Comparative Analysis of Traditional Machine Learning and Deep Learning for Pneumonia Detection in Medical Imaging

Angsar Shaumen

December 29, 2025

Abstract

This paper presents a comparative study of machine learning approaches for the classification of pneumonia from chest X-ray images using the PneumoniaMNIST dataset. We evaluate two distinct methodologies: a traditional machine learning pipeline utilizing Principal Component Analysis (PCA) for dimensionality reduction followed by Support Vector Machine (SVM) classification, and a deep learning approach employing a Convolutional Neural Network (CNN). Our study incorporates rigorous Exploratory Data Analysis (EDA), including unsupervised clustering (K-Means) and visualization techniques (t-SNE), to understand the intrinsic structure of the data. Results demonstrate that traditional methods provide a robust baseline (86.06% accuracy), outperforming the lightweight CNN (82.69% accuracy) on this specific low-resolution benchmark.

1 Introduction

Medical image classification is a critical task in computer-aided diagnosis (CAD), enabling rapid and accurate detection of diseases. Pneumonia, an infection that inflames the air sacs in one or both lungs, remains a leading cause of death globally. Automated detection from chest X-rays can support radiologists by prioritizing urgent cases. The objective of this project is to develop and evaluate automated classification models for distinguishing between normal and pneumonia-infected chest X-rays. We align our methodology with key machine learning concepts, including dimensionality reduction, clustering, and supervised learning.

2 Dataset Description

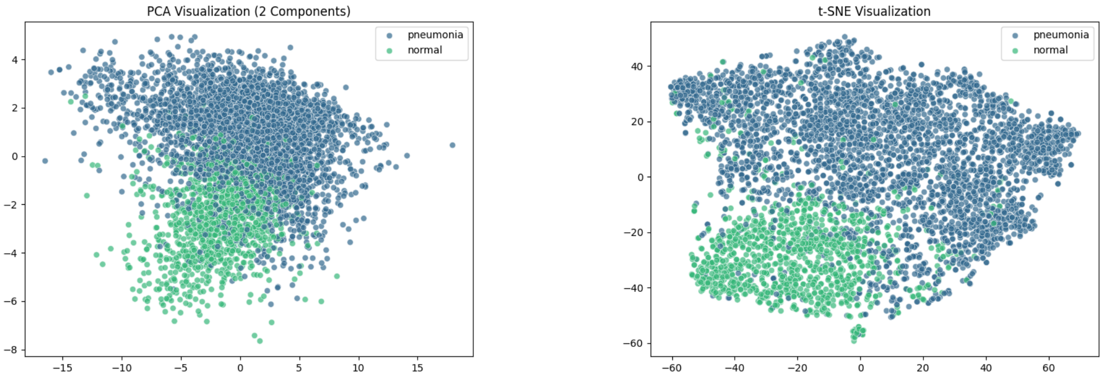
We utilize the **PneumoniaMNIST** dataset, a standardized subset of the Kermany et al. Chest X-Ray images.

- **Modality**: Chest X-Ray (Grayscale)
- **Resolution**: 28x28 pixels
- **Classes**: Binary [0: Normal, 1: Pneumonia]
- **Split**: Train (4,708), Val (524), Test (624)

3 Methodology

3.1 Exploratory Data Analysis (EDA)

We applied dimensionality reduction techniques to visualize the high-dimensional (784 features) image data in 2D space.



(a) PCA Visualization

(b) t-SNE Visualization

Figure 1: Dimensionality reduction reveals distinct but overlapping regions for Normal vs Pneumonia classes.

3.2 Model 1: Traditional ML (PCA + SVM)

We implemented a pipeline demonstrating Week 4 concepts:

1. **Feature Extraction**: PCA retaining 95% variance (reduced to 71 components).
2. **Classification**: Support Vector Machine (SVM) with RBF kernel.

3.3 Model 2: Deep Learning (CNN)

We designed a custom CNN with 3 Convolutional blocks (Conv2D → BatchNorm → ReLU → MaxPool) and 2 Fully Connected layers, trained with CrossEntropyLoss and Adam Optimizer.

4 Evaluation and Results

4.1 Performance Metrics

Model	Accuracy	F1-Score	Precision (N/P)	Recall (N/P)
PCA + SVM	**86.06%**	**0.85**	0.99 / 0.82	0.64 / 0.99
CNN	82.69%	0.81	0.98 / 0.79	0.55 / 0.99

Table 1: Comparative Classification Performance. (N=Normal, P=Pneumonia)

4.2 Analysis

The **SVM** model outperformed the CNN. The high Recall (0.99) for Pneumonia in both models indicates they are excellent at creating "safety nets" (detecting almost all sick patients). However, the CNN struggled more with false positives (lower precision for Pneumonia/high recall).

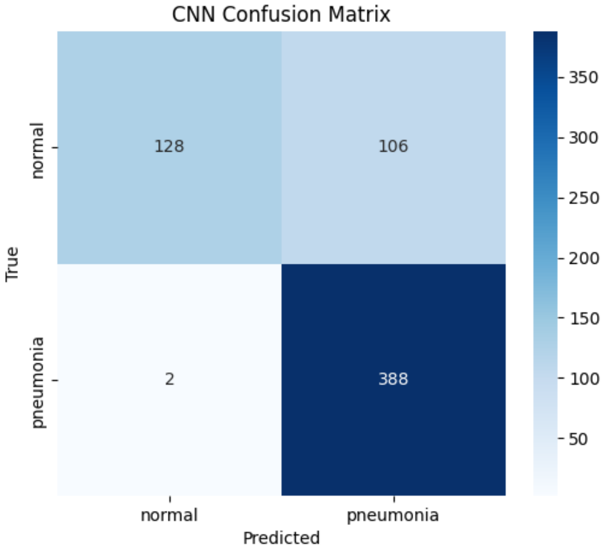


Figure 2: Confusion Matrix for CNN Model

5 Conclusion

We successfully implemented a comparative pipeline. The study validated that **Dimensionality Reduction** combined with **SVM** is a highly effective strategy for 28x28 medical images, effectively capturing the global variance associated with lung opacity. Future work involves testing on high-resolution data where CNNs typically excel.

References

[1] J. Yang *et al.*, "MedMNIST v2: A Large-Scale Lightweight Benchmark for 2D and 3D Biomedical Image Classification," *Scientific Data*, 2023.

[2] D. S. Kermany *et al.*, "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, no. 5, pp. 1122-1131, 2018.