Enhancing CNN Interpretability

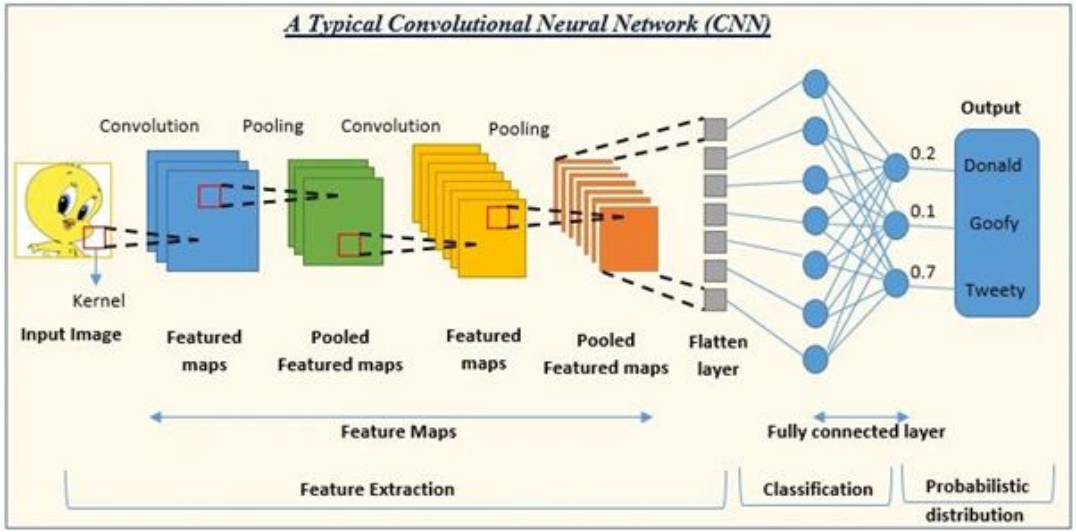
Using Multi-Layer Feature Extraction with Decision Trees

By: Kanishk Sivanandam and Chetan Maviti

Project Objective

CNNs are widely used in image classification, but their decision-making process is hard to interpret (black boxes)

Medical AI applications require interpretable results to ensure patient safety



Relevant Work/Info

1. Neural-Backed Decision Trees (NBDT) - ICLR 2021
    - Replaces CNN's fully connected layer with a hierarchical decision tree
    - Limitation: Uses only final-layer CNN features, disregarding lower/mid-level patterns
2. PCA for Dimensionality Reduction in CNNs
    - PCA is utilized for feature 'compression' while still preserving as much variance within the dataset as possible
    - PCA reduces the number of features, ultimately leading to faster training time and potentially less overfitting.
    - How is PCA relevant to our project?
        i. CNN features are extremely high-dimensional. PCA helps extract just the *most relevant* features through eigenvectors.
        ii. Leads to faster, more efficient decision tree creation

Dataset Overview

Public Dataset containing x-ray images for pneumonia classification

Dataset Size:

- 2186 instances
- 1562 Train, 624 Test (~70/30 Split)

Data is split into images for 2 classes:

- Normal or Pneumonia

➔ Dataset
- ◆ Train
  - Normal
  - Pneumonia
- ◆ Test
  - Normal
  - Pneumonia

Preprocessing

1. Image Resizing
    - Convert all images to 224 x 224 (required for ResNet50 input)
2. Color Conversion
    - Convert images from BGR (OpenCV) to RGB (PIL) for correct color representation
3. Normalization
    - Apply ImageNet mean and standard deviation for consistent feature scaling
        i. Mean: [0.485, 0.456, 0.406]
        ii. Std: [0.229, 0.224,0.225]
4. Label Encoding: (0 = Normal, 1 = Pneumonia)
5. Dataset Loading (32 images per batch for efficiency and shuffle data for training)

Methods (CNN+DT with Conv2, Conv4, Final)

1. Feature Extraction from CNN
    - X-Ray images are forward passed through the ResNet50 CNN model (pretrained on ImageNet)
    - Feature maps from Conv2, Conv4, and FC layers are extracted (output intercepted)
        i. Conv2 → Low layer (edges, corners, textures)
        ii. Conv4 → Mid layer (shapes, structures, regions)
        iii. FC (fully connected) → Deep layer (high level representations of xrays)
    - Extracted features are concatenated into a single high-dimensional feature vector
2. Decision Tree trained on high-dimensional feature vector and evaluated
3. Dimensionality Reduction using PCA
4. Decision Tree trained on PCA-transformed features and evaluated

Methods (CNN+DT Final)

1. Feature Extraction from CNN
    - X-Ray images are forward passed through the ResNet50 CNN model (pretrained on ImageNet)
    - Feature maps from FC layer is extracted (output intercepted)
        i. FC (fully connected) → Deep layer (high level representations of xrays)
    - Extracted features are concatenated into a single high-dimensional feature vector
2. Decision Tree trained on high-dimensional feature vector and evaluated
3. Dimensionality Reduction using PCA
4. Decision Tree trained on PCA-transformed features and evaluated

Methods (Standalone CNN)

1. Train ResNet50 on Pneumonia dataset
    - Replace the final fully connected layer with a 1-neuron ouptut + Sigmoid Activation for binary classification
    - Loss Function: Binary Cross-Entropy
    - Batch Size: 32
    - Training Time: 5 Epochs
2. Evaluate CNN performance with accuracy, precision and recall to compare with CNN+DT method

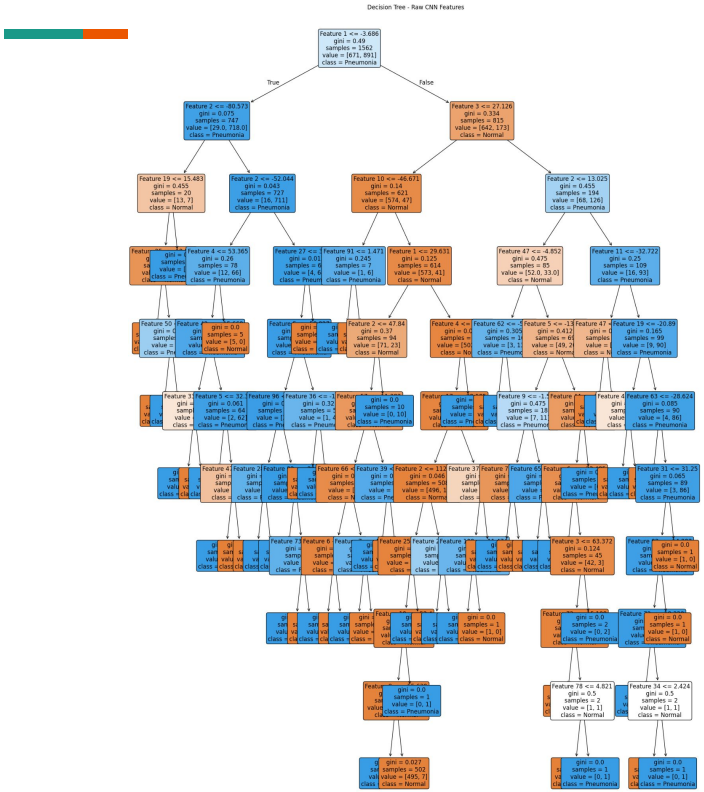
Results

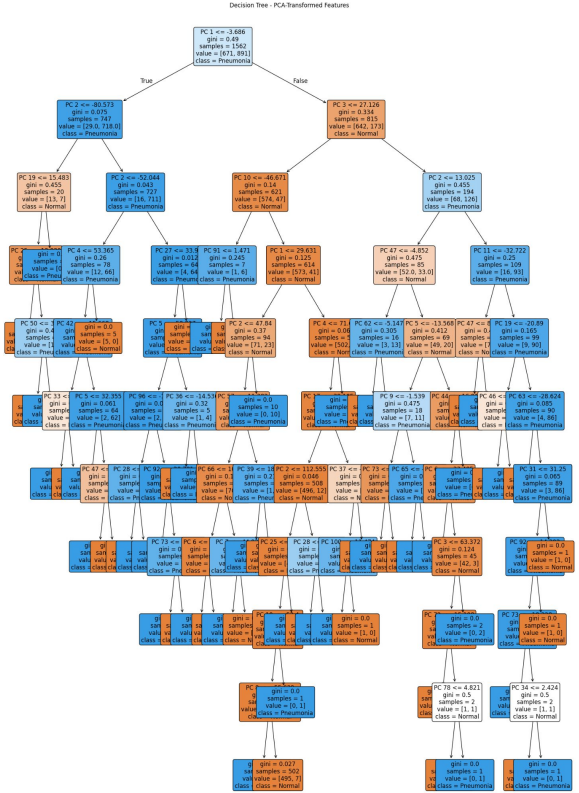
Model	Accuracy	Runtime	Precision	Recall
CNN Only (ResNet50)	83.81%	~2.5 hours	0.8	0.98
CNN → Decision Tree (Conv2, Conv4, Final)	79.65%	~11:10 (extract) ~8:25 (tree) = **19:36 mins**	0.77	0.95
CNN → Decision Tree (Final Only)	74.52%	~10:32 (extract) ~0:02 (tree) = **10:35 mins**	0.74	0.93

Results with PCA

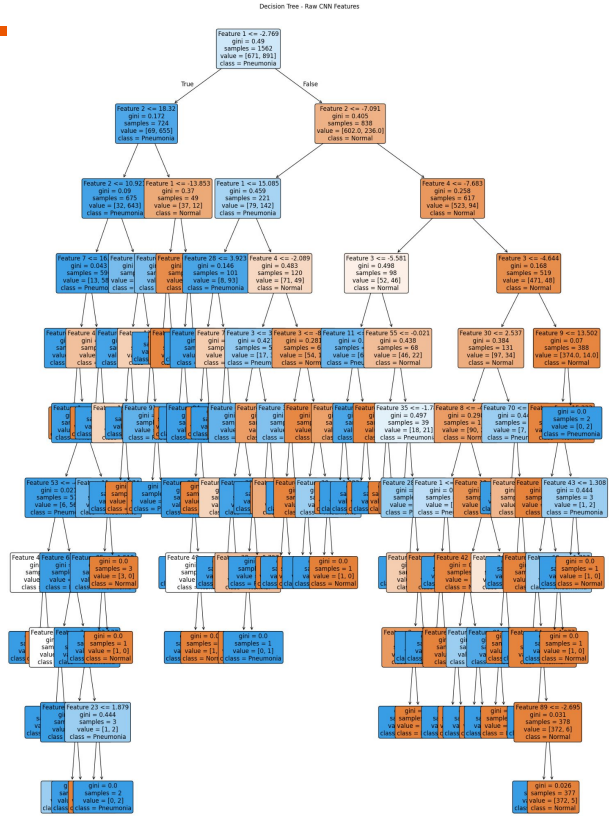
Model	Accuracy	Runtime	Precision	Recall
CNN → Decision Tree (Conv2, Conv4, Final)	79.65%	~11:10 (extract) ~8:25 (tree) = **19:36 mins**	0.77	0.95
+ PCA	79.17%	~11:10 (extract) ~1:43 (PCA+tree) = **12:54 mins**	0.79	0.92
CNN → Decision Tree (Final Only)	74.52%	~10:32 (extract) ~00:02 (tree) = **10:35 mins**	0.74	0.93
+ PCA	73.56%	~10:32 (extract) ~0:006 (PCA+tree) = **10:33 mins**	0.74	0.90

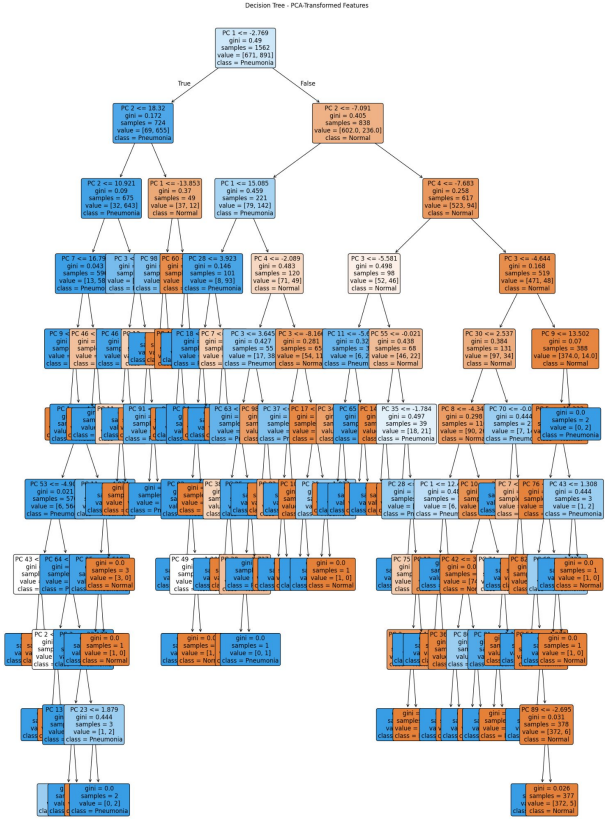
Interpretability (CNN+DT with Conv2, Conv4, Final)





Interpretability (CNN+DT with Final)





Conclusion

- CNN Only (ResNet50) had the highest accuracy (80.45%) BUT is a black-box model.
  - Best classification performance, but lacks transparency
- CNN + Decision Tree (Multi-Layer) retained strong accuracy while improving interpretability
  - Small trade-off in accuracy for better decision explainability
- CNN + Decision Tree (Final Layer Only) performed worse (74.52%)
  - Confirms that multi-layer features improve performance
- PCA slightly reduced accuracy but improved *decision tree* creation efficiency
  - PCA led to reduced runtimes to build Decision Trees because of reduced feature dimensionality

Precision - Recall for CNN + Decision Tree (Multi-Layer):

- Precision (0.79 for Pneumonia) → 79% of *predicted* pneumonia cases were correct.
- Recall (0.92 for Pneumonia) → 92% of *actual* pneumonia cases were detected.

Future Work

- Optimize PCA Components
  - Experiment with different numbers of features in PCA to find the best balance of interpretability vs. accuracy.
- Test with other classifiers
  - Random forests, for example, could prove to provide better accuracy at the cost of lowered interpretability. Would have been interesting to test out.
- Feature selection from CNN Layers
  - Currently using Conv2, Conv4, and FC (Final) Layer
  - Would testing Conv3, Conv5, or alternative layer combinations improve performance?
- Generalization to other datasets
  - Would this work for MRI scans, CT scans, etc.?
- Explainability Metrics
  - Apply SHAP or LIME to visualize the CNN and compare interpretability with Decision Tree