PNEUMONIACXR: AI-ENABLED PNEUMONIA DETECTION

Gary Kong
UC Berkeley

Drew Piispanen
UC Berkeley

Diqing Wu
UC Berkeley

April 19, 2024

ABSTRACT

Challenges in chest X-ray (CXR) interpretation, particularly the overlapping visual features of different respiratory diseases, can hinder accurate pneumonia diagnosis. To address this, we developed machine learning models capable of classifying CXR images as COVID-19 pneumonia, non-COVID pneumonia, or normal. Our approach combines edge-based features (HOG), textural patterns (Radiomics), and deep learning representations (ResNet) to capture diverse image characteristics. Following dimensionality reduction, we explored Logistic Regression, Support Vector Machines (SVM), Gradient Boosting, and Random Forest classifiers. Logistic Regression and SVM outperformed others with accuracies reaching as high as 0.90, with Logistic Regression showing remarkable efficiency. These results demonstrate the potential of machine learning to enhance pneumonia diagnosis. Future research should focus on model explainability, probabilistic outputs, and validation of generalizability across different platforms and healthcare settings.

1 Introduction

The global COVID-19 outbreak has emphasized the need for fast, accurate diagnostic methods. Chest X-ray (CXR) imaging is a widely accessible tool for diagnosing respiratory diseases. However, traditional reliance on radiologist expertise can lead to variability and create diagnostic bottlenecks. Automated systems could offer greater diagnostic consistency and alleviate pressure on medical professionals.

Pneumonia, a common respiratory condition manifests in chest X-rays (CXR) as airspace opacification, where normally dark areas turn white/grey due to alveoli filling with infectious materials. These changes, ranging from patchy to confluent, are key indicators for pneumonia, alongside air bronchograms [6]. For COVID-19, specific patterns include asymmetric airspace opacities, bilateral ground-glass opacities or consolidations with a peripheral and mid-to-lower lung distribution, and ill-defined margins of lesions [3, 9, 10]. However, it is important to note that while these visual characteristics are suggestive, there may be some overlap in the appearance of different respiratory diseases on chest X-rays. This diagnostic complexity, along with the potential for variability among radiologists, highlights the need for automated tools to aid in pneumonia diagnosis.

We propose a machine learning-based approach to classify chest X-ray images into COVID-19 pneumonia, non-COVID pneumonia, and normal cases. This approach leverages diverse features, including edge gradients (HOG), texture patterns (Radiomics), and deep learning representations (ResNet). Following feature extraction, we employ Principal Component Analysis (PCA) for dimensionality reduction and explore various classification models (i.e., Logistic Regression, SVM, Gradient Boost, and Random Forest). Our goal is to identify a model that balances strong performance with computational efficiency, enabling rapid training and diagnosis. This system has the potential to assist clinicians in making more informed and timely diagnosis, ultimately improving patient outcomes.

2 Data and Preprocessing

Our study leverages the COVID-QU-Ex Dataset, a publicly available collection of chest-X ray images compiled from 10 separate studies spanning diverse pneumonia severities, diagnostic platforms and geographies [12]. The dataset comprises 32,103 chest X-ray images categorized as COVID-19 positive, non-COVID infections, and normal cases. To

ensure a balanced representation of classes, we performed random undersampling, resulting in an even 33% split for each class. Table 1 summarizes the class distribution before and after applying random undersampling.

Data Set	COVID-19		non-COVID		Normal	
	Before	**After**	**Before**	**After**	**Before**	**After**
Training	7,658	6,849	7,208	6,849	6,849	6,849
Validation	1,903	1,712	1,802	1,712	1,712	1,712
Test	2,395	2,140	2,253	2,140	2,140	2,140
Total	11,956	10,701	11,263	10,701	10,701	10,701

Table 1: Number of Samples by Diagnostic Class and Split, Before and After Resampling

Each record included a chest X-ray image and a corresponding lung mask delineating the lung regions. The images retained their original grayscale intensities (0 to 255), while lung masks were binarized for effective segmentation. Fig. 1 highlights inconsistencies in the original dataset, including variations in contrast, the proportions of each image occupied by the lungs (with some images exhibiting black borders), and inconsistent image orientation.

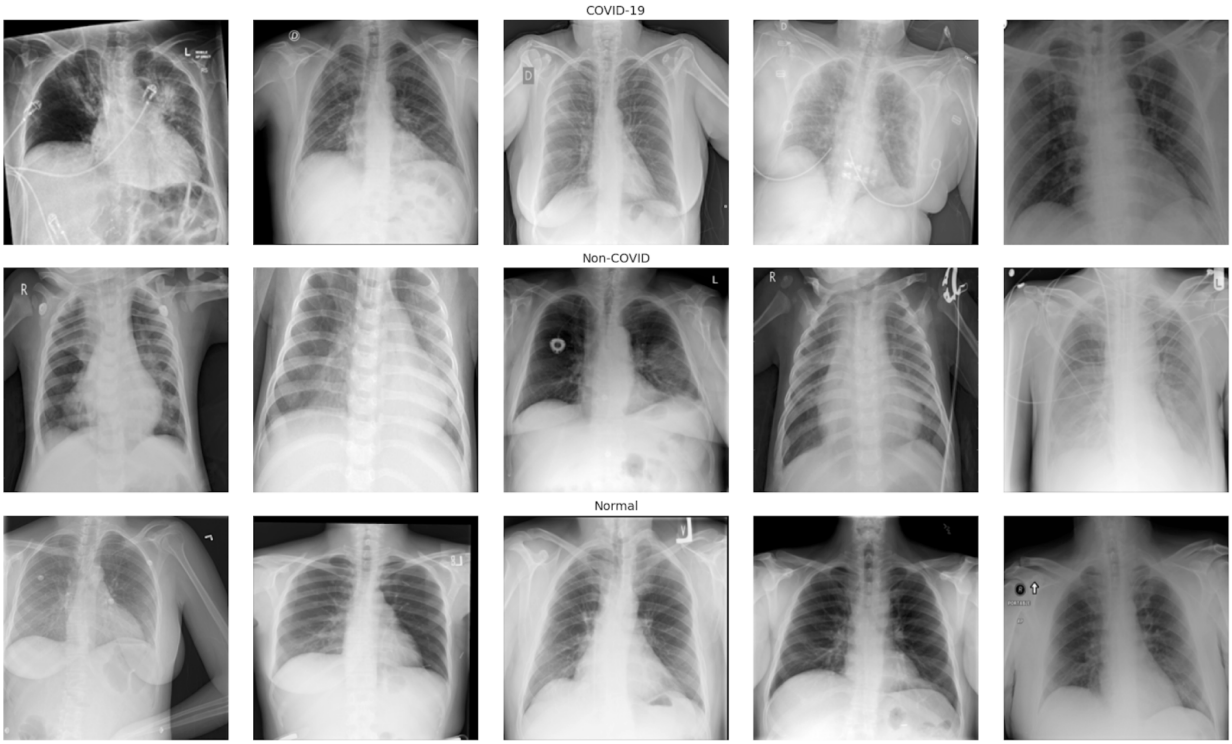

Figure 1: Unprocessed Sample Images

To address these inconsistencies, we performed several preprocessing steps. First, we applied z-score scaling to standardize the contrast range within each image. Next, we cropped images and their corresponding masks, focusing on lung regions identified by the masks. We added padding during cropping to preserve valuable anatomical features outside the lungs (e.g., the diaphragm). Finally, we used OpenCV to rotate images, achieving a standard perpendicular orientation based on spinal cord alignment.

3 Feature Engineering

To effectively analyze the complex visual patterns of pneumonia in chest X-rays, we employed a multi-modal feature engineering approach. This included Histogram of Oriented Gradients (HOG), Radiomics, and features extracted via a pre-trained ResNet model. This approach was designed to capture diverse aspects of chest X-ray images, with HOG emphasizing edge-based patterns of opacity, Radiomics quantifying textural variations, and ResNet extracting deeper patterns through its learned representations.

For HOG and Radiomics features, hyperparameter tuning was performed on a 10% subset of the training data using Optuna, a well-established optimization library. The objective function was set to maximize the sum of mutual information between the extracted HOG features and the disease labels (pneumonia or normal). Mutual information quantifies the amount of information one variable (features) provides about another (disease labels). In feature engineering, higher mutual information between features and the class labels suggests the features' potential usefulness for classification.

Following individual feature extraction, we adopted Principal Component Analysis (PCA) for dimensionality reduction, ensuring the most informative aspects of each feature set are retained while mitigating redundancy and noise, and reducing computational requirements during model training. The culmination of this process involved concatenating the principal components of the feature sets, thus creating a comprehensive representation of each image in the dataset.

3.1 Histogram of Oriented Gradients (HOG)

HOG is a feature descriptor well-suited for analyzing edge patterns in chest X-rays. By detecting the distribution of gradient orientations, HOG effectively highlights areas of consolidation, ground-glass opacities, and their relative locations within the lung field. HOG excels at capturing these edge-based attributes, which are crucial for differentiating between pneumonia types. For instance, consolidations would manifest as stronger edges compared to semi-transparent ground-glass opacities.

Parameters	**Initial Configuration**	**Search Values**	**Parameter Importance**	**Best Configuration**
Image Size	(128, 255)	(128,255), (255,128), (255,255)	30.9%	(255, 255)
Orientations	9	7, 8, 9	4.3%	9
Pixels per Cell	(16, 6)	(8,8), (12,12), (16,16)	58.7%	(16, 16)
Cells per Block	(2, 2)	(2,2), (3,3)	6.2%	(2, 2)

Table 2: HOG Hyperparameter Tuning Results

Key parameters for HOG include image size, pixels per cell, and cells per block. These parameters control the level of detail captured, the granularity of image partitioning, and the extent of local gradient patterns considered. Table 2 summarizes our hyperparameter tuning process for HOG, initially using configurations aligned with a similar study on pneumonia detection [8]. Hyperparameter tuning was conducted using 40 trials, in which trials leading to vector sizes exceeding 15,000 were discarded. L2 normalization was applied across all tested configurations.

HOG hyperparameter tuning (Table 2) reveals that pixels per cell is the most important parameter, followed by image size. This is likely because pixels per cell directly controls the fineness of edge detection. A smaller pixels per cell value leads to the detection of more subtle variations in edges, crucial for distinguishing different types of opacities. Image size is also important, as it determines the overall visual field considered by the HOG descriptor. Interestingly, cells per block and orientations had a relatively minor impact on mutual information.

Based on these findings, the configuration for HOG feature extraction was set with (255, 255) image size (maintaining the original image size), (16,16) pixels per cell, and (2,2) cells per block, with L2 normalization applied. This improved total mutual information in the validation set by 230% compared to the initial settings, demonstrating the value of higher-resolution detail for pneumonia classification.

Examination of sample HOG images 2 helps elucidate how this descriptor aids in pneumonia classification. The emphasis on edges allows HOG to delineate areas of consolidation. However, in cases of diffuse opacities, the edges become less distinct, reflecting HOG's primary focus on localized gradient patterns.

3.2 Radiomics

Radiomics, the process of extracting and analyzing high-dimensional quantitative features from medical images, offers valuable insights beyond traditional visual assessment. It analyzes subtle textural patterns correlated with underlying disease characteristics. Applied to CXR images, Radiomics can differentiate between normal tissue, non-COVID pneumonia and COVID-19 pneumonia based on variations in texture. This technique complements HOG, which primarily focuses on shape and edge patterns. Radiomics' potential in pneumonia classification is supported by recent studies. One study demonstrated its potential in differentiating COVID-19 pneumonia from influenza virus pneumonia based on CT scan images [1]. Another successfully used a Radiomics-based model to detect pneumonia in CXR images based on textural features [7].

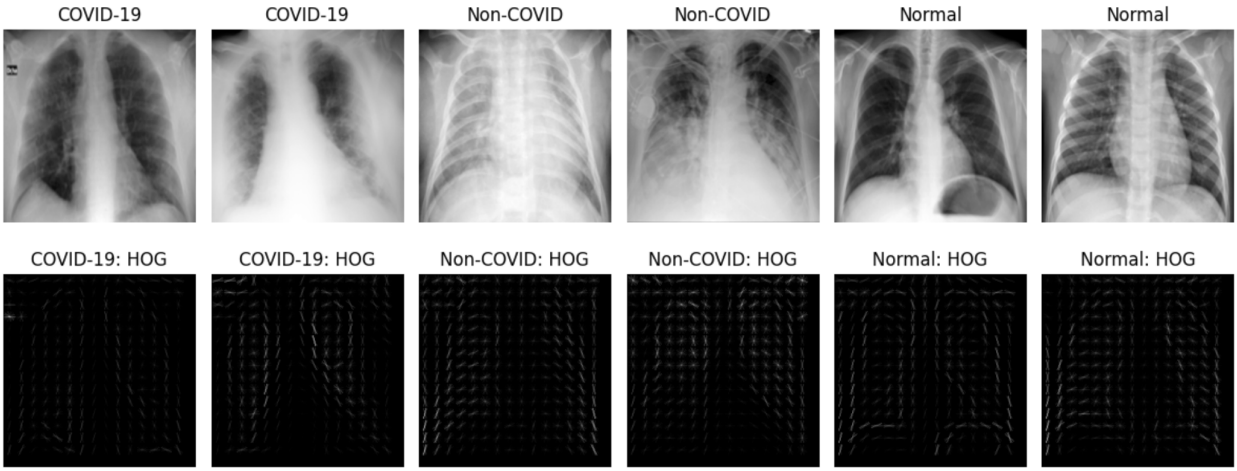

Figure 2: HOG Features for Example Images by Diagnosis Class with
(255, 255) Image Size, (16, 16) Pixels per cell, (2,2) Cells per Block, and L2 Normalization

We extracted Radiomics features from the segmented lung regions (defined by the provided masks) using the well-established PyRadiomics package [13]. Images were z-score normalized, but cropping and rotation adjustments were not used due to the invariance of the selected features to orientation and the use of lung segmentation.

Radiomics feature engineering followed a multi-stage process. We began by exploring a wide range of potential textural features on a 10% subset of the training data, using multiple image transformations and feature classes. Next, we optimized hyperparameters for the most promising transformations and feature classes based on mutual information analysis, again utilizing the 10% subset. Building on these insights, we performed feature selection on the complete training dataset to identify the most informative features while minimizing redundancy. Finally, we extracted the selected features with their optimal hyperparameter configurations across the full training, validation, and test data splits for use in downstream classification models.

3.2.1 Initial Feature Extraction

To thoroughly analyze textural changes relevant to pneumonia, we employed multiple image transformations. The original image provided a baseline, while the gradient image emphasized edges and transitions in intensity to highlight potential abnormalities by calculating the rate of change in pixel values. Local Binary Patterns (LBP2D), which compare a central pixel to its neighbors to create a binary code, captured microtextural variations that may indicate early signs of pneumonia. Finally, wavelet transform decomposed the image into various frequency scales, uncovering textural details potentially obscured at the original resolution.

Radiomics feature classes, including First Order Statistics, Gray Level Dependence Matrix (GLDM), Gray Level Co-occurrence Matrix (GLCM), Gray Level Size Zone Matrix (GLSZM), Gray Level Run Length Matrix (GLRLM), and Neighboring Gray Tone Difference Matrix (NGTDM), play crucial roles in characterizing the complexity of lung textures in pneumonia. These features capture various aspects of pixel intensity distribution, texture patterns, and spatial relationships, collectively differentiating between normal tissue, non-COVID pneumonia, and COVID-19 pneumonia. Further details of these features and their relevance to pneumonia are presented in Table 3.

The combination of image transformations and feature classes yielded a total of 651 Radiomics features initially extracted from a 10% subset of the training data. Analysis of the mean and sum of mutual information scores by feature transformation (Fig. 3) revealed several insights. Gradient and wavelet transformations generally enhanced mutual information compared to original images. Gradient-transformed features demonstrated the highest average mutual information, followed by wavelet-transformed features. Notably, while the wavelet transform generated many features (resulting in the highest sum of mutual information), linear binary patterns (LBP2D) displayed lower mean and sum mutual information, suggesting they may be less informative for this specific application. These findings informed our decision to prioritize gradient and wavelet transformations in hyperparameter tuning.

Analysis of mutual information scores also highlighted insights regarding feature classes (Fig. 3). The feature classes GLCM, GLDM, first-order statistics, and GLRLM emerged as the most promising based on their overall mutual information scores. Consequently, subsequent feature engineering efforts focused primarily on features derived from

Feature Class	# **Features**	**Description**	**Relevance to Pneumonia**
First Order Statistics	19	Measures pixel intensity distribution (e.g., mean, variance).	Detects variations in lung tissue density, useful for distinguishing between normal and pathological states.
Gray Level Dependence Matrix (GLDM)	14	Captures dependencies of pixel pairs, considering distance.	Highlights textural changes in the lung indicative of pneumonia, such as areas of consolidation.
Gray Level Co-occurrence Matrix (GLCM)	24	Analyzes spatial relationships of pixel intensities (contrast, homogeneity).	Useful for differentiating types of pneumonia by assessing changes in lung texture homogeneity.
Gray Level Size Zone Matrix (GLSZM)	16	Quantifies homogeneous zones by size and intensity distribution.	Identifies regions of lung consolidation, contributing to distinguishing pneumonia in lung images.
Gray Level Run Length Matrix (GLRLM)	16	Measures consecutive pixels of the same intensity in various directions.	Effective in analyzing vascularity by identifying continuous lines (blood vessels). Lesions may disrupt these patterns, indicating pathological changes.
Neighboring Gray Tone Difference Matrix (NGTDM)	5	Analyzes intensity differences between a pixel and its neighborhood.	Highlights local heterogeneity in lung texture, useful for identifying early signs of pneumonia.

Table 3: Description of Radiomics Feature Classes

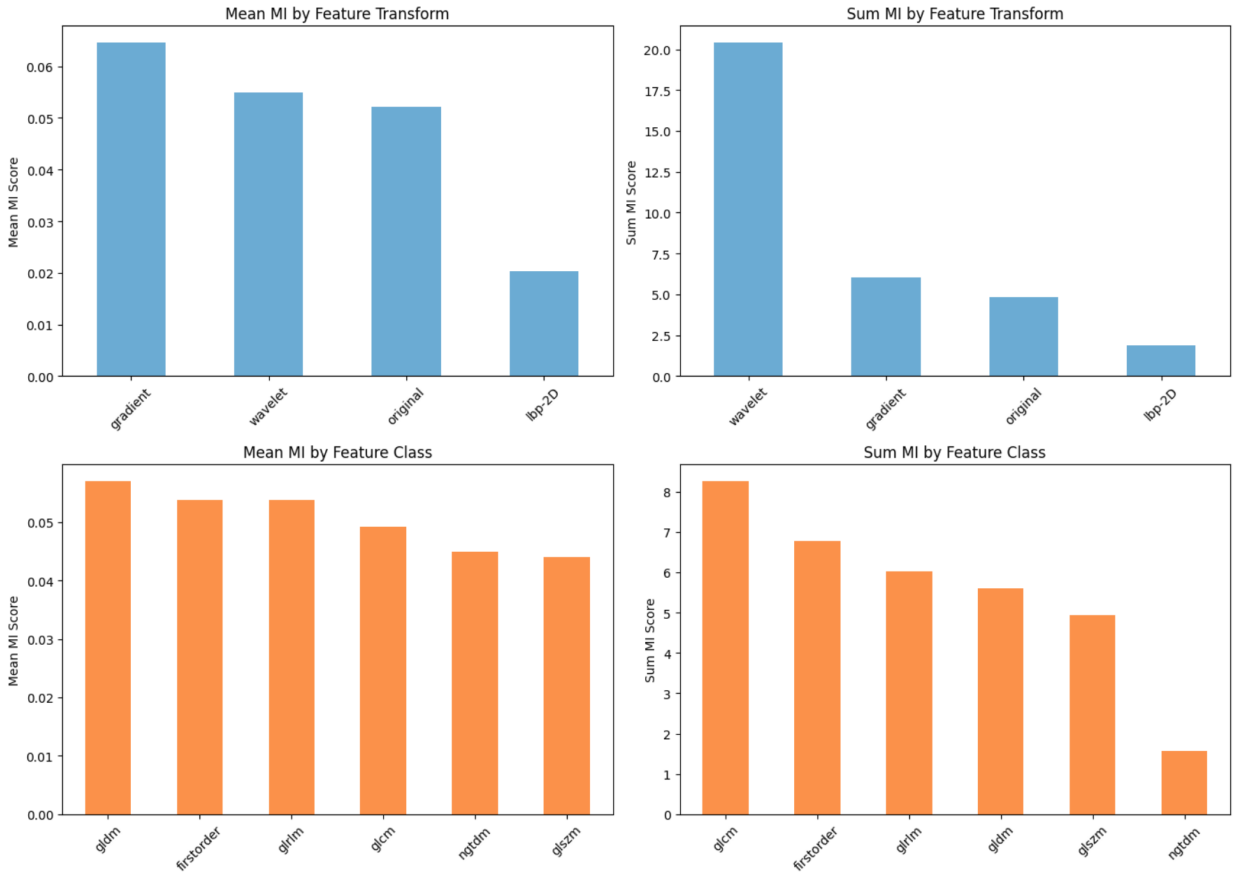

Figure 3: Mean and Sum Mutual Information, by Feature Transformation and Feature Class (10% Subset of Training Data)

gradient, wavelet, and original transformations in combination with the GLCM, GLDM, first-order statistics, and GLRLM feature classes.

3.2.2 Hyperparameter Tuning

To maximize the informativeness of our Radiomics features, we employed a sequential hyperparameter tuning strategy. Mutual information (MI) gain with the target class labels served as our primary optimization metric. We adopted the Optuna framework to efficiently explore the parameter space.

Prioritizing computational efficiency, we tuned hyperparameters associated with image sampling, transformations, and feature classes sequentially. We began by optimizing resampling and binning settings, considering that regions of interest (ROI) of images had varying image resolutions. Subsequent tuning focused on gradient and wavelet transformations, inheriting the optimized sampling settings. Finally, we optimized hyperparameters for the GLRLM and GLDM feature classes, which also inherited settings from prior stages. Initial iterations suggested that tuning GLCM parameters was unlikely to substantially improve MI. Due to the computationally intensive nature of Radiomics feature extraction, we performed tuning on a 1% subset of the original training data (76 Normal, 68 COVID-19, and 62 non-COVID images).

Transformation/ Class	**Parameters**	**Initial Configuration**	**Search Values**	**Best Configuration**
Sampling	**Resample**	False	True, False	False
	Bin Width	None	None, 5, 10, 15, 20	None
Wavelet Transformation	**Wavelet Type**	coif1	coif1, db1, sym2	coif1
	Start Level	0	0, 1	1
	Number of Levels	1	1, 2, 3	3
Gradient Transformation	**Gradient Use Spacing**	False	True, False	False
GLRLM Class	**Weighting Norm**	None	Manhattan, Euclidian, None	Manhattan
GLDM Class	**Distance**	1	1, 2, 3	1
	Alpha	0	0, 1, 2, 3, 4, 5	1

Table 4: Radiomics Hyperparameter Tuning Results

Table 4 summarizes the hyperparameter tuning process and outcomes. Importantly, this optimization led to a 146% increase in mutual information within our validation dataset. Our analysis of the tuning process reveals that several key parameters remained optimal at their initial configuration, including sampling settings (no resampling or binning). This suggests the robustness of our chosen Radiomics features to variations in ROI resolution. On the other hand, optimizing the wavelet transform and feature class parameters yielded the most significant MI gains. Interestingly, the optimal wavelet start level falls at the edge of our tested range, indicating the potential for further improvements by exploring even higher start levels in future investigations.

It is crucial to note that while maximizing the sum of MI led to aggregate feature set improvements, this approach has limitations. As each level of the wavelet transform generates a larger number of features, the sum of MI naturally increases. In future work, using mean MI or a composite metric might provide a more balanced assessment of feature informativeness, independent of sheer feature quantity.

3.2.3 Feature Selection and Final Extraction

Feature selection aimed to reduce redundancy and improve computational efficiency while minimizing loss of valuable information. Fig. 4a demonstrates the high correlation of the initial Radiomics features. To address this, we applied a two-step selection process. First, we removed features with mutual information scores in the bottom 5th percentile, reducing the feature set from 864 to 820. Next, we iteratively selected the feature with the highest mutual information score and removed all other features with a correlation above 0.95. This process further reduced the feature set from 820 to 241 features, as visually demonstrated by the decrease in correlation in Fig. 4b.

Visual inspection of the top 50 Radiomics features (ranked by mutual information) highlights the importance of employing multiple image transformations and feature classes (Fig. 5). Wavelet-transformed features dominate the top ranks, with various levels of decomposition represented (Fig. 5a), demonstrating their effectiveness in extracting valuable information for pneumonia classification. Furthermore, while first-order and GLCM features exhibit the highest mutual information, GLDM and GLRLM features are also well-represented (Fig. 5b). This diversity in feature classes ensures the Radiomics features captures a broad range of image characteristics.

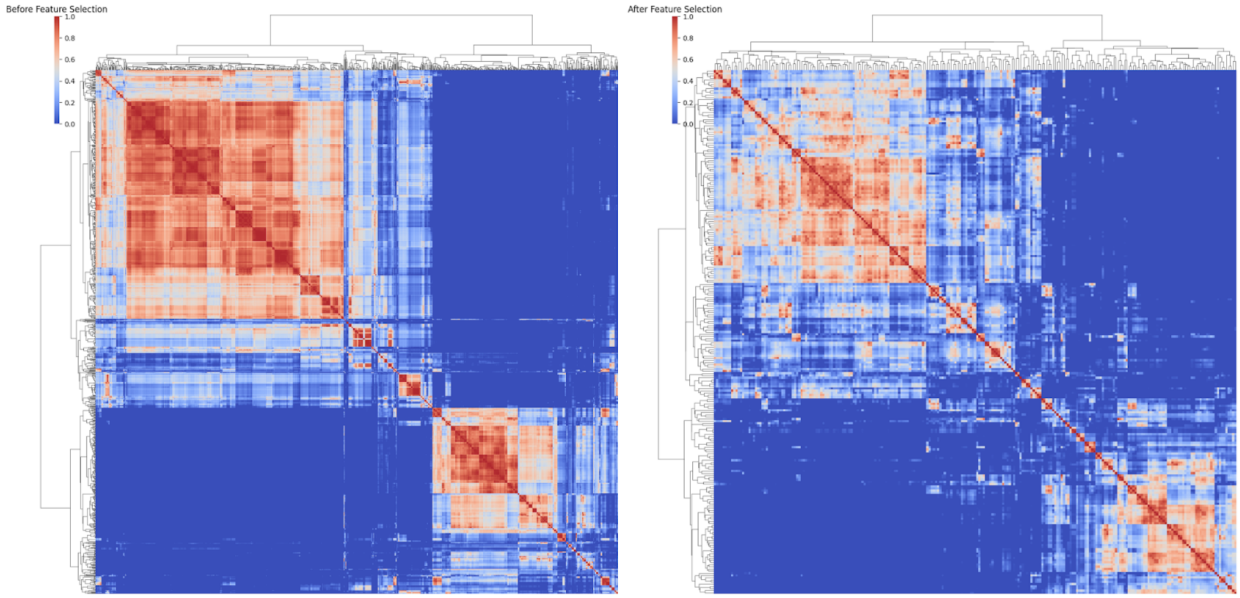

Figure 4: Radiomics Feature Correlations a) Before and b) After Feature Selection

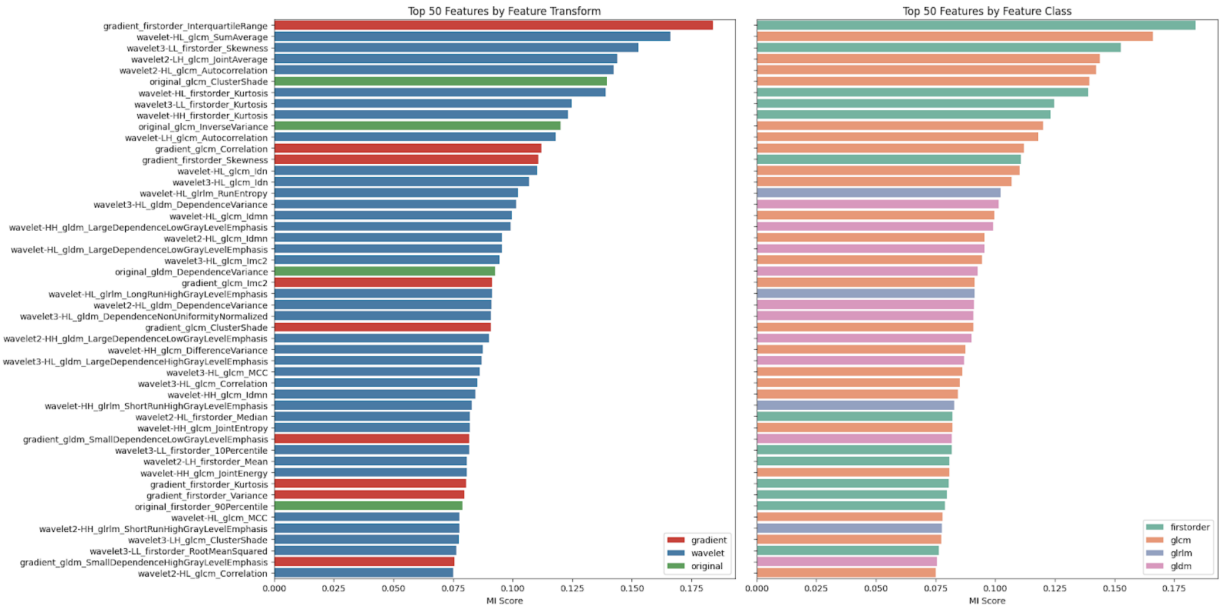

Figure 5: Top 50 Radiomics Features by a) Image Transformation Applied and b) Feature Class

3.3 RESNET

Residual Network (ResNet), a convolutional neural network architecture, addresses the problem of vanishing gradients encountered in deeper networks. Its core innovation, the residual block, introduces skip connections that allow for more efficient learning of deeper representations crucial for complex image analysis tasks [2]. This makes ResNet well-suited for medical image applications.

The effectiveness of ResNet for pneumonia classification is supported by recent research. For instance, in a study on viral, bacterial, and normal case classification using chest X-ray images, ResNet50 achieved an accuracy of 93.01%, while ResNet-101 demonstrated an accuracy of 97.78% in distinguishing between COVID-19 and other viral pneumonia [4]. We chose ResNet50 over other variants for its balance of depth and computational efficiency.

To extract meaningful features from CXRs, we employed a pre-trained ResNet50 model. Images were resized to (224, 224), converted to grayscale, and normalized using standard ImageNet statistics. We opted for direct resizing to preserve detail around the periphery (since images were pre-cropped). We flattened extracted embeddings into a 1D vector, providing us with the raw ResNet50 embeddings.

3.4 T-distributed Stochastic Neighbor Embeddings

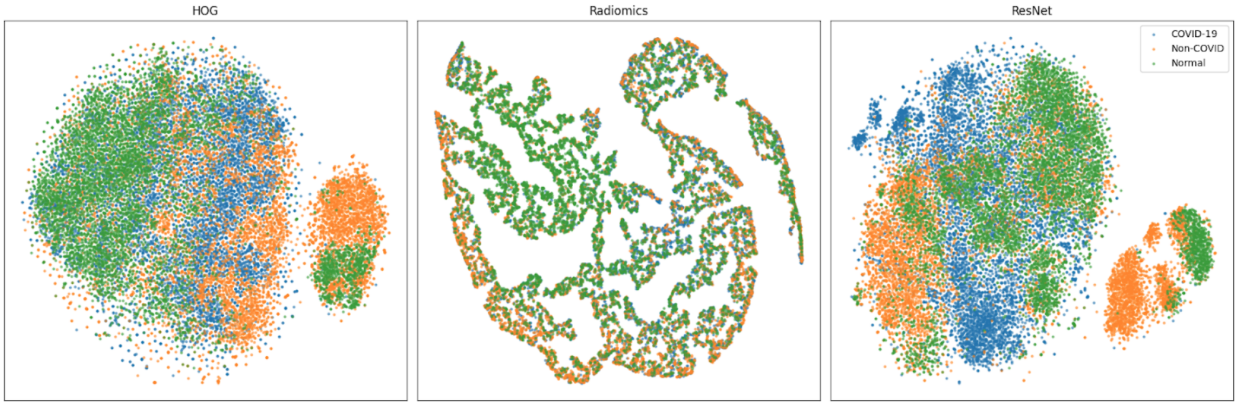

Figure 6: t-SNE Visualization of HOG, Radiomics and ResNet Features (Training Set)

T-distributed Stochastic Neighbor Embeddings (t-SNE) is a dimensionality reduction technique that projects how high-dimensional data points onto a lower-dimensional space (typically 2D or 3D) for easier visualization. In the context of machine learning and feature extraction, t-SNE helps us analyze how well the extracted features separate data points belonging to different classes.

As shown in Fig. 6, our t-SNE visualizations reveal that while perfect class separation is not achieved, certain methods demonstrate tendencies toward clustering. HOG and ResNet features exhibit some clustering behavior, suggesting that these feature extraction techniques might inherently capture characteristics that allow for partial differentiation between COVID-19 pneumonia, non-COVID pneumonia, and normal cases. Conversely, the lack of clustering observed with the Radiomics feature set could stem from the diverse nature of the extracted features. This highlights a limitation of t-SNE when visualization relationships between heterogenous features, where preserving global structure in a lower-dimensional space can be challenging, particularly if features are uncorrelated or orthogonal.

3.5 Principal Component Analysis

Principal Component Analysis (PCA) was employed to reduce the high dimensionality of our feature sets, specifically HOG (13,689 dimensions), Radiomics (241 dimensions), and ResNet (2,048 dimensions). This technique can mitigate the risk of overfitting and improves computational efficiency by creating a more manageable feature space.

Fig. 7 illustrates the cumulative explained variance captured by PCA for each feature set as the number of principal components increases. As expected, the inherent complexity of HOG features is reflected in the plot, with a relatively high number of components (607 for 80% variance, 1,065 for 90%) required to capture most of the variation. Conversely, Radiomics features exhibit a much steeper curve, requiring only 11 and 22 components to reach 80% and 90% explained variance, respectively. This suggests some redundancy within the Radiomics features, indicating potential benefit from

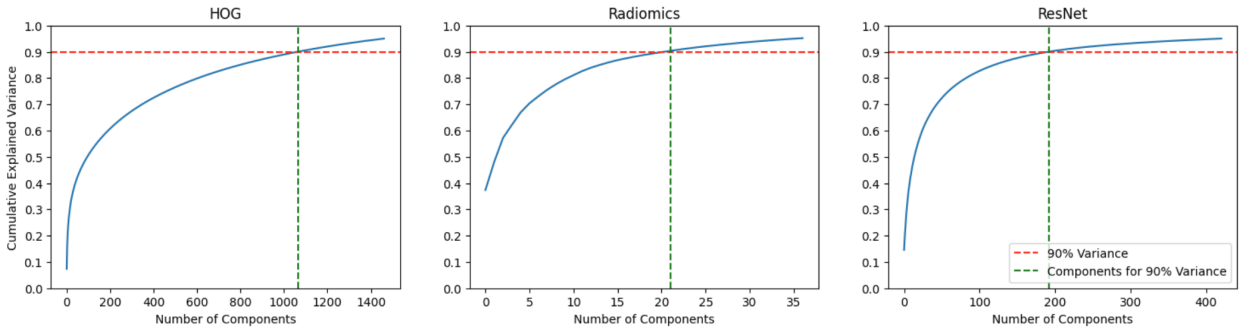

Figure 7: Cumulative Explained Variance vs. Number of PCA Components for HOG, Radiomics and ResNet Features (Training Set)

further feature selection techniques. ResNet features show a similar trend to HOG, also requiring a substantial number of components (84 for 80% variance, 193 for 90%) to capture important variations.

By applying PCA to each feature set and then concatenating the resulting principal components, we obtain a consolidated feature vector that is significantly lower in dimensionality (702 and 1280 features for 80% and 90% cumulative explained variance, respectively) while still retaining a high percentage of the original information. This transformation allowed us to train more complex models on this reduced feature space while minimizing overfitting issues.

4 Results

Four models are introduced for classification: Logistic Regression, a linear classifier known for its efficiency and interpretability; Support Vector Machine (SVM), which excels at finding optimal boundaries between classes even in high-dimensional spaces; Gradient Boosting, an ensemble method combining weak learners for improved accuracy; and Random Forest, another ensemble method that leverages multiple decision trees to mitigate overfitting and provide insights into feature importance.

4.1 Hyperparameter Optimization

As with feature engineering, we employed the Optuna framework to perform a search over the hyperparameter space for each of the four models, aiming to identify the configurations that yielded the best performance metrics. We used accuracy as the objective function and conducted the hyperparameter search using the validation set, to avoid overfitting while maximizing generalizability. The parameters investigated, their search values, and the best parameter values for each model are shown in Table 5 below.

Model	**Number of Trials**	**Search Time Per Trial (s)**	**Parameter**	**Search Values**	**Best Value**
Logistic Regression	25	63.36	C	[0.01, 0.1, 1.0, 10]	0.1
			Penalty	[‘l1’, l2’]	l1
SVM	25	63.01	C	[1e-4, 1e-3, 1e-2, 1e-1, 1, 1e1, 1e2]	1
			Kernel	’linear’, ’sigmoid’, ’rbf’	rbf
Gradient Boost	15	2027.40	Learning Rate	[0.01, 0.05, 0.1]	0.1
			N Estimators	[100,200]	100
			Subsample	[0.5, 1.0]	1.0
Random Forest	25	53.44	N Estimators	[50, 100, 200]	200
			Max Depth	[5, 10]	10
			Criterion	["gini", "entropy"]	“entropy"
			Min Samples Split	[2, 5]	5

Table 5: Model Hyperparameter Tuning Results

4.2 Classification Performance

As shown in Table 6, our models demonstrated robust classification performance, with Logistic Regression and SVM achieving exceptional accuracy and F1-scores exceeding 0.89 (specifically, Logistic Regression achieved 0.90 accuracy on the test set). This, combined with their balanced precision and recall scores (Table A1), highlights their potential for reliable pneumonia diagnosis in settings where minimizing both false positives and false negatives is critical. While all models performed well on the training set (accuracy and F1-scores above 0.85), a slight performance drop was observed on the validation and test sets (ranging from 0.81 to 0.88). This decrease, particularly for SVM and Random Forest, suggests a degree of overfitting, which will be further explored in the "Generalizability" section.

Model	**Accuracy**			**F1-Score**		
	Training Set	**Validation Set**	**Test Set**	**Training Set**	**Validation Set**	**Test Set**
Logistic Regression	0.91	0.88	0.90	0.91	0.88	0.90
SVM	0.97	0.87	0.89	0.97	0.87	0.89
Gradient Boost	0.86	0.85	0.85	0.86	0.85	0.85
Random Forest	0.92	0.81	0.82	0.92	0.81	0.82

Table 6: Tuned Model Accuracy and F1 Scores on Test Set (PCA: 90% Variance)

Confusion matrix analysis (Fig. 8) revealed a significant challenge: for logistic regression, 6.4% of COVID-19 and 6.7% of non-COVID pneumonia cases were misclassified as Normal. This highlights the difficulty of distinguishing early-stage or subtle pneumonia through chest X-ray analysis, a challenge further supported by the fact that even untrained observers might struggle to discern the correct label in many of the misclassified images (Fig. 9). This aligns with clinical findings that chest X-ray interpretation can be subjective, with subtle or atypical presentations leading to diagnostic uncertainty [11].

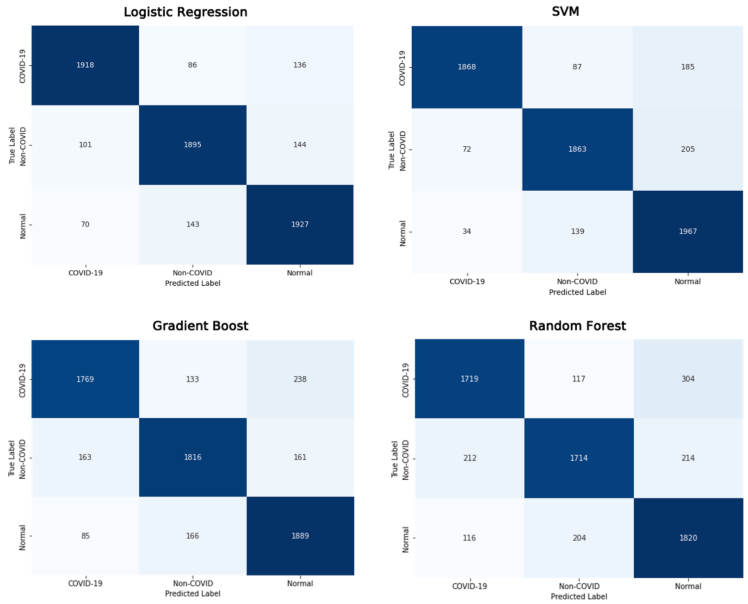

Figure 8: Confusion Matrices for a) Logistic Regression, b) SVM, c) Gradient Boost, and d) Random Forest on Test Set (PCA: 90%Variance)

There are three possible reasons for the model performance differences between classes across our models. First, there may be subtle differences between normal and early-stage pneumonia. All models might struggle to distinguish between normal and early-stage pneumonia (COVID-19 or non-COVID-19) due to their subtle visual similarities, on top of noise caused by different image quality (different chest X-ray platforms, experience of staff, etc). Second, despite balanced class distribution, some models seem more susceptible to overfitting. Random Forest and SVM, with high training accuracy and a drop in validation/test performance might be learning patterns specific to the training data that poorly generalize. Third, the feature set may not be equally effective for all categories. For instance, HOG features might be

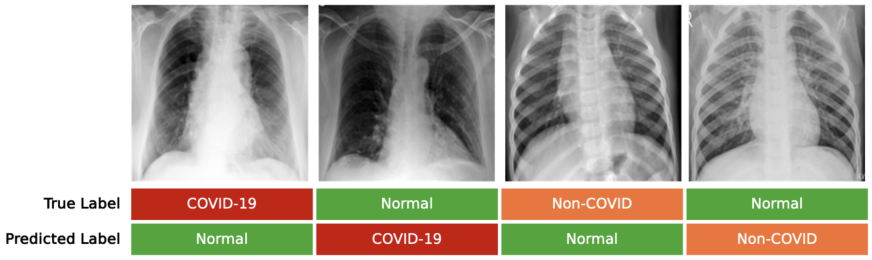

Figure 9: Example Misclassified Images for Logistic Regression

better suited for COVID-19 pneumonia due to their focus on edge and gradient patterns, potentially benefiting models like SVM that excel with clear feature separation (as observed with COVID-19 in the confusion matrices) over models like random forest.

These misclassifications further highlight limitations inherent in certain classification methods. For instance, SVM models employing the RBF kernel can be prone to overfitting, as they allow for highly complex decision boundaries that may become overly specific to the training dataset. While logistic regression demonstrated exceptional performance across our models, its assumption of linear feature relationships could hinder its ability to capture subtle, non-linear patterns in the data. While creating non-linear relationships through feature transformations is possible, the effectiveness of this approach might depend on the specific transformations chosen.

To further enhance our models' performance and robustness, future work should investigate ensemble methods that combine the predictions of multiple models, potentially leveraging their complementary strengths. Additionally, incorporating data augmentation techniques to enrich the dataset, particularly emphasizing early-stage pneumonia cases, could improve the models' ability to learn generalizable features.

4.3 Feature Importance

To examine factors influencing our models' decisions, we conducted Random Forest feature importance analysis to understand factors influencing our models' decisions. Random Forest feature importance analysis (Table 7, Table 8) underscores the value of our multi-modal feature approach. All three feature classes (HOG, ResNet, and Radiomics) contributed significantly to predictive power. While HOG features accounted for the largest overall importance (50.14%), ResNet (26.59%) and Radiomics (23.27%) features demonstrated higher average importance per feature. This highlights the distinct and complementary insights captured by each modality. The prominence of HOG features highlights the importance of edge and gradient patterns for identifying consolidations and opacities, which are hallmarks of pneumonia in chest X-rays. Radiomics features complement this by quantifying nuanced textural variations which may be especially useful for distinguishing pneumonia types. ResNet likely extracts additional complex patterns, encompassing both shape and texture, that are not fully captured by HOG or Radiomics alone. These insights, while derived from Random Forest, offer a valuable starting point for understanding the significance of our diverse feature set across all models.

	HOG	**Radiomic**	**Resnet**
Number of Features	1065	22	193
Sum of Feature Importance	50.14%	23.27%	26.59%
Average Importance per Feature	0.05%	1.06%	0.14%
Maximum Feature Importance	6.01%	2.91%	4.34%

Table 7: Summary of Random Forest Feature Importance, by Feature Set

The models' difficulty with cases where opacities have indistinct boundaries or atypical presentations, despite the presence of ResNet and Radiomics features, highlights the complexity of pneumonia diagnosis. This suggests that even with multi-modal features, there may be subtle visual cues that our current models are not fully capturing. Investigating the specific nature of these challenging cases could inspire the development of more refined features or the integration of additional modalities like clinical information. Exploring these avenues offers a promising direction for future research to enhance pneumonia classification models.

Rank of Feature	Feature Set	Feature Importance
1	HOG	6.0%
2	ResNet	4.3%
3	HOG	4.2%
4	ResNet	3.4%
5	Radiomics	2.9%
6	Radiomics	2.9%
7	Radiomics	2.2%
8	ResNet	2.2%
9	HOG	2.1%
10	HOG	2.1%

Table 8: Top 10 Important Random Forest Features

4.4 Generalizability

In image classification, a model's ability to correctly identify diseases or objects on new, unseen images, even those acquired differently or containing variations not in the training dataset, is termed generalizability. In our study, we evaluated generalizability primarily by assessing our models' performance on a held-out test dataset and measuring performance differences across training, validation, and test sets. A significant drop in performance on new data suggests overfitting, where the model has learned training-specific patterns rather than generalizable features.

Our models demonstrated promising generalization capabilities, evidenced by their strong performance on the held-out test set (Table 6). Logistic Regression and Gradient Boosting were particularly robust, maintaining stable accuracy across training, validation, and testing (with training-to-validation accuracy drops of 0.01). This suggests an ability to effectively classify new pneumonia cases, even with potential variations in image characteristics. While SVM and Random Forest exhibited greater differences between training and testing scores (with training-to-validation accuracy drops of 0.08 and 0.1, respectively) their overall performance on unseen data remains strong. These more significant drops suggest a greater degree of overfitting for SVM and Random Forest.

Despite these promising results, there is room to further improve generalizability, particularly for mitigating the overfitting observed in SVM and Random Forest. To further improve these models, future studies should prioritize several strategies. Data augmentation, by introducing artificial variations that mimic real-world image differences, can expand the training dataset and reduce overfitting. Additionally, hyperparameter tuning could benefit from incorporating an objective function that explicitly balances accuracy with minimizing the difference between training and validation scores. Nonetheless, even with these refinements, systematic validation of the models' generalizability across diverse settings is crucial.

The dataset used comprises 10 different studies, encompassing images from various platforms and healthcare systems (Qatar, Europe, etc.). This offers promising potential for generalizability across diverse settings. Manual investigation suggests representation of both genders. To validate generalizability and uncover potential biases, future studies should employ stratified cross-validation using metadata such as country, platform used, and patient demographics (gender, age, ethnicity). If, for instance, discrepancies arise between platforms, developing platform-specific models or integrating platform metadata into non-linear models could address such variations and improve performance.

4.5 Efficiency

Rapid and accurate diagnosis is essential in medical image analysis, especially for conditions like pneumonia where timely treatment influences outcomes. Efficiency is therefore crucial alongside accuracy for image classification models. This study, conducted in a standard Colab environment with CPU, highlights the importance of efficient models suitable for real-world deployment where computational resources might be limited.

Table 9 reveal significant insights into model efficiency across our dataset. Logistic Regression stands out, demonstrating the fastest tuning, training, and prediction times, while also achieving the highest accuracy and F1 scores. While Gradient Boost had considerably longer tuning and training times, its prediction speed was second only to Logistic Regression. Interestingly, PCA played a crucial role – without dimensionality reduction, Logistic Regression was unable to process the full feature set. Importantly, in the context of pneumonia detection, all prediction times across both 80% and 90% PCA variance levels fall within an acceptable range: the least efficient model, SVM at 90% PCA

variance achieves a prediction time of 7.7 milliseconds per image and non-SVM model predictions times are in the order of microseconds per image.

Model	Per Trial Tuning Time (s)	Training Time (s)	Prediction Time (s)			Accuracy	F1-Score
			Training Set	Validation Set	Test Set	Test Set	Test Set
Logistic Regression	63.36	16.97	0.02	0.00	0.01	0.90	0.90
SVM	63.08	20.89	50.10	11.79	16.20	0.89	0.89
Gradient Boost	2027.44	2089.13	0.30	0.06	0.08	0.85	0.85
Random Forest	53.44	111.38	0.60	0.14	0.17	0.82	0.85

Table 9: Tuned Model Efficiency, Accuracy and F1 Scores (PCA: 90% Variance).
Times shown are for 6,849 training images, 1,712 validation images, and 2,140 test images.

Table 9 shows the efficiency gains achieved through PCA. Moving from 90% to 80% explained variance resulted in faster training times for all models. Logistic Regression and Gradient Boosting exhibited the most significant reductions in training time, with Logistic Regression also demonstrating a slight improvement in prediction speed. SVM and Random Forest also benefited from PCA, with notable decreases in training and prediction times.

Model	Test Set Metrics	PCA Variance	
		90%, Tuned (1280 Dimensions)	**80%, Untuned (702 Dimensions)**
Logistic Regression	**Training Time (s)**	16.97	9.75
	Prediction Time (s)	0.01	0.01
	Accuracy	0.90	0.88
SVM	**Training Time (s)**	20.89	14.39
	Prediction Time (s)	16.20	10.45
	Accuracy	0.89	0.89
Gradient Boost	**Training Time (s)**	2089.13	1116.67
	Prediction Time (s)	0.08	0.05
	Accuracy	0.85	0.85
Random Forest	**Training Time (s)**	111.38	82.21
	Prediction Time (s)	0.17	0.15
	Accuracy	0.82	0.83

Table 10: Comparison of Model Efficiency and Accuracy at 90% and 80% PCA Variance.
Times shown are for 6,849 training images, 1,712 validation images, and 2,140 test images.

Interestingly, the impact of lowering PCA variance on accuracy varied (Table 10). While Logistic Regression experienced a minor decrease (from 0.90 to 0.88), SVM and Gradient Boosting maintained their accuracy levels. Random Forest saw a slight increase in accuracy (0.82 to 0.83), likely due to reduced overfitting with the lower-dimensional feature set.

Based on our efficiency experiments (Table 10), the most efficient model is Logistic Regression at 80% PCA variance, while the tuned Logistic Regression at 90% variance is the most accurate model. Overall, Logistic Regression at 90% PCA variance emerges as the optimal choice. Its fast prediction time (0.01s for 2,140 test images) and highest accuracy (0.90) make the minor efficiency trade-off worthwhile. Given these promising results, future studies should explore Logistic Regression with even higher PCA variance thresholds (e.g., 95%) to further optimize the balance between efficiency and accuracy.

5 Discussion and Conclusions

Our study demonstrates the potential of machine learning for accurate pneumonia classification using chest X-rays. Logistic Regression and SVM achieved particularly impressive results, with accuracies reaching as high as 0.90 across COVID-19, non-COVID pneumonia, and normal classes. However, the observed overfitting in SVM highlights the need to explore data augmentation and regularization techniques for improved generalizability if SVM were to be pursued.

Feature importance analysis underscores the value of our multi-modal approach. HOG effectively captured edge-based opacity patterns, while Radiomics quantified subtle textural variations, and ResNet likely extracted more complex patterns. This synergy suggests that diverse feature sets are crucial for robust pneumonia detection.

To further enhance performance, address misclassification challenges, and pave the way for real-world adoption, future work should prioritize several key strategies. Firstly, integrating readily available clinical data (e.g., patient symptoms and demographic factors) could provide the models with additional context for classifying challenging cases. Secondly, a thorough investigation of misclassified images is warranted. Analyzing the specific visual ambiguities that lead to errors would guide the development of more targeted feature engineering solutions. Additionally, we could explore higher levels of wavelet decomposition to capture more intricate visual cues, larger ResNet architectures for potentially richer representations, and higher PCA variance thresholds (e.g., 95%). These refinements could potentially boost model performance, especially in distinguishing between normal, early-stage pneumonia, and subtle atypical presentations.

Beyond these refinements to feature engineering, three key areas require further development for successful real-world adoption of such models. First, and perhaps most importantly, explainability is crucial for clinical adoption. Techniques like SHapley Additive exPlanations (SHAP) could provide valuable insights into the features driving predictions, fostering trust in the model's output and aiding clinical decision-making. SHAP analysis could reveal the specific image regions our models focus on, highlighting potential alignment or discrepancies with the visual cues used by clinicians [5]. Second, instead of binary classifications, our models could output probabilistic predictions for each class, empowering healthcare workers to make informed triage and referral decisions. Finally, despite encouraging initial findings regarding generalizability, rigorous validation across diverse settings remains crucial. Future studies should prioritize data augmentation, objective functions explicitly balancing accuracy and training-validation consistency, and stratified cross-validation using metadata (country, platform, demographics) to uncover biases. Addressing potential performance discrepancies across settings may necessitate platform-specific adaptations or metadata integration within the models.

In clinical practice, it is imperative to emphasize that such AI-powered models are intended to augment, not replace, the expertise of clinicians. Explainability plays a vital role in fostering this human-AI collaboration. For example, if the model outputs a near-certain probability for a particular pneumonia diagnosis, the clinician may confidently confirm it, streamlining the diagnostic process. Conversely, lower probability outputs can guide the clinician's decision-making, prompting them to consider additional investigations such as CT scans or second opinions. This collaborative approach leverages the strengths of both human judgment and the model's pattern recognition capabilities.

This work advances the field of AI-assisted pneumonia diagnosis, demonstrating its potential for improving patient outcomes. By further enhancing performance, addressing explainability, providing probabilistic outputs, and validating generalizability, future research can pave the way for the integration of such models into real-world clinical practice.

6 Appendix

Model	Precision			Recall		
	Training Set	Validation Set	Test Set	Training Set	Validation Set	Test Set
Logistic Regression	0.91	0.88	0.90	0.91	0.88	0.90
Random Forest	0.92	0.81	0.83	0.92	0.81	0.82
SVM	0.97	0.87	0.89	0.97	0.87	0.89
Gradient Boost	0.87	0.85	0.86	0.86	0.85	0.85

Table A1: Tuned Model Precision and Recall on Test Set (PCA: 90% Variance)

Model	Training Time (s)	Prediction Time (s)			Accuracy	F1-Score
		Training Set	Validation Set	Test Set	Test Set	Test Set
Logistic Regression	9.75	0.01	0.00	0.01	0.88	0.88
SVM	14.39	29.53	7.98	10.45	0.89	0.89
Gradient Boost	1116.67	0.15	0.03	0.05	0.85	0.85
Random Forest	82.21	0.50	0.11	0.15	0.83	0.83

Table A2: Untuned Model Efficiency, Accuracy and F1 Scores (PCA: 80% Variance).
Times shown are for 6,849 training images, 1,712 validation images, and 2,140 test images.

References

[1] Hui Juan Chen, Li Mao, Yang Chen, Li Yuan, Fei Wang, Xiuli Li, Qinlei Cai, Jie Qiu, and Feng Chen. Machine learning-based CT radiomics model distinguishes COVID-19 from non-covid-19 pneumonia. *BMC Infectious Diseases*, 21(1), 9 2021.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 6 2016.

[3] Michael Hollett and Daniel Bell. COVID-19. *Radiopaedia.org*, 1 2020.

[4] Govardhan Jain, Deepti Mittal, Daksh Thakur, and Madhup K. Mittal. A deep learning approach to detect covid-19 coronavirus with x-ray images. *Biocybernetics and Biomedical Engineering*, 40(4):1391–1405, 10 2020.

[5] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[6] David Luong and Craig Hacking. Pneumonia. *Radiopaedia.org*, 8 2015.

[7] César Ortiz-Toro, Angel García-Pedrero, Mario Lillo-Saavedra, and Consuelo Gonzalo-Martín. Automatic detection of pneumonia in chest x-ray images using textural features. *Computers in Biology and Medicine*, 145:105466, 6 2022.

[8] Mohammad Marufur Rahman, Sheikh Nooruddin, K. M. Azharul Hasan, and Nahin Kumar Dey. HOG + CNN net: Diagnosing COVID-19 and pneumonia by deep neural network from chest x-ray images. *SN Computer Science*, 2(5), 7 2021.

[9] J.C.L. Rodrigues, S.S. Hare, A. Edey, A. Devaraj, J. Jacob, A. Johnstone, R. McStay, A. Nair, and G. Robinson. An update on COVID-19 for the radiologist - A british society of thoracic imaging statement. *Clinical Radiology*, 75(5):323–325, 5 2020.

[10] Sana Salehi, Aidin Abedi, Sudheer Balakrishnan, and Ali Gholamrezanezhad. Coronavirus disease 2019 (COVID-19): A systematic review of imaging findings in 919 patients. *American Journal of Roentgenology*, 215(1):87–93, 7 2020.

[11] Wesley H. Self, D. Mark Courtney, Candace D. McNaughton, Richard G. Wunderink, and Jeffrey A. Kline. High discordance of chest x-ray and computed tomography for detection of pulmonary opacities in ED patients: implications for diagnosing pneumonia. *The American Journal of Emergency Medicine*, 31(2):401–405, 2 2013.

[12] Anas M. Tahir, Muhammad E.H. Chowdhury, Amith Khandakar, Tawsifur Rahman, Yazan Qiblawey, Uzair Khurshid, Serkan Kiranyaz, Nabil Ibtehaz, M. Sohel Rahman, Somaya Al-Maadeed, Sakib Mahmud, Maymouna Ezeddin, Khaled Hameed, and Tahir Hamid. COVID-19 infection localization and severity grading from chest x-ray images. *Computers in Biology and Medicine*, 139:105002, 12 2021.

[13] Joost J.M. van Griethuysen, Andriy Fedorov, Chintan Parmar, Ahmed Hosny, Nicole Aucoin, Vivek Narayan, Regina G.H. Beets-Tan, Jean-Christophe Fillion-Robin, Steve Pieper, and Hugo J.W.L. Aerts. Computational radiomics system to decode the radiographic phenotype. *Cancer Research*, 77(21):e104–e107, 10 2017.