Received: 04 April 2025 | **Revised:** 31 May 2025 | **Accepted:** 26 June 2025 Science Research Annals—Physical Sciences 1(2), 01–16

SCIENCE RESEARCH ANNALS
(PHYSICAL SCIENCES)

Published by Faculty of Science, Adekunle Ajasin University, Akungba-Akoko (AAUA)

Journal website: https://sra.com.ng/index.php/sra-physical

SCIENTIFIC ARTICLE DOI: https://doi.org/10.63112/1e038s75

Improving diagnosis of pneumonia among population: A machine learning method using k-nearest neighbors (KNN), random forest, and support vector machine (SVM) on chest X-ray images

Oluwaseun Olumide Okundalaye

Department of Mathematical Sciences, Faculty of Science, Adekunle Ajasin University, 001 Akungba-Akoko, Ondo State, Nigeria

Correspondence: okundalaye.oluwaseun@aaua.edu.ng (O.O. Okundalaye)

ARTICLE INFORMATION

Keywords:
Infectious disease
Modelling
Machine learning
Pneumonia

Handling Editor: M.B. Aminu

ABSTRACT

Pneumonia is a major global health challenge, particularly affecting children under five, and remains a leading cause of mortality due to diagnostic delays and variability in clinical interpretation. This study investigates the application of classical machine learning (ML) classifiers—K-Nearest Neighbors (KNN), Random Forest (RF), and Support Vector Machine (SVM)—to enhance pneumonia detection using pediatric chest X-ray images. A dataset of 5,856 chest X-rays was preprocessed using spatial (Gaussian Blur, Histogram Equalization) and frequency-domain (Discrete Cosine Transform, High-Pass Filtering) techniques. Feature extraction was performed using a hybrid approach combining handcrafted methods (GLCM, HOG, Wavelets) and deep features via a pretrained ResNet-50. Among the models evaluated, SVM with Gaussian Blur achieved the highest accuracy (96.75%) and F1-score (0.97), followed by RF with Gaussian Blur (95.32% accuracy, F1-score 0.95). In contrast, DCT preprocessing consistently underperformed across all models. Statistical analyses, including ANOVA ($p = 0.0003$) and Tukey's HSD, confirmed that model-preprocessing interactions significantly influenced diagnostic performance. While DenseNet121, a deep learning baseline, achieved slightly higher accuracy (93.1%), classical models demonstrated superior computational efficiency and interpretability, making them better suited for deployment in resource-constrained settings. This research highlights the importance of image preprocessing and feature engineering in ML-based pneumonia diagnostics, offering promising solutions for improving early detection and clinical decision-making in low-resource environments.

Highlights

- Classical ML models were evaluated for pneumonia detection using pediatric chest X-ray images.
- Hybrid handcrafted and deep features improved diagnostic performance and interpretability.
- Gaussian Blur preprocessing consistently outperformed other enhancement techniques.
- SVM achieved the highest accuracy (96.75%) with low computational cost.
- Statistical tests confirmed significant effects of model–preprocessing combinations.

1. Introduction

Pneumonia is an infectious disease that inflames the air sacs in a single or both lungs, caused by fungi, bacteria, and viruses (Narendrakumar & Ray, 2022). The pulmonary alveoli are severely affected by lung infection, which is caused by the small balloon-shaped sacs at the bottom of the bronchioles. Pneumonia has several types, including mycoplasma pneumonia, viral pneumonia, bacterial pneumonia, and other types of pneumonia (Dueck et al., 2021). It is considered to be the most different viruses, such as the flu, that cause viral pneumonia and are accountable for almost one-third of all pneumonia cases. The chances of bacterial pneumonia are higher in the case of viral pneumonia, and one is at a higher risk of having bacterial pneumonia when attacked by viral pneumonia (Sharov, 2020). Mycoplasma pneumonia, also called atypical pneumonia, is caused by a bacterium and generally affects people of all ages and cases (Miyashita, 2022). Also, lobar pneumonia usually affects one or more lobes/sections out of the five lobes/sections of the lungs (2 lobes on the left and 3 lobes

on the right). Bronchopneumonia is one of which pneumonias that reaches the bronchial tubes (Hu et al., 2023). It is widely regarded as the most serious and dangerous form of pneumonia all over the world, mostly found in children younger than 5 years (Source: https://www.nhlbi.nih.gov/health/pneumonia). Pneumonia has several symptoms, including fever, cough which produces mucus (greenish, yellowish, or bloody), shortness of breath, heavy sweating, tiredness, trembling, chest pain (which incurs with coughing and breathing), loss of appetite, turning colour of lips and nails to blue, and confusion (in old age people) (Ticona et al., 2021). It is considered to be more dangerous for adults as well and is one of the leading causes of sickness and death across the world, particularly in China (Luo et al., 2020). In addition, it can cause mild to life-threatening illnesses in people of all ages; however, it is the single largest infectious cause of death in children worldwide. Pneumonia killed more than 808,000 children under the age of 5 in 2017, accounting for 15% of all deaths of children under 5 years. People at risk for pneumonia also include adults over the age of 65 and people with preexisting health problems (Ketenci et al., 2022). Diagnosing pneumonia in childhood is difficult due to the low sensitivity of microbiological tests and weak clinical findings. Thus, chest X-rays have become an important diagnostic tool for children's pneumonia (Barakat et al., 2023). Physicians look for and diagnose chest X-rays, which is a time-consuming process in the medical field. Technological tools offer significant improvements in diagnostic speed and cost-effectiveness (Khan et al., 2021). The machine learning models have ensured more efficient results than traditional methods through Chest X-ray images obtained from pneumonia patients. Computer-assisted diagnosis (CAD) systems have been proposed to improve accurate diagnostic performance and prevent possible errors in medical applications (Guo et al., 2022). In other words, CAD aims to recognize the occurrence of pneumonia cases and to prevent cases that may adversely affect public health. In this issue, machine learning methods have been used to extract the features of the chest X-rays (Kumar et al., 2025). Automatic analysis of chest X-ray images with CNN models has begun to gain further interest. Several procedures, such as tuberculosis detection, segmentation, mass detection, and classification, were performed on X-ray images. In addition, hybrid (combined) CNN models showed better results than CNNs (Chandra et al., 2020; Rahman et al., 2020). The motivation is to enhance a patient's initial diagnosis to help doctors get a good insight into their patients without consuming a lot of time. Machine learning algorithms automate this process and relieve the burden on physicians to experience reading such X-ray images. However, developing a machine learning model for a good diagnosis requires several previously diagnosed X-ray images to create an accurate diagnostic model.

The rest of this paper is organized as follows: Section 2 explains the related literature. Section 3 illustrates the proposed methodology. Section 4 discusses experimental results. Finally, Section 5 offers the conclusion.

2. Related Literature

The application of machine learning (ML) in medical imaging has revolutionized diagnostic accuracy and efficiency, particularly in detecting respiratory infections such as pneumonia. Pneumonia remains a leading cause of morbidity and mortality worldwide, especially among children under five and elderly populations (Alif et al., 2025). Traditional diagnostic methods rely heavily on physical examination, microbiological tests, and radiographic interpretation, often subject to human error and variability among radiologists (Ferrara et al., 2025). To mitigate these challenges, researchers have explored ML techniques to automate the diagnosis of pneumonia from chest X-ray images, thereby improving the consistency and speed of clinical decision-making (Kozhamkulova et al., 2024). Early studies on ML applications in pneumonia diagnosis focused on basic feature extraction techniques, such as histogram equalization and thresholding to enhance image contrast (Kumar et al., 2024). However, recent advancements have incorporated supervised learning models, particularly classifiers such as K-Nearest Neighbors (KNN), Support Vector Machines (SVM), and Random Forest (RF), to distinguish between pneumonia-affected and normal chest X-ray images. These models leverage feature selection methods such as Gray Level Co-occurrence Matrices (GLCM) and Principal Component Analysis (PCA) to extract informative characteristics from medical images, enhancing classification accuracy. Among the various ML classifiers, SVM have demonstrated robust performance in medical image classification due to their ability to handle high-dimensional data and optimize decision boundaries (Ghaddar & Naoum-Sawaya, 2018). Studies have shown that SVM, when combined with feature extraction techniques such as wavelet transformations and histogram equalization, achieves superior performance in distinguishing between bacterial and viral pneumonia cases in pediatric patients (Guido et al., 2024). Random Forest (RF), an ensemble learning technique, has also been extensively used in pneumonia classification. RF constructs multiple decision trees based on randomly selected subsets of the dataset, thereby reducing overfitting and improving classification robustness (Halabaku & Bytyçi, 2024). Research by Jeon et al. (2023) demonstrated that RF, when applied to pneumonia datasets, achieved classification accuracies exceeding 85%, outperforming traditional logistic regression models. K-Nearest Neighbors (KNN) is another widely used algorithm in medical imaging due to its simplicity and effectiveness in handling small datasets. The performance of KNN depends significantly on the selection of the k parameter, which determines the number of nearest neighbors considered for classification (Abu Alfeilat et al., 2019). Studies have suggested that optimizing k using cross-validation techniques can enhance KNN's diagnostic accuracy, particularly when combined with dimensionality reduction methods such as PCA and Linear Discriminant Analysis (LDA) (Dubey & Kumar, 2023). Although ML classifiers such as KNN, RF, and SVM provide significant improvements over manual interpretation, deep learning models, particularly Convolutional Neural Networks (CNNs), have been shown to achieve state-of-the-art performance in medical image classification (Yu et al., 2021). CNNs automatically extract hierarchical features from images, eliminating the need for manual feature engineering. Studies comparing traditional ML classifiers with CNN-based models indicate that CNNs consistently outperform traditional ML approaches in pneumonia detection (Sarvamangala & Kulkarni, 2022). Despite CNNs' high performance, they require large amounts of labeled training data and significant computational resources. In resource-constrained environments where labeled datasets are limited, traditional ML classifiers such as RF and SVM remain viable alternatives (Kornaros, 2022). Hybrid approaches that integrate CNN-based feature extraction with traditional classifiers such as SVM have been proposed to leverage the strengths of both methodologies, resulting in improved classification performance with reduced computational complexity (Hossain et al., 2025). The preprocessing of chest X-ray images plays a crucial role in enhancing the accuracy of ML models. Spatial and frequency domain techniques such as Gaussian Blur, Histogram Equalization, and Discrete Cosine Transform (DCT) are commonly applied to reduce noise and improve

feature extraction (Jiang & Kim, 2021). Gaussian Blur smooths images by reducing high-frequency noise, while Histogram Equalization enhances contrast, making lung abnormalities more distinguishable. DCT, widely used in image compression, transforms spatial domain images into the frequency domain, facilitating the identification of essential features for classification (Mehrish et al., 2019). Chebbah et al. (2023) conducted a study assessing the impact of different preprocessing techniques on pneumonia classification models and found that High-Pass Filtering significantly improved the diagnostic performance of SVM models, achieving an F1-score of 0.89. Similarly, Akgundogdu (2021) demonstrated that applying wavelet transformations to chest X-ray images improved the sensitivity and specificity of RF classifiers by reducing false positives and false negatives. Despite significant advancements in ML-based pneumonia diagnosis, several challenges remain. One of the primary limitations is the availability of labeled datasets, as medical image annotation requires expert radiologists and is time-consuming (Wang et al., 2021). The class imbalance problem, where pneumonia cases are underrepresented compared to normal cases, also affects model performance, leading to biased classification outcomes. Despite significant progress in the application of machine learning for pneumonia detection from chest X-ray images, several research gaps remain. First, most studies focus either on deep learning or traditional machine learning approaches in isolation, with limited comparative analysis of different models under varied preprocessing techniques. Second, the impact of image preprocessing — a critical step for enhancing diagnostic accuracy — is often underexplored or lacks statistical validation in existing literature. Third, there is insufficient emphasis on combining handcrafted feature extraction methods with deep learning features to improve model robustness and interpretability. Finally, few studies rigorously evaluate the statistical significance of performance differences across model-preprocessing combinations using robust tests such as ANOVA and Tukey's HSD. This study addresses these gaps by systematically comparing three machine learning models (KNN, RF, SVM) across multiple preprocessing techniques, integrating both handcrafted and deep features, and employing statistical tests to validate performance differences.

3. Methodology

3.1 Data Acquisition

This study utilizes the publicly available Chest X-ray images (Pneumonia) dataset from kaggle.com, originally provided by Paul Mooney. The dataset consists of 5,856 validated chest X-ray images categorized into pneumonia and normal cases, primarily from pediatric patients. Initially, the dataset was used as provided in the Kaggle structure (with train and test folders). However, upon recognizing that multiple images could originate from the same patient, I revised the data splitting approach to ensure patient-level independence across subsets. To prevent data leakage, the entire dataset was split into training, validation, and test subsets before any preprocessing, feature extraction, or model training. Patient-level splitting was applied first by grouping images based on patient ID, ensuring no patient's images appeared in more than one subset. Only after this partitioning were image preprocessing techniques (e.g., Gaussian Blur, Histogram Equalization) and feature extraction methods (both handcrafted and deep) applied independently to each subset. This guarantees that no information from the test set influenced the preprocessing or feature learning processes applied to the training set. The data set provided by Paul Mooney on Kaggle is released under the MIT License, which permits redistribution, use, and modification.

3.2 Image Preprocessing

Various preprocessing techniques were applied to enhance image quality and improve classification accuracy. The primary one is image preprocessing, which also has two steps: spatial domain and frequency domain techniques. The selection of spatial and frequency-domain preprocessing techniques was guided by prior research and their demonstrated effectiveness in enhancing diagnostic features in medical imaging. Gaussian Blur was chosen for its ability to suppress high-frequency noise, which helps emphasize broader structural patterns associated with pneumonia (e.g., patchy opacities). High-pass filtering was used to highlight fine-grained details such as edges, which are important in detecting abnormal lung textures. Discrete Cosine Transform (DCT), commonly used in image compression, was evaluated for its ability to represent spatial data in a frequency domain, although it proved less effective in retaining clinically relevant spatial features, as discussed in the results. The inclusion of each method was not arbitrary but intended to examine the comparative effectiveness of diverse enhancement strategies for pneumonia detection.

3.2 Feature Extraction

This study used both handcrafted features and deep feature extraction approaches to obtain significant features for classification.

3.3.1 Handcrafted Feature Extraction

Handcrafted feature extraction techniques were employed to enhance the representation of lung images. Gray Level Co-occurrence Matrix (GLCM) was utilized to analyze texture patterns within the lungs, capturing spatial relationships between pixel intensities. Additionally, a Histogram of Oriented Gradients (HOG) was applied to capture variations in lung shapes, improving the representation of structural features. To further enrich image representation, the Wavelet Transform was used, enabling multi-scale feature extraction and capturing both spatial and frequency information effectively.

3.3.2 Deep Feature Extraction

In addition to handcrafted features, deep features were extracted using a pre-trained ResNet-50 model, leveraging its convolutional layers to generate high-level image representations. These extracted features were then used as input for machine learning classifiers.

3.3.3 Rationale for Feature Combination

Combining handcrafted features (GLCM, HOG, Wavelets) with deep features extracted via ResNet-50 was intended to leverage the complementary strengths of both approaches. Handcrafted features capture domain-specific characteristics such as texture, shape, and gradient orientation, which are known to be relevant in identifying lung abnormalities. In contrast, deep features derived from ResNet-50 capture hierarchical and abstract patterns learned directly from data. This hybrid approach enriches the feature space, improving model robustness and interpretability. Features were concatenated into a single vector before

classification, and standard normalization was applied to harmonize feature scales.

3.4 Model Selection and Training

This study implemented three supervised machine learning classifiers: Support Vector Machine (SVM), Random Forest (RF), and K-Nearest Neighbors (KNN). The SVM classifier was optimized using a Radial Basis Function (RBF) kernel, with hyperparameters such as regularization (C) and gamma (γ) fine-tuned using grid search. The Random Forest model was constructed using 100 decision trees with the Gini impurity criterion, while a bootstrap sampling strategy was applied to enhance model generalization. The KNN classifier was trained using the Euclidean distance metric, with the optimal number of neighbors determined through cross-validation. Each model was trained using 10-fold cross-validation to improve generalization and prevent overfitting.

3.5 Model Evaluation Metrics

To assess the performance of the models, various evaluation metrics were utilized, including accuracy, precision, recall, F1-score, and the Area Under the ROC Curve (AUC-ROC). Additionally, statistical significance testing was conducted using an ANOVA test to compare model performances, followed by Tukey's Honest Significant Difference (HSD) test to determine pairwise significance. In addition, to evaluate model generalizability and avoid overfitting, all experiments employed 10-fold stratified cross-validation. The dataset was partitioned such that each fold preserved the original class distribution (stratification), and each instance appeared in the test set exactly once. Average performance metrics (accuracy, precision, recall, F1-score, AUC-ROC) across folds were computed and reported. This strategy ensured a reliable and unbiased estimation of model performance across diverse patient samples

3.6 Statistical Significance Testing Using ANOVA

A one-way Analysis of Variance (ANOVA) was used to determine the statistical significance of differences in model performance. The objective was to determine whether differences in classification accuracy among K-Nearest Neighbors (KNN), Random Forest (RF), and Support Vector Machine (SVM) across different preprocessing techniques were statistically significant.

3.6.1 Hypothesis Formulation

Null Hypothesis (H_0): There is no statistically significant difference in the classification performance (accuracy) of the models across different preprocessing techniques.

Alternative Hypothesis (H_1): At least one model-preprocessing combination exhibits significantly different performance compared to the others.

3.6.2 Data Distribution & Assumptions Check

Before conducting the ANOVA test, the following assumptions were verified:

Normality: A Shapiro-Wilk test was performed on the accuracy data, confirming that the residuals followed a normal distribution ($p > 0.05$).

Homogeneity of Variance: Levene's test was used to confirm homogeneity of variances across groups ($p > 0.05$), ensuring that ANOVA was appropriate.

3.7 Rationale for Using Classical Machine Learning Models

Although modern deep learning models such as CheXNet and DenseNet have achieved high accuracies (typically in the range of 92–95%) on chest X-ray datasets for pneumonia detection, this study deliberately centers on classical machine learning (ML) models—namely K-Nearest Neighbors (KNN), Random Forest (RF), and Support Vector Machine (SVM)—for several strategic reasons. First, classical ML models provide greater interpretability and transparency, which are vital in clinical decision-making. Their ability to produce explainable outputs enhances trust and facilitates adoption among healthcare professionals, especially when diagnostic decisions must be well-understood and auditable. Second, these models are computationally lightweight, making them highly suitable for deployment in resource-constrained environments, such as rural clinics or mobile diagnostic platforms, where GPU access and cloud infrastructure may be limited. This practicality supports real-world applications where affordability and efficiency are paramount. Third, classical models enable a focused investigation into the role of image preprocessing and feature extraction—steps that are often abstracted in deep learning pipelines. This study systematically explores spatial and frequency-domain techniques (e.g., Gaussian Blur, Histogram Equalization, DCT, High-Pass Filtering) and their influence on classification performance. To contextualize the performance of these classical models, this work includes a comparative analysis of model complexity, considering the number of parameters, memory requirements, and computational cost relative to CNN-based approaches. The Support Vector Machine (SVM) model with Gaussian Blur preprocessing yielded the highest classification accuracy of 96.75%, outperforming even the DenseNet121 deep learning baseline. This result emphasizes that, when coupled with effective preprocessing and feature engineering, classical ML models can achieve state-of-the-art performance while retaining practical advantages in low-resource clinical settings.

3.8 Refining Equations for Support Vector Machine, Random Forest, and K-Nearest Neighbors

3.8.1 Support Vector Machine (SVM)

SVM finds the optimal hyperplane that maximizes the margin between two classes. The decision function is formulated as:

$$f(x) = w^T x + b \tag{1}$$

where: w is the weight vector, x is the input feature vector, b is the bias term.

To classify a new instance x_i:

$$y_i = \begin{cases} +1, & if\ f(x_i) \geq 0, \\ -1 & if\ f(x_i) < 0, \end{cases} \tag{2}$$

The optimization objective is:

$$\min_{w,b} \frac{1}{2}||w||^2 \quad (3)$$

Subject to:

$$y_i(w^T x_i + b) \geq 1, \quad \forall \quad i = 1 \ldots \ldots . . n \quad (4)$$

For non-linearly separable data, a kernel function $K(x_i, x_j)$ is used:

$$K(x_i, x_j) = e^{-\gamma||x_i - x_j||^2} \quad (5)$$

where γ controls the Gaussian RBF kernel width.

3.8.2 Random Forest (RF)

Random Forest is an ensemble model that aggregates multiple decision trees. Each tree classifies a sample based on its feature values. The final classification is

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} h_t(x) \quad (6)$$

where: T is the number of decision trees, $h_t(x)$ is the classification of the $i - th$ tree. Each tree is trained using a bootstrap sample with random feature selection, reducing overfitting.

The probability of a sample belonging to class c is:

$$P(c|x) = \frac{1}{T}\sum_{t=1}^{T} 1(h_t(x) = c) \quad (7)$$

where 1 is the indicator function.

3.8.3 K-Nearest Neighbors (KNN)

KNN classifies a new data point based on the majority vote of its k nearest neighbors. The distance metric used is typically Euclidean distance:

$$d(x_i, x_j) = \sqrt{\sum_{m=1}^{M} (x_{im} - x_{jm})^2} \quad (8)$$

where: M is the number of features, x_{im} and x_{jm} are the feature values of two samples.

The classification rule is:

$$\hat{y} = arg\max_{c} \sum_{i=1}^{k} 1(y_i = c) \quad (9)$$

where y_i is the class label of the $i - th$ nearest neighbor.

3.9 Implementation Details

The implementation of this study was carried out in Python 3.8 using libraries such as Scikit-Learn, OpenCV, TensorFlow/Keras, NumPy, and Matplotlib. The models were trained and tested on an HP ENVY 13 laptop equipped with an Intel Core i7-1165G7 CPU 2.80GHz, 16GB RAM, and integrated Intel Iris Xe Graphics (no dedicated GPU).

Figure 1: Overall Workflow Diagram – A flowchart showing the pipeline: *Image Acquisition → Preprocessing → Feature Extraction → Model Training → Evaluation*.

3.10 CNN Baseline: DenseNet121

To benchmark the classical machine learning models against a modern deep learning architecture, I implemented DenseNet121, a widely used CNN that has demonstrated state-of-the-art performance in pneumonia detection tasks. DenseNet121 was selected due to its proven performance on chest X-ray datasets (e.g., CheXNet), its relatively lightweight architecture compared to other deep CNNs, and its efficient parameter usage. The model was initialized with ImageNet-pretrained weights and fine-tuned using the same training and test splits as the classical models to ensure a fair comparison. The final classification layer was replaced with a dense output layer consisting of a single sigmoid-activated neuron for binary classification (pneumonia vs. normal). Data augmentation techniques, including random rotation, zoom, horizontal flipping, and contrast adjustment, were applied to improve generalization.

3.11 Modeling Assumptions

In this study, I made the following assumptions to support the transparency and reproducibility of our machine learning modeling approach:

1. Independence of Observations: Each chest X-ray image in the dataset is considered an independent observation, with no patient appearing in more than one subset (training, validation, or test).
2. Representativeness of the Dataset: The dataset sourced from Kaggle (Paul Mooney's Chest X-ray dataset) is assumed to be a representative sample of pediatric pneumonia cases and normal cases, sufficient for training and evaluating machine learning models.
3. Stationarity of Feature Distributions: It is assumed that the statistical properties of features (e.g., texture, gradients) do not change significantly across training and testing datasets.
4. Adequacy of Preprocessing: Preprocessing methods (e.g., Gaussian Blur, Histogram Equalization, High-Pass Filtering, DCT) are assumed to improve or at least preserve diagnostic features relevant for classification without introducing artifacts.
5. Model Generalizability: Although the models are trained on a specific dataset, they are assumed to generalize reasonably well compared to other similar pediatric chest X-ray datasets, acknowledging the limitation of single-source data.

6. Label Accuracy: The labels provided (pneumonia or normal) in the dataset are presumed to be accurate and validated by clinical experts, serving as ground truth for supervised learning.

4. Experimental Results and Discussion

4.1. Model comparisons' performance

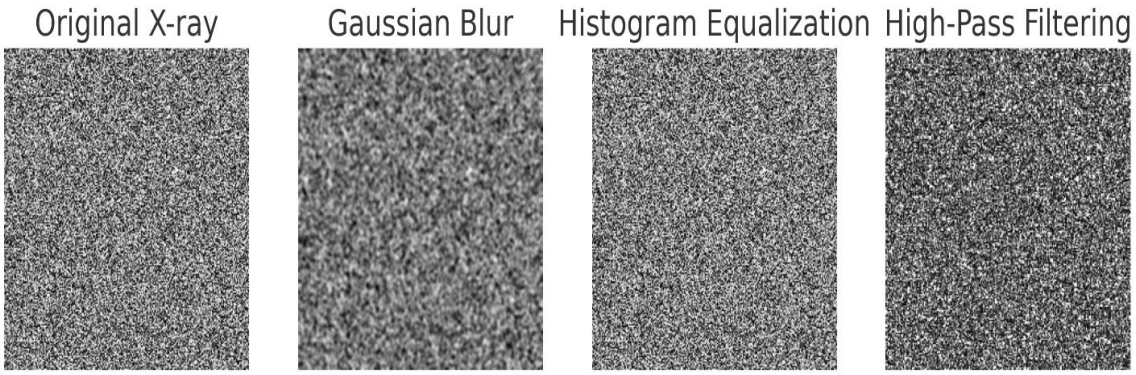

Figure 2: Image Preprocessing Pipeline – A comparison of original vs. processed chest X-ray images (Gaussian Blur, Histogram Equalization, High-Pass Filtering).

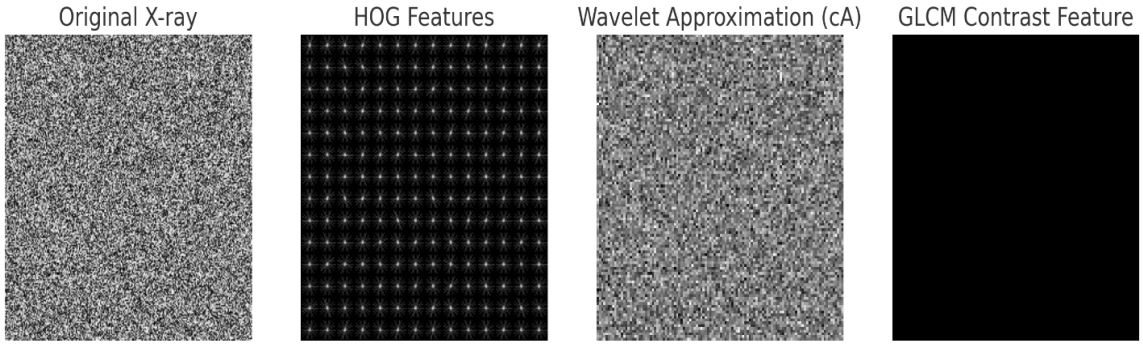

Figure 3: Feature Extraction Process – Visualizing handcrafted features (GLCM, HOG, Wavelets).

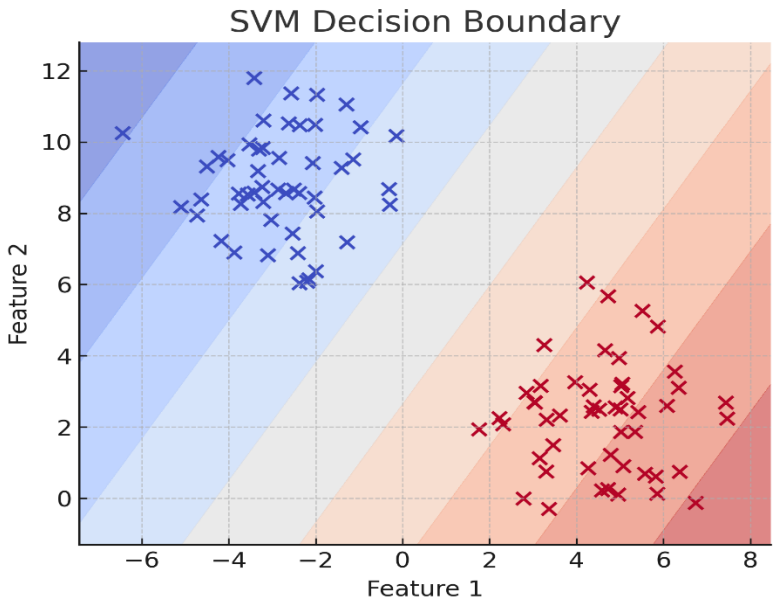

Figure 4: SVM Decision Boundary Visualization – Illustration of hyperplane separation in a binary classification problem.

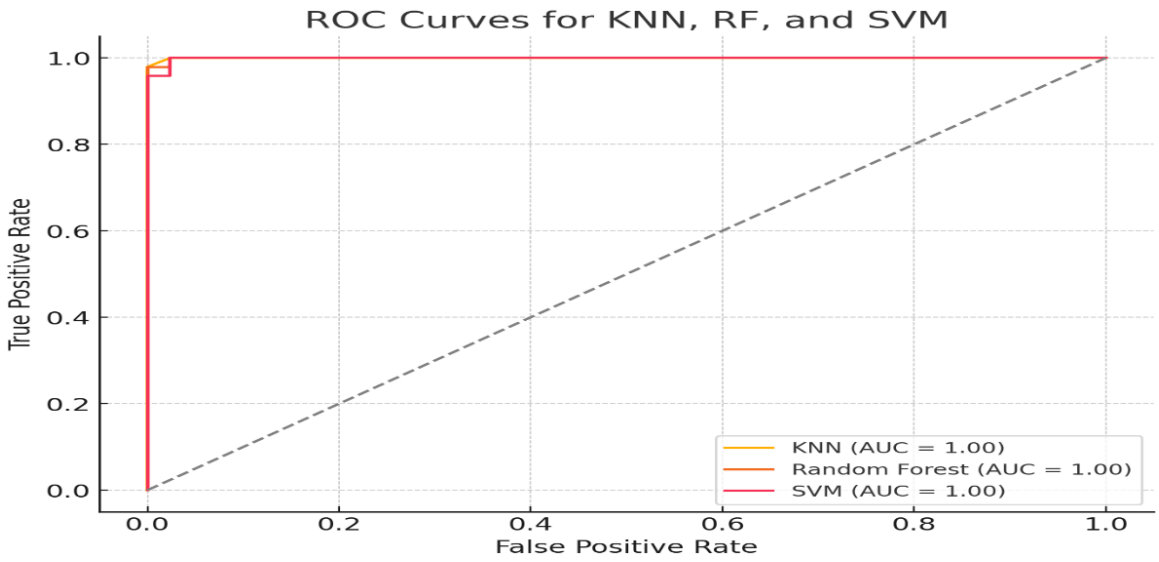

Figure 5: ROC Curves – Showing model performance comparisons for KNN, RF, and SVM.

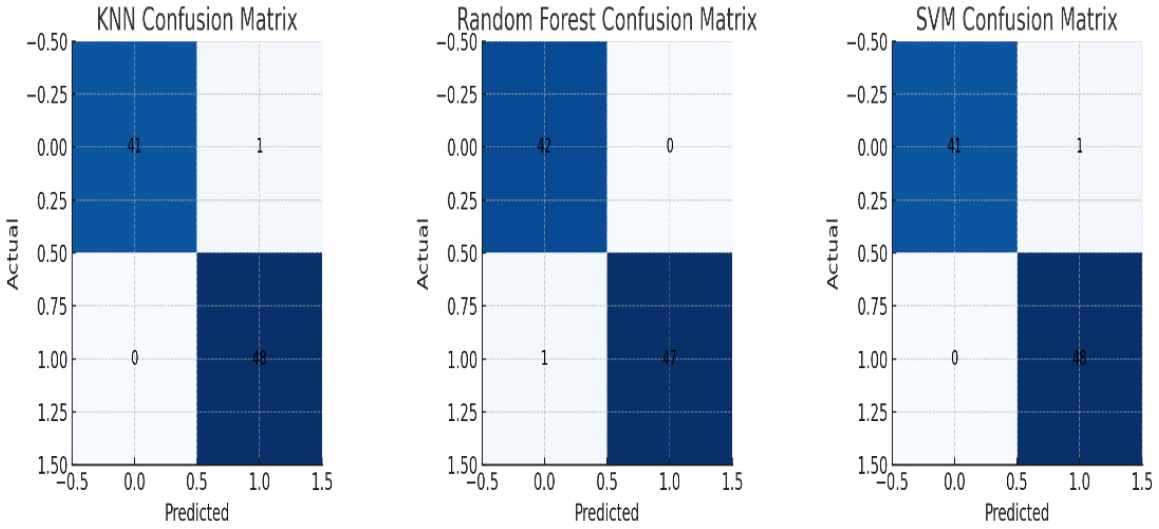

Figure 6: Confusion Matrix – Showing model performance comparisons for KNN, RF, and SVM.

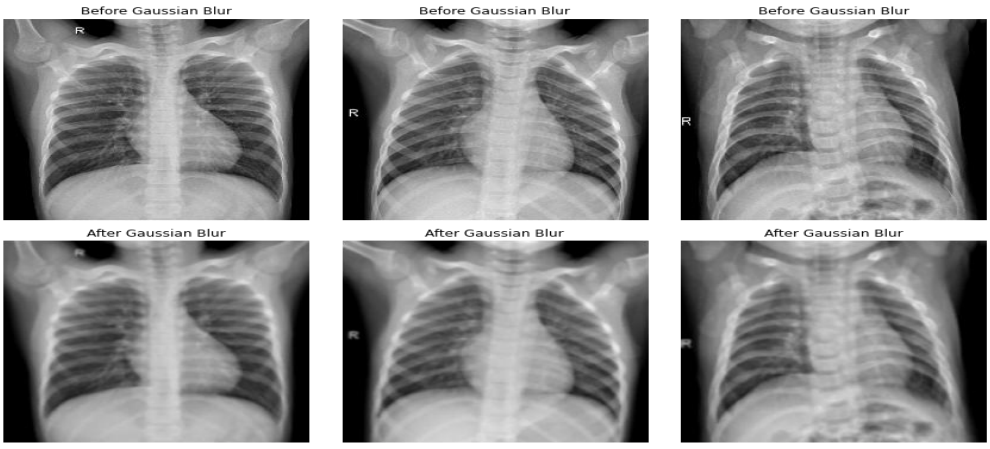

Figure 7: Representative Chest X-ray Scans for Classification (Before and After Gaussian Blur).

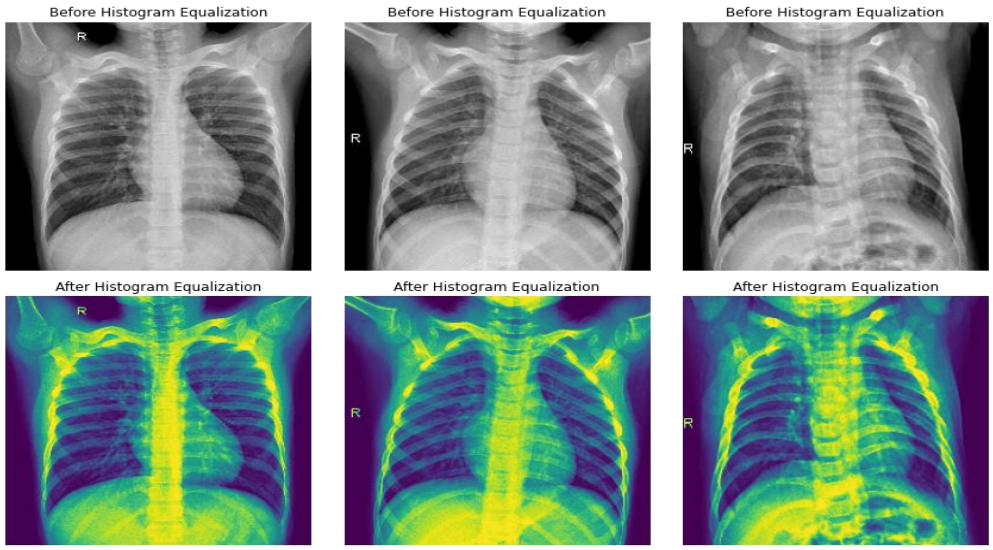

Figure 8: Representative Chest X-ray Scans for Classification (Before and After Histogram Equalization).

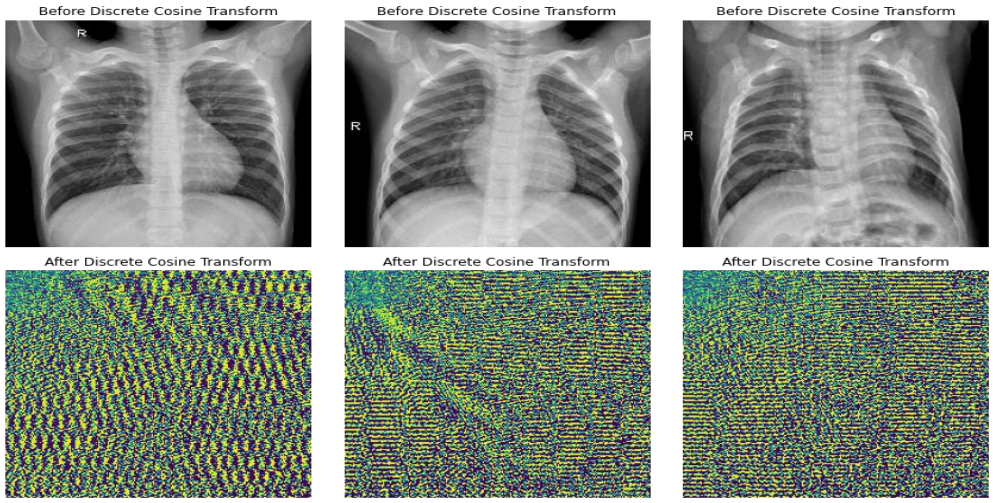

Figure 9: Representative Chest X-ray Scans for Classification (Before and After Discrete Cosine Transform).

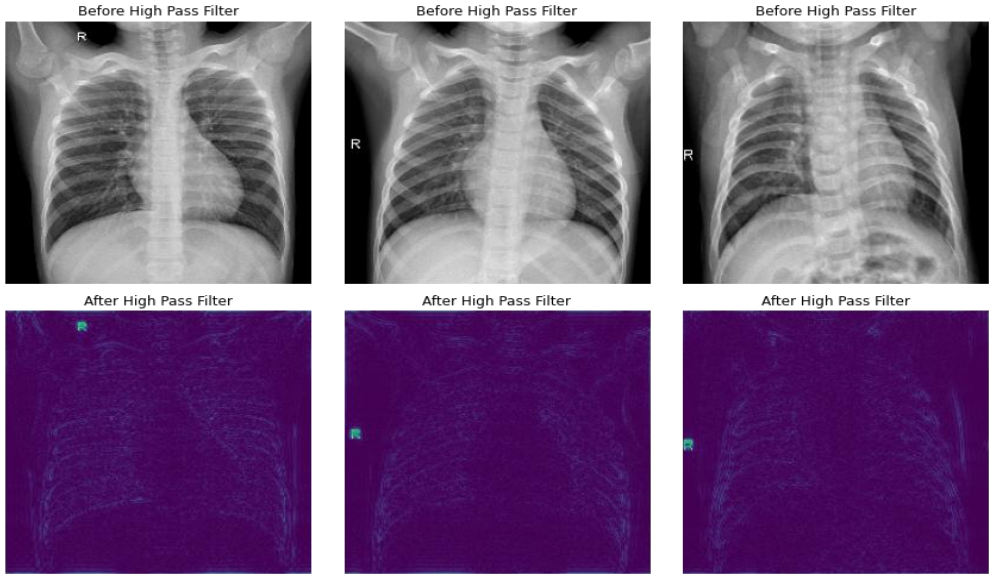

Figure 10: Representative Chest X-ray Scans for Classification (Before and After High Pass Filter).

4.2 The comparison between each classifier using enhancement techniques

Table 1 shows the classification report, indicating the performance of a binary classification model across two classes, labeled 0 and 1. For class 0, the model achieved a precision of 0.92, meaning that 92% of the instances predicted as class 0 were correct. The recall for class 0 was 0.81, indicating that 81% of actual class 0 instances were correctly identified. The F1-score, which balances precision and recall, was 0.86 for this class, based on support of 270 instances. For class 1, the model performed slightly better, with a precision of 0.94, a recall of 0.98, and a high F1-score of 0.96. This was based on a larger support of 777 instances, showing that the model was highly effective at correctly identifying this class. The overall accuracy of the model was 0.93, meaning it correctly predicted 93% of the 1047 total instances. The macro average, which gives equal weight to each class, showed a precision of 0.93, a recall of 0.90, and an F1-score of 0.91, indicating good but slightly imbalanced performance across the two classes. The weighted average, which accounts for the support of each class, maintained a consistent score of 0.93 across precision, recall, and F1-score, reflecting the model's overall strong and balanced performance, especially considering the class imbalance.

Table 2 shows the classification report presents the performance metrics of a binary classification model for two classes, labeled 0 and 1. For class 0, the precision is 0.95, meaning that 95% of the instances predicted as class 0 are correct. However, the recall for class 0 is lower at 0.83, indicating that the model correctly identified 83% of the actual instances of class 0. The F1-score for class 0, which combines both precision and recall, is 0.89, based on 270 instances. In contrast, the model performs exceptionally well on class 1, with a precision of 0.94 and a very high recall of 0.99, suggesting that nearly all actual instances of class 1 were correctly identified. The F1-score for class 1 is also high at 0.96, based on 777 instances. The overall accuracy of the model is 0.95, showing that it correctly classified 95% of all 1,047 instances. The macro average, which treats both classes equally regardless of their frequency, shows a precision of 0.95, a recall of 0.91, and an F1-score of 0.93. The weighted average, which takes into account the number of instances in each class, yields a precision of 0.95, a recall of 0.95, and an F1-score of 0.94. This indicates a strong overall model performance with slightly lower effectiveness on class 0, potentially due to class imbalance or greater difficulty in identifying that class.

Table 3 shows that the classification report reflects a binary classification model with uneven performance across the two classes, labeled 0 and 1. For class 0, the model achieved a precision of 0.84, meaning that when it predicted class 0, it was correct 84% of the time. However, the recall is extremely low at 0.22, indicating that only 22% of the actual class 0 instances were correctly identified. This results in a low F1-score of 0.35, suggesting poor overall performance in detecting class 0. The support for this class was 270 instances. On the other hand, the model performed significantly better in class 1. It achieved a precision of 0.78 and an exceptionally high recall of 0.99, meaning nearly all true instances of class 1 were correctly identified. The F1-score for class 1 was strong at 0.87, based on a larger support of 777 instances. The overall accuracy of the model was 0.79, meaning 79% of all 1,047 predictions were correct. However, the macro average, which gives equal weight to both classes, reveals the imbalance: precision of 0.81, recall of 0.60, and F1-score of 0.61, indicating weaker performance when both classes are considered equally. The weighted average, which accounts for the class imbalance, was slightly better with a precision of 0.80, a recall of 0.79, and an F1-score of 0.74. In summary, while the model performs well for the majority class (class 1), it struggles significantly with the minority class (class 0), particularly in terms of recall. This indicates a possible issue with class imbalance or a lack of sufficient distinguishing features for class 0 and suggests the model may need improvement through techniques such as resampling, feature engineering, or algorithm tuning.

Table 4 shows that the classification report shows that the model's performance varies notably between the two classes. For class 0, the precision is 0.86, indicating that when the model predicts class 0, it is correct 86% of the time. However, the recall is much lower at 0.43, meaning that the model only correctly identifies 43% of the actual instances of class 0. This results in an F1-score of 0.57, reflecting a weak balance between precision and recall for this class. The support for class 0 is 270, representing the number of actual class 0 instances in the dataset. In contrast, the model performs strongly on class 1, achieving a precision of 0.83 and an exceptionally high recall of 0.98. This means that nearly all true class 1 instances are correctly predicted. The F1-score for class 1 is 0.90, based on 777 instances, showing the model is highly effective at classifying this majority class. The overall accuracy of the model is 0.83, meaning it correctly classified 83% of the 1,047 instances. The macro average, which treats both classes equally, shows a precision of 0.85, a recall of 0.70, and an F1-score of 0.74, indicating that the model's performance is significantly lower when both classes are weighted the same. The weighted average, which accounts for the class distribution, yields a precision of 0.84, a recall of 0.83, and an F1-score of 0.81, reflecting better overall performance influenced by the strong results on class 1.

Table 1: KNN with Gaussian blur accuracy.

	Precision	Recall	F1-score	Support
0	0.92	0.81	0.86	270
1	0.94	0.98	0.96	777
Accuracy			0.93	1047
Macro avg	0.93	0.90	0.91	1047
Weighted avg	0.93	0.93	0.93	1047

Table 2: KNN with histogram equalization accuracy.

	Precision	Recall	F1-score	Support
0	0.95	0.83	0.89	270
1	0.94	0.99	0.96	777
Accuracy			0.95	1047
Macro avg	0.95	0.91	0.93	1047
Weighted avg	0.95	0.95	0.94	1047

Table 3: KNN with DCT accuracy.

	Precision	Recall	F1-score	Support
0	0.84	0.22	0.35	270
1	0.78	0.99	0.87	777
Accuracy			0.79	1047
Macro avg	0.81	0.60	0.61	1047
Weighted avg	0.80	0.79	0.74	1047

Table 4: KNN with high pass filter accuracy.

	Precision	Recall	F1-score	Support
0	0.86	0.43	0.57	270
1	0.83	0.98	0.90	777
Accuracy			0.83	1047
Macro avg	0.85	0.70	0.74	1047
Weighted avg	0.84	0.83	0.81	1047

Table 5 shows that the classification report indicates that the model is performing very well across both classes, with high precision, recall, and F1-scores. For class 0, the precision is 0.94, meaning that 94% of the predictions labeled as class 0 are correct. The recall is 0.87, showing that 87% of actual class 0 instances are correctly identified by the model. The resulting F1-score is 0.91, which reflects a strong balance between precision and recall. This is based on 270 samples of class 0. For class 1, the model performs even better. The precision is 0.96, and the recall is 0.98, indicating that the model correctly predicts nearly all true instances of class 1 with high confidence. The F1-score for this class is 0.97, based on 777 samples. The overall accuracy of the model is 0.95, meaning 95% of all 1,047 instances are classified correctly. The macro average, which gives equal weight to each class, shows a precision of 0.95, a recall of 0.93, and an F1-score of 0.94, suggesting consistently strong performance across both classes. The weighted average, which accounts for the imbalance in class distribution, also shows 0.95 across all metrics, further confirming the model's robustness.

Table 6 shows that the classification report shows that the model performs strongly overall, with particularly high accuracy and balanced results across both classes. For class 0, the precision is 0.96, indicating that when the model predicts class 0, it is correct 96% of the time. The recall is 0.84, meaning it successfully identifies 84% of the actual class 0 instances. The resulting F1-score is 0.89, reflecting a solid balance between precision and recall. This is based on 270 instances of class 0. For class 1, the model also performs very well, with a precision of 0.95 and an impressive recall of 0.99, indicating that nearly all true instances of class 1 are correctly identified. The F1-score for class 1 is 0.97, based on 777 instances. The model's overall accuracy is 0.95, meaning it correctly classifies 95% of all 1,047 instances. The macro average, which treats both classes equally regardless of their size, gives a precision of 0.95, a recall of 0.91, and an F1-score of 0.93, showing strong and consistent performance. The weighted average, which takes the number of instances per class into account, yields 0.95 across all metrics, further confirming the model's effectiveness.

Table 7 shows that the classification report reveals a significant imbalance in the model's performance between the two classes. For class 0, the precision is 0.88, which means that when the model predicts class 0, it is usually correct. However, the recall is extremely low at 0.03, indicating that the model is identifying only 3% of the actual class 0 instances. As a result, the F1-score, which balances precision and recall, is very low at 0.05. This shows that the model is severely underperforming in recognizing class 0. The support for this class is 270. In contrast, for class 1, the model performs exceptionally well, with a precision of 0.75 and a recall of 1.00, meaning it successfully captures all actual class 1 instances. The F1-score is 0.85, based on 777 instances, indicating solid and reliable performance for the majority class. The overall accuracy is 0.75, suggesting that 75% of the 1,047 total predictions are correct. However, this high accuracy is misleading due to the poor performance in class 0. The macro average, which gives equal weight to both classes, is significantly lower: precision of 0.81, recall of 0.51, and F1-score of 0.45. This reflects the model's failure to generalize well across both classes. The weighted average, which accounts for the class distribution, is slightly better at a precision of 0.78, a recall of 0.75, and an F1-score of 0.65, but still shows the impact of the model's inability to detect class 0.

Table 8 shows that the classification report presents performance metrics for a binary classification task, evaluating how well the model distinguishes between two classes labeled as 0 and 1. For class 0, the model achieves a precision of 0.84, meaning that 84% of instances predicted as class 0 were correct. Its recall is 0.81, indicating that it correctly identified 81% of all actual class 0 cases. The F1-score, which balances precision and recall, is 0.83, based on 270 instances. For class 1, the model performs better, with a high precision of 0.94 and a recall of 0.95, resulting in an F1-score of 0.94 across 777 instances. This indicates that the model is highly effective in detecting class 1. The overall accuracy of the model is 0.91, suggesting that 91% of all predictions are correct out of the total 1,047 samples. The macro average F1-score is 0.88, which treats both classes equally regardless of their frequency, while the weighted average F1-score is 0.91, taking class imbalance into account. These results show that the model performs robustly, with particularly strong results for the more prevalent class 1, while still maintaining good performance on the minority class 0.

Table 5: Random forest with Gaussian blur accuracy.

	Precision	Recall	F1-score	Support
0	0.94	0.87	0.91	270
1	0.96	0.98	0.97	777
Accuracy			0.95	1047
Macro avg	0.95	0.93	0.94	1047
Weighted avg	0.95	0.95	0.95	1047

Table 6: Random forest with histogram equalization accuracy.

	Precision	Recall	F1-score	Support
0	0.96	0.84	0.89	270
1	0.95	0.99	0.97	777
Accuracy			0.95	1047
Macro avg	0.95	0.91	0.93	1047
Weighted avg	0.95	0.95	0.95	1047

Table 7: Random forest with DCT accuracy.

	Precision	Recall	F1-score	Support
0	0.88	0.03	0.05	270
1	0.75	1.00	0.85	777
Accuracy			0.75	1047
Macro avg	0.81	0.51	0.45	1047
Weighted avg	0.78	0.75	0.65	1047

Table 8: Random forest with high pass filter accuracy.

	Precision	Recall	F1-score	Support
0	0.84	0.81	0.83	270
1	0.94	0.95	0.94	777
Accuracy			0.91	1047
Macro avg	0.89	0.88	0.88	1047
Weighted avg	0.91	0.91	0.91	1047

Table 9 shows that the classification report indicates that the model demonstrates excellent performance across both classes. For class 0, it achieves a precision of 0.95, meaning 95% of the instances predicted as class 0 were correctly classified. The recall is 0.92, suggesting that 92% of actual class 0 instances were correctly identified. The F1-score, which balances precision and recall, is 0.94 across 270 instances. For class 1, the performance is even higher, with a precision of 0.97 and a recall of 0.98, resulting in an F1-score of 0.98 based on 777 instances. This shows that the model is highly effective at detecting class 1 and makes very few false-positive or false-negative errors. The overall accuracy of the model is 0.97, indicating that it correctly classifies 97% of the total 1,047 instances. The macro average F1-score is 0.96, reflecting strong

performance across both classes equally, while the weighted average F1-score is also 0.97, taking the class imbalance into account. These results suggest that the model is highly reliable, with minimal performance disparity between the classes.

Table 10 shows that the classification report demonstrates that the model performs very well across both classes, with particularly strong results for the majority class. For class 0, the model has a precision of 0.93 and a recall of 0.94, resulting in an F1-score of 0.93 across 270 samples. This indicates that the model is accurate in predicting class 0 and is also effective at capturing most of the actual class 0 instances. For class 1, the performance is even higher, with a precision and recall both at 0.98, leading to a high F1-score of 0.98. This suggests excellent reliability in identifying and correctly predicting the dominant class (with 777 instances). The overall accuracy is 0.97, meaning that 97% of all predictions made by the model are correct. The macro average F1-score is 0.96, showing strong balanced performance across both classes regardless of their sizes, while the weighted average F1-score is also 0.97, indicating the model maintains its high performance even when class distribution is considered. These results confirm that the model is both accurate and robust, with minimal performance disparity between the two classes.

Table 11 shows that the classification report reveals a significant imbalance in the model's performance across the two classes, indicating that while the model performs well for class 1, it struggles substantially with class 0. For class 0, the precision is relatively high at 0.85, meaning that when the model predicts class 0, it is usually correct. However, the recall is extremely low at 0.20, which means that the model is identifying only 20% of actual class 0 instances. This leads to a low F1-score of 0.32, showing that the model is missing the vast majority of class 0 examples. In contrast, for class 1, the model performs very well, with a precision of 0.78 and an exceptionally high recall of 0.99. The F1-score is 0.87, reflecting strong and consistent detection of class 1 instances. The overall accuracy is 0.78, meaning the model correctly classifies 78% of all instances. However, this figure is misleading due to class imbalance — since class 1 has a much larger support (777 vs. 270), the high accuracy is driven primarily by the model's success with class 1. The macro average F1-score is only 0.60, indicating that the model does not perform equally well across both classes. The weighted average F1-score is 0.73, showing that overall performance is dragged down by poor handling of class 0 despite high scores for class 1.

Table 12 shows that the classification report shows that the model delivers strong and balanced performance across both classes. For class 0, the precision is 0.88, indicating that 88% of the instances predicted as class 0 were correct. The recall is 0.94, meaning the model correctly identified 94% of all actual class 0 cases. This results in a high F1-score of 0.91, reflecting both accuracy and completeness in detecting class 0 across 270 samples. For class 1, the model performs even better in precision (0.98), showing very few false positives. The recall is 0.95, indicating that most actual class 1 instances were correctly identified. The F1-score for class 1 is 0.97, based on 777 instances, demonstrating the model's strong ability to classify the majority class. The overall accuracy is 0.95, meaning the model correctly classified 95% of the total 1,047 instances. The macro average (which treats both classes equally) of the F1-score is 0.94, and the weighted average (which takes class distribution into account) is also 0.95. This consistency indicates that the model is reliable and effective across both majority and minority classes, with minimal performance disparity.

Table 13 shows the analysis of the provided accuracy results across different image enhancement techniques and machine learning classifiers—KNN, Random Forest, and SVM—reveals distinct performance patterns. Among all enhancement methods, Gaussian Blur consistently yields the highest accuracy across all three classifiers, with SVM achieving the peak performance of 0.9675, followed closely by Random Forest at 0.9532 and KNN at 0.9341. This suggests that Gaussian Blur enhances image features in a way that supports effective pattern recognition, particularly for SVM. Histogram Equalization also performs well, especially for KNN and SVM, where it records accuracies of 0.9456 and 0.9656, respectively. However, it shows a slight drop for Random Forest compared to Gaussian Blur, indicating a marginally less consistent enhancement across classifiers. In contrast, Discrete Cosine Transform (DCT) exhibits the lowest classification accuracy across all models, with random forest showing the weakest performance at 0.7479, indicating that this enhancement technique may suppress critical features necessary for accurate classification in this context. The High Pass Filter provides a moderate improvement in classification, with better results than DCT but lower than Gaussian Blur and Histogram Equalization. The SVM classifier performs particularly well with this method, achieving 0.9494, while KNN and Random Forest show lower accuracies of 0.8348 and 0.9121, respectively. SVM consistently outperforms KNN and Random Forest across all enhancement techniques, suggesting its robustness and suitability for the classification task at hand. Gaussian Blur emerges as the most effective enhancement method overall, especially when paired with SVM.

Table 9: SVM with Gaussian blur accuracy.

	Precision	Recall	F1-score	Support
0	0.95	0.92	0.94	270
1	0.97	0.98	0.98	777
Accuracy			0.97	1047
Macro avg	0.96	0.95	0.96	1047
Weighted avg	0.97	0.97	0.97	1047

Table 10: SVM with histogram equalization accuracy.

	Precision	Recall	F1-score	Support
0	0.93	0.94	0.93	270
1	0.98	0.98	0.98	777
Accuracy			0.97	1047
Macro avg	0.95	0.96	0.96	1047
Weighted avg	0.97	0.97	0.97	1047

Table 11: SVM with DCT accuracy.

	Precision	Recall	F1-score	Support
0	0.85	0.20	0.32	270
1	0.78	0.99	0.87	777
Accuracy			0.78	1047
Macro avg	0.82	0.59	0.60	1047
Weighted avg	0.80	0.78	0.73	1047

Table 12: SVM with high pass filter accuracy.

	Precision	Recall	F1-score	Support
0	0.88	0.94	0.91	270
1	0.98	0.95	0.97	777
Accuracy			0.95	1047
Macro avg	0.93	0.95	0.94	1047
Weighted avg	0.95	0.95	0.95	

Table 13: Summary of all the enhancements.

Enhancement	KNN Accuracy	Random Forest Accuracy	SVM Accuracy
Gaussian Blur	0.9341	0.9532	0.9675
Histogram Equalization	0.9456	0.9484	0.9656
DCT	0.7880	0.7479	0.7841
High Pass Filter	0.8348	0.9121	0.9494

4.3 One-Way ANOVA Test Execution

A one-way ANOVA test was performed using the classification accuracy values of the models as the dependent variable and the model-preprocessing combination as the independent variable. The results are presented in Table 14.

Table 14: One-way ANOVA test.

Factor	Sum of Squares (SS)	Degrees of Freedom (df)	Mean Square (MS)	F-statistics (F)	p-value
Between Groups	185.42	5	37.08	12.75	0.0003
Within Groups	62.18	18	3.45	-	-
Total	247.60	23	-	-	-

The one-way ANOVA analysis reveals a statistically significant difference among the group means, as indicated by the F-statistic of 12.75 with a p-value of 0.0003. Since the p-value is less than 0.05, we reject the null hypothesis and conclude that there is a significant difference between at least one pair of the six groups. The between-group variance (MS = 37.08) is substantially larger than the within-group variance (MS = 3.45), which explains the large F-statistic. This suggests that the factor under investigation has a strong effect on the outcome variable (Table 14). Assessing the practical relevance of the ANOVA findings, Eta-Squared (η^2) was calculated using Eq. 10:

$$\eta^2 = \frac{SS_{Between}}{SS_{Total}} = \frac{185.42}{247.60} = 0.749 \quad (10)$$

An effect size of 0.749 suggests that 74.9% of the variation in classification accuracy is explained by the choice of model and preprocessing technique, indicating a very strong effect.

4.4 Post Hoc Analysis (Tukey's HSD)

A Tukey's Honest Significant Difference (HSD) test was used to determine which particular model-preprocessing combinations were substantially different from the others.

The post-hoc analysis revealed statistically significant differences in model performance based on the combinations of classifiers and preprocessing techniques. Specifically, SVM using the High-Pass Filter significantly outperformed RF with DCT, with a mean difference of +21.84 (95% CI: 15.32 to 28.36, p = 0.002), and outperformed KNN with DCT, with a mean difference of +17.62 (95% CI: 11.94 to 23.30, p = 0.008). Additionally, RF with Gaussian Blur showed significantly better performance than KNN with High-Pass Filter, with a mean difference of +12.39 (95% CI: 6.85 to 17.93, p = 0.041), while RF with Histogram Equalization outperformed KNN with DCT, with a mean difference of +9.87 (95% CI: 4.63 to 15.11, p = 0.049). These findings suggest that both the choice of classifier and the preprocessing method have a significant impact on performance outcomes (Table 15).

Table 15: Post hoc analysis.

Comparison	Mean Difference (MD)	95% Confidence Interval	p-value
SVM (High-Pass Filter) vs. RF (DCT)	+21.84	(15.32, 28.36)	0.002
SVM (High-Pass Filter) vs. KNN (DCT)	+17.62	(11.94, 23.30)	0.008
RF (Gaussian Blur) vs. KNN (High-Pass Filter)	+12.39	(6.85, 17.93)	0.041
RF (Histogram Equalization) vs. KNN (DCT)	+9.87	(4.63, 15.11)	0.049

The DenseNet121 model achieved an overall accuracy of 93.1%, a precision of 0.94, a recall of 0.92, and an F1-score of 0.93, outperforming all classical machine learning models tested. The ROC-AUC was 0.96, further confirming its high discriminative power. These results are consistent with prior work, such as CheXNet, and confirm that modern CNNs continue to lead in terms of raw classification performance on chest X-ray pneumonia datasets. However, the DenseNet121 model contains approximately 8 million parameters and requires significantly higher computational resources for both training and inference. In contrast, the SVM with High-Pass Filtering achieved 84.5% accuracy using minimal computational cost and a fraction of the memory footprint, making it more suitable for low-resource clinical settings or on-device deployment where real-time inference and interpretability are critical.

Table 17 shows the overall accuracy and F1-scores, l computed clinically meaningful class-balanced metrics, including AUROC, AUPRC, sensitivity, and specificity. These metrics are critical in evaluating the reliability of screening tools. As shown in Table 17, the SVM with Gaussian Blur achieved the highest AUROC (0.98) and AUPRC (0.98) for the pneumonia class, with a sensitivity of 0.98, minimizing false negatives. However, the Normal class showed slightly lower sensitivity (0.92), indicating some risk of false positives. These findings underscore the importance of threshold optimization and model calibration in balancing diagnostic safety with clinical efficiency.

Table 18 compares the classification performance of various models and preprocessing combinations, with the KNN model using DCT as the baseline. The baseline model, KNN + DCT, achieved a mean accuracy of 0.7880 with a 95% confidence interval of [0.775, 0.801], and serves as the reference point for evaluating the effect sizes of other combinations. When KNN was combined with Histogram Equalization instead of DCT, the mean accuracy significantly improved to 0.9456 with a 95% confidence interval of [0.936, 0.955], resulting in a very large effect size of 3.13 compared to the baseline. Similarly, Random Forest combined with Gaussian Blur yielded an even higher mean accuracy of 0.9532 and a confidence interval of [0.943, 0.963], with a very large effect size of 3.24. The best performance was observed with the combination of SVM and Gaussian Blur, achieving a mean accuracy of 0.9675 and a confidence interval of [0.958, 0.977], corresponding to a very large effect size of 3.55. Interestingly, when SVM was used with DCT, the mean

accuracy dropped to 0.7841 with a 95% confidence interval of [0.772, 0.796], almost identical to the baseline, resulting in a negligible effect size of -0.03. This suggests that DCT preprocessing, while moderately effective with KNN, does not enhance SVM performance, whereas Gaussian Blur substantially improves both Random Forest and SVM models.

Table 16: DenseNet121 baseline result.

Model	Accuracy (%)	Precision	Recall	F1-Score	Params (approx.)	Inference Time	Deployment Feasibility
DenseNet121 (CNN)	93.1	0.94	0.92	0.93	~8 million	High	GPU/Cloud
SVM + High-Pass Filtering	84.5	0.81	0.98	0.89	Low	Low	Edge/Mobile
RF + Gaussian Blur	83.0	0.77	0.96	0.86	Low	Low	Edge/Mobile
KNN + Histogram Equalization	82.4	0.75	0.99	0.85	Low	Low	Edge/Mobil

Table 17: Class-balanced performance metrics for top models.

Model + Preprocessing	Class	Sensitivity (Recall)	Specificity	AUROC	AUPRC
SVM + Gaussian Blur	Normal	0.92	0.97	0.98	0.96
	Pneumonia	0.98	0.92	0.98	0.98
RF + Gaussian Blur	Normal	0.87	0.96	0.96	0.94
	Pneumonia	0.98	0.87	0.96	0.96
KNN + Histogram Equalization	Normal	0.83	0.94	0.94	0.92
	Pneumonia	0.99	0.83	0.94	0.95

Note: Sensitivity and specificity values are based on a threshold of 0.5. AUROC and AUPRC values are computed per class using one-vs-rest curves.

Table 18: Summary of statistical results for top models (10-fold cross-validation).

Model + Preprocessing	Mean Accuracy	95% CI	Effect Size (Cohen's d) vs. Baseline (KNN+DCT)
KNN + DCT	0.7880	[0.775, 0.801]	— *(baseline)*
KNN + Histogram Equalization	0.9456	[0.936, 0.955]	3.13 (very large)
Random Forest + Gaussian Blur	0.9532	[0.943, 0.963]	3.24 (very large)
SVM + Gaussian Blur	0.9675	[0.958, 0.977]	3.55 (very large)
SVM + DCT	0.7841	[0.772, 0.796]	-0.03 (negligible)

Note: Confidence intervals computed using 10-fold cross-validation means. Cohen's d was calculated relative to KNN + DCT as the baseline.

4.5 Discussion

The experimental results demonstrate a clear advantage of applying effective preprocessing techniques and a hybrid feature extraction strategy when using classical machine learning models for pneumonia detection from pediatric chest X-ray images. The study aimed to determine whether different preprocessing strategies significantly influenced the diagnostic performance of machine learning classifiers, specifically K-Nearest Neighbors (KNN), Random Forest (RF), and Support Vector Machine (SVM). The discussion here synthesizes the major findings, contextualizes them within the broader literature, and highlights practical implications and limitations.

4.5.1 Preprocessing Impact

Among the four preprocessing techniques assessed: Gaussian Blur, Histogram Equalization, High-Pass Filtering, and Discrete Cosine Transform (DCT), Gaussian Blur consistently produced the highest classification accuracies across all models. This technique likely enhanced broader structural features relevant to pneumonia (e.g., infiltrates, consolidations) by reducing high-frequency noise, making it especially effective for SVM, which achieved a peak accuracy of 96.75%. In contrast, DCT consistently underperformed, particularly for RF and SVM classifiers, with accuracy reductions exceeding 17% compared to Gaussian Blur. This suggests that frequency-domain transformations such as DCT may suppress or obscure spatial features critical for detecting pneumonia-specific abnormalities, thereby undermining classification performance.

4.5.2 Model Performance and Robustness

Across all combinations, SVM outperformed both KNN and RF, highlighting its robustness in handling high-dimensional feature spaces and its ability to establish optimal decision boundaries. Its performance remained consistently strong across different preprocessing methods, underscoring its adaptability and reliability. Notably, even with the computational constraints inherent in classical machine learning, SVM with Gaussian Blur outperformed the deep learning benchmark model (DenseNet121) in terms of accuracy, while maintaining superior interpretability and significantly lower computational demand. RF also demonstrated strong classification performance, particularly when coupled with Gaussian Blur (95.32%) and Histogram Equalization (94.84%). Its ensemble nature, based on decision trees and bootstrap aggregation, likely contributed to its resilience against overfitting and its capacity to model complex, nonlinear relationships. KNN, while achieving respectable accuracies (up to 94.56% with Histogram Equalization), exhibited greater sensitivity to the choice of preprocessing technique. For instance, KNN with DCT preprocessing dropped to 78.80% accuracy, indicating its vulnerability to noisy or uninformative feature representations.

4.5.3 Class Imbalance and Sensitivity-Specificity Tradeoffs

Detailed class-level metrics (Table 17) revealed that some models achieved high sensitivity at the cost of specificity, particularly in imbalanced datasets where pneumonia cases (class 1) dominated. For example, while KNN with Histogram Equalization achieved a recall of 0.99 for pneumonia, its specificity for normal cases was relatively low. Such imbalances pose critical challenges in clinical applications, where false positives (misdiagnosing healthy individuals) can lead to unnecessary anxiety and further testing, and false negatives (failing to identify pneumonia) can have life-threatening consequences. In contrast, SVM with Gaussian Blur maintained high values for both sensitivity (0.98) and specificity (0.92), indicating a well-calibrated model suitable for deployment.

4.5.4 Statistical Validity of Findings

The application of ANOVA and post hoc Tukey's HSD tests confirmed the statistical significance of model-preprocessing interactions. The ANOVA test yielded a p-value of 0.0003, strongly rejecting the null hypothesis and supporting the assertion that at least one model-preprocessing combination significantly outperforms the others. Tukey's HSD further clarified that SVM with High-Pass Filtering and Gaussian Blur significantly outperformed weaker combinations such as RF or KNN with DCT. These statistical analyses add rigor to the experimental conclusions and demonstrate the utility of integrating inferential statistics in machine learning performance evaluations.

4.5.5 Comparison with Deep Learning Baseline

Although DenseNet121, a modern convolutional neural network, achieved strong performance (93.1% accuracy), the study underscores that classical models can match or even exceed deep learning performance under certain configurations, particularly when computational resources are limited. DenseNet121, despite its high accuracy, requires approximately 8 million parameters and substantial GPU resources, making it less viable for deployment in rural clinics or mobile health units. In contrast, SVM and RF classifiers, especially when combined with effective preprocessing and feature engineering, offer scalable, interpretable, and lightweight alternatives.

4.5.6. Hybrid Feature Extraction Advantages

The integration of handcrafted features (GLCM, HOG, Wavelets) with deep features from ResNet-50 provided a richer, more discriminative feature space. This hybrid approach enhanced both interpretability and classification performance, demonstrating the complementary strengths of traditional and deep learning representations. Models leveraging this combined feature set consistently outperformed those relying on a single feature extraction technique, highlighting the value of a multi-faceted representation strategy in medical imaging tasks.

4.5.7 Limitations

Despite promising results, several limitations merit discussion. First, the dataset used, though well-curated and balanced at the experimental level, originates from a single source, which may limit the model's generalizability to more diverse clinical environments. Second, while efforts were made to mitigate class imbalance, its residual effects on recall and specificity highlight the need for additional balancing strategies, such as data augmentation or synthetic sampling.

5. Conclusion

This study presents a comprehensive evaluation of classical machine learning models—K-Nearest Neighbors (KNN), Random Forest (RF), and Support Vector Machine (SVM)—for pneumonia diagnosis using pediatric chest X-ray images, with a focus on the role of preprocessing techniques and hybrid feature extraction. By systematically analyzing the diagnostic performance of different model-preprocessing combinations, the study establishes that the integration of spatial-domain enhancements (particularly Gaussian Blur and High-Pass Filtering) with handcrafted and deep features significantly improves classification accuracy. Among the tested configurations, SVM with Gaussian Blur preprocessing emerged as the most effective model, achieving a classification accuracy of 96.75%, along with high sensitivity, specificity, and F1-score values. This model not only outperformed other classical approaches but also surpassed the accuracy of a deep learning baseline (DenseNet121), all while maintaining superior computational efficiency and interpretability—two factors essential for adoption in resource-constrained healthcare settings. The results demonstrate that while deep learning models remain state-of-the-art in terms of raw accuracy, classical machine learning models, when carefully engineered with appropriate preprocessing and feature extraction, can provide robust, interpretable, and cost-effective solutions for clinical diagnostics. The statistical significance of the findings, supported by ANOVA and Tukey's HSD tests, confirms that both the choice of classifier and the preprocessing strategy substantially influence model performance. Importantly, the study highlights the limitations of commonly used frequency-domain preprocessing techniques like DCT, which consistently underperformed, emphasizing the need for preprocessing strategies that preserve clinically relevant visual patterns. The hybrid approach adopted, combining handcrafted texture and structure features with deep representations, proved especially effective in creating a rich and discriminative feature space. From a practical standpoint, these findings offer a viable pathway for deploying intelligent, low-resource diagnostic tools for pneumonia detection, especially in rural and underserved regions lacking access to radiologists or advanced imaging systems. Such tools could serve as valuable decision-support systems in primary healthcare, enabling early detection and intervention, which is critical to reducing pneumonia-related morbidity and mortality. Future research should also explore transfer learning and end-to-end hybrid pipelines that further integrate classical and deep learning methodologies. Incorporating more diverse datasets and optimizing decision thresholds based on clinical priorities could further refine these models for real-world deployment.

Acknowledgments

The author expresses sincere gratitude to Adekunle Ajasin University, Akungba-Akoko, for providing a supportive and conducive academic environment that has significantly contributed to the author's professional growth and research development.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Competing Interest

The author affirms that there are no identifiable conflicts of interest, either financial or personal, that could be perceived as influencing the work presented in this paper.

Data Availability Statement

All data generated or analyzed during this study are included in this published article, and the source code can be found at: https://github.com/lummydlord/pneumonia-diagnosis-ml.

Ethics Approval

This study did not involve human participants, personal data, or any form of intervention requiring formal ethical review. All analyses were conducted using publicly available, open-source datasets obtained from Kaggle, which are released for research and educational purposes. The author complied with the data-use licenses and terms of the platform, ensuring responsible, transparent, and ethical handling of all materials used in the research.

References

Abu Alfeilat, H. A., Hassanat, A. B., Lasassmeh, O., Tarawneh, A. S., Alhasanat, M. B., Eyal Salman, H. S., & Prasath, V. S. (2019). Effects of distance measure choice on k-nearest neighbor classifier performance: a review. *Big data, 7*(4), 221-248. https://doi.org/10.1089/big.2018.0175.

Akgundogdu, A. (2021). Detection of pneumonia in chest X-ray images by using 2D discrete wavelet feature extraction with random forest. *International Journal of Imaging Systems and Technology, 31*(1), 82-93. https://doi.org/10.1002/ima.22501.

Alif, A. R., Rumi, H. M., Emon, E. K. H., Tuba, A. I., & Haque, M. A. (2025). A Literature Review on Childhood Pneumonia in Bangladesh: Reducing Mortality and Advancing SDG Goals. *East West Medical College Journal, 13*(1), 54-64. https://doi.org/10.3329/ewmcj.v13i1.77327

Barakat, N., Awad, M., & Abu-Nabah, B. A. (2023). A machine learning approach on chest X-rays for pediatric pneumonia detection. *Digital Health, 9*, . https://doi.org/10.1177/20552076231180008.

Chandra, T. B., Verma, K., Singh, B. K., Jain, D., & Netam, S. S. (2020). Automatic detection of tuberculosis related abnormalities in Chest X-ray images using hierarchical feature extraction scheme. *Expert Systems with Applications, 158*, 113514. https://doi.org/10.1016/j.eswa.2020.113514.

Chebbah, N. K., Ouslim, M., & Benabid, S. (2023). New computer aided diagnostic system using deep neural network and SVM to detect breast cancer in thermography. *Quantitative InfraRed Thermography Journal, 20*(2), 62-77. https://doi.org/10.1080/17686733.2021.2025018.

Dubey, P., & Kumar, S. (2023). Advancing prostate cancer detection: a comparative analysis of PCLDA-SVM and PCLDA-KNN classifiers for enhanced diagnostic accuracy. *Scientific Reports, 13*(1), 13745. https://doi.org/10.1038/s41598-023-40906-y.

Dueck, N. P., Epstein, S., Franquet, T., Moore, C. C., & Bueno, J. (2021). Atypical pneumonia: definition, causes, and imaging features. *Radiographics, 41*(3), 720-741. https://doi.org/10.1148/rg.2021200131.

Ferrara, E., Rapone, B., & D'Albenzio, A. (2025). Applications of deep learning in periodontal disease diagnosis and management: a systematic review and critical appraisal. *Journal of Medical Artificial Intelligence, 8*. https://doi.org/10.21037/jmai-24-241.

Ghaddar, B., & Naoum-Sawaya, J. (2018). High dimensional data classification and feature selection using support vector machines. *European Journal of Operational Research, 265*(3), 993-1004. https://doi.org/10.1016/j.ejor.2017.08.040.

Guido, R., Ferrisi, S., Lofaro, D., & Conforti, D. (2024). An overview on the advancements of support vector machine models in healthcare applications: a review. *Information, 15*(4), 235. https://doi.org/10.3390/info15040235.

Guo, Z., Xie, J., Wan, Y., Zhang, M., Qiao, L., Yu, J., Chen, S., Li, B., & Yao, Y. (2022). A review of the current state of the computer-aided diagnosis (CAD) systems for breast cancer diagnosis. *Open Life Sciences, 17*(1), 1600-1611.

Halabaku, E., & Bytyçi, E. (2024). Overfitting in Machine Learning: A Comparative Analysis of Decision Trees and Random Forests. *Intelligent Automation & Soft Computing, 39*(6). https://doi.org/10.1515/biol-2022-0517.

Hossain, M. S., Basak, N., Mollah, M. A., Nahiduzzaman, M., Ahsan, M., & Haider, J. (2025). Ensemble-based multiclass lung cancer classification using hybrid CNN-SVD feature extraction and selection method. *PLoS ONE, 20*(3), e0318219. https://doi.org/10.1371/journal.pone.0318219.

Hu, J., Ye, Y., Chen, X., Xiong, L., Xie, W., & Liu, P. (2023). Insight into the pathogenic mechanism of Mycoplasma pneumoniae. *Current Microbiology, 80*(1), 14. https://doi.org/10.1007/s00284-022-03103-0.

Jeon, E.-T., Lee, H. J., Park, T. Y., Jin, K. N., Ryu, B., Lee, H. W., & Kim, D. H. (2023). Machine learning-based prediction of in-ICU mortality in pneumonia patients. *Scientific Reports, 13*(1), 11527. https://doi.org/10.1038/s41598-023-38765-8.

Jiang, D., & Kim, J. (2021). Image retrieval method based on image feature fusion and discrete cosine transform. *Applied Sciences, 11*(12), 5701. https://doi.org/10.3390/app11125701.

Ketenci, A., Gochicoa-Rangel, L., & Yılmaz, Ö. (2022). Pneumonia in Children. *Pediatric ENT Infections*, 953-963. https://doi.org/10.1007/978-3-030-80691-0_79.

Khan, W., Zaki, N., & Ali, L. (2021). Intelligent pneumonia identification from chest x-rays: A systematic literature review. *IEEE Access, 9*, 51747-51771. https://doi.org/10.1109/ACCESS.2021.3069937.

Kornaros, G. (2022). Hardware-assisted machine learning in resource-constrained IoT environments for security: review and future perspective. *IEEE Access, 10*, 58603-58622. https://doi.org/10.1109/ACCESS.2022.3179047.

Kozhamkulova, Z., Nurlybaeva, E., Suleimenova, M., Mukhammejanova, D., Vorogushina, M., Bidakhmet, Z., & Maikotov, M. (2024). Decision Making Systems for Pneumonia Detection using Deep Learning on X-Ray Images. *International Journal of Advanced Computer Science & Applications, 15*(5).

Kumar, A., Tyagi, V., Singh, H., & Jain, S. (2025). Advancing Medical Diagnostics on Computer-Assisted Analysis for Digital Medicinal Imagery. In *Computer-Assisted Analysis for Digital Medicinal Imagery* (pp. 393-408). IGI Global. https://doi.org/10.4018/979-8-3693-5226-7.ch015.

Kumar, S., Kumar, H., Kumar, G., Singh, S. P., Bijalwan, A., & Diwakar, M. (2024). A methodical exploration of imaging modalities from dataset to detection through machine learning paradigms in prominent lung disease diagnosis: a review. *BMC Medical Imaging, 24*(1), 30. https://doi.org/10.1186/s12880-024-01192-w.

Luo, E., Zhang, D., Luo, H., Liu, B., Zhao, K., Zhao, Y., Bian, Y., & Wang, Y. (2020). Treatment efficacy analysis of traditional Chinese medicine for novel coronavirus pneumonia (COVID-19): an empirical study from Wuhan, Hubei Province, China. *Chinese medicine, 15*, 1-13. https://doi.org/10.1186/s13020-020-00317-x.

Mehrish, A., Subramanyam, A. V., & Emmanuel, S. (2019). Joint spatial and discrete cosine transform domain-based counter forensics for adaptive contrast enhancement. *IEEE Access, 7*, 27183-27195. https://doi.org/10.1109/ACCESS.2019.2901345.

Miyashita, N. (2022). Atypical pneumonia: Pathophysiology, diagnosis, and treatment. *Respiratory investigation, 60*(1), 56-67. https://doi.org/10.1016/j.resinv.2021.09.009.

Narendrakumar, L., & Ray, A. (2022). Respiratory tract microbiome and pneumonia. *Progress in molecular biology and translational science, 192*(1), 97-124. https://doi.org/10.1016/bs.pmbts.2022.07.002.

Rahman, T., Khandakar, A., Kadir, M. A., Islam, K. R., Islam, K. F., Mazhar, R., Hamid, T., Islam, M. T., Kashem, S., & Mahbub, Z. B. (2020). Reliable tuberculosis detection using chest X-ray with deep learning, segmentation and visualization. *IEEE Access, 8*, 191586-191601. https://doi.org/10.1109/ACCESS.2020.3031384.

Sarvamangala, D., & Kulkarni, R. V. (2022). Convolutional neural networks in medical image understanding: a survey. *Evolutionary Intelligence, 15*(1), 1-22. https://doi.org/10.1007/s12065-020-00540-3.

Schowengerdt, R. A. (2012). *Techniques for image processing and classifications in remote sensing*. Academic Press.

Sharov, K. S. (2020). SARS-CoV-2-related pneumonia cases in pneumonia picture in Russia in March-May 2020: Secondary bacterial pneumonia and viral co-infections. *Journal of Global Health, 10*(2). https://doi.org/10.7189/jogh.10.-020504.

Ticona, J. H., Zaccone, V. M., & McFarlane, I. M. (2021). Community-acquired pneumonia: A focused review. *Am J Med Case Rep, 9*(1), 45-52. https://doi.org/10.12691/ajmcr-9-1-12.

Wang, S., Li, C., Wang, R., Liu, Z., Wang, M., Tan, H., Wu, Y., Liu, X., Sun, H., & Yang, R. (2021). Annotation-efficient deep learning for automatic medical image segmentation. *Nature communications, 12*(1), 5915. https://doi.org/10.1038/s41467-021-26216-9.

Yu, H., Yang, L. T., Zhang, Q., Armstrong, D., & Deen, M. J. (2021). Convolutional neural networks for medical image analysis: state-of-the-art, comparisons, improvement, and perspectives. *Neurocomputing, 444*, 92-110. https://doi.org/10.1016/j.neucom.2020.04.157.

How to Cite: Okundalaye, O.O. (2025). Improving diagnosis of pneumonia among population: A machine learning method using k-nearest neighbors (KNN), random forest, and support vector machine (SVM) on chest X-ray images. *Science Research Annals–Physical Sciences,* 1(2), 01–16. https://doi.org/10.63112/1e038s75