Deep Learning Project Proposal
CSC 871 - Fall 2025

Fayeeza Shaikh Part's Info

Project Title: Automated Pneumonia Detection from Chest X-ray Images

---

DATASET INFORMATION

Dataset Name: Chest X-Ray Images (Pneumonia) **Source:** Kaggle - Guangzhou Women and Children's Medical Center **Total Images:** 5,856 chest X-ray images **Classes:** NORMAL (0) and PNEUMONIA (1) **Format:** JPEG grayscale images, various sizes

Final Data Split:

Split	Images	Purpose
Training	5,216	Model training
Validation	782	Hyperparameter tuning
Test	624	Final evaluation

Note: Original validation set had only 16 images - too small for proper validation. Created new validation set with 15% of training data (782 images).

WORK COMPLETED

Tasks Accomplished:

✓ Downloaded and organized dataset (5,856 images) ✓ Explored data and counted images per class ✓ Visualized sample X-rays from both classes ✓ Fixed validation set (increased from 16 to 782 images) ✓ Implemented data preprocessing pipeline ✓ Created PyTorch DataLoader with augmentation ✓ Tested data pipeline - confirmed working ✓ Created documentation for team (README)

Files Delivered to Team:

• data_module.py - Main data loading code with PyTorch DataLoaders • create_val_split.py - Script that created improved validation set • explore_data.py - Data exploration script • visualize_samples.py - Visualization script • README_DATA.md - Complete documentation and usage instructions • requirements.txt - Python dependencies list • NORMAL_samples.png -

Visual samples of normal X-rays • PNEUMONIA_samples.png - Visual samples of pneumonia X-rays

---

TECHNICAL IMPLEMENTATION

Data Preprocessing:

Image Transformations: • Resize: All images standardized to 224×224 pixels • Normalization: ImageNet mean and standard deviation • Color space: Converted to RGB (3 channels)

Data Augmentation (Training Only): • Random horizontal flip (probability = 0.5) • Random rotation (±10 degrees) • Color jitter (brightness ±20%, contrast ±20%)

Key Findings:

Class Imbalance Detected: • PNEUMONIA images: 3,875 (74.3%) • NORMAL images: 1,341 (25.7%) • Ratio: Approximately 3:1 (pneumonia:normal)

Recommendation: Use weighted loss function during training to address class imbalance and prevent model bias toward pneumonia class.

USAGE INSTRUCTIONS FOR TEAM

Quick Start Code:

```
from data_module import get_data_loaders

train_loader, val_loader, test_loader = get_data_loaders(
    batch_size=32,
    image_size=224
)

for images, labels in train_loader:
    # images: shape [32, 3, 224, 224]
    # labels: 0=NORMAL, 1=PNEUMONIA
    # Your training code here
    pass
```

Output Specifications:

• Batch shape: [batch_size, 3, 224, 224] • Label encoding: 0 = NORMAL, 1 = PNEUMONIA • Pixel value range: Normalized (approximately -2.12 to 2.64)

SUMMARY

Successfully prepared a production-ready data pipeline for the chest X-ray pneumonia detection project. The pipeline includes proper train/validation/test splits, data augmentation for training, and is fully documented for team use. All data loading code is tested and working correctly with PyTorch. Team can now proceed with model development and training.

Status: ✅ Data preparation phase complete **Ready for:** Model development (Person B) and experiments (Person C)