E-ISSN: 2962-6110
P-ISSN: 2964-7843
JSRET
jsret.knpub.com
Journal of Scientific Research, Education, and Technology

Optimizing Pneumonia Detection in X-Ray Using Binary Statistical Image Features

Moch Daffa Yudis Averill, Muhammad Zakariyah
Universitas Teknologi Yogyakarta, Yogyakarta, Indonesia

ABSTRACT

Manual detection of pneumonia from X-ray images still faces challenges due to the long processing time, high cost, and strong dependence on radiologist expertise. This dependence increases the risk of delayed diagnosis and interpretation errors, potentially worsening patient conditions. To address these issues, this study proposes optimizing pneumonia detection using deep learning through the application of Binary Statistical Image Feature (BSIF) feature extraction. BSIF highlights important texture patterns in X-ray images to enhance the model's ability to recognize pneumonia affected lung areas. The dataset consists of 2,239 chest X-ray images divided into two categories: normal lungs and pneumonia. The research stages include image preprocessing, BSIF feature extraction, model training using Convolutional Neural Network (CNN) and Vision Transformer (ViT) architectures, and performance evaluation based on precision, recall, f1-score, specificity, and ROC AUC. The results show that the CNN+BSIF combination achieved the best performance with 99.69% training accuracy and 79.17% validation accuracy, precision 87%, recall 72%, f1-score 74%, specificity 45.30%, and ROC AUC 94.08%. Meanwhile, ViT+BSIF reached 99.35% accuracy, CNN without BSIF 98.24%, and ViT without BSIF 90.16%. Therefore, CNN+BSIF proved to be the most optimal method for fast and accurate pneumonia detection.

Keywords: *CNN, ViT, BSIF, Medical Image Detection, Deep Learning, Image Segmentation.*

Corresponding author
Name: Moch Daffa Yudis Averill
Email: daffayudis1019@gmail.com

INTRODUCTION

Pneumonia is a respiratory infection that causes inflammation of the lung parenchyma and may lead to serious complications such as respiratory failure, sepsis, or even death, particularly in vulnerable populations such as children, the elderly, and individuals with compromised immune systems (Nabuasa et al., 2024; Rachman et al., 2024). According to the World Health Organization (WHO), pneumonia remains one of the leading causes of mortality from infectious diseases worldwide, contributing to millions of deaths each year. In Indonesia, pneumonia consistently ranks among the top ten causes of hospitalization, reflecting its significant public health burden. The early detection and prompt treatment of pneumonia are therefore essential to reduce morbidity and mortality

rates, prevent complications, and improve overall healthcare outcomes(Cahyanto et al., 2023).

Conventionally, pneumonia diagnosis relies on clinical examination supported by radiographic imaging, particularly chest X-rays. However, radiological interpretation depends heavily on the expertise and experience of medical practitioners. This subjectivity, combined with varying image quality and the presence of overlapping anatomical structures, can lead to diagnostic inconsistencies and human error. In many healthcare facilities, particularly in developing countries, the limited availability of radiologists exacerbates this challenge, leading to delays in diagnosis and treatment. Consequently, there is a growing need for automated and intelligent diagnostic systems capable of assisting medical professionals by providing accurate and efficient analysis of chest radiographs.

In recent years, Artificial Intelligence (AI) and Deep Learning (DL) have shown remarkable progress in the field of medical image analysis. Among these, the Convolutional Neural Network (CNN) has emerged as one of the most widely adopted architectures due to its effectiveness in automatically learning hierarchical visual features from raw image data (Nova et al., 2025). CNNs operate through a sequence of convolutional, pooling, and fully connected layers that extract spatial patterns ranging from simple edges to complex shapes representing anatomical structures. Their ability to learn discriminative representations without manual feature engineering makes CNNs particularly powerful for medical imaging applications, including pneumonia detection. Numerous studies have demonstrated that CNN-based models can achieve performance comparable to, and in some cases exceeding, that of experienced radiologists when properly trained on large, high-quality datasets.

Despite these achievements, CNNs have inherent limitations. The convolutional operation primarily focuses on local receptive fields, restricting the model's ability to capture long-range dependencies and global contextual information within an image (Jiangtao et al., 2025). This limitation can result in the loss of subtle, yet clinically important, spatial relationships between different lung regions. To address this shortcoming, researchers have turned to transformer-based architectures, leading to the development of the Vision Transformer (ViT) a model adapted from the transformer mechanism originally used in Natural Language Processing (NLP) (Li et al., 2023). ViT divides an image into non-overlapping patches, converts them into embedding vectors, and processes these embeddings using a self-attention mechanism that enables the model to learn inter-patch dependencies across the entire image. This allows ViT to capture both local and global contextual features more effectively than traditional CNNs, resulting in improved performance in various computer vision tasks, including medical image classification.

Several studies have highlighted the advantages of deep learning models in enhancing pneumonia detection. (Yopento & Coastera, 2022)proposed a CNN combined with Sobel edge detection, achieving 91% precision, 92.8% recall, and 91.54% accuracy. (Ifayatin et al., 2024) employed the Faster R-CNN architecture to perform lung segmentation and region-based classification, successfully improving diagnostic efficiency.

Similarly, (Satria Wiratama et al., 2023) utilized ResNet50V2 with transfer learning to optimize feature extraction and enhance classification accuracy. These findings confirm the strong potential of deep learning models for automated pneumonia detection; however, challenges remain in improving model robustness, particularly in recognizing fine texture variations, minimizing false positives, and ensuring model generalization across different datasets.

To enhance the discriminative power of deep learning models, researchers have explored the integration of texture-based feature extraction methods. One promising technique is the Binarized Statistical Image Features (BSIF) algorithm, which captures detailed texture information through statistically independent filters generated via Independent Component Analysis (ICA). When applied to medical images, BSIF encodes local texture patterns into binary feature maps, which are then represented as histograms. This process effectively captures subtle differences in lung textures that may correspond to infection-related anomalies. Compared to conventional descriptors such as Local Binary Pattern (LBP), BSIF has demonstrated superior robustness against illumination variations and noise while maintaining high sensitivity to fine-grained textural changes (Rubio & Magnier, 2024). Empirical evidence shows that domain-specific BSIF filters, when combined with deep learning classifiers, can achieve accuracy levels exceeding 99% in texture recognition tasks.

Integrating BSIF with advanced deep learning models such as ViT can therefore offer complementary advantages. While CNNs and ViTs excel at spatial and contextual feature extraction, BSIF enriches the representation by embedding statistical texture information. This hybrid approach combines the strengths of local pattern recognition, global contextual awareness, and texture sensitivity, forming a more comprehensive representation of pneumonia features in chest X-ray images. Moreover, incorporating BSIF preprocessing can improve the interpretability of the learned features, which is crucial for clinical adoption and trust in AI-assisted diagnostic tools.

Based on these considerations, this study proposes a hybrid pneumonia detection framework that integrates Convolutional Neural Networks (CNN), Vision Transformers (ViT), and Binarized Statistical Image Features (BSIF) for enhanced classification accuracy and robustness. The primary objectives are to improve the precision and efficiency of pneumonia detection through multi-level feature fusion, validate the model's performance using publicly available chest X-ray datasets, and demonstrate its potential applicability in real-world clinical settings. This research contributes to the growing body of literature on AI-assisted diagnosis by combining spatial, contextual, and texture-based feature extraction into a unified model. In the long term, the proposed approach is expected to support radiologists in early diagnosis, reduce diagnostic errors, and contribute to the broader adoption of intelligent healthcare systems in Indonesia and beyond.

METHOD

This research methodology is systematically designed to ensure that the pneumonia detection model development process is well structured and produces

scientifically accountable outcomes. The research stages include data collection, image annotation, feature extraction, model design based on Convolutional Neural Network (CNN) and Vision Transformer (ViT), as well as model performance evaluation.

1. DATA SOURCE

This study utilizes a chest X-ray image dataset obtained from the Chest X-Ray Pneumonia Dataset available on the Kaggle platform This study utilizes a chest X-ray image dataset obtained from the Chest X-Ray Pneumonia Dataset available on the Kaggle platform (https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia). This dataset was selected due to its high image quality, well structured organization, and widespread use in previous studies, allowing the research results to be compared with existing works.

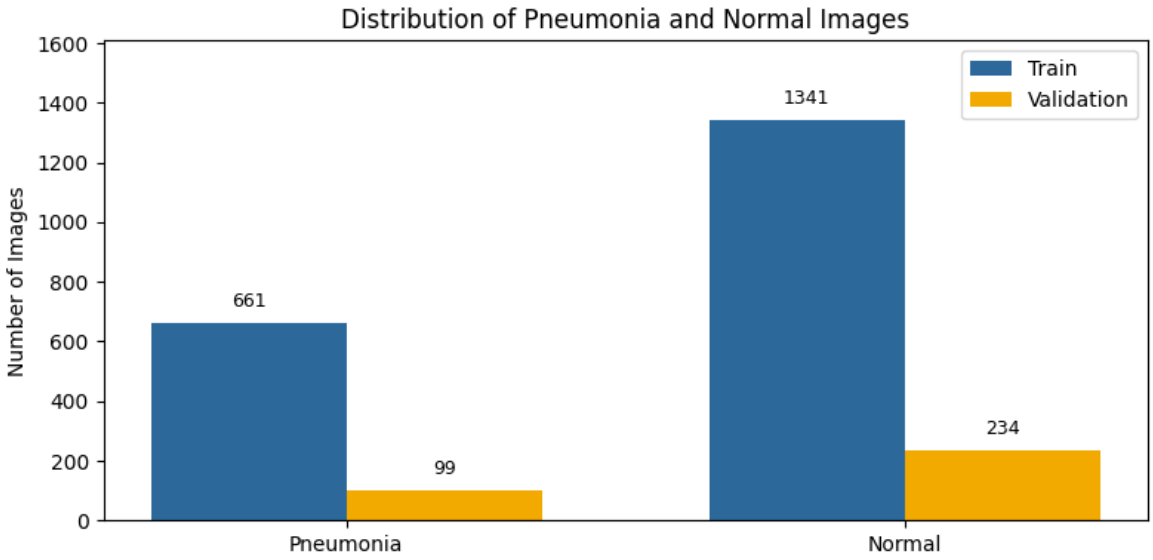

Figure 1: Dataset Distribution

Figure 1 shows that the total number of X-ray images used in this study is 2,335. Of these, 2,002 images are used as the training set, while 333 images are allocated as the validation set. The data were divided proportionally to ensure that the model could be trained and evaluated systematically.

The dataset distribution indicates that the Pneumonia category consists of 661 images in the training set and 99 images in the validation set, whereas the Normal category includes 1,341 images in the training set and 234 images in the validation set. With this composition, the dataset effectively represents the two main lung conditions under investigation the normal condition and pneumonia infected condition.

2. CLAHE

The preprocessing stage includes contrast enhancement utilizing the Contrast Limited Adaptive Histogram Equalization (CLAHE) technique. Unlike conventional

Histogram Equalization (HE), which applies global histogram equalization to the entire image, CLAHE performs the process locally by dividing the image into several small regions (tiles) and applying histogram equalization to each region independently (Lorinez S et al., 2025).

The main advantage of CLAHE is its ability to enhance image details while preventing excessive noise amplification, a limitation often encountered in global HE. By restricting contrast amplification within each region, CLAHE effectively improves the visibility of fine structures, particularly in medical images or those with non uniform illumination.

3. BSIF Feature Extraction

Feature extraction is the process of transforming raw data into a simplified representation by emphasizing key characteristics that are relevant for analysis or decision making (Wibisono et al., 2025). In this study, the Binarized Statistical Image Features (BSIF) method is used as an approach to extract local texture patterns from chest X-ray images. BSIF operates by convolving an image with binary filters learned independently using Independent Component Analysis (ICA). The result of this convolution is a set of binary values arranged into a histogram, representing the texture characteristics of the corresponding image.

The BSIF method combines the image with specific filters or kernels to generate binary values for each pixel. The binary values within each kernel are summed to produce a pixel intensity value. After the convolution process is applied to all pixels in the image, the final output is represented as a histogram that records the frequency of specific texture patterns. This histogram serves as the feature representation for the pattern recognition process. Previous studies have shown that BSIF outperforms the Local Binary Patterns (LBP) method in several recognition tasks, and its variants, such as Domain Specific BSIF (DS-BSIF) and combinations with Discrete Wavelet Transform (DWT), have been shown to further improve detection accuracy in specific domains.

4. CNN Modeling

The Convolutional Neural Network (CNN) is a deep learning approach designed to automatically recognize visual patterns from image data (Arrofiqoh & Harintaka, 2018). Unlike traditional classification methods that require manual feature extraction (Intyanto, 2021), CNN can learn hierarchical image representations — starting from simple patterns such as edges and textures to complex structures that represent organs. This advantage makes CNN widely used in medical image analysis, including the detection of pneumonia in chest X-ray images.

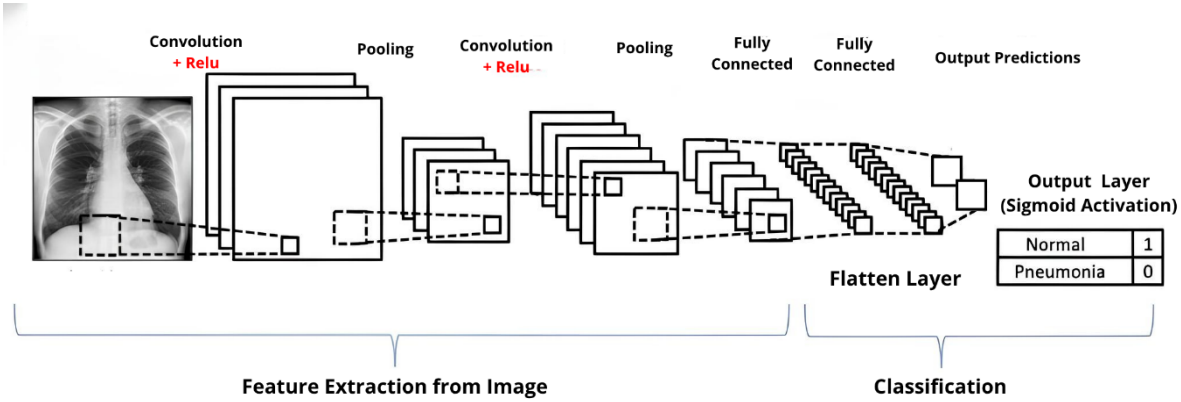

Figure 2: CNN Architecture

The CNN architecture in this study is composed of convolutional layers that extract image features, followed by non linear activation functions to enhance representation, and pooling layers that reduce dimensionality without losing essential information. The extracted features are then processed through a fully connected layer that connects to the classification stage. In this study, CNN is designed to classify two main categories: normal lungs and pneumonia affected lungs, using a sigmoid activation function in the output layer.

The training process employs binary cross entropy loss as the loss function and the Adam optimizer to update network weights stably and efficiently. To minimize the risk of overfitting, regularization strategies such as dropout and data augmentation are applied (Pellicer et al., 2023). These techniques enable the model to learn from more diverse image variations without losing focus on pneumonia specific patterns.

Although CNN excels in automatic feature extraction, it still has limitations including the requirement for a large amount of training data and difficulty in interpreting the learned internal features. Therefore, in this study, the CNN model not only serves as a primary model but also as a baseline for evaluating the effectiveness of other approaches, such as Vision Transformer (ViT) and the integration of CNN with Binarized Statistical Image Features (BSIF).

5. ViT Modeling

The Vision Transformer (ViT) is a deep learning architecture based on the transformer framework, originally developed for Natural Language Processing (NLP) but successfully adapted for image analysis with competitive performance (Utami et al., 2021). Unlike CNN, which relies on convolution operations to extract local features, ViT employs a self attention mechanism to model global relationships between image regions (Yan et al., 2025). This approach enables ViT to understand the overall patterns in X-ray images, making it particularly relevant for detecting pneumonia that can spread across various lung areas.

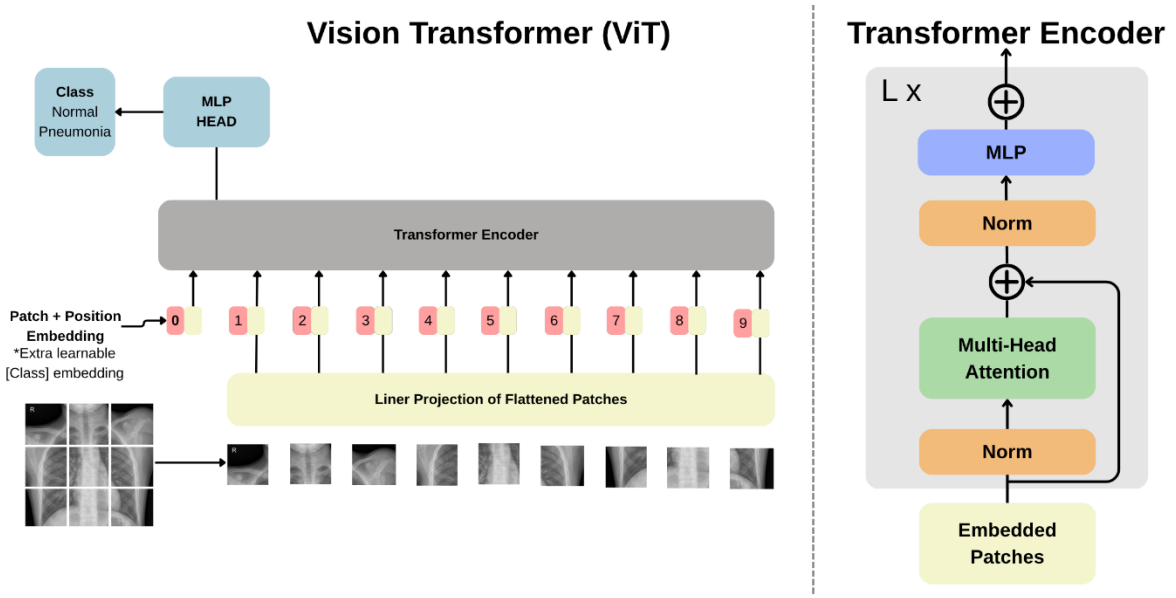

Figure 2: ViT Architecture

In this study, the ViT-B/16 architecture is utilized with pretrained weights from ImageNet-1K. This strategy belongs to the transfer learning category, where a model trained on a large general dataset is repurposed for a more specific domain—in this case, chest X-ray images. Through this approach, the fundamental representation of general visual patterns is retained, while the model is fine tuned to adapt to the specific characteristics of pneumonia.

To enable ViT to process inputs derived from Binarized Statistical Image Features (BSIF), modifications were made to the initial patch embedding layer (conv_proj) so that the number of input channels could be adjusted (e.g., to 8 channels). This allows the texture information produced by BSIF to be processed together with the spatial structure of the image by ViT. Furthermore, the final classification head was replaced with a linear layer producing two outputs for binary classification normal lungs and pneumonia affected lungs.

The training process uses the Cross Entropy Loss function and the Adam optimizer with a carefully tuned learning rate to ensure stable model weight updates. All parameters in ViT are retrained, making this approach a full fine tuning process rather than merely replacing the classification layer. With this design, ViT is expected to combine the strengths of BSIF's texture extraction with the global representation capability of self attention, thereby improving pneumonia detection accuracy compared to purely convolution based architectures.

6. Research Stages

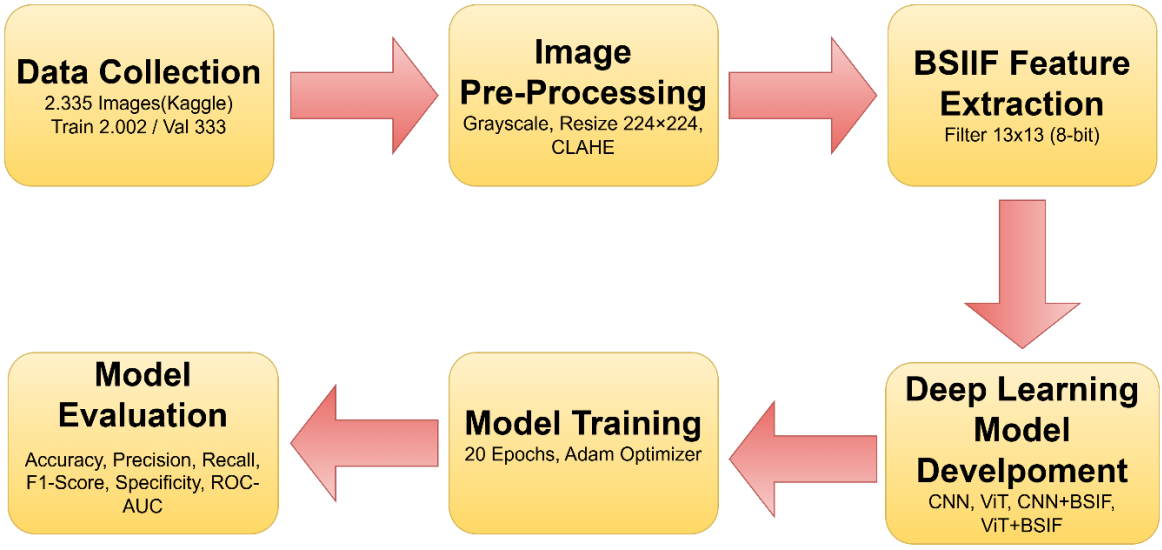

Figure 4: Research Framework

The figure above illustrates the research framework, which consists of six main stages. The first stage is Data Collection, involving the acquisition of chest X-ray images categorized as normal and pneumonia from public datasets such as the Kaggle Chest X-Ray Dataset. The next stage is Image Pre-processing, which includes converting the images to grayscale, resizing them to 224×224 pixels, and applying the Contrast Limited Adaptive Histogram Equalization (CLAHE) method to enhance image contrast and clarify lung tissue structures, making the visual patterns more informative for the model.

The next stage is BSIF Feature Extraction, which involves extracting features using Binary Statistical Image Features (BSIF). In this study, the BSIF filter used is a domain specific BSIF with a size of 13×13 bits, specifically trained on medical images to enhance sensitivity to fine texture patterns in lung tissue. The features obtained from this process are then integrated with two different deep learning approaches: Convolutional Neural Network (CNN) and Vision Transformer (ViT).

In the first approach, a pure CNN model is employed, utilizing convolutional operations to extract spatial features from X-ray images and perform binary classification (normal and pneumonia). CNN is used as the baseline model because it has proven effective in various medical image classification tasks. Subsequently, a CNN + BSIF approach is developed by providing the BSIF feature extraction results as additional input. The purpose of this integration is to enable CNN to capture not only global spatial features but also local texture patterns identified by BSIF.

In the second approach, the Vision Transformer (ViT) architecture processes chest X-ray images by dividing them into small patches, where each patch is projected into a vector space and learned through a self attention mechanism to capture global relationships between different regions of the image. To enhance performance in pneumonia detection, BSIF is integrated as an additional channel in this architecture (ViT

+ BSIF). This integration enables the model not only to understand the global representations derived from the transformer mechanism but also to capture important local texture information that is crucial for identifying subtle patterns in lung tissue.

The next stage is Model Training, where all the designed architectures CNN, CNN + BSIF, ViT, and ViT + BSIF are trained using the dataset that has undergone preprocessing and feature extraction. The data is divided into three subsets: a training set for model training, a validation set for monitoring during training to prevent overfitting, and a testing set for final model evaluation. The training process involves optimizing the loss function using optimization algorithms such as Adam and applying early stopping to maintain model performance stability. The main objective of this stage is for the model to learn to recognize texture patterns and distinctive characteristics in lung X-ray images, both in normal and pneumonia conditions.

After the training process is completed, the Model Evaluation stage is conducted to assess how well the deep learning–based pneumonia detection system performs in classifying X-ray images according to their true labels (Ningsih et al., 2024). Evaluation is carried out using a confusion matrix as the basis for calculating various key performance metrics, including accuracy, precision, recall, F1-score, specificity, and the Receiver Operating Characteristic – Area Under Curve (ROC-AUC).

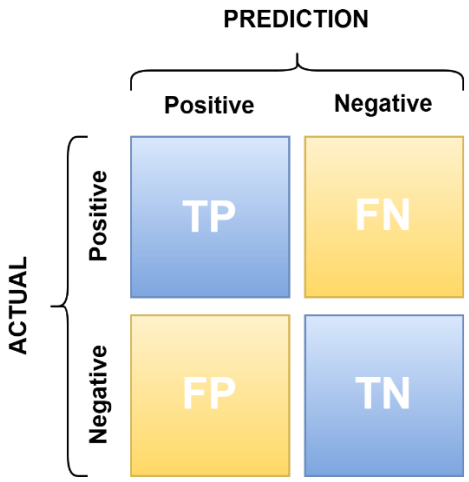

Figure 5: Confusion Matrix

A confusion matrix is a tabular representation comparing the model's predicted results with the actual data (ground truth). This matrix helps identify the number of correct and incorrect predictions for each class (Inonu et al., 2025). Its main components include: True Positive (TP), the number of X-ray images that actually contain pneumonia and are correctly classified; False Positive (FP), normal images incorrectly detected as pneumonia; False Negative (FN), pneumonia images that were not detected; and True Negative (TN), normal images that were correctly classified.

The model's performance is then evaluated using several key metrics as follows:

1. **Accuracy**
   Measures how many of the model's predictions are correct compared to the total number of predictions (Inonu et al., 2025).
   $$Accuracy: \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

2. **Precision**
   Measures the proportion of positive predictions that are truly positive (Inonu et al., 2025).
   $$Precision: \frac{TP}{TP+FP} \quad (3)$$

3. **Recall**
   Measures the model's ability to identify all positive (pneumonia) cases (Inonu et al., 2025).
   $$\text{Recall: } \frac{\text{TP}}{\text{TP+FN}} \quad (4)$$

4. **F1-Score**
   The harmonic mean between precision and recall, providing a balanced measure of both (Inonu et al., 2025).
   $$F1 - Score: 2 \times \frac{Precision \times Recall}{Precision+Recall} \quad (5)$$

5. **Specificity**
   Measures the model's ability to correctly recognize normal (negative) images (Inonu et al., 2025).
   $$\text{Specifity: } \frac{\text{TN}}{\text{TN+FP}} \quad (6)$$

6. **ROC-AUC**
   The Receiver Operating Characteristic (ROC) curve illustrates the relationship between the True Positive Rate (TPR) and the False Positive Rate (FPR) at various classification thresholds (Utami et al., 2021). The Area Under the Curve (AUC) value is used to assess the model's ability to distinguish between positive and negative classes; the closer the value is to 1, the better the model's performance.
   $$\text{FPR: } \frac{\text{FP}}{\text{FP+TN}} \quad (7)$$
   $$\text{TPR: } \frac{\text{TP}}{\text{TP+FN}} \quad (8)$$

RESULTS

1. Preprocessing

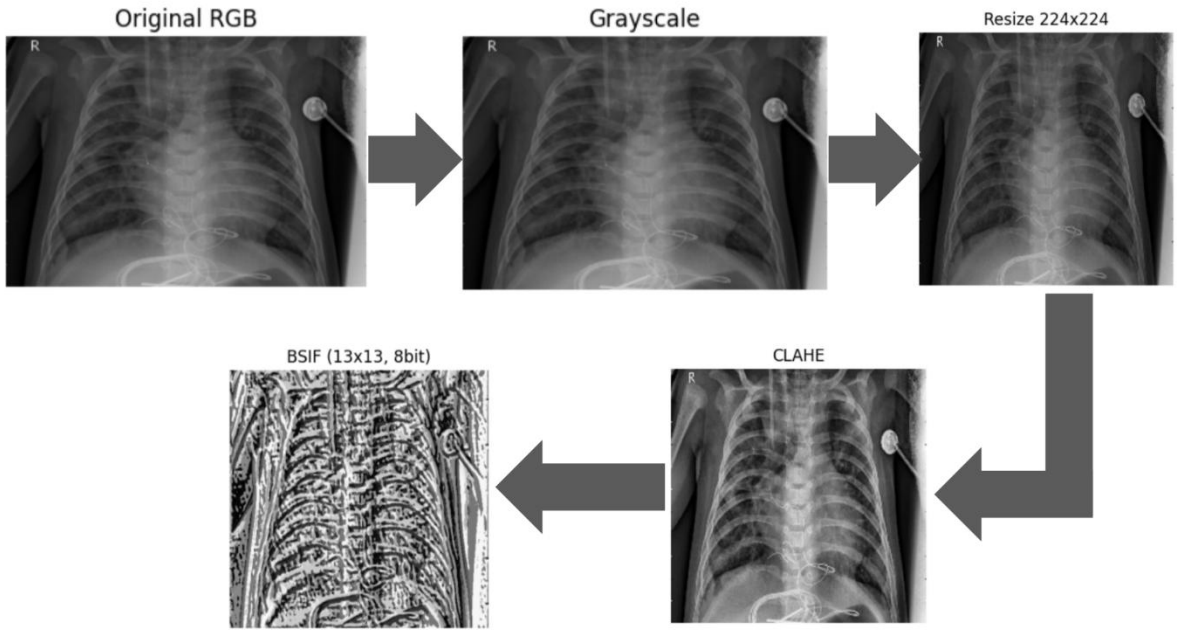

Figure 5: Image Preporcessing Workflow

Image preprocessing is an essential stage before data can be used for machine learning model training. In this study, the preprocessing stages included image conversion to grayscale, image resizing, and image enhancement using the Contrast Limited Adaptive Histogram Equalization (CLAHE) method.

The conversion of X-ray images to grayscale was performed to simplify the visual information contained within the images. Since radiographic images inherently do not require color information, this transformation reduces data complexity without losing essential structural information of the lungs. Consequently, the computational process becomes more efficient as the model analyzes only one intensity channel.

Next, all X-ray images were resized to 224 × 224 pixels. This resizing process is crucial to adjust the image dimensions according to the requirements of the Convolutional Neural Network (CNN) architecture used. Moreover, uniform sizing ensures consistent spatial representation, enabling the model to perform more stable and accurate feature extraction.

The subsequent stage was applying CLAHE (Contrast Limited Adaptive Histogram Equalization), a method for local contrast enhancement. Unlike standard histogram equalization, which may excessively amplify noise, CLAHE limits the contrast amplification within each local block. As a result, detailed lung structures—such as spots or infiltrates indicative of pneumonia become more visible without excessive distortion. This contrast enhancement is expected to improve the model's ability to distinguish between normal and pneumonia affected X-ray images.

2. Binary Statistical Image Features (BSIF)

The image resulting from BSIF feature extraction displays distinct texture patterns compared to the original image. Rib lines, lung contours, and tissue structures become more prominent as variations in grayscale and black white textures. This highlights local patterns that may be difficult to observe in the original X-ray image.

As shown in Figure 6, BSIF effectively captures fine textural details from medical images, such as more defined rib patterns and intensity variations within lung areas. This representation is then utilized as an input feature for deep learning models, enhancing their capability to differentiate between healthy lungs and those affected by pneumonia.

Through the integration of BSIF feature extraction, the system does not rely solely on raw pixel intensity but also considers richer texture variations. This approach has proven beneficial in improving the accuracy of both classification and detection processes.

3. CNN Model Training

The CNN model was trained for 20 epochs using a sigmoid activation function in the output layer to perform binary classification (normal lungs vs. pneumonia lungs). The model's performance evaluation is presented in two main plots: the training loss curve and the accuracy curve (Figure 7).

The training loss graph shows a significant reduction in loss values starting from the early epochs, decreasing from approximately 0.35 at epoch 1 to 0.0484 at epoch 20. This consistent decline indicates that the model gradually learned from the training data, reducing prediction errors over time. The stable downward trend also suggests that the optimization process proceeded effectively without signs of divergence.

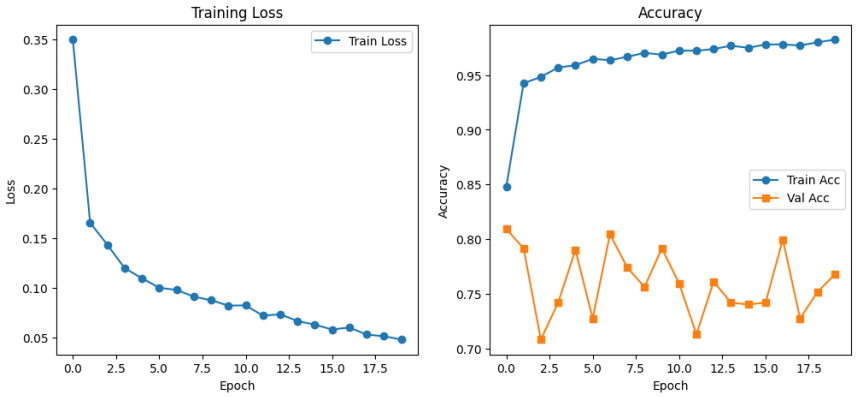

Figure 7: training performance graphs CNN

Meanwhile, the accuracy graph shows a different trend between the training and validation datasets. The training accuracy increased sharply from 0.8480 in the first epoch to 0.9824 at epoch 20, indicating that the CNN successfully captured the visual patterns in the training data.

However, the validation accuracy exhibited noticeable fluctuations. It started at 0.8093, dropped to 0.7083 at epoch 3, and then oscillated between 0.71–0.81 throughout the remaining epochs. At the final epoch, the validation accuracy reached 0.7676. This pattern indicates potential overfitting, where the model becomes overly specialized to the training data, leading to unstable generalization performance on unseen data.

Despite this, the validation accuracy remaining above 0.70 suggests that the model still achieved a reasonable level of classification performance, although further improvement is needed to enhance generalization.

4. CNN + BSIF Model Training

The integration of CNN with Binary Statistical Image Features (BSIF) was implemented to enhance the model's ability to recognize fine texture patterns in chest X-ray images. BSIF provides detailed textural representations that serve as additional input features for the CNN architecture.

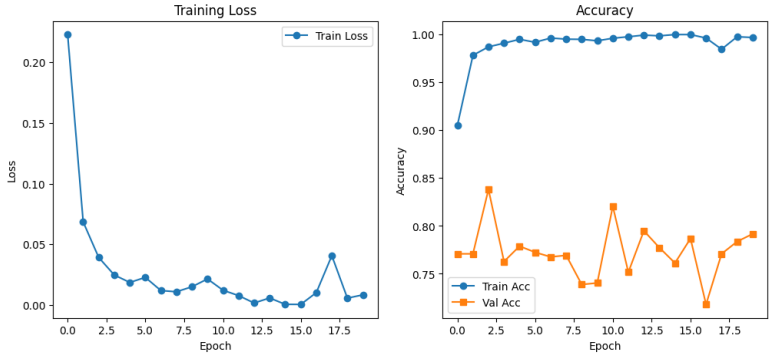

Figure 8: training performance graphs CNN+BSIF

Figure 8 illustrates the training loss and accuracy curves for the CNN+BSIF model over 20 epochs. The training loss decreased rapidly from 0.2228 at epoch 1 to 0.0085 at epoch 20, with a notably low loss of 0.0006 achieved as early as epoch 15. This demonstrates that the combination of CNN and BSIF accelerated the learning process and produced a model with very low prediction error on the training data.

The accuracy graph shows that the CNN+BSIF model outperformed the pure CNN model. Training accuracy increased from 0.9051 in the first epoch to 1.0000 at epoch 15, remaining above 0.9960 through the end of training. This result indicates that the model learned the training image patterns exceptionally well.

Although the validation accuracy still fluctuated, it performed better and more stably than the pure CNN model. Validation accuracy ranged between 0.71–0.84, peaking at 0.8381 in epoch 3 and ending at 0.7917 in epoch 20. While minor overfitting indications persisted, the validation trend was relatively more stable.

These results demonstrate that incorporating BSIF provided a substantial improvement in the model's generalization ability. By integrating additional texture

information from BSIF, the CNN model could more effectively distinguish between normal and pneumonia affected lungs, resulting in better validation accuracy and overall performance enhancement.

5. Vision Transformer (ViT) Model Training

The Vision Transformer (ViT) was employed as a transformer based deep learning approach designed to recognize visual patterns in chest X-ray images. Unlike CNNs, which extract local features through convolution operations, ViT divides each image into small patches and processes them using a self-attention mechanism to capture global dependencies between patches.

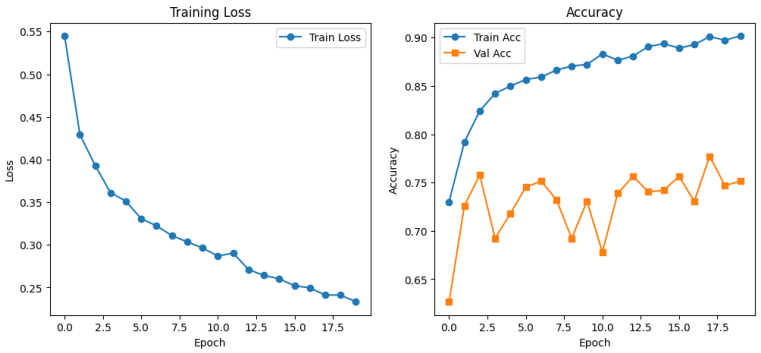

Figure 9: training performance graphs ViT

Figure 9 illustrates the training loss and accuracy curves of the ViT model over 20 epochs. The training loss gradually decreased from 0.5447 at epoch 1 to 0.2333 at epoch 20. This consistent downward trend indicates that the model effectively learned from the training data and progressively reduced prediction errors, although at a slower rate than CNN or CNN+BSIF.

From the accuracy graph, it is observed that ViT performed reasonably well on the training data, with accuracy increasing from 0.7301 in the first epoch to 0.9016 at epoch 20. This result suggests that ViT successfully captured the underlying data representation, even though its accuracy did not reach the level achieved by CNN+BSIF.

On the other hand, the validation accuracy displayed noticeable fluctuations, ranging between 0.62–0.78, peaking at 0.7772 during epoch 18 and ending at 0.7516 in epoch 20. This pattern indicates limited generalization capability, suggesting that the ViT model experienced mild overfitting, though not as severe as that observed in the pure CNN model.

Nevertheless, the application of ViT in pneumonia classification remains promising, as its self attention mechanism allows the model to capture global contextual information that conventional CNNs may overlook. With further optimization techniques such as regularization, fine tuning, or data augmentation the generalization performance of ViT can be significantly improved.

6. Vision Transformer (ViT) + BSIF Model Training

The integration of ViT with Binary Statistical Image Features (BSIF) was designed to optimize the model's ability to capture both global contextual relationships and fine grained local texture patterns in chest X-ray images. BSIF generates binary texture descriptors that enrich the visual representation before being processed by ViT's self-attention mechanism.

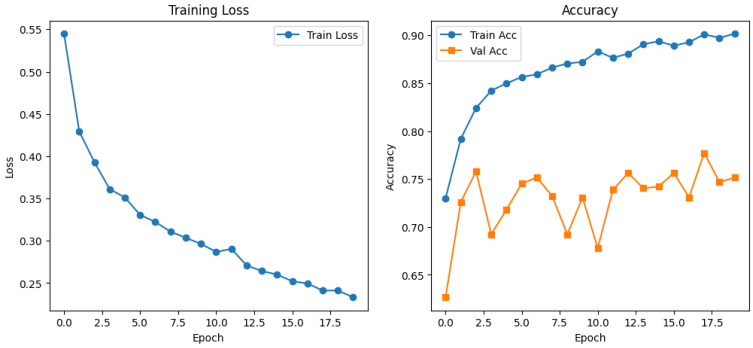

Figure 10: training performance graphs ViT+BSIF

Figure 10 presents the training loss and accuracy curves for the ViT+BSIF model across 20 epochs. The training loss shows a consistent decline from 0.3278 at epoch 1 to 0.0172 at epoch 20, indicating that the combination of ViT and BSIF accelerated the learning process and achieved a low prediction error on the training data.

The accuracy graph demonstrates that the ViT+BSIF model achieved strong performance. Training accuracy increased sharply from 0.8524 in the first epoch to 0.9935 at epoch 20, showing that the model effectively learned and adapted to the training data, approaching near perfect performance.

Meanwhile, validation accuracy fluctuated within the 0.68–0.82 range, reaching its peak of 0.8173 at epoch 19 and concluding at 0.7580 in epoch 20. Compared with the pure ViT model, ViT+BSIF achieved more stable validation performance and a higher average accuracy, demonstrating that the inclusion of BSIF features improved the model's generalization ability.

Overall, this hybrid approach enables ViT to utilize both global spatial relationships and local texture information, leading to a more comprehensive feature representation. Consequently, ViT+BSIF becomes more effective in distinguishing between normal lungs and those affected by pneumonia, outperforming its standalone counterpart.

6. Model Evaluation

Figure 11 presents the confusion matrices of the four models used in this study: (a) CNN, (b) CNN + BSIF, (c) ViT, and (d) ViT + BSIF.

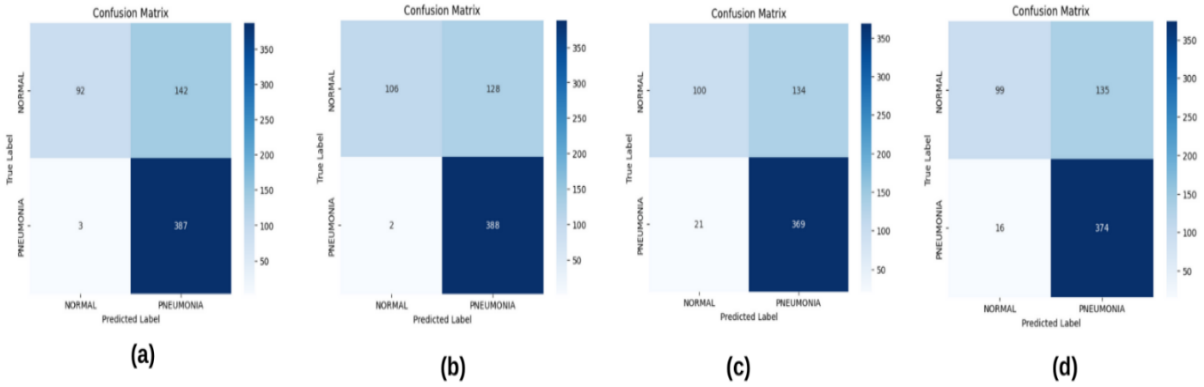

Figure 11: Confusion Matrix

For the CNN model (Figure 11a), the number of correctly predicted Pneumonia samples is quite high (387 samples). However, a significant number of Normal images were misclassified as Pneumonia (142 samples). This indicates that CNN is highly sensitive in detecting pneumonia but less accurate in recognizing normal lung images, leading to a high false positive rate.

A more balanced result is achieved by the CNN + BSIF model (Figure 11b). The integration of BSIF improves the distribution of predictions, increasing the number of correctly classified Normal images (106 samples) while reducing misclassification in the Pneumonia class (2 samples). This finding demonstrates that BSIF enriches the textural information of the input images, allowing the model to more effectively distinguish between normal and pneumonia affected lungs.

For the ViT model (Figure 11c), the number of correctly classified Pneumonia samples remains high (369 samples), but many Normal images are still misclassified (134 samples). This pattern suggests that ViT tends to be biased toward the pneumonia class, although it maintains good sensitivity in detecting infected lungs.

The ViT + BSIF model (Figure 11d) shows improved performance compared to the pure ViT. The number of correctly identified Normal samples increases (99 samples), and the number of misclassified Pneumonia images decreases (16 samples). The integration of BSIF provides additional local texture information, enabling ViT to not only leverage global contextual representations from image patches but also to better distinguish subtle texture variations within lung structures.

Model	Akurasi	Recall	Precision	F1-Score	Spesifisity
CNN	77	69	85	70	39,32
CNN+BSIF	79	72	87	74	45,30
VIT	75	69	78	69	42,74
VIT+BSIF	76	69	80	70	42,31

The CNN model achieved an accuracy of 77%, precision of 85%, recall of 69%, F1-score of 70%, and specificity of 39.32%. By incorporating BSIF, the CNN + BSIF model showed a significant performance improvement, achieving 79% accuracy, 87% precision, 72% recall, 74% F1-score, and 45.30% specificity, making it the best performing model among all tested approaches.

The ViT model recorded 75% accuracy, 78% precision, 69% recall, 69% F1-score, and 42.74% specificity. After integrating BSIF, the ViT + BSIF model exhibited moderate improvement with 76% accuracy, 80% precision, 69% recall, 70% F1-score, and 42.31% specificity.

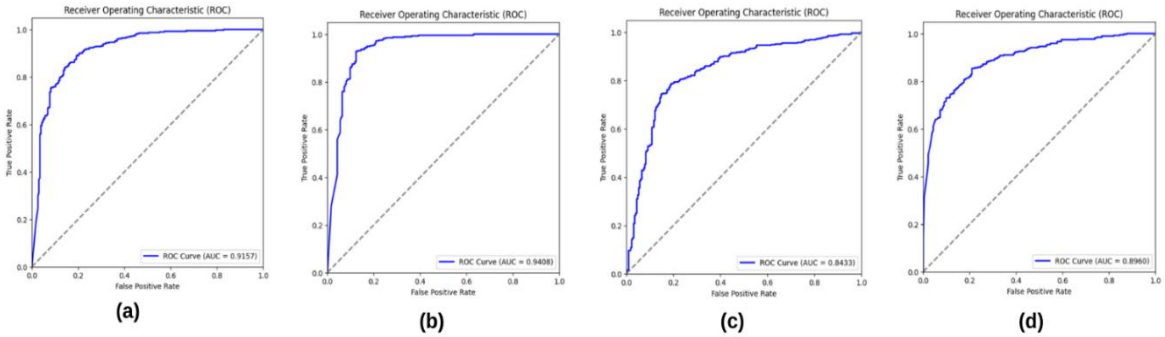

Figure 12: ROC Curve

The CNN model (Figure 12a) achieved an Area Under the Curve (AUC) value of 0.9157, indicating good classification performance. When BSIF was incorporated, the CNN + BSIF model attained the highest AUC of 0.9408 (Figure 12b), demonstrating a notable improvement. The ViT model (Figure 12c) obtained an AUC of 0.8433, which is lower than that of CNN, but after integrating BSIF, its performance improved with an AUC of 0.8960 (Figure 12d).

Overall, the ROC curve results and AUC values strengthen the findings of the previous quantitative evaluation, namely that BSIF contributes significantly to improving model performance, with a greater impact on CNN than ViT.

DISCUSSION

1. CNN Performance and the Effect of BSIF Integration

The pure Convolutional Neural Network (CNN) model initially demonstrated relatively high performance, achieving an accuracy of 77% and a precision of 85%. This result indicates that the CNN is capable of recognizing key visual patterns in X-ray images,

particularly in lung areas where structural changes occur due to infection. However, fluctuations in validation accuracy and the considerable gap between training and validation accuracy suggest indications of overfitting (Anissa et al., 2025). This implies that the model has adapted too closely to the training data and lacks the ability to generalize effectively to new, unseen data.

After integrating the Binary Statistical Image Features (BSIF), there was a significant improvement in performance. The CNN+BSIF model achieved an accuracy of 79%, a precision of 87%, and an F1-score of 74%. This improvement demonstrates that BSIF successfully contributes fine-texture information that was previously difficult for CNN to capture. BSIF emphasizes local texture patterns such as variations in lung tissue density, rib line structures, and infiltration regions, which are key indicators of pneumonia.

These findings are consistent with the study conducted by (Freitas et al., 2020), which compared several texture descriptors, including BSIF, Completed Local Binary Pattern (CLBP), Local Contrast Pattern (LCP), and Local Phase Quantization (LPQ). The study reported that BSIF provides a more effective and stable texture representation when analyzing micro-patterns in images, particularly in cases that require high sensitivity to subtle texture variations. Therefore, the integration of BSIF into CNN can be considered to enrich the visual feature representation, enabling a more balanced model performance in distinguishing between normal and pneumonia-affected X-ray images.

2. Performance of Vision Transformer (ViT) and the Influence of BSIF

The ViT model achieved reasonably good performance, with an accuracy of 75% and an F1-score of 69%. The main advantage of ViT lies in its self-attention mechanism, which allows the model to learn global relationships between different regions of an image useful for detecting pneumonia patterns that may spread across multiple lung areas. However, the pure ViT model performed slightly below CNN, likely due to the relatively small dataset size. According to (Utami et al., 2021), transformer-based architectures generally require a large amount of data to reach optimal performance because of their high model complexity.

After integrating BSIF, the performance of the ViT+BSIF model improved to an accuracy of 76% and a precision of 80%. Although this improvement was not as pronounced as that observed in CNN+BSIF, the results still demonstrate that the textural features provided by BSIF successfully enriched ViT's visual representation. This integration allows ViT to capture both the global information from the self-attention mechanism and the local texture details emphasized by BSIF. Therefore, this hybrid approach can be considered an effective strategy for medical image classification tasks that require high sensitivity to subtle structural patterns.

3. Comparative Analysis Between Models

When comparing all models, CNN+BSIF achieved the best overall performance among the evaluated approaches. This superiority is reflected not only in its higher accuracy and precision values but also in its AUC score of 0.9408, which is the highest among all models. The AUC value indicates the model's strong ability to discriminate between pneumonia and normal chest X-ray classes. In contrast, ViT+BSIF achieved an AUC of 0.8960, showing a notable improvement compared to the base ViT model, but still not surpassing the CNN+BSIF combination.

These findings reinforce the notion that combining traditional texture-based methods with modern deep learning architectures yields better performance than using a single model alone. BSIF acts as a complementary component that enhances the model's sensitivity to micro-level texture variations often overlooked by convolutional or attention-based mechanisms.

4. Limitations and Future Research Directions

Despite the promising results obtained, this study still has several limitations. First, the dataset size used in this study was relatively limited only 2,335 chest X-ray images making it difficult to fully assess the model's generalization capability across data from different hospitals or radiographic devices. Second, the fluctuations in validation accuracy observed in some models indicate the need for additional regularization strategies, such as early stopping, more diverse data augmentation, or further fine-tuning in the ViT layers.

For future work, it is recommended that the model be tested on external and more heterogeneous datasets, and that clinical validation be performed to assess its effectiveness in real-world diagnostic environments. Furthermore, integrating BSIF with other attention mechanisms, such as spatial attention or channel attention, may further enhance the model's ability to detect complex pneumonia patterns in chest X-ray images.

CONCLUSION

This study comprehensively evaluated four models namely CNN, CNN combined with Binary Statistical Image Features (BSIF), Vision Transformer (ViT), and ViT combined with BSIF for pneumonia classification based on chest X-ray images. The experimental results demonstrate that the hybrid CNN+BSIF model achieved the most optimal and stable performance among all tested architectures. Specifically, this model reached an accuracy of 79%, a recall of 72%, a precision of 87%, an F1-score of 74%, and a specificity of 45.30%. These metrics indicate that the integration of BSIF effectively enhances the CNN's capability to capture fine-grained texture details, thereby improving the discriminative power of the model in differentiating between normal and pneumonia-affected lungs.

In comparison, the standalone CNN model obtained 77% accuracy, 69% recall, 85% precision, 70% F1-score, and 39.32% specificity, while the ViT model achieved 75% accuracy, 69% recall, 78% precision, 69% F1-score, and 42.74% specificity. The integration of BSIF into

the ViT architecture slightly improved the model's overall performance, achieving 76% accuracy, 69% recall, 80% precision, 70% F1-score, and 42.31% specificity. These findings suggest that while both deep learning architectures benefit from texture-based feature augmentation, the CNN framework demonstrates higher adaptability and stability when combined with BSIF.

Overall, this research highlights that incorporating traditional texture descriptors such as BSIF into modern deep learning architectures can significantly enhance feature representation and classification accuracy in medical imaging applications. The CNN+BSIF model, in particular, exhibits superior performance by effectively balancing sensitivity and specificity, making it a promising approach for automated pneumonia detection from chest radiographs.

Nevertheless, several limitations were identified in this study. The relatively small and homogeneous dataset may constrain the model's generalization capability across various imaging conditions or populations. Therefore, future work should focus on expanding the dataset with multi-center X-ray collections, implementing advanced regularization and augmentation strategies, and validating the proposed method in real clinical settings. Additionally, exploring other hybrid combinations such as integrating BSIF with attention-based mechanisms or transformer backbones could further enhance detection accuracy and robustness in medical image classification tasks.

REFERENCES

Anissa, T., Ita Mubarokah, & Eneng Susilistia Agustini. (2025). Klasifikasi Rhinosinusitis Menggunakan Modifikasi VGG16. *Jurnal RESTIKOM : Riset Teknik Informatika Dan Komputer*, *7*(2), 253–263. https://doi.org/10.52005/restikom.v7i2.470

Arrofiqoh, E. N., & Harintaka, H. (2018). Implementasi Metode Convolutional Neural Network Untuk Klasifikasi Tanaman Pada Citra Resolusi Tinggi. *GEOMATIKA*, *24*(2), 61. https://doi.org/10.24895/JIG.2018.24-2.810

Cahyanto, H. N., Zulkarnain, O., Farida, D., Kesehatan, I., & Surabaya, B. (2023). Pengembangan Deteksi Dini Dan Penanganan Pneumonia Menggunakan Expert System Berbasis Web. . . *Jurnal Kesehatan Tambusai*, *4*(4), 5182–5187.

Freitas, P. G., da Eira, L. P., Santos, S. S., & Farias, M. C. Q. (2020). Image quality assessment using BSIF, CLBP, LCP, and LPQ operators. *Theoretical Computer Science*, *805*, 37–61. https://doi.org/10.1016/j.tcs.2019.10.038

Ifayatin, H. N., Sarita, I., & Saputra, R. A. (2024). Sistem Deteksi Penyakit Pneumonia Menggunakan Algoritma Faster R-CNN Berbasis Citra Digital Rontgen Dada. *JUSTIN (Jurnal Sistem Dan Teknologi Informasi)*, *12*(4), 645–652. https://doi.org/10.26418/justin.v12i4.81304

Inonu, O. Y., Magda, K., & Amarudin, A. (2025). Analisis Kinerja Algoritma Random Forest Dengan Model Machine Learning Pada Dataset Penyakit Diabetes. *EXPERT: Jurnal Manajemen Sistem Informasi Dan Teknologi*, *15*(1), 1. https://doi.org/10.36448/expert.v15i1.4312

Intyanto, G. W. (2021). Klasifikasi Citra Bunga dengan Menggunakan Deep Learning: CNN (Convolution Neural Network). *Jurnal Arus Elektro Indonesia*, *7*(3), 80. https://doi.org/10.19184/jaei.v7i3.28141

Jiangtao, W., Ruhaiyem, N. I. R., & Panpan, F. (2025). A Comprehensive Review of U-Net and Its Variants: Advances and Applications in Medical Image Segmentation. *IET Image Processing*, *19*(1). https://doi.org/10.1049/ipr2.70019

Li, J., Chen, J., Tang, Y., Wang, C., Landman, B. A., & Zhou, S. K. (2023). Transforming medical imaging with Transformers? A comparative review of key properties, current progresses, and future perspectives. *Medical Image Analysis*, *85*, 102762. https://doi.org/10.1016/j.media.2023.102762

Lorinez S, Y., Yusuf Al Hafiz, A., Khoiri Nasution, A., Denil Sitepu, A., & Syahputra, H. (2025). Implementasi Teknik Histogram Equalization Untuk Meningkatkan Kualitas Citra Pada Foto Lama Yang Pudar. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *9*(5), 7524–7530. https://doi.org/10.36040/jati.v9i5.14723

Nabuasa, M., Turu Allo, D., Reuben Suwitono, M., Studi Farmasi, P., & Matematika dan Ilmu Pengetahuan Alam, F. (2024). Analisis Efektivitas Biaya (Cost Effectiveness Analysis) Penggunaan Antibiotik Pada Pasien Pneumonia Di Rumah Sakit X Bandung. *INNOVATIVE: Journal Of Social Science Research*, *4*, 7690–7705.

Ningsih, N., Ramadhani, A., Santoso, D., Ramadhani, B. D., & Ghofiqi, I. A. El. (2024). Penggunaan Metode Deep Learning untuk Pengembangan Sistem Komunikasi Cerdas bagi Penyandang Disabilitas. *MIND Journal*, *9*(2), 206–219. https://doi.org/10.26760/mindjournal.v9i2.206-219

Nova, N., Mulyanti, A., Burhanie, C. S. A. P., Mulyani, L., Nurjanah, R. G., Utami, W., & Sukaesih, N. S. (2025). Systematic Review: Pemanfaatan Deep Learning untuk Diagnosis Penyakit Menggunakan MRI. *Jurnal Penelitian Inovatif*, *5*(2), 839–852. https://doi.org/10.54082/jupin.1336

Pellicer, L. F. A. O., Ferreira, T. M., & Costa, A. H. R. (2023). Data augmentation techniques in natural language processing. *Applied Soft Computing*, *132*, 109803. https://doi.org/10.1016/j.asoc.2022.109803

Rachman, H., Nuradi, N., Nasir, M., Nurdin, N., & Patricia Hopwood Pasauran, J. (2024). Penentuan Spesies Jamur Pada Sampel Sputum Pasien Pneumonia di UPF BBKPM RSUP Dr. Tadjuddin Chalid Makassar. *Jurnal Media Analis Kesehatan*, *15*(2), 140–146. https://doi.org/10.32382/jmak.v15i2.1200

Rubio, A., & Magnier, B. (2024). Preprocessing of Iris Images for BSIF-Based Biometric Systems: Binary Detected Edges and Iris Unwrapping. *Sensors*, *24*(15), 4805. https://doi.org/10.3390/s24154805

Satria Wiratama, A., Rifqi, M., Maesaroh, S., & Mercubuana, U. (2023). Efektivitas Transfer Learning Dalam Pendeteksian Penyakit Pneumonia Melalui Citra X-Ray Paru Manusia. *Jurnal Ilmiah Sains Dan Teknologi*, *7*(1), 43–52.

Utami, D. Y., Nurlelah, E., & Hasan, F. N. (2021). Comparison of Neural Network Algorithms, Naive Bayes and Logistic Regression to predict diabetes. *Journal of Informatics And Telecommunication Engineering*, *5*(1), 53–64. https://doi.org/10.31289/jite.v5i1.5201

Wibisono, A. D. R., Mandyartha, E. P., & Al Haromainy, M. M. (2025). Klasifikasi Penyakit Kulit Berbasis Support Vector Machine Dengan Ekstraksi Fitur Abcd Rule. *JIPI (Jurnal Ilmiah Penelitian Dan Pembelajaran Informatika)*, *10*(1), 686–698. https://doi.org/10.29100/jipi.v10i1.6039

Yan, H., Mubonanyikuzo, V., Komolafe, T. E., Zhou, L., Wu, T., & Wang, N. (2025). Hybrid-RViT: Hybridizing ResNet-50 and Vision Transformer for Enhanced Alzheimer's disease detection. *PLOS ONE*, *20*(2), e0318998. https://doi.org/10.1371/journal.pone.0318998

Yopento, J., & Coastera, F. (2022). Identifikasi Pneumonia Pada Citra X-Ray Paru-Paru Menggunakan Metode Convolutional Neural Network (Cnn) Berdasarkan Ekstraksi Fitur Sobel. *Jurnal Rekursif*, *10*(1). http://ejournal.unib.ac.id/index.php/rekursif/40