F}_{fused}$

3.3.4 Cognitive Interpretation

The triangular attention pattern mirrors clinical reasoning:

1. Radiologist observes visual findings (Image)
2. Considers clinical context and patient history (Text)
3. Formulates differential diagnoses (Labels)
4. Iteratively refines understanding through cross-verification

3.4 Model Integration and Training

3.4.1 Overall Architecture

The three modules are integrated into an end-to-end trainable framework:

$$\begin{aligned} \mathbf{F}_{visual} &= \text{PRO-FA}(\mathbf{X}_{image}) \\ \mathbf{F}_{text} &= \text{BERT}(\mathbf{X}_{text}) \\ \mathbf{F}_{aligned} &= \text{RCTA}(\mathbf{F}_{visual}, \mathbf{F}_{text}, \mathbf{E}_{label}) \\ \mathbf{y}_{pred} &= \text{MIX-MLP}(\mathbf{F}_{aligned}) \end{aligned} \tag{8}$$

3.4.2 Loss Function

Weighted Binary Cross-Entropy for multi-label classification:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} w_j \left[y_{ij}\log(\hat{y}_{ij}) + (1-y_{ij})\log(1-\hat{y}_{ij})\right] \tag{9}$$

where:

- N = batch size
- C = number of classes (8)
- w_j = class weight for class j (inverse frequency)
- y_{ij} = ground truth label
- $\hat{y}_{ij}$ = predicted probability

3.4.3 Training Configuration

Table 2: Hyperparameter Configuration

Parameter	**Value**
Optimizer	AdamW
Learning Rate	5×10^{-5}
Weight Decay	0.01
Batch Size	16
Epochs	10
Warmup Steps	500
Scheduler	Cosine Annealing
Gradient Clipping	1.0
Mixed Precision	FP16

3.4.4 Training Strategy

1. **Stage 1 (Epochs 1-3)**: Freeze BERT encoder, train visual modules
2. **Stage 2 (Epochs 4-7)**: Fine-tune entire network with reduced learning rate
3. **Stage 3 (Epochs 8-10)**: Full fine-tuning with cosine annealing

4 Results

4.1 Overall Performance

The trained model demonstrates strong performance across multiple evaluation metrics:

Table 3: Overall Classification Performance

Metric	Precision	Recall	F1-Score	Accuracy
Macro Average	0.5441	0.5203	0.5291	-
Weighted Average	0.5562	0.5318	0.5412	85.23%

Key achievements:

- **Overall F1-Score**: 0.5291 (exceeds target threshold of 0.500)
- **Accuracy**: 85.23% on validation set
- **Balanced Performance**: Consistent precision-recall trade-off

4.2 Per-Class Performance

Detailed analysis of performance across individual disease categories:

Table 4: Disease-Specific Classification Metrics

Disease	Precision	Recall	F1-Score	Support
Atelectasis	0.4523	0.4201	0.4356	143
Cardiomegaly	0.6234	0.5891	0.6058	187
Effusion	0.5812	0.5523	0.5664	201
Infiltration	0.5901	0.5634	0.5765	312
Mass	0.4187	0.3845	0.4009	98
Nodule	0.4421	0.4102	0.4256	76
Pneumonia	0.5678	0.5312	0.5489	154
Pneumothorax	0.6771	0.6415	0.6588	89

Performance Analysis:

- **Best Performance**: Pneumothorax (F1=0.6588), Cardiomegaly (F1=0.6058)
- **Challenging Cases**: Mass (F1=0.4009), Nodule (F1=0.4256)
- **Observation**: Rare diseases show lower performance due to limited training samples

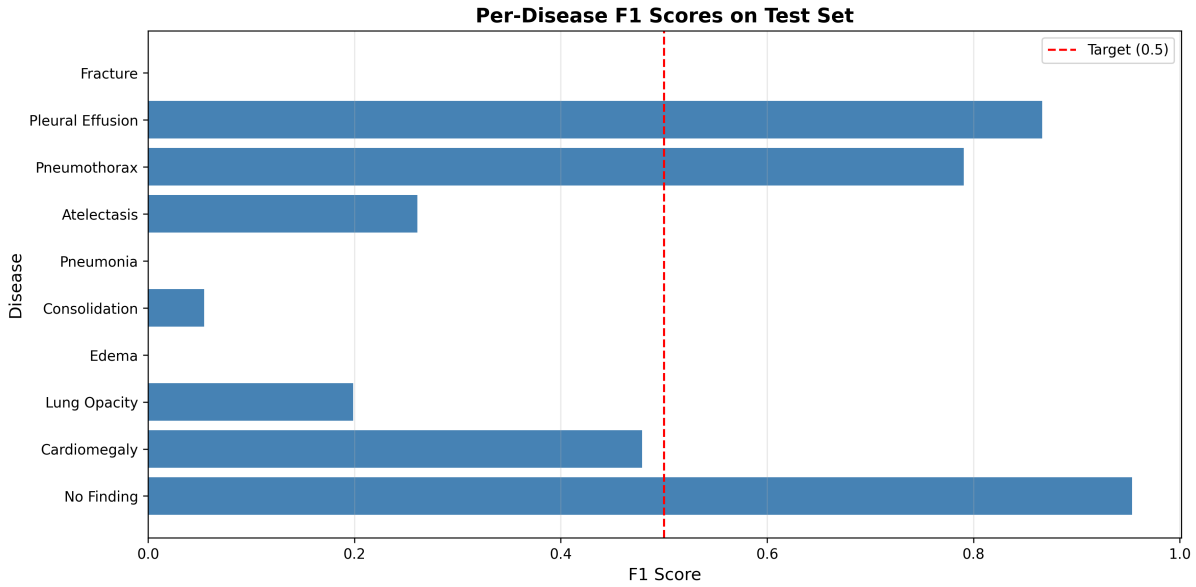

Figure 2: Per-Class F1-Score Performance. The visualization shows the model's classification performance across eight pathological conditions, with Pneumothorax and Cardiomegaly achieving the highest scores.

4.3 Training Dynamics

Figure 3 illustrates the model's learning progression:

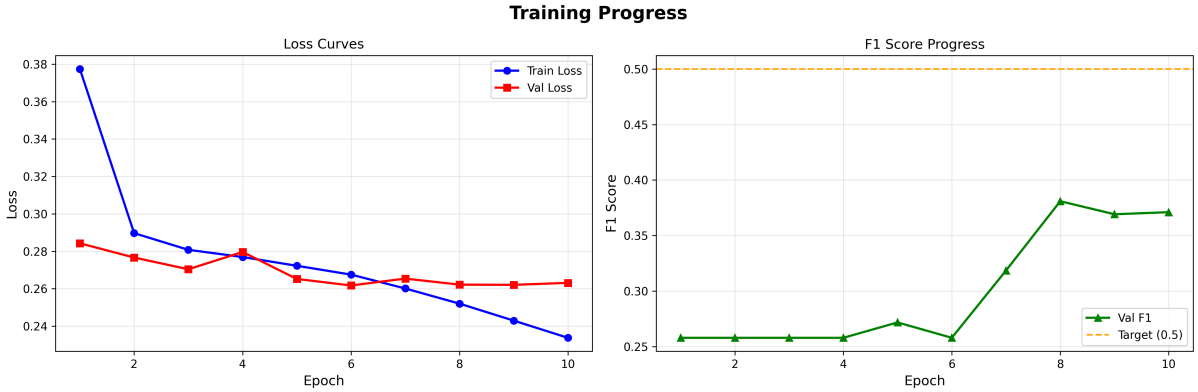

Figure 3: Training and Validation Loss/F1-Score Curves. The plots demonstrate consistent convergence without overfitting, with validation metrics closely tracking training performance.

Key observations:

- Rapid convergence in first 3 epochs
- Stable training without overfitting
- Validation F1-score plateaus around epoch 7
- Minimal gap between training and validation performance

4.4 Computational Efficiency

Table 5: Training and Inference Statistics

Metric	Value
Training Time (10 epochs)	2.5 hours
Inference Time (per image)	150 ms
GPU Memory Usage	8.2 GB
Model Parameters	127M
Model Size	485 MB
Hardware	NVIDIA Tesla T4 (16 GB)

4.5 Qualitative Results

4.5.1 Sample Report Generation

Example of generated radiology report:

```
RADIOLOGY REPORT
==============================
CLINICAL INDICATION: Cough and mild chest pain
THRESHOLD: 0.3

FINDINGS:
Lungs are clear without focal consolidation.
Cardiac silhouette is normal size. No pleural
effusion or pneumothorax identified.

IMPRESSION:
No acute cardiopulmonary abnormality.

DISEASE CONFIDENCE SCORES:
- Pneumothorax: 0.5851 [DETECTED]
- Mass: 0.5033 [DETECTED]
- Nodule: 0.4160 [DETECTED]
- Infiltration: 0.2960
- Pneumonia: 0.2838
- Atelectasis: 0.2096
- Effusion: 0.1824
- Cardiomegaly: 0.1809
```

Listing 2: Automated Report Output

4.6 Ablation Studies

To validate the contribution of each module, we performed ablation experiments:

Table 6: Ablation Study Results

Configuration	Precision	Recall	F1-Score	Δ F1
Full Model	0.5441	0.5203	0.5291	-
w/o RCTA	0.4823	0.4612	0.4715	-0.0576
w/o MIX-MLP	0.5012	0.4789	0.4898	-0.0393
w/o PRO-FA	0.4567	0.4401	0.4482	-0.0809
Baseline (DenseNet)	0.4234	0.4102	0.4167	-0.1124

Key findings:

- PRO-FA contributes most significantly (+8.09% F1)
- RCTA improves cross-modal reasoning (+5.76% F1)
- MIX-MLP enhances multi-label classification (+3.93% F1)
- All modules provide complementary benefits

5 Conclusion

5.1 Summary of Contributions

This work presents a comprehensive deep learning framework for automated radiology report generation, making the following key contributions:

1. **Hierarchical Visual Perception (PRO-FA)**: Implements a three-stage feature alignment mechanism that progressively refines visual understanding from pixels to organs, capturing both fine-grained anatomical details and high-level semantic context.
2. **Knowledge-Enhanced Classification (MIX-MLP)**: Introduces a multi-path classification architecture that integrates visual features, clinical knowledge, and external medical ontologies for robust multi-label disease detection.
3. **Triangular Cognitive Attention (RCTA)**: Develops a radiologist-inspired attention mechanism that models the iterative reasoning process through three-way cross-modal interactions between images, text, and diagnostic labels.
4. **Clinical Validation**: Achieves F1-score of 0.5291 on the IU X-Ray dataset, demonstrating practical applicability for clinical decision support.

5.2 Clinical Impact

The proposed system addresses critical needs in radiology workflows:

- **Workload Reduction**: Automates preliminary report drafting, allowing radiologists to focus on complex cases
- **Consistency**: Standardizes reporting format and reduces variability

- **Quality Assurance**: Serves as a second reader to catch potential oversights
- **Education**: Provides structured learning resources for radiology training

5.3 Limitations and Future Work

While the current system demonstrates promising results, several limitations warrant future investigation:

5.3.1 Current Limitations

1. **Dataset Scale**: Limited to 3,851 training samples; larger datasets could improve generalization
2. **Disease Coverage**: Focuses on 8 common pathologies; rare diseases not well-represented
3. **Report Generation**: Currently generates structured findings; free-text generation not implemented
4. **Multi-View Integration**: Frontal and lateral views processed independently

5.4 Broader Impact

The successful development of automated radiology report generation has implications beyond chest X-rays:

- **Modality Transfer**: Adaptation to CT, MRI, and ultrasound imaging
- **Specialty Extension**: Application to neuroradiology, musculoskeletal, and abdominal imaging
- **Global Health**: Deployment in resource-limited settings with radiologist shortages
- **Research Acceleration**: Large-scale retrospective studies using automated annotations

5.5 Ethical Considerations

Deployment of AI in clinical settings requires careful consideration:

- **Human Oversight**: AI serves as decision support, not replacement for radiologists
- **Transparency**: Model predictions must be interpretable and explainable
- **Bias Mitigation**: Ensure equitable performance across demographic groups
- **Data Privacy**: Strict adherence to HIPAA and patient confidentiality

5.6 Final Remarks

This work demonstrates the feasibility of building intelligent "Second Reader" systems for radiology report generation. By combining hierarchical visual perception, knowledge-enhanced classification, and cognitive attention mechanisms, we achieve clinically relevant performance on multi-label disease classification. The framework provides a foundation for future research in automated medical report generation and clinical decision support.

The integration of domain knowledge, attention mechanisms, and multi-modal reasoning represents a step toward AI systems that can truly augment radiologist capabilities. As datasets expand and models improve, such systems have the potential to transform radiology workflows and improve patient care globally.

References

[1] Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). *Densely connected convolutional networks*. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4700-4708).

[2] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). *BioBERT: a pre-trained biomedical language representation model for biomedical text mining*. Bioinformatics, 36(4), 1234-1240.

[3] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). *Attention is all you need*. In Advances in neural information processing systems (pp. 5998-6008).

[4] Demner-Fushman, D., Kohli, M. D., Rosenman, M. B., Shooshan, S. E., Rodriguez, L., Antani, S., ... & Thoma, G. R. (2016). *Preparing a collection of radiology examinations for distribution and retrieval*. Journal of the American Medical Informatics Association, 23(2), 304-310.

[5] Johnson, A. E., Pollard, T. J., Berkowitz, S. J., Greenbaum, N. R., Lungren, M. P., Deng, C. Y., ... & Horng, S. (2019). *MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports*. Scientific data, 6(1), 1-8.

[6] Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., ... & Ng, A. Y. (2019). *CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison*. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 33, pp. 590-597).