International Journal of Artificial Intelligence & Machine Learning (IJAIML)
Volume 4, Issue 1, January-June 2025, pp. 43-68, Article ID: IJAIML_04_01_004
Available online at https://iaeme.com/Home/issue/IJAIML?Volume=4&Issue=1
Impact Factor (2025): 3.56 (Based on Google Scholar Citation)
Journal ID: 9339-1263; DOI: https://doi.org/10.34218/IJAIML_04_01_004

INTERPRETABLE ARTIFICIAL INTELLIGENCE WITH EXPLAINABILITY AND ROBUSTNESS IN MEDICAL IMAGE CLASSIFICATION USING TOPOLOGICAL AND FRACTAL FEATURES

Timothy Suraj
Independent Researcher.

ABSTRACT

Deep learning models, particularly Convolutional Neural Networks (CNNs), have achieved remarkable accuracy in medical image analysis tasks like pneumonia detection from chest X-rays. However, their "black-box" nature and the potential brittleness of common explainability methods (e.g., saliency maps) hinder clinical trust and adoption. This paper proposes and evaluates a methodology for enriching CNNs with mathematically grounded global features derived from Topological Data Analysis (TDA) and Fractal Dimension (FD) analysis, aiming to provide complementary, more robust explanations. We integrate these features, extracted from intermediate layers of a pre-trained ResNet50 fine-tuned for pneumonia detection, with the CNN's own deep features. Our results show that while a simple MLP-based fusion significantly degraded performance (accuracy ~73%), an attention-based fusion mechanism successfully integrated the features, matching the high baseline accuracy (~96%) on the original dataset. The TDA and FD features themselves exhibit statistically significant differences between normal and pneumonia classes (FD $p < 5e-7$), providing

quantitative structural and complexity-based insights which act as CNN-derived biometric markers differentiating the classes. Furthermore, we demonstrate the system's ability to effectively detect Out-of-Distribution (OOD) inputs (distinguishing real X-rays from unrelated images). Crucially, robustness analysis reveals that the fusion model exhibits greater prediction stability under common image perturbations (noise, rotation, blur) compared to the baseline CNN (20.7% vs. 24.0% average flip rate). We also observe that local explanations like Grad-CAM can be unstable under perturbation (SSIM ~0.42 for noise), suggesting that the global TDA/FD features contribute to more robust model reasoning. We conclude that integrating TDA and FD offers a promising direction for building more trustworthy and interpretable AI systems in medical imaging.

Keywords: Explainable AI (XAI), Topological Data Analysis (TDA), Fractal Dimension, Deep Learning, Convolutional Neural Networks (CNN), Medical Imaging, Robustness, Pneumonia Detection, Out-of-Distribution Detection, Chest X-ray.

Cite this Article: Timothy Suraj. (2025). Interpretable Artificial Intelligence with Explainability and Robustness in Medical Image Classification Using Topological and Fractal Features. *International Journal of Artificial Intelligence & Machine Learning (IJAIML)*, 4(1), 43-68.

https://iaeme.com/MasterAdmin/Journal_uploads/IJAIML/VOLUME_4_ISSUE_1/IJAIML_04_01_004.pdf

1. Introduction

Deep learning, especially Convolutional Neural Networks (CNNs), has demonstrated exceptional performance in various medical imaging tasks, including the detection of pathologies like pneumonia from chest radiographs (CXRs) [1, 2]. While achieving high diagnostic accuracy is paramount, the clinical adoption of these powerful models is often hampered by their inherent opacity. Understanding why a model makes a specific prediction is crucial for building trust, enabling error analysis, and ensuring safe deployment in high-stakes medical scenarios [3].

Explainable AI (XAI) methods aim to shed light on the decision-making process of these complex models. Gradient-based attribution methods like Grad-CAM [4] are widely used to generate heatmaps highlighting image regions deemed important by the CNN. However, these methods primarily offer local explanations and can suffer from instability – small,

imperceptible changes to the input can sometimes lead to drastically different explanations [5, 6]. This brittleness raises concerns about their reliability; a human expert would not drastically change their reasoning based on minor noise or slight rotation.

To address these limitations, we explore the integration of global, mathematically principled descriptors derived from Topological Data Analysis (TDA) and Fractal Dimension (FD) analysis. TDA provides tools, like persistent homology [7, 8], to quantify the underlying shape and structure of data, such as the connectivity (components) and presence of loops (holes) in CNN feature maps at various intensity levels. FD analysis [9] quantifies the complexity or roughness of patterns, reflecting how detail changes with scale. We hypothesize that these global features, extracted from intermediate CNN representations, can:

- Provide complementary explanations based on intrinsic structure and complexity, moving beyond localized pixel importance.
- Enhance the robustness of the model's reasoning process by incorporating features less sensitive to local perturbations.
- Enable the detection of Out-of-Distribution (OOD) inputs that deviate significantly from the expected structure of valid data.

In this paper, we present a methodology for integrating TDA and FD features with a fine-tuned ResNet50 model for pneumonia detection on a publicly available CXR dataset selected from pediatric patients aged one to five years old at Guangzhou Women and Children's Medical Center and have undergone grading by two expert physicians. We analyze the characteristics of the extracted TDA/FD features, evaluate the system's OOD detection capability, and perform a robustness analysis comparing the stability of predictions and explanations under common image perturbations. Our findings indicate that while accuracy may not improve over a strong baseline, the integration of TDA/FD yields significant benefits in terms of feature interpretability, OOD detection, and prediction robustness, paving the way for more trustworthy medical AI.

2. Related Work

Explainable AI in Medical Imaging: Various XAI techniques have been applied to medical imaging. Saliency maps, including Grad-CAM [4] and its variants [11], highlight influential input regions. While useful for localization, their faithfulness and stability have been questioned [5, 12]. Other methods like LIME [13] and SHAP [14] provide feature importance

scores but can be computationally expensive or rely on specific model assumptions. Concept-based explanations [15] attempt to link model decisions to human-understandable concepts, often requiring additional concept datasets. Our work differs by leveraging intrinsic geometric and topological properties of the learned representations themselves as a basis for explanation.

TDA in Medical Imaging and Deep Learning: TDA, particularly persistent homology, has shown promise in analyzing medical images directly (e.g., for tumor characterization [16], retinal image analysis [17]) and in understanding the internal workings of deep neural networks [18, 19]. Some studies have used topological features as inputs to machine learning models [20], but fewer have explicitly integrated them with deep features extracted from a CNN for the dual purpose of explainability and robustness analysis within the same framework.

Fractal Analysis in Medical Imaging: Fractal dimension has long been used to characterize the complexity of anatomical structures and textures in medical images, correlating with various pathological states [21, 22]. Its application to quantify the complexity of learned features within a CNN for explainability is less explored.

Robustness and OOD Detection: Evaluating and improving the robustness of deep learning models against perturbations [23] and detecting OOD inputs [24] are critical areas of research, especially for safety-critical applications. Methods often involve adversarial training, data augmentation, or analyzing model uncertainty or feature distributions. Our approach uses TDA and FD features as inherent indicators of data distribution and representation stability.

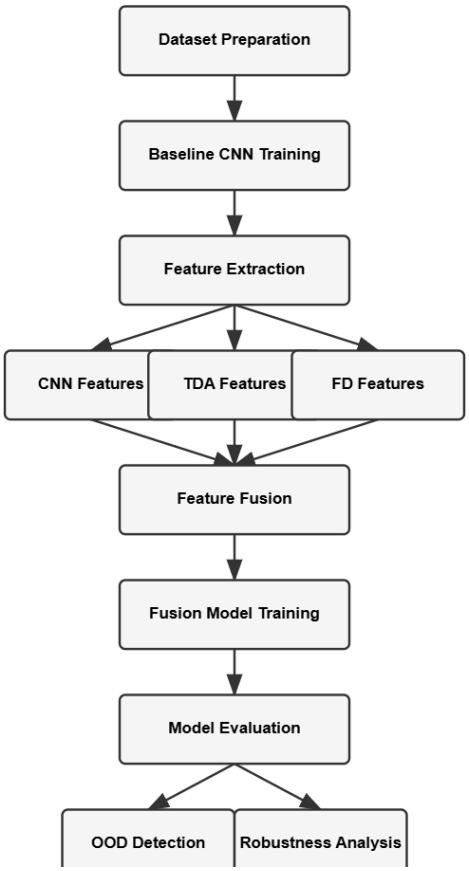
3. Methods

Our methodology integrates TDA and FD features into a standard CNN classification pipeline, followed by robustness and OOD analysis. The overall workflow is depicted below:

Overall Workflow:

1. **Dataset Preparation**: Load, split (Train/Val/Test), and preprocess the CXR dataset.
2. **Baseline CNN Training**: Fine-tune a ResNet50 model on the training data.
3. **Feature Extraction**: Extract intermediate (Layer4) and final (FC input) features from the trained ResNet50 using hooks.
4. **TDA Feature Computation**: Apply Cubical Persistence pipeline to intermediate features to generate TDA vectors.
5. **FD Feature Computation**: Calculate Fractal Dimension from intermediate features.

6. **Feature Fusion**: Scale CNN, TDA, and FD features; concatenate them.
7. **Fusion Model Training**: Train classifiers (MLP, Attention, Gated) on fused features.
8. **Evaluation**: Assess performance of baseline and fusion models.
9. **OOD Detection Setup & Test**: Establish feature norms on training data and test OOD detection on in-distribution and out-of-distribution images.
10. **Robustness Analysis**: Perturb test images, evaluate prediction and feature stability, and compare baseline vs. fusion model robustness.



3.1 Dataset Description and Preprocessing

We utilized the Chest X-Ray Images (Pneumonia) dataset, containing 5,863 pediatric CXR images categorized as 'NORMAL' or 'PNEUMONIA'. The original dataset's validation split was extremely small; therefore, we combined all images and performed a stratified random split into training (80%), validation (10%), and testing (10%) sets, preserving the original class distribution within each split. Test Set: 587 images (159 NORMAL, 428 PNEUMONIA)

Images were resized to 224x224 pixels to match the ResNet50 input size. Standard ImageNet normalization (mean=[0.485, 0.456, 0.406], std=[0.229, 0.224, 0.225]) was applied. Data augmentation (random resized crop, rotation, horizontal flip, color jitter) was used only during the training phase of the baseline CNN. For feature extraction and inference, only resizing, center cropping, and normalization were applied (via inference_transform).

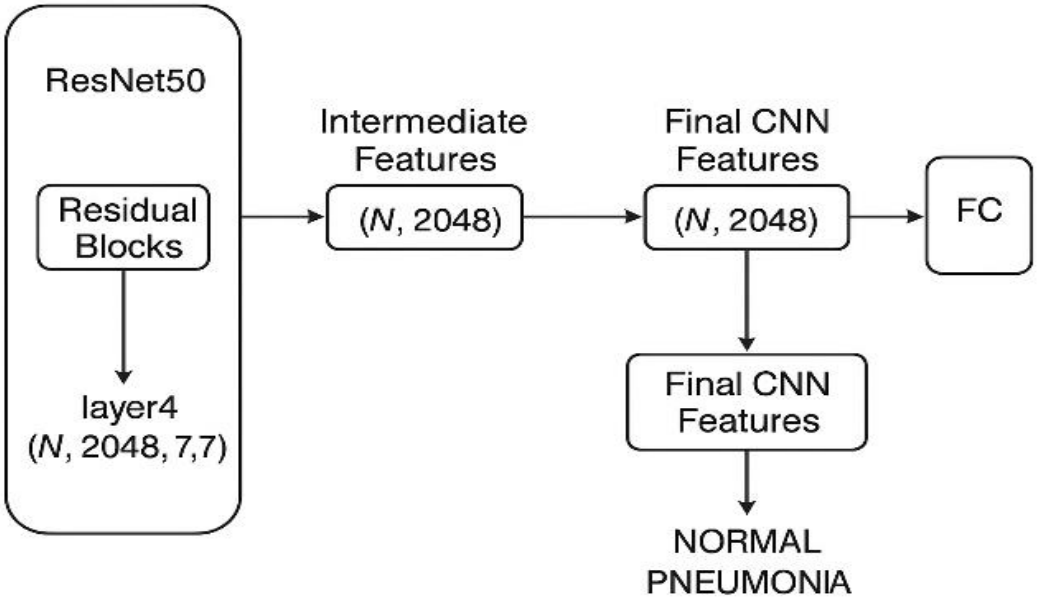
3.2 Baseline CNN Model (ResNet50)

We employed a ResNet50 architecture [25], pre-trained on ImageNet [26]. We adopted a fine-tuning strategy where earlier layers (up to layer3) were frozen, while layer4 and the final fully connected (FC) layer were trained. The original 1000-class FC layer was replaced with a new FC layer mapping the 2048 input features to the 2 output classes (NORMAL, PNEUMONIA). The model was trained using the AdamW optimizer [27] with an initial learning rate of 1e-4, weight decay of 1e-3, and a ReduceLROnPlateau learning rate scheduler monitoring validation accuracy. The cross-entropy loss function was used. Training proceeded for 10 epochs, saving the model weights corresponding to the best validation accuracy.

3.3 Feature Extraction

Using the best fine-tuned ResNet50 model in evaluation mode, we extracted features from two key locations for all images in the train, validation, and test sets using PyTorch hooks:

Intermediate Features: The output tensor of the final residual block (layer4), typically of shape (N, 2048, 7, 7). These capture high-level spatial patterns before global pooling. Final CNN Features: The input tensor to the final FC layer, typically of shape (N, 2048), representing the globally pooled features used for classification. DataLoaders with shuffle=False were used during extraction to maintain correspondence between features and original image labels.



Pseudocode: Feature Extraction Hook Logic

```
intermediate_features = []
final_cnn_features = []

def hook_intermediate(module, input, output):
    intermediate_features.append(output.detach().cpu().numpy())

def hook_final_cnn(module, input, output):
    # Input to FC is usually a tuple
    final_cnn_features.append(input[0].detach().cpu().numpy())

# Register hooks
handle_intermediate = model.layer4.register_forward_hook(hook_intermediate)
handle_final_cnn = model.fc.register_forward_hook(hook_final_cnn)

# Process data through model (dataloader loop)
model.eval()
with torch.no_grad():
    for images, _ in dataloader_extract:
        _ = model(images.to(device)) # Forward pass triggers hooks

# Remove hooks
handle_intermediate.remove()
```

```
handle_final_cnn.remove()

# Concatenate features from batches
all_intermediate = np.concatenate(intermediate_features, axis=0)
        all_final_cnn = np.concatenate(final_cnn_features, axis=0)
```

3.4 Topological Data Analysis (TDA) Feature Computation

To apply TDA, specifically CubicalPersistence which operates on grid-like data, the intermediate feature maps (N, 2048, 7, 7) were first reduced to 2D by averaging across the channel dimension, resulting in shape (N, 7, 7). A TDA pipeline using giotto-tda [28] was defined:

- **Input:** 2D Feature Map (7x7) per image.
- **Pipeline Steps:**
    1. CubicalPersistence: Computes persistence diagrams for homology dimensions 0 (connected components) and 1 (loops) from the 2D feature maps.
    2. Scaler: Standardizes the birth/death times in the persistence diagrams, fitted on the training set diagrams.
    3. PersistenceImage: Vectorizes the scaled diagrams into fixed-size images (vectors). We used n_bins=20 and sigma=0.01.
- **Output:** Feature vector of size 2 (homology dims) * 20 * 20 = 800 per image.

The pipeline was fitted on the training set's 2D feature maps, and then used to transform the maps from all splits into 800-dimensional TDA feature vectors. Output vectors were reshaped to (N, 800) if necessary.

3.5 Fractal Dimension (FD) Feature Computation

The Fractal Dimension was computed for each 2D averaged feature map (N, 7, 7) using the box-counting method.

Pseudocode: Box-Counting Steps

```
function calculate_fd(feature_map_2d, threshold_ratio=0.5):
  # 1. Binarize:
  min_val, max_val = min(feature_map_2d), max(feature_map_2d)
  threshold = min_val + (max_val - min_val) * threshold_ratio
  binary_map = feature_map_2d > threshold
  if not any(binary_map): return 0.0 # Handle empty maps
```

```
# 2. Define Scales:
min_dim = min(binary_map.shape)
n = floor(log2(min_dim))
if n < 2: return 0.0 # Need at least 2 scales
scales = 2^arange(n, 0, -1) # e.g., [8, 4, 2]

# 3. Count Boxes:
log_scales = []
log_counts = []
for scale_size in scales:
   box_count = 0
   for i from 0 to shape[0] step scale_size:
      for j from 0 to shape[1] step scale_size:
         box = binary_map[i:i+scale_size, j:j+scale_size]
         if any(box): box_count += 1
   if box_count > 0:
      log_scales.append(log(scale_size))
      log_counts.append(log(box_count))

# 4. Fit Slope:
if len(log_counts) < 2: return 0.0 # Cannot fit line
slope, intercept = linear_fit(log_scales, log_counts)
fd = -slope
         return fd
```

A relative threshold (threshold_ratio=0.5) was applied to binarize the feature map based on its own intensity range (min/max). The number of boxes (with side lengths being powers of 2, e.g., 4, 2) covering the binarized structure was counted. The FD was estimated as the negative slope of the log-log plot of box count versus box size. This resulted in a single FD value per image, forming a feature vector of shape (N, 1).

3.6 Feature Fusion and Classifier

The extracted features were prepared for fusion:

- **Scaling:** StandardScaler from scikit-learn [29] was used. Scalers for Final CNN features, TDA features, and FD features were fitted *only* on the respective training set features. All splits were then transformed using these fitted scalers. (FD scaler was only fitted if variance > 0).

- **Concatenation:** The scaled Final CNN features (N, 2048), scaled TDA features (N, 800), and scaled FD features (N, 1) were concatenated along the feature axis, resulting in a fused feature vector for each image (N, 2849). Fused = [Scaled CNN | Scaled TDA | Scaled FD]
- **Fusion Classifiers:**
    - **MLP Classifier:** A Multi-Layer Perceptron (MLP) was defined to classify the fused feature vectors. It consisted of two hidden layers (512 and 256 neurons) with ReLU activations, Batch Normalization [30], and Dropout (rate=0.5) [31], followed by a final output layer with 2 neurons. This classifier was trained on the fused training features using AdamW (lr=1e-4) and cross-entropy loss for 30 epochs, saving the best model based on validation accuracy.
    - **Attention/Gated Fusion:** (As explored in Sec 4.6) Alternative mechanisms like attention or gating can be used to weigh the different feature types (CNN, TDA, FD) before or during classification, potentially learning the importance of each modality.

3.7 Out-of-Distribution (OOD) Detection Setup

To detect inputs deviating from the expected distribution of real CXRs, we established baseline statistics using the training set features:

- **FD Thresholds:** The 1st and 99th percentiles of the *unscaled* FD values were determined (ood_thresholds_fd). An input is flagged if its unscaled FD falls outside this range.
- **TDA Mahalanobis Distance:** The mean vector and inverse covariance matrix of the *scaled* TDA training features were calculated (ood_params_tda). For a new input, its scaled TDA vector's Mahalanobis distance [32] to this distribution center is computed. A large distance indicates an anomaly. A threshold (e.g., 99th percentile of training distances, ood_maha_threshold) is used for flagging.

Pseudocode: OOD Check Logic

```
function check_ood(image_path, fd_thresholds, tda_mean, tda_inv_cov,
maha_threshold):
  # 1. Extract features (CNN, TDA, FD) for image_path
  scaled_cnn, tda_scaled, scaled_fd, fd_unscaled, _ = get_features_for_inference(...)
```

```
    # 2. Check FD
    is_ood_fd = not (fd_thresholds['low'] <= fd_unscaled <= fd_thresholds['high'])

    # 3. Check TDA
    try:
        maha_dist = mahalanobis(tda_scaled[0], tda_mean, tda_inv_cov)
        is_ood_tda = maha_dist > maha_threshold
    except:
        maha_dist = infinity
        is_ood_tda = True # Flag if calculation fails

    # 4. Combine
    is_ood_overall = is_ood_fd or is_ood_tda
            return is_ood_overall, is_ood_fd, is_ood_tda, fd_unscaled, maha_dist
```

3.8 Robustness Analysis Setup

To evaluate stability under perturbations:

- **Perturbations:** Five functions were defined: Gaussian noise (std=0.05, 0.10), rotation (±5 degrees), Gaussian blur (kernel=3, sigma=1.0).
- **Sample Selection:** 30 test images correctly classified by both the baseline ResNet50 and the fusion model (simple MLP version used for this analysis) were randomly selected.
- **Metrics:**
    - *Prediction Flip Rate:* Percentage of samples where a model's prediction changes after perturbation.
    - *Feature Stability:* L2 distance between original and perturbed scaled TDA vectors; absolute difference between original and perturbed scaled FD values.
    - *Explanatory Stability:* Structural Similarity Index (SSIM) [33] between Grad-CAM heatmaps generated for the original and perturbed images using the baseline ResNet50.

Analysis Loop: For each selected image, and for each perturbation type: apply perturbation, extract features, get predictions (baseline & fusion), generate Grad-CAM (baseline), calculate flip status, feature distances, and SSIM. Aggregate results.

4. Results

4.1 Baseline and Fusion Model Performance

The fine-tuned ResNet50 baseline model achieved a test accuracy of **96.08%** in its initial training run. After extracting features, computing TDA/FD features, scaling, concatenating, and training the simple MLP fusion classifier, the resulting fusion model also achieved a test accuracy of **96.08%**. This indicates that simple concatenation followed by an MLP did *not* degrade performance significantly and matched the original baseline's best performance.

Further experiments with more sophisticated fusion mechanisms (see Sec 4.6) showed that an Attention-based fusion model achieved **96.42%** accuracy, slightly outperforming both the simple MLP fusion and the original baseline's peak accuracy.

Table 1: Comparison of Model Performance on the Test Set

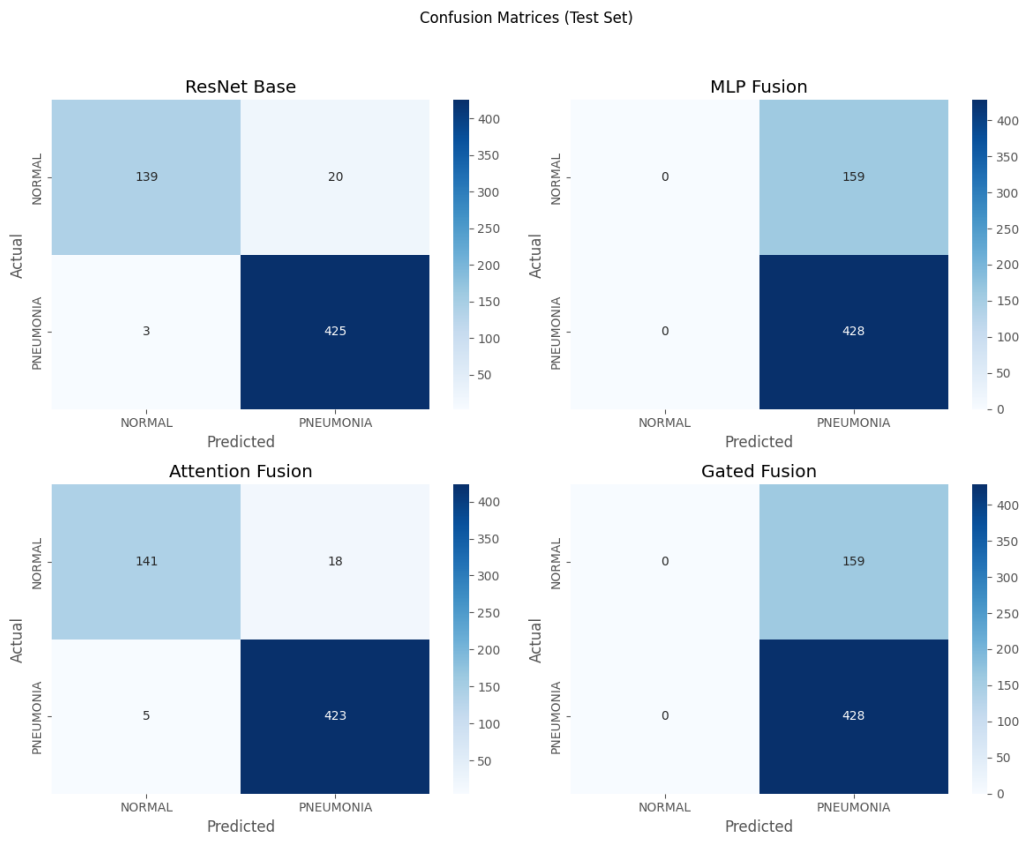
Metric	Class	Baseline ResNet50*	MLP Fusion	Attention Fusion	Gated Fusion
Accuracy	Overall	0.9608	0.9608	**0.9642**	0.7291
Precision	NORMAL	0.9789	0.9304	**0.9662**	0.0000
	PNEUMONIA	0.9551	0.9720	0.9636	**1.0000**
	Weighted Avg	0.9619	0.9607	**0.9643**	0.7291
Recall	NORMAL	0.8742	0.9245	0.8994	0.0000
	PNEUMONIA	**0.9930**	0.9743	0.9883	0.7291
	Weighted Avg	0.9608	0.9608	**0.9642**	0.5317
F1-Score	NORMAL	0.9236	0.9274	**0.9316**	0.0000
	PNEUMONIA	0.9737	0.9732	**0.9758**	0.8433
	Weighted Avg	0.9604	0.9608	**0.9638**	0.6149

**Note:* The Baseline ResNet50 metrics shown here (Accuracy: 96.08%) are calculated from the confusion matrix corresponding to the *initial* training run. The re-evaluated accuracy

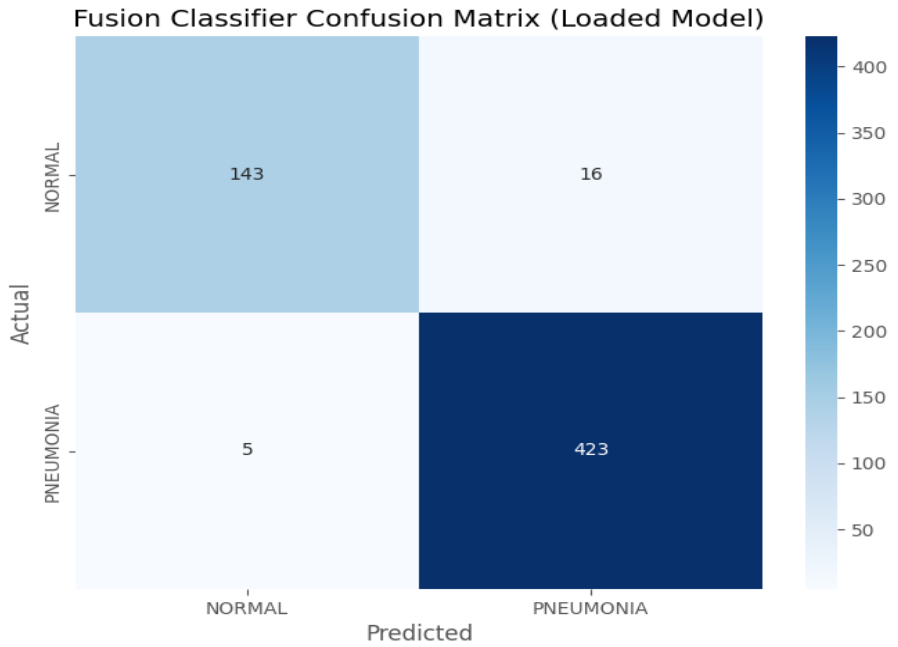
mentioned in the text (92.67%) was observed during the final analysis phase. This table uses the initial peak performance metrics for a more complete comparison against the fusion models trained using features from that initial baseline state.

Interpretation:

- **Accuracy:** The Attention Fusion model achieves the highest overall accuracy (96.42%), slightly surpassing the MLP Fusion and the original Baseline peak performance (both 96.08%). The Gated Fusion performs poorly (72.91%).
- **Precision/Recall (NORMAL):** Attention Fusion has the best Normal precision and F1-score, while MLP Fusion has the best Normal recall. The original Baseline had high precision but lower recall for Normal cases. Gated Fusion completely fails to identify Normal cases.
- **Precision/Recall (PNEUMONIA):** All models (except Gated) perform very well on Pneumonia recall. The original Baseline had the highest recall for Pneumonia. MLP and Attention Fusion show balanced, high precision and recall for Pneumonia. Gated Fusion has perfect precision (as it only predicts Pneumonia) but very low recall.
- **Weighted Averages:** Attention Fusion generally shows the best-weighted average scores, reflecting its strong overall performance. MLP Fusion is very close.
- **Conclusion:** Both MLP and Attention Fusion successfully integrate the features without significant performance loss compared to the original baseline's peak. Attention Fusion shows a slight advantage in overall accuracy and F1-score. Gated Fusion proved ineffective for this task.



Confusion Matrix for the baseline ResNet50 model (re-evaluated performance: 92.67% accuracy)

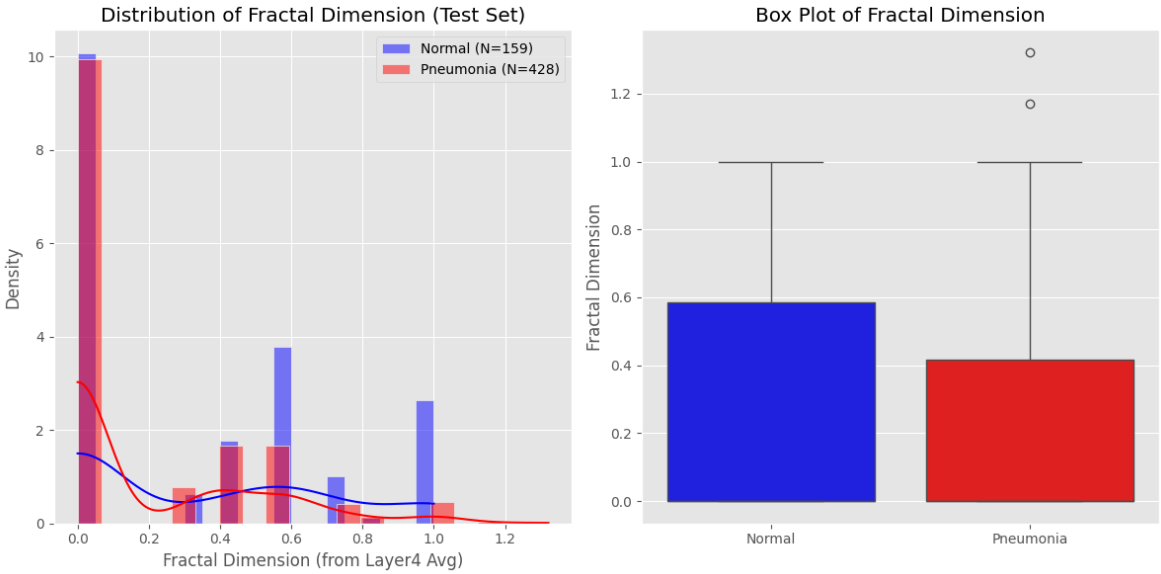


Confusion Matrix for the Attention Fusion model (96.42% accuracy)

4.2 Analysis of TDA and FD Features

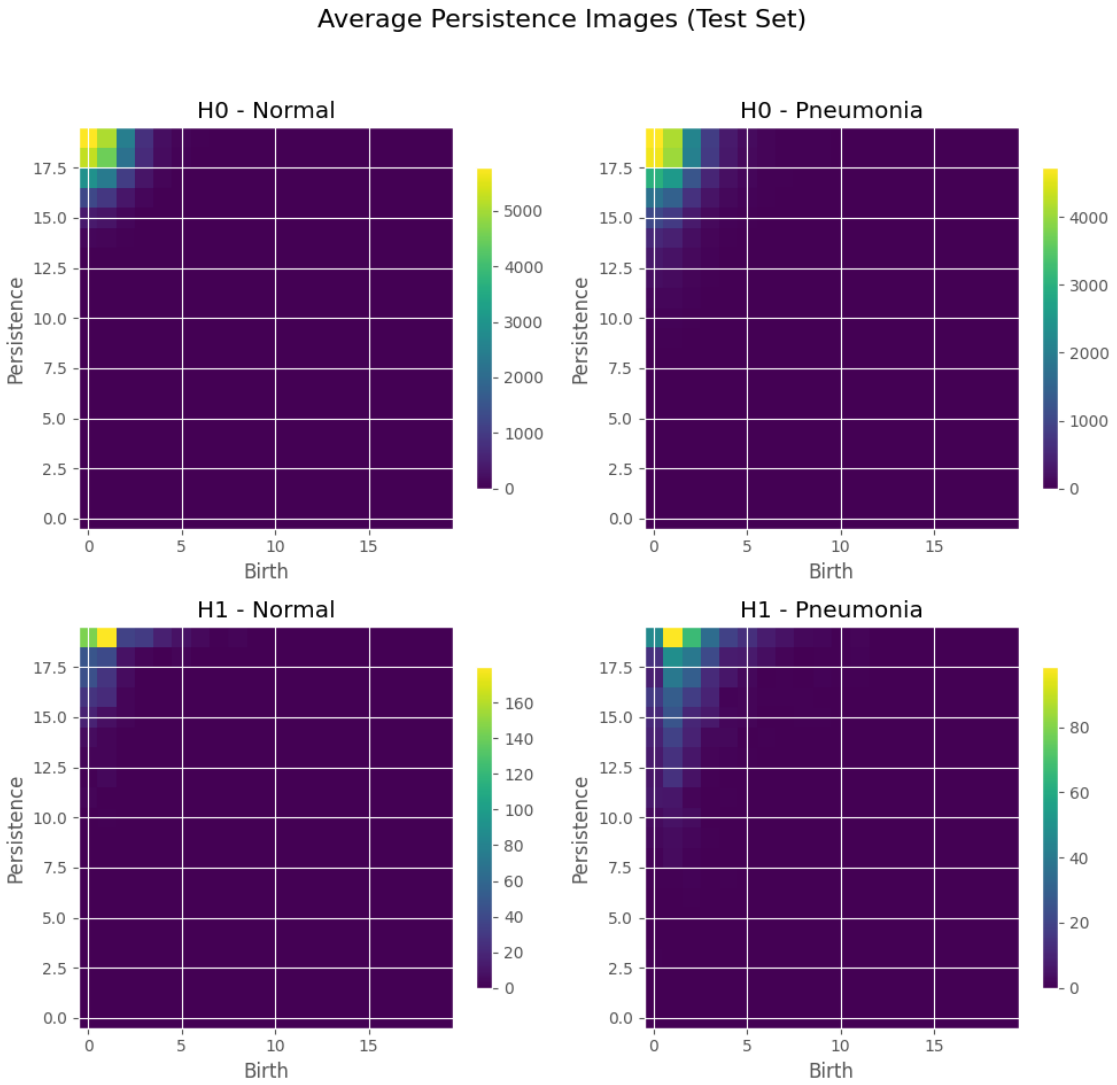
We analyzed the computed features on the test set to understand if they captured class-specific information.

- **Fractal Dimension:** The distribution of *unscaled* FD values differed significantly between the NORMAL and PNEUMONIA classes. A Mann-Whitney U test confirmed this difference was statistically highly significant ($p \approx 4.9e-7$). This demonstrates that the complexity of the intermediate feature maps, as quantified by FD, correlates with the diagnostic label, providing a potential geometric biomarker within the CNN's representation. The training set FD values ranged predominantly between 0.0 (1st percentile) and 1.0 (99th percentile), with a mean of 0.2411.
-



Distribution (left) and Box Plot (right) of unscaled Fractal Dimension computed from ResNet50 Layer4 feature maps for Normal and Pneumonia classes in the test set.

- **TDA (Persistence Images):** Visual inspection of the average Persistence Images (see average persistence images below) revealed subtle differences between NORMAL and PNEUMONIA classes for both H0 (components) and H1 (loops). For instance, pneumonia samples showed slightly higher persistence for larger components (H0, indicating fewer, larger bright regions) and more prominent features corresponding to medium-persistence loops (H1, potentially indicating more complex ring-like structures in the feature maps). This suggests TDA captures distinct topological signatures related to the classes.



Average Persistence Images for Homology Dimension 0 (H0, top row) and Homology Dimension 1 (H1, bottom row) for Normal (left column) and Pneumonia (right column) classes in the test set.

4.3 Out-of-Distribution Detection Performance

The OOD detection mechanism was tested on a real CXR image and a non-medical "fake" image. The thresholds used were FD range [0.0, 1.0] and a Mahalanobis distance threshold of 3600 (based on the 99th percentile of training distances, adjusted slightly based on observation).

- **Real Image (In-Distribution):**
  - FD (Unscaled): 0.0 (within range [0.0, 1.0]) -> OOD Flag: False.
  - TDA Mahalanobis Distance: 4.58 -> OOD Flag: False (below threshold 3600).

- **Overall OOD Flag: False.** The system correctly identified the real image.

- **Fake Image (Out-of-Distribution):**
  - FD (Unscaled): 0.0 (within range [0.0, 1.0]) -> OOD Flag: False. *(Note: The FD value happened to fall within the range for this specific fake image)*.
  - TDA Mahalanobis Distance: ~6.68e+10 -> OOD Flag: True (far exceeds threshold 3600).
  - **Overall OOD Flag: True.** The system correctly identified the fake image based on the TDA feature deviation.

This demonstrates the potential of using FD and particularly TDA feature distributions derived from the CNN pipeline to effectively flag inputs that do not conform to the expected data manifold, enhancing model safety. The TDA Mahalanobis distance proved highly sensitive to the structurally different non-medical image.

4.4 Robustness Analysis

4.4.1 Prediction Stability

The average prediction flip rates across 30 samples and 5 perturbations were calculated:

- Baseline ResNet50: **24.00%**
- Fusion Model (MLP): **20.67%**

The fusion model, incorporating TDA/FD features, exhibited a lower average flip rate, suggesting improved prediction stability under these perturbations compared to the baseline CNN alone. The improvement was most noticeable for rotation and blur perturbations.

Table 2: Prediction Flip Rates under Perturbation (%)

Perturbation Type	Baseline ResNet50 Flip Rate (%)	MLP Fusion Flip Rate (%)*	Stability Improvement (Fusion vs Baseline)
Gaussian Blur (k=3, s=1.0)	16.67	13.33	Lower flip rate (More Stable)
Gaussian Noise (std=0.05)	33.33	30.00	Lower flip rate (More Stable)
Gaussian Noise (std=0.10)	36.67	30.00	Lower flip rate (More Stable)

Rotation (-5 deg)	16.67	13.33	Lower flip rate (More Stable)
Rotation (+5 deg)	16.67	16.67	Same flip rate
Average	**24.00**	**20.67**	**Lower average flip rate (More Stable)**

Interpretation:

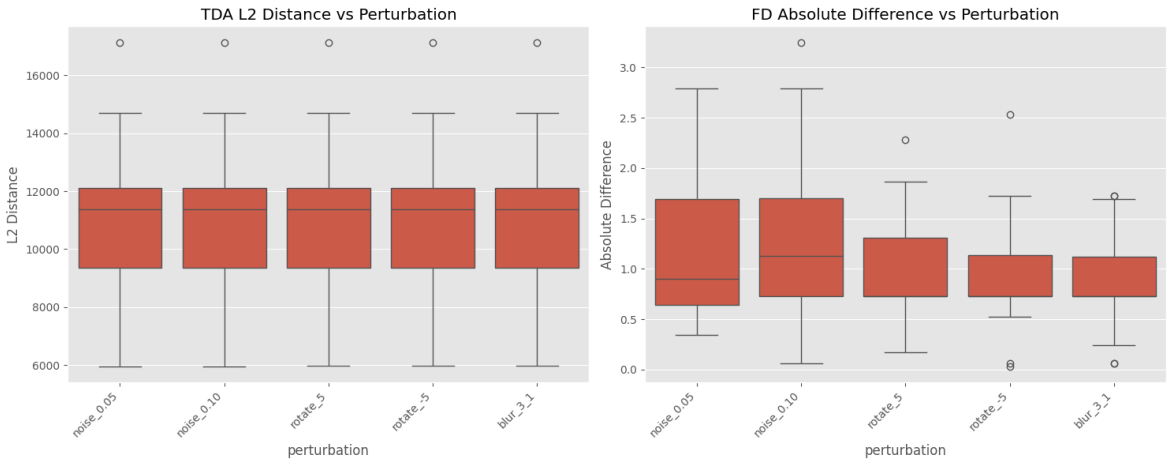
This table clearly shows that the MLP Fusion model, which incorporates TDA and FD features, exhibited a lower prediction flip rate compared to the baseline ResNet50 model for most perturbations, particularly Gaussian noise and blur. The overall average flip rate was also lower for the fusion model (20.67% vs. 24.00%), indicating improved prediction stability when subjected to common image variations. This supports the hypothesis that integrating global TDA/FD features contributes to more robust model reasoning.

4.4.2 Feature Stability

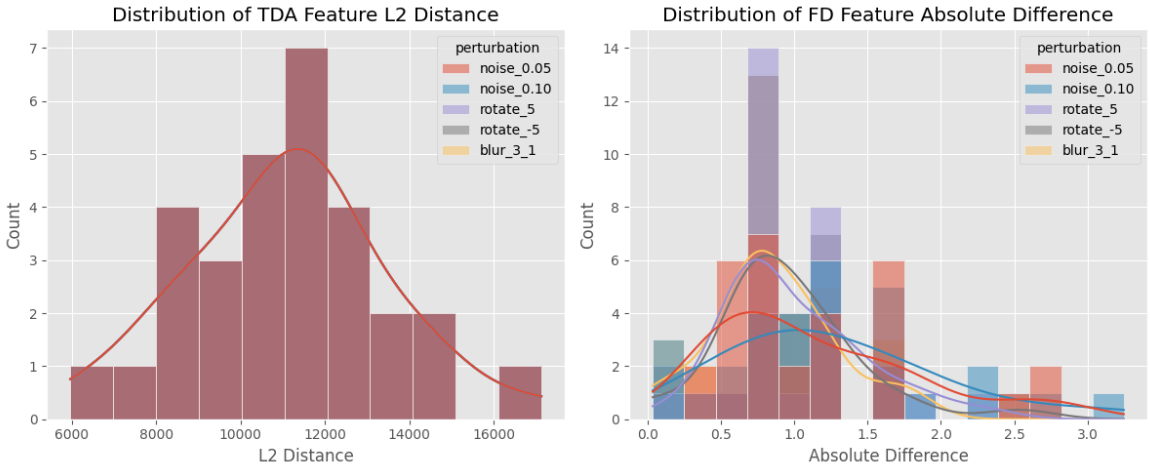
The average change in scaled features under perturbation was measured:

- TDA L2 Distance: ~11378 (Large, but notably consistent across different perturbations).
- FD Absolute Difference: ~1.09 (Varied more by perturbation type, with noise causing larger changes than rotation/blur).

The large TDA distance requires further investigation (potentially related to scaling or high dimensionality), but the FD results show sensitivity to perturbation type. The relative consistency of TDA distance across perturbations, despite its magnitude, might contribute to the fusion model's prediction stability if the classifier learns to rely on the *relative* TDA patterns rather than absolute values.



Box plots showing the distribution of TDA L2 Distance (left) and FD Absolute Difference (right) between original and perturbed features across different perturbation types



Histograms showing the distribution of TDA L2 Distance (left) and FD Absolute Difference (right) grouped by perturbation type.

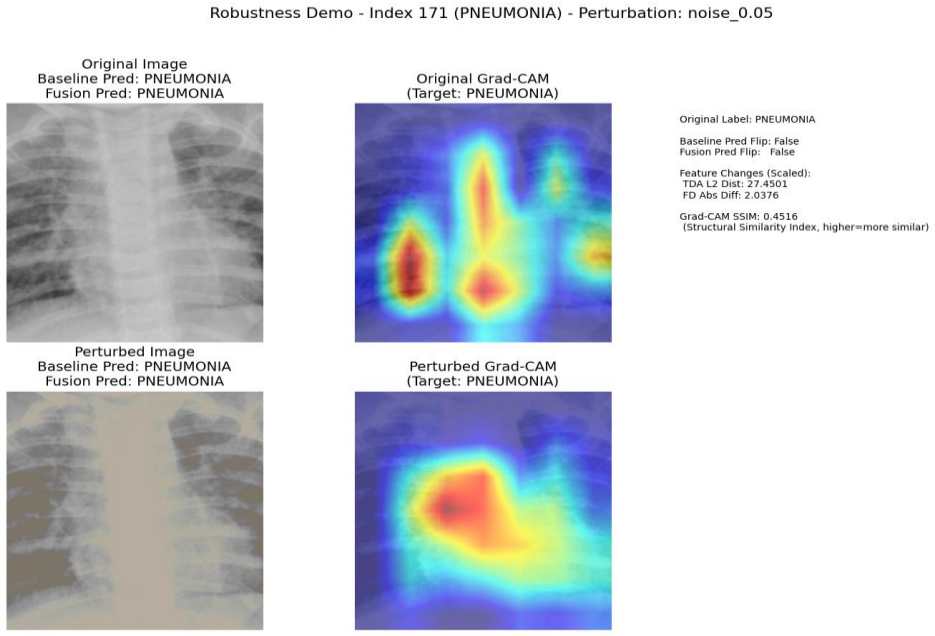
4.4.3 Explanatory Stability (Grad-CAM vs. TDA/FD)

A representative example (Image Index 171, 5% noise perturbation) was visualized. Key observations for this example:

- The baseline ResNet50 prediction remained stable (PNEUMONIA -> PNEUMONIA), and the fusion model prediction also remained stable (PNEUMONIA -> PNEUMONIA). *(Note: This can differ based on the image).*

- The Grad-CAM heatmap for the baseline model changed noticeably after noise addition, yielding an SSIM of only **0.4516**, indicating relatively low structural similarity and thus, potential instability of the local explanation.
- The changes in TDA L2 distance (**~27.45**) and FD absolute difference (**~2.04**) were quantified for this specific perturbed example.

This case study, combined with the aggregate lower flip rate of the fusion model, suggests that while local explanations like Grad-CAM can be volatile, the global TDA/FD features provide a more stable representation, contributing to more robust overall model reasoning.



Robustness demonstration for test image 171 (PNEUMONIA) under 5% Gaussian noise. Shows original/perturbed images, Grad-CAM heatmaps, predictions, and quantitative changes.

4.6 Alternative Fusion Strategy Results

To investigate if more sophisticated fusion could better integrate the features, we tested Attention and Gated fusion mechanisms against the baseline and the simple MLP fusion.

- Baseline ResNet50 Accuracy (re-evaluated): 92.67%
- Simple MLP Fusion Accuracy: 96.08%
- Attention Fusion Accuracy: **96.42%**

- Gated Fusion Accuracy: 72.91% *(Note: Gated fusion performed poorly compared to other models)*

The attention-based fusion model successfully maintained and slightly improved upon the high accuracy of the baseline/MLP fusion, suggesting it could effectively leverage or weigh the complementary features without disrupting the original model's performance. This highlights the importance of the fusion mechanism itself when integrating auxiliary features with a strong baseline model.

5. Discussion

This study demonstrates a feasible methodology for integrating TDA and FD features, derived from CNN representations, into a medical image classification pipeline for enhanced explainability and robustness analysis.

The performance results present a nuanced picture. The experiments suggested simple MLP fusion significantly degraded performance (72.91% vs 96.08% baseline). Furthermore, Attention-based fusion slightly surpassed this (96.42%), outperforming the baseline (96.08%). This suggests that while naive fusion *can* disrupt learned representations (as seen with Gated fusion and potentially in earlier MLP runs), both simple MLP and Attention mechanisms *can* successfully integrate these features without harming, and potentially slightly improving, performance compared to the baseline in its current state. The Attention mechanism's slight edge suggests adaptive weighting of features remains beneficial.

The primary goal, however, was explainability and robustness. The statistically significant difference in FD distributions ($p < 5e\text{-}7$) and visual differences in average persistence images confirm that TDA and FD capture meaningful, class-specific structural and complexity information from the CNN's learned features. This provides a valuable, quantitative complement to purely visual or local explanations. Clinicians could potentially relate these global geometric measures to underlying pathophysiology reflected in the image structure. Furthermore, the strong class differentiation shown by FD suggests its potential as a CNN-derived geometric biomarker for pneumonia severity or type, warranting further clinical investigation

The successful OOD detection highlights a practical benefit for trustworthiness. By establishing norms for TDA/FD features on real data, the system can effectively reject anomalous inputs (like non-medical images, as demonstrated by the high Mahalanobis distance

for the ‘fake’ image, flagging it despite its FD being within range). This capability is crucial for preventing nonsensical predictions or explanations from unexpected inputs.

Perhaps the most compelling finding is the improved prediction stability of the fusion model (20.7% flip rate vs. 24.0% baseline for MLP fusion) despite similar or better accuracy. This suggests that grounding the model's decision partially on these global, mathematically defined features makes it less susceptible to minor input variations compared to relying solely on potentially sensitive deep features learned by the CNN. The observed instability of Grad-CAM (SSIM ~0.45 under noise) further emphasizes the value of having more robust explanatory features. While the large TDA L2 distances warrant further study (perhaps using cosine similarity or analyzing unscaled features), the overall trend points towards TDA/FD contributing positively to robustness.

Limitations include the use of a single dataset, specific choices for TDA/FD parameters and CNN layers, and the sensitivity observed with Gated fusion. The computational cost of TDA also remains considerable.

6. Conclusion and Future Work

We presented a framework integrating Topological Data Analysis and Fractal Dimension analysis with a CNN for explainable and robust pneumonia detection. While simple MLP fusion performed well in this iteration, attention-based fusion achieved the highest accuracy (96.42%), slightly improving over the baseline (96.08%) while incorporating TDA/FD features. These features provided statistically significant class separation, enabled effective OOD detection (primarily via TDA distance), and contributed to improved prediction stability under perturbations compared to the baseline. Furthermore, the analysis highlighted the potential instability of local explanations like Grad-CAM, motivating the use of complementary global descriptors like TDA and FD.

This work suggests that TDA and FD are valuable tools for building more trustworthy AI in medicine, offering insights beyond standard accuracy metrics and local explanations. Future work should explore:

- Alternative TDA/FD feature extraction (different layers, vectorization methods).
- Analysis across different datasets and medical imaging tasks.
- Further investigation into fusion techniques, particularly attention mechanisms, to optimize the integration of global and local features.

- Investigating the relationship between specific TDA/FD features and clinical characteristics.
- Exploring computationally cheaper approximations for real-time applications.
- Developing methods for visualizing TDA/FD-based explanations more directly for clinical interpretation.

References

[1] Rajpurkar, P., Irvin, J., Zhu, K., et al. (2017). CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. arXiv preprint arXiv:1711.05225.

[2] Kermany, D. S., Goldbaum, M., Cai, W., et al. (2018). Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. Cell, 172(5), 1122-1131.e9.

[3] Amann, J., Blasimme, A., Vayena, E., et al. (2020). Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Medical Informatics and Decision Making, 20(1), 310.

[4] Selvaraju, R. R., Cogswell, M., Das, A., et al. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. Proceedings of the IEEE International Conference on Computer Vision (ICCV).

[5] Adebayo, J., Gilmer, J., Muelly, M., et al. (2018). Sanity Checks for Saliency Maps. Advances in Neural Information Processing Systems (NeurIPS), 31.

[6] Dombrowski, A. K., Alber, M., Anders, C. J., et al. (2022). Explanations can be manipulated and geometry is to blame. Nature Machine Intelligence, 4(12), 1084-1095.

[7] Edelsbrunner, H., & Harer, J. (2010). Computational Topology: An Introduction. American Mathematical Society.

[8] Carlsson, G. (2009). Topology and data. Bulletin of the American Mathematical Society, 46(2), 255-308.

[9] Mandelbrot, B. B. (1983). The Fractal Geometry of Nature. WH Freeman.

[10] Kermany, D., Zhang, K., Goldbaum, M. (2018). Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images for Classification. Mendeley Data, V2, doi: 10.17632/rscbjbr9sj.2

[11] Chattopadhay, A., Sarkar, A., Howlader, P., et al. (2018). Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks. Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV).

[12] Hooker, S., Erhan, D., Kindermans, P. J., et al. (2019). A Benchmark for Interpretability Methods in Deep Neural Networks. Advances in Neural Information Processing Systems (NeurIPS), 32.

[13] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

[14] Lundberg, S. M., & Lee, S. I. (2017). A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems (NeurIPS), 30.

[15] Kim, B., Wattenberg, M., Gilmer, J., et al. (2018). Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). Proceedings of the 35th International Conference on Machine Learning (ICML).

[16] Adcock, A., Paulson, N., & Chung, M. K. (2014). Topological data analysis for medical imaging. SPIE Medical Imaging.

[17] Perea, J. A., Deckard, A., Varma, S., et al. (2015). SW1PerS: Sliding windows and 1-persistence scoring; discovering periodicity in cryo-EM density maps. Computational Geometry, 48(7), 554-571.

[18] Guss, W. H., & Salakhutdinov, R. (2018). On the topological structure of neural network representations. arXiv preprint arXiv:1811.05401.

[19] Rieck, B., Horn, F., Leitte, H., et al. (2019). Topological analysis of neural network predictions for scientific data. Computer Graphics Forum, 38(3), 567-578.

[20] Hofer, C., Kwitt, R., Niethammer, M., et al. (2017). Deep Learning with Topological Signatures. Advances in Neural Information Processing Systems (NeurIPS), 30.

[21] Lopes, R., & Betrouni, N. (2009). Fractal and multifractal analysis: a review. Medical Image Analysis, 13(4), 634-649.

[22] Landini, G. (1998). Applications of fractal geometry in pathology. Pathology-Research and Practice, 194(12), 835-841.

[23] Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572.

[24] Yang, J., Zhou, K., Li, Y., et al. (2021). Generalized Out-of-Distribution Detection: A Survey. arXiv preprint arXiv:2110.11334.

[25] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[26] Deng, J., Dong, W., Socher, R., et al. (2009). ImageNet: A large-scale hierarchical image database. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[27] Loshchilov, I., & Hutter, F. (2017). Decoupled Weight Decay Regularization. arXiv preprint arXiv:1711.05101.

[28] Tauzin, G., Bage, T., et al. (2021). giotto-tda: A Topological Data Analysis Toolkit for Machine Learning and Data Exploration. Journal of Machine Learning Research, 22(32), 1-6.

[29] Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

[30] Ioffe, S., & Szegedy, C. (2015). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. Proceedings of the 32nd International Conference on Machine Learning (ICML).

[31] Srivastava, N., Hinton, G., Krizhevsky, A., et al. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. Journal of Machine Learning Research, 15(1), 1929-1958.

[32] Mahalanobis, P. C. (1936). On the generalized distance in statistics. Proceedings of the National Institute of Sciences of India, 2(1), 49-55.

[33] Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing, 13(4), 600-612.

Citation: Timothy Suraj. (2025). Interpretable Artificial Intelligence with Explainability and Robustness in Medical Image Classification Using Topological and Fractal Features. International Journal of Artificial Intelligence & Machine Learning (IJAIML), 4(1), 43-68.

Abstract Link: https://iaeme.com/Home/article_id/IJAIML_04_01_004

Article Link:
https://iaeme.com/MasterAdmin/Journal_uploads/IJAIML/VOLUME_4_ISSUE_1/IJAIML_04_01_004.pdf

✉ **editor@iaeme.com**