Multi-Label Chest X-Ray Classification Report

A. Methodology

This project implemented a comprehensive pipeline for multi-label classification of chest X-ray images using both deep learning and hybrid approaches. The methodology consisted of several key steps:

1. Data Preparation and Augmentation

The pipeline began with robust data preparation:

- Loaded and filtered metadata to match available images
- Used MultiLabelBinarizer to convert text labels to binary vectors
- Split data into training (80%), validation (10%), and test sets (10%)
- **Implemented extensive image augmentation techniques:**
  - A. Random rotations (±15 degrees)
  - B. Random shifts (up to 10% of image dimensions)
  - C. Random zoom (90-110% scaling)
  - D. Brightness and contrast adjustments
  - E. Gaussian noise addition
  - F. CLAHE histogram equalization

2. Feature Extraction

Two parallel approaches were implemented:

1. Deep Learning Features:

- Used ResNet50 as backbone with frozen initial layers
- Added GlobalAveragePooling and dense layers
- Output layer with sigmoid activation for multi-label classification

2. Traditional Computer Vision Features:

- Extracted ORB features (500 keypoints per image)
- Extracted SIFT features (500 keypoints per image)

- Visualized feature points for quality assessment

3. Model Architectures

Two primary models were developed:

Standard ResNet Model:

- ResNet50 base (pre-trained on ImageNet)
- Global average pooling layer
- 1024-unit dense layer with ReLU activation
- Dropout layer (0.5 rate)
- Output layer matching number of classes

Hybrid Model:

- Combined ResNet50 features with ORB features
- **Separate processing branches:**
  - Image branch: ResNet50 → GlobalAveragePooling → Dense(1024)
  - ORB branch: Dense(256) processing of ORB features
- Concatenated branches with additional dense layers
- Final output layer with sigmoid activation

4. Training Process

Both models were trained with:

- Adam optimizer (learning rate 0.0001)
- Binary cross-entropy loss
- Key metrics: Accuracy, AUC, Precision, Recall
- **Callbacks:**

A. - ModelCheckpoint (saving best model by val_auc)

B. - EarlyStopping (patience=10)

C. - ReduceLROnPlateau (factor=0.1, patience=3)

5. Evaluation

Comprehensive evaluation included:

- Classification reports (precision, recall, f1-score)
- ROC AUC scores (per-class and micro/macro averages)

- Confusion matrices for key classes
- Comparative performance analysis
- Grad-CAM visualizations for model interpretability

B. Model Performance Comparison

Quantitative Results

Standard ResNet Model:

- Micro-average ROC AUC: 0.8284
- Macro-average ROC AUC: 0.6049
- Best class performance (Edema): 0.9146 AUC
- Worst class performance (Fibrosis): 0.4975 AUC

Hybrid Model:

- Micro-average ROC AUC: 0.8167
- Macro-average ROC AUC: 0.5836
- Best class performance (Hernia): 0.8324 AUC
- Worst class performance (Edema): 0.3191 AUC

Class-wise Performance Highlights:

- The standard model performed better on 10/15 classes
- The hybrid model showed particular strength on Hernia (0.8324 vs 0.6047)
- Both models struggled with Fibrosis and Pneumonia detection
- The "No Finding" class had moderate performance in both models

Training Dynamics

- Both models showed similar training curves
- The hybrid model required more training time per epoch
- Neither model showed signs of severe overfitting
- Learning rate reduction was triggered for the hybrid model

Confusion Matrix Analysis

For the sample class "Atelectasis":

- Standard model achieved 344 true negatives
- Both models showed room for improvement in positive case detection
- High specificity but low sensitivity patterns observed

C. Insights and Conclusions

Key Findings

1. Feature Effectiveness:

- The traditional ORB features did not provide significant complementary information to the deep learning features in this application
- The hybrid model showed marginally worse performance overall
- Certain classes (like Hernia) may benefit from feature fusion approaches

2. Class Imbalance Challenges:

- Rare classes (Emphysema, Hernia, Pneumonia) showed poor performance
- The "No Finding" class dominated predictions due to its frequency
- Many classes showed 0 precision due to no positive predictions

3. Model Behavior:

- Both models tended to be conservative, prioritizing specificity
- The Grad-CAM visualizations showed reasonable focus on lung areas
- Some misclassifications appeared to focus on irrelevant image regions

Recommendations for Improvement

1. Data-Level Solutions:

- Implement more aggressive class balancing techniques
- Add additional augmentation specifically for rare classes
- Consider collecting more samples for under-represented conditions

2. Model Architecture:

- Experiment with different feature fusion approaches
- Try attention mechanisms to better focus on relevant regions
- Test other backbone architectures (DenseNet, EfficientNet)

3. Training Process:

- Implement class-weighted loss functions
- Add more sophisticated learning rate scheduling
- Experiment with different optimizer configurations

4. Evaluation:

- Develop more comprehensive visual explanation tools
- Create case-based analysis for error patterns
- Implement clinician-in-the-loop validation

Final Conclusion

While both models showed reasonable performance on this multi-label classification task, the standard ResNet approach outperformed the hybrid model in most metrics. The project demonstrated that:

1. Deep learning alone can achieve good performance on medical image classification
2. Traditional feature fusion requires careful implementation to provide benefits
3. Class imbalance remains a significant challenge in medical imaging
4. There is substantial room for improvement, particularly for rare conditions

Future work should focus on addressing the class imbalance issue and exploring more sophisticated architectures that can better handle the multi-label nature of the task while maintaining clinical relevance.