Multi-Label Medical Diagnosis Classification

Project Report Submitted

to the

SRM University-AP, Andhra Pradesh

for the partial fulfillment of the requirements to award the degree of

Bachelor of Technology

in

Computer Science & Engineering

School of Engineering & Sciences

submitted by

Kondeti Sai Lohitaksh(AP22110010480)

Under the Guidance of

Dr. Isnuri Bala Venkateswarlu

Department of Computer Science & Engineering SRM University-AP

Neerukonda, Mangalgiri, Guntur

Amaravati, Andhra Pradesh - 522 240

Dec 2025

Abstract

This project explores automated multi-label classification for medical diagnosis using chest X-ray imagery from the ChestMNIST dataset. The study implements and compares three problem transformation approaches—Binary Relevance, Classifier Chains, and Label Powerset—to predict the presence of 14 different thoracic diseases from low-resolution grayscale images. The dataset comprises 78,468 training samples and 22,433 test samples, with each image potentially associated with multiple disease labels. Binary Relevance achieved the best performance with an F1-Micro score of 0.1635, though all models struggled with the severe class imbalance inherent in medical datasets.

Introduction

Background

Medical image analysis represents a critical application area for machine learning, particularly in diagnosing conditions from chest radiographs. Traditional diagnostic workflows require expert radiologists to manually review images, a time-intensive process that can delay treatment. Automated classification systems hold promise for preliminary screening and triage, potentially improving healthcare accessibility in resource-limited settings.

Problem Statement

The ChestMNIST dataset presents a challenging multi-label classification problem where a single chest X-ray may exhibit multiple pathological findings simultaneously. Unlike traditional single-label classification, this scenario requires models capable of predicting independent probabilities for each of 14 disease categories: Atelectasis, Cardiomegaly, Effusion, Infiltration, Mass, Nodule, Pneumonia, Pneumothorax, Consolidation, Edema, Emphysema, Fibrosis, Pleural Thickening, and Hernia.

Objectives

This work aims to evaluate three multi-label learning strategies on the ChestMNIST dataset and identify the most effective approach for automated thoracic disease detection. Secondary goals include analyzing per-disease performance characteristics and understanding the impact of class imbalance on model effectiveness.

Dataset Description

Data Source

The ChestMNIST dataset was sourced from Zenodo (record 10519652) and contains preprocessed chest X-ray images standardized to 28×28 pixels in grayscale format. The dataset is partitioned into 78,468 training images, 11,219 validation images, and 22,433 test images.

Label Distribution

The dataset exhibits substantial class imbalance, a common characteristic of medical datasets. Analysis of the training set reveals significant variability in disease prevalence:

- Most frequent: Infiltration (13,914 instances)

- Moderately common: Effusion (9,261), Atelectasis (7,996)
- Rare conditions: Hernia (144 instances)

The average number of labels per training sample is approximately 0.74, indicating that most images show either no pathology or a single condition.

Label Co-occurrence

Examination of label co-occurrence patterns through heatmap visualization revealed that certain disease combinations appear more frequently than others. This inter-label dependency motivates the use of Classifier Chains, which can potentially capture these relationships.

Methodology

Data Preprocessing

Images were reshaped from 28×28 pixel arrays into 784-dimensional feature vectors using flattening operations. Pixel intensities were not explicitly normalized, as logistic regression models can handle the original intensity range reasonably well. To manage computational constraints, training was conducted on a stratified subsample of 40,000 images selected to preserve the original label distribution.

Multi-Label Learning Strategies

Three problem transformation approaches were implemented using the scikit-multilearn library:

Binary Relevance (BR): This baseline approach trains 14 independent binary classifiers, one per disease label. Each classifier uses logistic regression with L2 regularization (C=1.0, max_iter=500). Binary Relevance ignores label correlations but offers computational efficiency and interpretability.

Classifier Chains (CC): This method arranges classifiers in a sequential chain where each model receives predictions from preceding classifiers as additional features. The chain order follows the original label sequence. This architecture enables the model to capture label dependencies at the cost of increased complexity.

Label Powerset (LP): This approach treats each unique label combination as a distinct class, transforming the multi-label problem into multi-class classification. While theoretically capable of modeling all label interactions, Label Powerset suffers from exponential growth in the number of possible label combinations and severe data sparsity for rare combinations.

Base Classifier Configuration

All three methods employed logistic regression as the base classifier with the following hyperparameters:

- Regularization: L2 penalty (C=1.0)
- Solver: lbfgs
- Maximum iterations: 500
- Convergence tolerance: 1e-4

These settings were selected to balance model complexity with training efficiency on the large-scale dataset.

Threshold Optimization

Since multi-label classification requires converting probability outputs to binary predictions, threshold selection significantly impacts performance. Experiments were conducted with thresholds ranging from 0.2 to 0.6 to identify the value that best balances precision and recall. A threshold of 0.6 was ultimately selected, yielding an average of 2.47 predicted labels per sample compared to the ground truth average of 0.74. While this produces more predictions than the true distribution, it was found to optimize F1-score by improving recall.

Experimental Results

Overall Performance Comparison

Evaluation on the 22,433-sample test set using a 0.6 decision threshold produced the following results:

Model	Hamming Loss	Jaccard Index	F1-Micro	F1-Macro	Precision	Recall
Binary Relevance	0.1913	0.0579	0.1635	0.1345	0.1062	0.3557
Classifier Chain	0.326	0.058	0.1505	0.129	0.0872	0.5494
Label Powerset	0.0638	0.0171	0.0736	0.0419	0.1558	0.0482

Binary Relevance achieved the highest F1-Micro score (0.1635) and was identified as the best-performing model. Label Powerset showed the lowest Hamming loss but poor F1 scores, suggesting it makes fewer predictions overall but misses many true positives. Classifier Chains demonstrated the highest recall (0.5494) at the expense of precision.

Per-Disease Performance Analysis

Detailed analysis of Binary Relevance performance across individual disease categories revealed substantial variability:

Top Performers:

- Effusion: F1=0.342, Precision=0.262, Recall=0.492
- Infiltration: F1=0.287, Precision=0.280, Recall=0.296
- Atelectasis: F1=0.262, Precision=0.194, Recall=0.405

Poorest Performers:

- Hernia: F1=0.012, Precision=0.006, Recall=0.167
- Pneumonia: F1=0.035, Precision=0.019, Recall=0.306
- Fibrosis: F1=0.059, Precision=0.032, Recall=0.348

The strong correlation between training set size and F1-score suggests that rare diseases suffer from insufficient training examples. Hernia, with only 144 training instances, achieved the lowest performance.

Discussion

Model Performance Insights

The modest overall performance metrics reflect the inherent difficulty of the task. Several factors contribute to these challenges:

1. Low image resolution: 28×28 pixels may not preserve sufficient diagnostic detail
2. Severe class imbalance: Rare diseases have inadequate representation
3. Feature representation: Simple pixel intensity features lack the discriminative power of deep convolutional features

Binary Relevance's superior performance despite its simplicity suggests that the label independence assumption, while theoretically limiting, provides sufficient modeling capacity for this particular problem. The poor performance of Label Powerset indicates that the exponential growth in label combinations leads to extreme data fragmentation.

Clinical Relevance

The higher recall values (e.g., 0.492 for Effusion) suggest the model could potentially serve as a screening tool to flag cases requiring expert review. However, the low precision (e.g., 0.262 for Effusion) would result in numerous false alarms, potentially overwhelming radiologists. Further optimization would be necessary before clinical deployment.

Limitations

Several limitations constrain the generalizability of these findings:

- Training on only 40,000 of 78,468 available samples due to computational constraints
- Use of basic logistic regression rather than deep learning architectures
- Limited hyperparameter tuning
- Absence of external validation on independent datasets

Conclusion

This project successfully implemented and compared three multi-label classification approaches for automated chest X-ray diagnosis using the ChestMNIST dataset. Binary Relevance emerged as the most effective strategy, achieving an F1-Micro score of 0.1635. Performance varied substantially across disease categories, with common conditions like Effusion and Infiltration showing reasonable detection rates while rare diseases like Hernia remained challenging.

Future Work

Several directions could improve upon these baseline results:

1. Deep learning: Implementing convolutional neural networks to automatically learn hierarchical image features
2. Data augmentation: Generating synthetic training examples for rare disease categories
3. Ensemble methods: Combining predictions from multiple model architectures
4. Attention mechanisms: Focusing on relevant image regions for specific diagnoses

5. Full dataset training: Utilizing all available training samples with distributed computing resources

The findings demonstrate both the potential and current limitations of machine learning for automated medical image analysis.